Clemens Meinhart

Studying User Submissions and Content on Reddit

Master Thesis

Graz University of Technology

Knowledge Technologies Institute
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Supervisor: Prof. Dr. Markus Strohmaier
Advisor: Dipl.-Ing. Philipp Singer

Graz, April 2014

Clemens Meinhart

Studie von Benutzereinträgen und Inhalten auf Reddit

Masterarbeit

Technische Universität Graz

Institut für Wissenstechnologien
Vorständin: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Begutachter: Prof. Dr. Markus Strohmaier
Betreuer: Dipl.-Ing. Philipp Singer

Graz, April 2014

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, _____ _____

Datum Unterschrift

1 Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Acknowledgements

This Master Thesis has been developed at the Knowledge Technologies Institute at Graz University of Technology.

Many people were associated in the process of research and writing, and I wish to express my sincere gratitude to all of them.

First and foremost, I wish to thank my supervisor Prof. Dr. Markus Strohmaier, for his help, advice, ideas and inspirations, for all the conversations despite long distances, and for talking sense into a blogger. This work would not have been possible without his guidance.

Further on, I would like to thank my academic advisor Dipl.Ing. Philipp Singer, who answered questions even in the middle of the night, counseled me and gave feedback on almost every step, and who quickly straightened out the writings as well as the opinions.

Special thanks to Elias Zeitfogel for his assistance in idea generation, research and writing, for long days and nights of programming, debugging and computing, and most of all for the camaraderie during our studies.

I would also like to thank Philipp Singer, Florian Flöck, Markus Strohmaier and Elias Zeitfogel for the close collaboration on the research paper *volution of Reddit: From the Front Page of the Internet to a Self-referential Community?*, which was published during the course of this work.

I want to thank Jason Baumgartner for the provision of the data set, and Deni Obid for all the proofreading, corrections and grammar lessons.

My heartfelt thanks also to my girlfriend Anna for the emotional support, and to my family for the encouragement, backing, and the opportunity to carry out my studies in the first place.

Thank you!

Graz, April 2014 Clemens Meinhart

Abstract

Reddit is a very popular website that combines many features, such as social web, link aggregation, democratic voting methods and more. As opposed to Facebook, Twitter or Wikipedia, it is, in relation to its size, an almost blank area on the map of scientific studies. Because of this, the definition of what reddit really is, is ambiguous, and reddit is referred to in many different ways. The purpose of this work is to clarify and unify the definition of reddit by pointing out what reddit is primarily used for. In order to do so, this thesis grants a view into the structure, content and features of reddit. It provides visualized analysis of the growth and evolution of reddit in terms of submissions, of the composition of its content, which sources are used, and how it has changed over time. A categorization of domains is introduced to generalize and summarize the submissions to reddit and its subsections in six categories. That way, the development of the content is manageable and more easily comprehensible. Statistics on the extent of moderation, or in other words the rate of deletions of certain terms and sources, in the political parts of reddit are featured as well. Finally, topic modeling is utilized to find the core topics users are writing about in the subsections of reddit and investigate how well these topics mirror real world events. The results give an insight into the clockwork that drives reddit as well as what kind of content can be expected, embodying a starting point for deeper analysis.

Zusammenfassung

Reddit ist eine sehr populäre Website, welche die Eigenschaften von Social Web, Link Sammlung, demokratischer Wahlmethodik und vielem mehr kombiniert. Im Gegensatz zu Facebook, Twitter oder Wikipedia ist reddit, im Verhältnis zu seiner Größe, beinahe ein weißer Fleck auf der Landkarte der Wissenschaft. Daher ist auch die Definition, was reddit genau ist, nicht eindeutig, und in schriftlichen Quellen werden oft sehr unterschiedliche Bezeichnungen gewählt. Die Aufgabe dieser Arbeit ist es, hier Klarheit zu schaffen und die Definition von reddit zu vereinheitlichen, indem gezeigt wird, wofür reddit hauptsächlich verwendet wird. Diese Arbeit bietet einen Einblick in den Inhalt und die Eigenschaften von reddit anhand visualisierter Analysen des Wachstums in Form monatlicher Einträge, der Entwicklung und der Zusammensetzung von Inhalten und deren Quellen. Eine Kategorisierung von Domains wird eingeführt, die die Inhalte der Einträge generalisieren und in sechs Bereiche zusammenfassen soll. Auf diese Weise wird die Inhaltsentwicklung überschaubar und verständlich. Statistiken über den Ausmaß der Moderierung im Hinblick auf Löschungsraten spezieller Begriffe oder Quellen werden in politischen Abschnitten reddits ebenfalls angeführt. Abschließend wird Topic Modeling angewendet um die Kernthemen, über die die Nutzer in den größten Sektionen reddits schreiben, zu identifizieren, und es wird untersucht, inwiefern diese Themen die Geschehnisse der realen Welt widerspiegeln. Die Ergebnisse bieten einen Einblick in das Uhrwerk, das reddit antreibt, was von der Website erwartet werden kann, und bilden eine breite Basis für weitere, vertiefende Analysen.

Contents

Abbreviations

API	**A**pplication **P**rogramming **I**nterface
AWS	**A**mazon **W**eb **S**ervices
CSS	**C**ascading **S**tyle **S**heet
GIF	**G**raphics **I**nterchange **F**ormat
hLDA	**h**ierarchical **L**atent **D**irichlet **A**llocation
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
IP	**I**nternet **P**rotocol
JSON	**J**ava**S**cript **O**bject **N**otation
LDA	**L**atent **D**irichlet **A**llocation
LSA	**L**atent **S**emantic **A**nalysis
LSI	**L**atent **S**emantic **I**ndexing
MIT	**M**assachusetts **I**nstitute of **T**echnology
NSFW	**N**ot **S**afe **F**or **W**ork
PCA	**P**rincipal **C**omponent **A**nalysis
pLSI	**p**robabilistic **L**atent **S**emantic **I**ndexing
PRAW	**P**ython **R**eddit **A**PI **W**rapper
Reddit	**Re**a**d** **It**
Redditor	**R**eddit **E**ditor
sLDA	**s**upervised **L**atent **D**irichlet **A**llocation
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
TIL	**T**oday **I** **L**earned
URL	**U**niform **R**esource **L**ocator
VSM	**V**ector **S**pace **M**odel

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Reddit, the self-appointed *Front Page of the Internet*, was founded by Alexis Ohanian and Steve Hufman in 2005. Since then it has raised to one of the most popular social community websites, especially in the United States of America where about 6% of all adult Internet users are consuming reddit's[1] services (Duggan and Smith, 2013). Today the website has, by its own admission, more than 100 million unique users from over 196 countries each month[2].

In short, the idea and functionality of reddit is that people submit links or textual content and vote or comment on these submissions. The website ranks the submissions depending on votes and elapsed time since posting, resulting in a top list on what is popular or interesting at the moment.

Zuckerman (2013), director of the Center for Civic Media at Massachusetts Institute of Technology (MIT) and principal research scientist at MIT's Media Lab, wrote in his article *Reddit: A Pre-Facebook Community in a Post-Facebook World* for the *The Atlantic*:

> "Reddit, which calls itself 'The Front Page of the Internet,' is more influential in shaping Internet culture than its comparatively small reach would lead you to believe. Content featured on Reddit frequently 'goes viral,' spreading to other websites, including Facebook. As a result, it's become a popular

1 Although *reddit* is a name, this work follows the trademark guidelines of reddit, which require it to be written in lowercase.

2 Figure 2.3 shows statistics for December 2013 taken from `http://reddit.com/about` and accessed on January 12, 2014.

> destination for politicians and other public figures, including President Obama, to meet their online audiences [...]"

Although reddit can count itself among the largest social online communities by now and represents a driving force of Internet culture, it drew little scientific attention in contrast to other popular representatives of the area, such as Twitter, Facebook or Flickr. Nonetheless, being such a rapidly growing phenomenon on the World Wide Web, reddit makes itself a paramount example for a defining analysis on its own content, structure and temporal changes.

The lack of a full definition engenders a variety of (sometimes even conflicting) statements on what reddit is, both in the press and scientific works. It has become customary to rather describe it repeatedly from scratch without giving an explicit definition. Even the few currently existing papers on reddit either avoided defining it at all, or made statements on content and website-type without justification or reference and defined it rather vaguely. The most common definitions appear to be *"Social news web site"* or *"Social news aggregator"*, as reddit was called by Weninger, Zhu, and Han (2013), Jakić (2012), Lerman (2006) and Mieghem (2011). None of these papers substantiated this statement nor referred to each other or a third external source. But is it really a news website? To the casual visitor, reddit's front page of today is often studded with content for purposes of entertainment, linking to funny images or videos rather than news articles and broadcasts.

Reddit itself stays cautious and simply calls itself *"a source for what's new and popular on the web."*[3]

1.2. Research Questions and Assumptions

Reddit started in 2005 and is now one of the most popular websites worldwide. Within the last eight years it evolved to an established web portal, which suggests that it has experienced a great process of growth. It is unlikely that this was a steady, linear progress, but rather some form of exponential increase. In the beginning, there were some early testers, lured by the announcement and recommendation

3 `http://reddit.com/wiki/faq`

by Paul Graham and the feigned activity through the fake accounts of the developers (Johnston, 2012), building a slow early gain in users as a basis. However, reddit soon got viral when subreddits and other features were introduced.

Kwak et al. (2010) asked whether Twitter is social news media or not. Reddit might be a social news media website as well, and many peers already assume that this is the case. As people post links where original and new content seems to be most interesting, and several subreddits are dedicated to recent events in news and politics, reddit could be a social news aggregator. On the other hand, if image submissions are prevalent, it is more likely that reddit is an image board dedicated to amusement. For the same reason, it could be an aggregator for whatever the dominant content is on reddit, or it could be balanced and merely collect and propagate everything that is new and popular on the Web.

If reddit is, to some point, news media and people use it for information retrieval, it is also powerful in opinion forming, simply because of its wide area of influence given by the number of users within its reach. There are no editors on reddit who direct what gets published. Every user may submit whatever he or she wants, as long as it follows the rules of the subreddit. There are moderators who monitor the adherence of these rules and control what stays on reddit and what is deleted. In the end this should result in a uniform distribution of opinions. There should not be a strong bias in one way or the other. In this matter, the moderators' authority to control the information flow is precarious in subreddits like r/politics, because it could be abused to suppress certain views or links. Due to suppression, it would be possible to steer the public opinion of reddit's community, and with that, probably the electoral behavior and mindset of many unguarded users. If the moderators in certain subreddits favor one direction, reddit is subject to censorship.

With the analysis of content being a major focus of this work, it should be possible to find core topics within each subreddit that mirror its subject. There are two categories of topics to be expected: First, topics that are invariable, constant over time, forming what the subreddit is always about. Second, there should be short term trends of topics, that come up, rise high and disappear again. The second kind of topics are

triggered by newsworthy events and occurrences, or ephemeral hypes and passing fads.

Summarizing the research questions:

1. What model fits reddit's growth best?
2. What kind of content is submitted to reddit and what is the dominant media?
3. Is reddit a social news aggregator, or rather an image board?

Further inquiries revolve around the issues if reddit is subject to biased moderation in the field of politics, if subjects or characteristics of subreddits can be identified using topic modeling, and whether short term trends of topics in subreddits mirror real world events. These inquiries add more insights into the content of reddit and support the answering of the research questions.

1.3. Contributions

The contributions of this master thesis and its results are the following:

1. Similar to the strategy applied by Suh et al. (2009) this master thesis demonstrates the application of growth models known from ecology on the growth of reddit and highlights the best fitting model.
2. A statistical and longitudinal evaluation of content on reddit is provided to settle the question on the existence of a predominant information medium on reddit.
3. A definition of reddit based on statistical analyses of growth, the evolution of subreddits and content is introduced to unify the designations in scientific literature.

These contributions are accompanied by a comprehensive literature synopsis, covering a variety of papers and works on social media websites. Previous publications about reddit are introduced and the approaches and methods utilized by researchers are described.

As basis for many statistical approaches in this work a categorization of content of all domains submitted to reddit, namely the categories *self, text, image, video, audio and miscellaneous*, is introduced. Furthermore, an

extraction of latent topics from submission titles and a trend analysis on genuine topics in subreddits are presented.

1.4. Thesis Outline

This thesis grants a view into the structure, content and features of reddit, inspired by the papers *What is Twitter, a Social Network or a News Media* by Kwak et al. (2010) and *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community* by Bernstein et al. (2011), among others. All of the influencing and related works are described in chapter 3. After giving an overview of the data set in chapter 4, the methods and measures used for the analytical parts are described and explained (chapter 5). The thesis provides visualized analysis of the growth of reddit from January 2007 to December 2012, a categorization of the submissions to reddit and its subsections, researches the development over time, and utilizes topic modeling to find the core topics redditors are writing about. The results are depicted in chapter 6 and further discussed in chapter 7. Throughout this work several subreddits are mentioned by name. For a brief explanation on what their name stands for and what they are about, consult table A.1 in the appendix.

1.5. Collaborations

This work profits from collaborations with and contributions of several colleagues.

First of all Jason Baumgartner, software engineer at the National Democratic Institute in Washington D.C., USA, has to be mentioned, who collected and delivered the data set which is the foundation of all statistics and analysis done here.

The normalization and categorization of domains were cooperatively executed by Elias Zeitfogel and the author of this work. The domain normalization was created collectively in terms of process development and debugging. The author of this work came up with the idea of compiling a reasonably sized sample list of normalized domains. This list was

categorized individually. A comparison script by Elias Zeitfogel marked differences between the categorizations, which were equalized after consultation with advisor Philipp Singer. Elias Zeitfogel investigated attention patterns on reddit in the same time based on the same data set for his master thesis. Both works use the combined categorization and the normalized domains, but for different purposes. Furthermore, both works utilize modified term frequency - inverse document frequency approaches in different environments.

Some results of this work have already been published in the paper *Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?*[4], where my co-authors Philipp Singer, Fabian Flöck, Elias Zeitfogel and Markus Strohmaier are co-responsible for gained insights. The conclusions drawn and represented in this paper had considerable impact on the measuring methods and conclusions that are delineated in this master thesis, as ideas for research objectives and analyses have been cooperatively generated in the process of publication. The paper, however, focuses on questioning reddit's bold claim to be *the front page of the Internet*, thus it depicts and reviews results on structure and content analysis from a different point of view than this thesis.

Aside from these collaborations, this work has been created and its studies conducted independently by its author, Clemens Meinhart.

4 Singer et al., 2014.

2. Introducing Reddit

2.1. History and General Information

What is reddit?[1] The website itself states that the name is a contraction of *Read it*, referring to a usage such as *I already read it on reddit*, which foreshadows what reddit is all about. The term *reddit* interpreted as a Latin word translates to *render*, which fits well, but by coincidence, because it was not intended by the founders. The idea was it to create a list of everything that is currently interesting on the web. This list forms, as reddit claims, the *Front Page of the Internet*.

Alexis Ohanian and Steve Hufman met in college and laid the foundation of reddit right after graduation in 2005 (ATD, 2012). They managed to get funded by Y Combinator[2], a new company by Paul Graham to support start-ups. The idea was to build the website as a combination of Delicious and Slashdot. Before long an early version of reddit went online. The first content was posted by the founders themselves, and to create the illusion of high popularity they used fake accounts for the initial submissions. Shortly thereafter reddit ran on its own, and a year later in 2006 it was bought by Condè Nast Publications, from which it became independent again in 2011 (Martin, 2011a).

The timeline in figure 2.1 grants a good overview on some of the important events in reddit's history. It lists technical advancements of the system, such as the introduction of comments or the launch of subreddits, as well as events of the community like the first presidential *Ask Me Anything* by United States President Barack Obama.

1 Parts of this description of reddit is inspired by an eponymous introductory video on Youtube by an anonymous artist under the alias CGPGrey (2013).

2 `http://ycombinator.com`

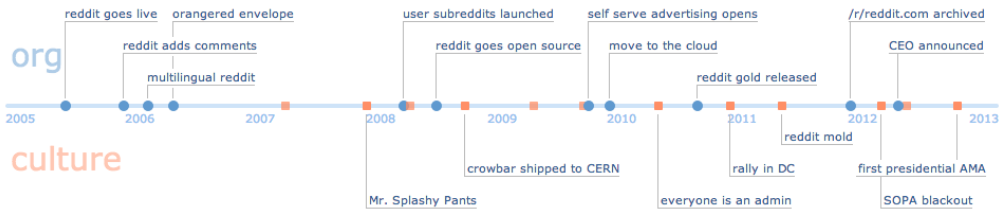


Figure 2.1.: **This is a timeline of reddit history.** Blue circles represent events that were important for the organization and the website itself, like changes to service, technology, policies, or the organization itself. Orange squares are for cultural events that crystallized out of the community on reddit, such as the first presidential Ask Me Anything by President Barack Obama. Source: `http://reddit.com/about`, accessed on January 12, 2014

Regarding technology, reddit nowadays runs on Python (Huffman, 2005) on Amazon Web Services servers (Edberg, 2009) and since 2008 most of its code has become open source (Huffman, 2008).

Revenue is generated with banner advertisement and *reddit Gold*. *Reddit Gold* is a premium membership program that adds extra features for a monthly fee, and can be given away to other users. However, reddit is still not profitable (reddit, 2013).

2.2. Functionality of Reddit

2.2.1. A Quick Overview

For its users, reddit is the gateway to everything interesting going on in the world, as intended by its founders. It is the entry point to the rest of the internet. Reading reddit is comparable to the experience of reading the daily newspaper, except that reddit is supposedly timely, interactive, personalized and participatory (CGPGrey, 2013). In short, the services it offers are that people submit texts or links to external websites on reddit (e.g. links to articles, or images, or videos), and other users vote those links up or down.

This simple mechanism makes reddit into a list of the most popular things that its users are consuming at the time. The reddit homepage,

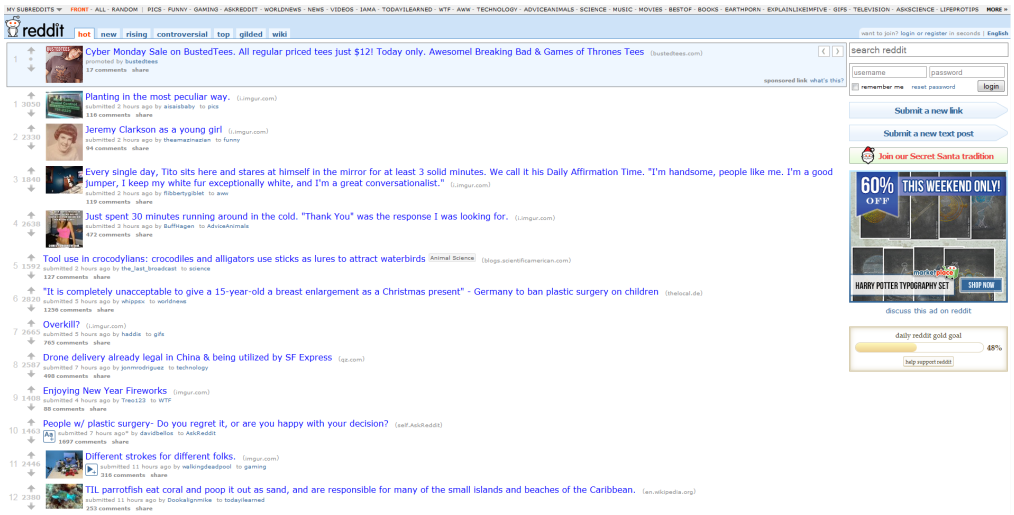


Figure 2.2.: **The reddit front page:** a compilation of the currently best submissions of the subreddits a user is subscribed to.

as depicted in figure 2.2, compiles and sorts this list and makes idle exploration of the internet easy. The design pattern of online communities that gather links for their navigation purposes has been around for quite some time now. It is applied in many web portals, and is known under the term *social navigation*. (Dourish and Chalmers, 1994) However, reddit's good design and many features enhance the navigation experience.

Reddit's content is sorted via a voting process. The higher the difference between up- and downvotes (called *score*), the higher is the ranking in the list. The unofficial goal for a submission is to reach the *front page*, the first 25 submissions ranked, and thus being among the currently most interesting or popular submissions. However, the highest score does not implicate that it is ranked first on the top-list. A submission on the front page is more visible, draws more votes, and its score grows eventually even more (Mieghem, 2011). This by itself would result in a static list of old submissions. Thus, a decay is periodically added to the plain voting scores to prevent ever growing numbers and a nearly unvarying ranking. It ensures that newer submissions have a chance to compete against the ones that already have been on reddit for some time, and that the older content is vanishing continuously. A submission

has a maximum lifespan of about 24 hours. Afterwards the submission is not deleted - it can still be found, but is no longer competing in the ranking process.

Reddit is not only a list, it is a compilation of lists. Each list can be described as a section, channel or community and is called *subreddit*. Subreddits are dedicated to a certain topic, such as politics, images or programming. These channels are identified and linked via the prefix *r/* and their name (e.g. *r/programming* for a subreddit about programming or *r/politics* for a subreddit about politics), which will be used in this thesis to mark a subreddit as such. In a similar manner the prefix *u/* will be used to mark user names. For both subreddits and users the link address is built from the reddit domain *reddit.com/*, followed by the prefix and the name. Reddit also parses its comments for these prefixes and automatically links to the respective page. Subreddits are created by users, therefore there are subreddits on almost every topic imaginable. While r/politics and r/worldnews are what the name suggests, there are also odd examples like r/birdswitharms, a subreddit for edited pictures of birds with human arms attached to them, or subreddits competing for the most disgusting and distasteful images such as r/WTF, or subreddits competing for the cutest animal picture such as r/aww. Each subreddit works the same way the main page does - a constantly updating list of interesting posts according to the people interested in that topic, and it looks just like the front page. A user can subscribe to subreddits he or she is interested in. The aforementioned front page of reddit is the toplist of submissions from certain default-subreddits, which are the largest by number of subscriptions. But by subscribing to or unsubscribing from subreddits, registered users alter the sources and content that compose their personal front page.

Aside from posting links, reddit also features so-called self-posts and a comment system. Self-posts are submissions that do not contain a link to an interesting resource but rather a user-created text, intended for starting debates and discussions, asking the huge community for information or help, or otherwise offering content. The comment system encourages users to have progressive discussions and sorts via a ranking algorithm based on voting to get the allegedly most interesting comments and contributions on top, where they are more likely to be read. This offers the possibility of very popular subreddits like r/IAmA,

an abbreviation for *I am a* combined with *Ask Me Anything*, where users state what or who they are, and then answer questions of other users.

The voting system that distinguishes between uprising, interesting submissions or comments and those that are not is also a driving factor in motivating people to submit content. For the score a submission or comment gets, the authors earn *karma*, imaginary internet points, that have no value at all. Neither can the points be exchanged for anything, nor are there rankings or other benefits from high scores. Nonetheless, watching the numbers in ones profile grow larger and larger is enough to satisfy users, even enough to get people hunting for karma, and other people defending the righteousness of earned points. Karma turns reddit into a game, and games call for rules and fair play. These rules are community-made, sometimes described in the rules for a subreddit, sometimes just an unwritten agreement between users. It is frowned upon resubmitting links that have already been posted before, or to pose as somebody else, or to wrongly claim to be the origin of a submission. Subreddits like r/KarmaConspiracy or r/KarmaCourt watch over submissions and denounce authors that infringe the moral rules of earning karma on reddit.

Reddit features friend-relations as well, which is a unidirectional following system, similar to the one Twitter established. The submissions of befriended users are then listed in its own r/friends subreddit.

There are no dedicated editors on reddit, so its content is essentially unpredictable. Every user is consumer, author and editor in one, which is why users call themselves *redditor*, a contraction of reddit and editor. But each subreddit has moderators, whose objective is to preserve the structure and purpose of a subreddit. Moderators decide whether a submission does or does not fit into the subreddit it is posted to based on the rules that are defined for it. Another cornerstone of reddit is anonymity. Registration is optional, but the features of subscribing to subreddits, voting, commenting and submitting are only available when registered and logged in. Only nickname and password are necessary for registration.

Nonetheless, having an account is not necessary. As it is the case in most user based collaboration websites the *one percent rule*, as defined and illustrated by Nielsen (2006) and Nonnecke and Preece (2000), also applies to reddit. This rule is based on an analysis of online communities

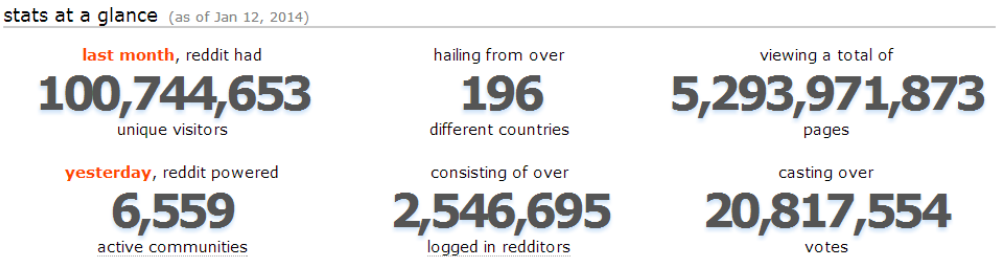


Figure 2.3.: **The user statistics of reddit** states by its own account that the website had 100,744,653 unique visitors from 196 different countries that viewed a total of 5,293,971,873 pages in December 2013. Source: `http://reddit.com/about`, accessed on January 12, 2014

and states that the majority of users (90%) are so called *lurkers*, people who never make accounts and never contribute. A small part (9%) contributes a little and very few, namely 1% of users, are responsible for almost all the content and action in the online community. Because it is not necessary to have an account on reddit to enjoy it as a source for information, most users do not have an account and therefore do not contribute. Figure 2.3 shows that of all unique visitors (who are not all necessarily users of course) only 2.5% actually have an account. Even of those that have accounts, the majority never votes or contributes in other ways. Only very few users contribute on a regular basis.[3]

2.2.2. Detailed Description of Core Features

The Submission

In figure 2.4, a typical submission is depicted with all its elements. A submission gets a title (e) of 300 characters maximum, a subreddit (i) to which the submission is posted and either a link (f) or in case of a self-post a plain text, as illustrated by figure 2.5, from its author (h). The plain text can be a maximum of 10,000 characters in length, but can also be empty - which is often the case when users start a discussion

3 Further detailed statistics of reddits visitor counts in 2013 have been published at `http://redditblog.com/2013/12/top-posts-of-2013-stats-and-snoo-years.html`.

Figure 2.4.: **A typical submission** to reddit as it is listed on the front page. It features information about

a) ranking on one's front page (*1*)
b) current obfuscated score (*3050*)
c) up- and downvote buttons (*arrows*)
d) a thumbnail which links to the submission if it is an image or video
e) the title of the submission (*Planting in the most peculiar way*), which links to the submission
f) the domain it originates from (*i.imgur.com*), which links to a list of recent submissions linking to this domain
g) the time passed since the submission was posted (*2 hours*)
h) the author (*aisaisbaby*), which is also the link to the author's profile
i) the subreddit it was submitted to (r/*pics*), which is also the link to the subreddit
j) the number of comments (*116*), which is also the link to the comments section
k) the share link to email the submission to somebody

with their submission title or as a question to the community. Once posted, the submission appears on the subreddit's list of submissions (best visible when it is sorted by newest entries first, instead of *hot*, which lists the currently best performing submissions first) and users can vote (c) and comment (j) on it.

Comments work much the same way self-posts do. The user submits a text with a maximum of 10,000 characters either directly to the submission itself as a top-level comment or in response to another comment, resulting in a hierarchical comment tree. Users can also vote on the comments and the reddit ranking algorithm then sorts the comments similarly to submissions.

The first votes on the submission are especially important, as early upvotes cause the submission to bubble up in the ranking (a) and thus direct more attention towards the submission, while early downvotes cause the opposite. The exact score (b) is not visible to users until the voting process is halted, when the submission gets archived. Reddit takes effort to obfuscate the current score of submissions and comments for each user to prevent abuse, spam and cheating by randomly varying the exact score. The resulting score is credited to the author as *karma*,

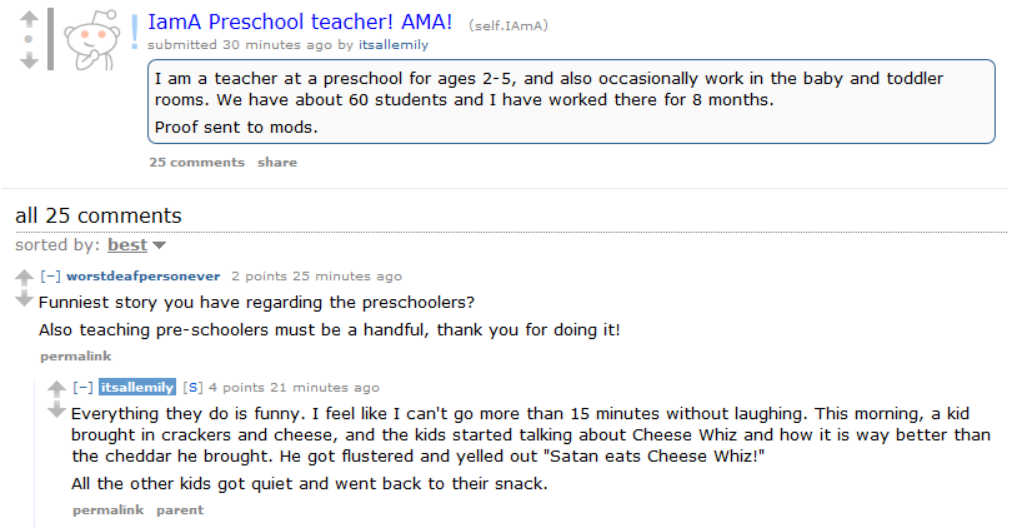


Figure 2.5.: **A self-post** contains text written by the author of the submission and aims to start discussions or pose questions to the community. The maximum length for the text is 10,000 characters. In this particular example, there is a person who claims to be a preschool teacher and offers people to answer their questions. Beneath, there is the comments section, where user *u/worstdeafpersonever* already asked a question and the author answered it.

points without value or meaning, but only if it is a comment (for *comment karma*) or a link submission (for *link karma*). A user cannot gain karma from self-posts. Negative scores are possible as well.

A user can submit almost anything, as long as it fits into the rules of reddit itself and the rules of the particular subreddit. Submissions can be deleted by its author or by a moderator of the respective subreddit.

The Subreddits

As mentioned earlier, reddit is a collection of communities called subreddits. Any user can create a subreddit. A subreddit is always dedicated to a certain theme. A subreddit can also have its own logo, an altered CSS styling and presentation of itself, and different *flair* labels and styling for users, like special batches or titles that appear next to a user's name in the subreddit.

Each subreddit should feature a description and a tag whether it is considered Not Safe For Work (NSFW, or *Over 18* content). The latter is necessary because some subreddits revolve around explicit content, such as nudity, erotica, violence, and gore, which may not be appropriate for all audiences. Furthermore, it can define its own set of rules on how submissions have to be created, what is allowed within it and what not.

Each subreddit also has its own moderators, users who monitor the adherence to these rules and configure all the previously mentioned settings. The first moderator of a subreddit is its creator, additional moderators can be appointed by already declared moderators. Moderators can also mark their own posts as the community moderator's submission (this adds an "[M]" prefix to the title and marks their user name green), delete submissions and ban users from submitting to their community. Outside of their community, moderators have no special powers[4].

Users can subscribe to or unsubscribe from subreddits and thereby define the content of their personal front page. In the data set for this work there are 125,662 unique subreddits, but most of them are not very

4 http://reddit.com/wiki/faq#wiki_moderators

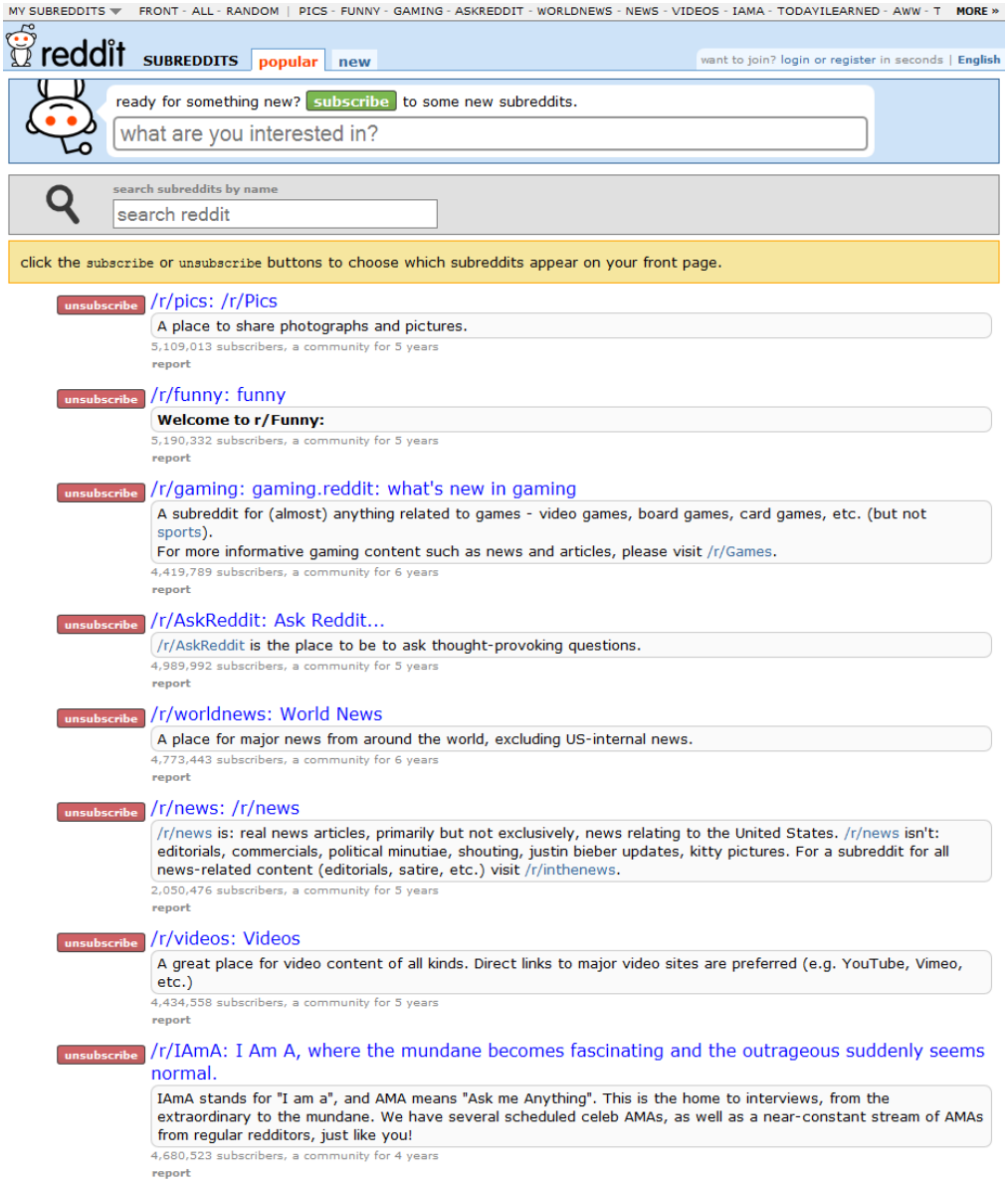


Figure 2.6.: **The subreddits:** different channels for different topics. The list features information about whether or not the user is already subscribed to the subreddit, title and a short description of the subreddit, how many subscribers the subreddit already has, for how long the subreddit has already existed, and options to create so called *multireddits* - custom channels for a user to combine the lists of posts of selected subreddits. The list can be filtered using the search bar, and ordered by popularity and age (*new*).

active. Figure 2.6 shows the interface that provides a search mechanism to find the subreddits a user is interested in. Alternatives to find new subreddits are for example the subreddit r/newreddits, where people promote newly created subreddits, and r/random, which automatically forwards the user to a random subreddit. Until October 2011, a user who was not registered or was registered, but had not yet modified subscriptions, was subscribed to the subreddit r/reddit.com, which cumulated submissions of all manner of content.

In October 2011, r/reddit.com was closed, archived and replaced with so called *default subreddits*. (Martin, 2011b) The initial default subreddits were

- r/pics
- r/gaming
- r/worldnews
- r/videos
- r/todayilearned
- r/IAmA
- r/funny
- r/atheism
- r/politics
- r/science
- r/AskReddit
- r/technology
- r/WTF
- r/blog
- r/announcements
- r/bestof
- r/AdviceAnimals
- r/Music
- r/aww
- r/askscience
- r/movies

A little later r/news was added to this list. Being default subreddits increased the traffic and attention on them, which resulted in most of them being among the top 20 largest subreddits by submissions.

Two years later, the list of default subreddits was changed again, adding r/books, r/earthporn, r/explainlikeimfive, r/gifs and r/television, while dropping r/politics and r/atheism. However, this happened outside of the data set of this thesis.

The Front Page

The front page, as depicted in 2.2, is the first thing a user sees when connecting to reddit. Typically, it contains the currently top rated submissions (dependent on their age) of the default subreddits. The front

page of a registered user does the same for submissions of subreddits the user is subscribed to. Mieghem (2011) proved that a successful post gets more attention due to this sorting, thus generating more votes and becoming even more successful, until the penalty that the sorting algorithm calculates from the submission's age overcomes the gain of upvotes and newer submissions overtake it in the ranks.

With this functionality, each user can create and customize his personal reddit front page, and basically exfiltrate which content is presented.

3. Related Work

To the best of my knowledge, there is no previous work on the definition of reddit. However, there are published studies about related topics and methods, some with partly similar motivations and ideas used on various social networks, that served as an inspiration (chapter 3.1.1). Reddit itself is almost a blank spot on the map of scientific research. Only a few pioneers conducted studies on this website until now, but they grant great insights on several aspects of reddit (chapter 3.1.2). Works about the research methodology of this thesis are introduced in chapter 3.2.

3.1. Social Media Research

3.1.1. General

There are lots of impressive and fascinating works about several different online communities, especially about Twitter[1]. Most of these online communities have at least some similar features to reddit. Twitter, for example, is based on users following the short messages of each other, a functionality that reddit provides too, although it does not publish the relationship data of its users. The titles of submissions, the links and self-texts could be analyzed in similar ways.

The paper *What is Twitter, a Social Network or a News Media?* by Kwak et al. (2010) is about a search for the features that make Twitter what it is. To accomplish this, Kwak et al. managed to crawl an enormous data set of user profiles, social relationships, hashtag topics and messages from Twitter. In contrast to reddit, Twitter concentrates on communication

1 https://twitter.com/

and user relationships, and these factors were interpreted as complex networks and reviewed in this paper. The hashtag streams of Twitter relate to certain topics (comparable to subreddits), and Kwak et al. executed a trend analysis by measuring the activity on those streams. Classifying the topics and comparing them to the time when they had been trending, Kwak et al. discovered that over 85% of these topics are news related. Another characteristic of Twitter is the retweet mechanic, the act of simply forwarding a received message to one's own followers. Kwak et al. found out that this functionality, if applied once, almost always triggers chain reactions of multiple further retweets, often leading to a fast and wide propagation of those tweets.

As retweets are a form of attention a message gets, it is important how tweets are displayed, sorted and ranked. This is a subject matter that is relevant to reddit as well. The paper *An Empirical Study on Learning to Rank of Tweets* by Duan et al. (2010) deals with that matter. It states that Twitter already has a ranking method implemented to find popular tweets in terms of retweets as alternative to chronological ordering, but improvements would be feasible. The authors came up with a ranking strategy for tweets using a Rank Support Vector Machine algorithm, based on various features of tweets. The feature set was constructed to describe three factors:

- Content relevance between queries and tweets
- Twitter specific characteristics like retweet count
- Account authority features that represent the influence of authors of tweets

Duan et al. concluded that if a tweet contained a link, this would be the most effective feature for their ranking purposes to identify popular tweets. Again there are striking parallels to reddit.

The paper *Everyone's an Influencer: Quantifying Influence on Twitter* by Bakshy et al. (2011) confirms this conclusion. They discovered that links in messages, aside from the number of followers, are a feature that increases the popularity and likelihood of diffusion as well. The authors of this work used the link-feature to predict attention, whether a tweet with a link in it would generate a large number of retweets or not. The motivation of this analysis was to identify prominent individuals in the network and to understand how much influence users need to have in order to be most cost-effective for seeding information to a

wide audience - an interesting trait when it comes to opinion forming or advertising. But the experiments showed that a link in a tweet was not a reliable feature for this purpose. Another finding of this work revealed that the few exceedingly prominent individuals of Twitter with lots of followers do not embody optimal starting points for spreading information. Instead, the opposite holds true: Average users with average influence are most cost-effective, according to Bakshy et al. The influence, aside from follower counts, comes from discussions and conversations that go back and forth.

Discussions on Twitter can set off an avalanche of tweets and retweets, leading to new trends and forming public opinion if influential users are involved. Identification of possible discussion starters in advance might be valuable for preventative measures or tracking purposes. Hereby motivated, Rowe, Angeletou, and Alani (2011) presented a method to anticipate if a message would trigger a discussion in *Predicting Discussions on the Social Semantic Web*. The authors examined tweets and looked for prominent features in terms of user influence as well as content of the message that distinguish discussion starters from other tweets. After receiving satisfying results in the training of the classifiers, the best features for their predictions are listed and ranked by their Information Gain Ratio. This substantiated that a central position in the network marks the user as influential and of high reputation, which turned out to be the best feature for prediction of discussion starters as well. Regarding the content, the authors revealed that informativeness, point in time of the post and polarity are the most important features.

The previously mentioned works concentrated on user relationships, network structures, influence or structural conversation analysis, but mostly omitted extraction of information (latent or literally) from the messages. An approach to this is presented in *Twitter as a Corpus for Sentiment Analysis and Opinion Mining* by Pak and Paroubek (2010). Employing linguistic analysis, a recognition of crowd sentiment towards topics or products was viable. The authors used messages with emoticons[2], assuming they mirror the true emotion behind the message, as a pre-labeled training set for a multinomial Naive Bayes classifier. In the process, Pak and Paroubek observed clear distinctions in syntactic

2 An emoticon is a representation of a facial expression by using punctuation marks, numbers and letters to illustrate the author's mood.

structures that allowed conclusions upon the sentiment of a Twitter message. Applying the gained knowledge, their classifier was able to determine positive, negative and neutral sentiments within messages.

Twitter is well covered with scientific attention, but there are thought-provoking and engaging works on other online communities as well. Facebook[3], for example, features functions that have similar counterparts on reddit. Posting of status messages, optionally with links, resembles a submission to reddit. Users can also comment on these status messages. Facebook's well known *Like* is an upvote-only voting system. As described by Rowe, Angeletou, and Alani (2011), discussions are a major driving factor for motivating people to use a social network. Spiliotopoulos and Oakley (2013) looked into the motivations of a Facebook user in *Understanding Motivations for Facebook Use: Usage Metrics, Network Structure, and Privacy* and combined network analysis with interviews of users. Spiliotopoulos and Oakley listed seven motives, namely

- Social Connection
- Shared Identities
- Photos
- Content
- Social Investigation
- Social Network Surfing
- Newsfeed

The authors took the content people posted in status messages into account by differentiating between plain status messages, links, questions, activity references, location check-ins, photos, media clips and others. The motives, merged with content measures and background information about the user (such as age, gender or nationality) extracted from an online survey, were tested for correlations. The paper's results demonstrate that the gender of a user has significant impact on the prediction of the motives *Social Connection* (associated with females) and *Social Network Surfing* (associated with males). Users from outside the USA and older participants show correlations with the *Shared Identities* motivation, indicating that their strongest motive is to be connected with like-minded individuals. These participants have higher numbers of links posted in their status messages, demonstrating that it is important for them to share information with like-minded people. Ties

3 `https://facebook.com/`

were found between the location *USA* and *Photos* (both as motive and content in messages), a coherence that the authors linked to the high distribution of smartphones with build-in cameras on America's market. Additionally, the gathered information and motivations were used to predict a user's answers on privacy related questions. In this experiment the nationality of a user turns out to have significant influence on the question *Generally, how concerned are you about your privacy on Facebook?*, leading to the conclusion that

> "participants from the USA [are] less concerned about their privacy on Facebook"

(Spiliotopoulos and Oakley, 2013)

For reddit however, anonymity and privacy are important features. But there is an online community website that even surpasses reddit in this context: 4chan[4] is a bulletin board with many sub-boards dedicated to certain topics, just like subreddits. In contrast to reddit, submissions that lose attention are automatically deleted. The 4chan community developed its own culture and vocabulary among its users, but its content stream has a similar feel to it as the front page of reddit. The anonymity is even more central to 4chan than it is to reddit. Users of reddit need accounts (with an open profile containing a submission and comment history) to submit. On 4chan, however, accounts are not necessary and users can contribute without an them under the alias *Anonymous* (abbreviated as *anon* by 4chan users).

The effects of the absolute anonymity were explained by Bernstein et al. (2011) in *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community*. Examining the content of posts, it turns out that 4chan users show status and affiliation to the community with slang and system relevant knowledge. Bernstein et al. (2011) stated that

> "anonymity is likely shaping a strong communal identity among a very large set of individuals."

Furthermore, an insight into the ephemerality of 4chan is provided in the paper. Submitted threads are listed chronologically on 4chan, but deleted when pushed to the end of the list by newer submissions. Comments in the thread set the thread back to the front of the list.

4 `http://4chan.org/`

Bernstein et al. observed threads from the moment they were posted until their deletion, and compiled statistics on life time, comments and content of comments. The authors illuminated that some posts had been kept alive for some time by repeatedly posting comments to them, while others dropped almost instantly off the first page (a median thread was about five seconds on the first page) and deleted soon afterwards. The shortest life time of a recorded thread was 28 seconds, from creation to deletion. Manual categorisation of content within the posts to 4chan represent a core element of the text analysis in this paper. The authors defined nine categories for this purpose:

- Themed - posts that start a theme and users answered in respect to the theme
- Sharing content - posts aiming to receive feedback from the community
- Question - posts that ask for advice or suggestions
- Sharing personal information
- Discussion
- Request for item - posts that ask for images or information
- Request for action - posts where users instigate real-life actions
- Meta - posts about the site itself
- Other

Statistically, the *Themed* type of threads have the largest share of 28%, seconded by *Sharing Content* with 19% - together almost half of the posts in their data sample. Both categories often involve an attached image as a central element.

Digg[5] and Slashdot[6] are very similar systems to reddit. Lerman (2006) examined Digg in *Social Networks and Social Information Filtering on Digg* and compared it to reddit. At the time the paper was written, Digg was one of the largest competitors to reddit. This is no longer the case, since Digg lost a large part of its community and market share in 2010, when unpopular business plan changes were executed. Back in 2006, however, Digg was larger and more established than reddit. It featured more social network functionalities. The front page of Digg was more encapsulated, because submissions needed to pass a minimum of upvotes before appearing there, which made it considerably slower than reddit.

5 `http://digg.com/`
6 `http://slashdot.org/`

This paper specifically portrays the social elements of Digg in 2006 and their influence on collaborative ranking of information. It references characteristics of Digg that are probably no longer existent. Lerman tracked submissions over their life span from posting to reaching the front page of Digg for a week in May 2006 and a second time for comparison in November 2006. Observations in the first data set pointed out how social relationships pushed the submissions by members of those connections on Digg, which is called *social filtering*. As a consequence, the majority of successful posts on Digg originated repetitively from the same few users who upvoted each other, a phenomenon Lerman called *tyranny of the minority*.

The paper presents the other side of the coin as well: The social upvoting effects were noticed and opposed by the community. Digg responded, and the algorithm that selected the front page submissions was altered. Lerman showed that in November 2006, after the modification of the ranking algorithm, social relationships had no impact anymore, because the upvotes originating from friendships were ignored. According to the author, this only discouraged users from creating social relationships on the website.

Slashdot is a technology-focused news website. Apart from the predefined theme it is well comparable to reddit and Digg. Slashdot aggregates links to news articles about technology submitted by users, who evaluate them and discuss the subject in a section for comments. In contrast to reddit and Digg, Slashdot features a voting system where only moderators can vote on comments, but not on submissions, in order to encourage and focus on discussions. In *Statistical Analysis of the Social Network and Discussion Threads in Slashdot* Gómez, Kaltenbrunner, and López (2008) analyzed the discussions statistically, and interpreted them as a network. Relations between the author of a comment and the author of a response were understood as edges, the authors as nodes. Investigations of the network graph enabled statements on the topology of the network, the discussion structures and the community structure. The results led to the conclusion that discussions on Slashdot commonly arise when the topic is controversial and many different opinions collide. The controversy of a discussion was then measured via classification methods with features that combined semantic and structural information. The classifier could rank discussions on Slashdot and monitor them while continuously receiving new comments. Controversy, however, is

often dependent on subjective perception, which is why the classifier alone was not enough and human validation still necessary.

There are studies that gather and use data from multiple websites simultaneously. *Characterizing User Behavior in Online Social Networks* by Benevenuto et al. (2009), for example, analyzes the click-streams that are collected by an online social network aggregator system located in Brazil that aggregates content from multiple social networks for its users. The monitored social networks in this paper are Orkut[7], Myspace[8], Hi5[9], and LinkedIn[10]. Orkut is a social network website that commenced in 2004 and is owned by Google. It is very popular in Brazil and India. MySpace was the largest social network website from 2005 to 2008, when Facebook (Schonfeld, 2008) surpassed it. The speciality of LinkedIn is job service. It is a social network to share ones profession and occupation or to search for new employments and business contacts. Hi5 was the third largest social network in 2008 (Schonfeld, 2008), but shifted its focus on gaming and entertainment in 2009. Benevenuto et al. (2009) categorized the user interactions on these social networks in two groups: Publicly visible activities and silent activities. Silent activities (such as browsing profiles and pictures), which are not visible to other users are stated to be the most dominant behavior on all these websites with an overall share of 92%, dwarfing the share of publicly visible activities like writing status messages.

A different approach was illustrated by Leskovec, Backstrom, and Kleinberg (2009) in *Meme-tracking and the Dynamics of the News Cycle*. Popular websites with a large number of visitors are often the origins and distributors of so-called *memes*, basically images, symbols, behaviors or short instances of writings that are popular themselves, repetitively reused and spread quickly, and reddit is no exception to this. Leskovec, Backstrom, and Kleinberg (2009) concentrated on topical memes from the news cycle, the daily rhythms in news media, and tracked short distinctive text phrases from a large number of news media sites and blogs instead of social networks. Their findings outlined how these phrases propagated among news media websites. Moreover, persistent temporal patterns in the propagation of memes were perceptible.

7 `http://orkut.com`
8 `https://myspace.com`
9 `http://hi5.com`
10 `https://linkedin.com`

3.1.2. Reddit Related

Reddit's probably most salient feature is its voting system, with its orange-red (for upvote) and periwinkle-blue (for downvote) design. Mieghem (2011) devoted his research in *Human Psychology of Common Appraisal: The Reddit Score* to finding a mathematical representation of a typical score distribution. The author defined a reddit score probability density function, and illustrated a proportionate effect regarding the scores. Results pointed out that the reddit score resembles a *general random walk*, delimited by the number of users that are able to vote. Its distribution corresponded to a power law form in its intermediate region with an exponentially decreasing tail. The quintessence of the study can be summarized in the following way: The more score a submission already has, the more up- as well as down-votes are received additionally, concluding that the strong grow stronger and the weak stay weak.

The voting system of reddit aims to select interesting submissions from those that are not. A good presentation of the content is critical to get enough attention. *What's in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media* by Lakkaraju, McAuley, and Leskovec (2013) is a study on how the factors *title*, *submission time* and *community choice* of image submissions affect the success of submitted content. To grasp the effects of the title correctly, exclusively image submissions that had been resubmitted multiple times with multiple titles to multiple subreddits were investigated. In the process, two models were created to evaluate the impact and interactions of the factors:

- A community model, containing factors such as number of resubmissions, time of day of the submission and the subreddits it was posted to.
- A language model that measures the quality of the title.

Several methods were applied to quantify the influence of these two models on the success of a submission, like an extension of the supervised topic model framework by Blei and McAuliffe (2007) and linguistic feature analysis. Lakkaraju, McAuley, and Leskovec predicated that good content can speak for itself, although a good title has a positive effect on popularity. This conclusion is rather unsurprising and simultaneously daring, because the paper does not involve a measurement

of content quality, which is arguably difficult for image contributions. Furthermore, the sample is very selective, and submissions that are successful on their first posting and not yet resubmitted are ignored. Nonetheless, the applied methods and resulting findings are alluring. Lakkaraju, McAuley, and Leskovec found various features of titles that had impact on the popularity but depended on the subreddit and time of the day it was posted. For example, words that have a high likelihood to occur in popular posts to the subreddit r/pics are *brilliant*, *optical* or *worth*, while *interesting* or *googled* are considered to be *bad* words to get attention.

Reddit's democratic aspect in ordering things not only applies to submissions, but also to comments. This motivates continuous discussions, as the best statements and most active conversations emerge up on top, and contributors earn comment karma from popular comments. Weninger, Zhu, and Han focused on that characteristic of reddit in *An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community*. Topic models based on LDA with and without non-parametric/hierarchical extension (HLDA) were applied to find topics in discussion threads. It was observed that the hierarchical comment threats generally get started by a top level comment that revolves around a subtopic to the original submission. The earlier a comment is submitted in the course of a discussion, the higher is the chance for it to gather high scores. More subtopics arise out of further sub-level comments as a natural part of the online discourse.

All these features of reddit, examined by previously mentioned scientific works, only operate properly if there is enough content submitted by users and there are users that vote and comment on it. Gilbert (2013) referred to a problem in *Widespread Underprovision on Reddit* that arises when too few users contribute. The provision of content and the filtering by voting was interpreted as work done for free by the author. If too many users *rely on others to contribute without doing so themselves*, underprovision occurred, a problem that is called the *Tragedy of the Commons* (Hardin, 1968). Like Lakkaraju, McAuley, and Leskovec (2013), Gilbert observed submissions that had been added to reddit multiple times as well, and compared their achieved voting score. The second method combined the resubmitted images with statistic reviews of page view data. Reddit, however, does not provide page view statistics, so the author came up with a workaround. Only the subreddit r/pics was

used for this analysis, because the majority of the submissions to r/pics in the data set originated from the website Imgur[11], where page view data is available. It is not possible to find out how many of the views truly referred from reddit and how many originated from other sites or from direct navigation through the browser, Gilbert admitted. This was treated as a tolerable inaccuracy.

While it is true that Imgur was specifically built for reddit's needs, it has its own community as well, its own voting and commenting system. It features accounts and galleries, internal browsing, listing of, for example, the currently most viral images, or the overall highest scoring ones. It bears a close resemblance to reddit, but exclusively for images. All of this was not mentioned by the author, hinting that the part of page views that did not come from reddit were probably considerably larger than assumed in this paper.

The page view data was used to quantify the differences in attention received by submissions that were popular and ascended to the front page of reddit, and those that only appeared on the newest-first sorted list of reddit. The results showed that

> "On average, the most popular images received two orders of magnitude more page views than images on the new queue."

(Gilbert, 2013)

It is suggested in the paper, that there is a widespread underprovision of votes happening on reddit, which means that potentially popular links are often ignored by the voting community on reddit and only achieve high scores on repetitive resubmissions of the same content. Many of the successful posts in the data set were reposted at least two or three times before performing well. More than 52% of all submissions were ignored the first time they were posted.

Repeatedly posted content on reddit is called a *repost* by the community, and redditors usually do not appreciate reposts if noticed. But it seems like these reposts are a driving factor behind reddit.

In his master thesis *Predicting Sentiment of Comments to News on Reddit*, Jakić (2012) applied classification methods to predict sentiment polarity

11 http://imgur.com/

in reactions to news articles posted to reddit. Instead of exploiting emoticons as a labeled training data like Pak and Paroubek (2010), the comments for the training data were manually classified. This approach was compared to a sentiment prediction based on a Twitter corpus. The author used domain-knowledge transfer methods to adapt polarity knowledge from tweets in order to classify the comments on reddit. In review of the results, the author stated that the prediction of the general sentiment is possible, but its outcome strongly depends on the audience and its demographics. Furthermore, if the content of the news articles was not politically motivated, but instead about entertainment for example, the performance of the prediction would be much lower.

Olson (2013) created rudimentary statistics of the relative size of subreddits for each year individually from 2005 to 2012 and published the results on his blog. For every year, the author listed the foundings or closings of noteworthy subreddits, and identified certain events and trends, such as the diversification of subreddits, or the continual descent of r/reddit.com. He noticed that in 2012 the image focussed subreddits became more and more popular and predominant, and concluded that reddit would become an image board in the near future. The reddit user needs to look for appealing content in the every day increasing number of subreddits. In order to minimize this effort, reddit needs to improve its supporting functionalities to find and promote subreddits a user might be interested in, according to the author. Both the statistics as well as the visualizations using stack plots inspired several statistical approaches in this thesis, which deepen and expand the first insights on the evolution of reddit by Olson.

3.1.3. Social Navigation

Reddit implements the mechanics of social navigation in the information space of the World Wide Web. Social navigation is a model where *navigable information systems are extended to support collaborative activity* (Dourish and Chalmers, 1994). The following works describe proper design and application of this model:

Running Out of Space: Models of Information Navigation by Dourish and Chalmers (1994),

Social Navigation - Techniques for Building More Usable Systems by Dieberger et al. (2000), and
Designing Information Spaces: The Social Navigation Approach by Höök, Benyon, and Munro (2003).

3.2. Analysis Methods

3.2.1. Growth Models

One of the research questions of this thesis asks for the best fitting growth model. Aside from pure statistics, an attempt is made to find a mathematical representation for it by fitting it to established models. This approach is inspired by *The Singularity is Not Near: Slowing Growth of Wikipedia* by Suh et al. (2009). The paper aims to show that Wikipedia once grew exponentially, but the increase slowed down and no longer fits the exponential model. In order to prove and find a mathematical representation of Wikipedia's evolution, the authors came up with an interesting idea. Encouraged by an argument from Kurzweil (2005) that biological and technological evolution follow similar rules, they suggested that Wikipedia in fact mirrors typical growth patterns of populations. Growth models known from ecology, usually used to describe population growth that depends on the presence and limits of natural resources, were adapted and fitted to Wikipedia's data. The paper focuses on monthly new articles, edits and active editors (for an analysis on growth rates). Further, extending results with statistics of and fits to the development of Wikipedia's growth and size in articles are published online[12], as the paper itself references.

Results demonstrated that Wikipedia indeed slowed down its growing process both in editor population and in creation and edits of articles. Moreover, the extended results on the Website showed that a fitted logistic extrapolation, at the time the studies had been conducted, predicted a maximum of three to four million articles for the English version of Wikipedia in the future. In the paper, the authors compared Wikipedia's growth to a hypothetical logistic Lotka-Volterra population growth

12 Results and updates can be found at `http://en.wikipedia.org/wiki?curid=244139`, visited on 04/12/2014

model that assumed a limitation of 3.5 million articles. Nevertheless, the knowledge as well as the article count on Wikipedia would not converge to a hard limit. Instead, a continuously decreasing gain of articles was deemed to be likely.[13] Suh et al. (2009) concluded in the paper, that Wikipedia slowed down and developed an increasing resistance to new edits (in terms of reverted edits).

3.2.2. Topic Models

Another objective for this thesis is to discover abstract topics within the submissions of subreddits to specify the content in them in greater detail. This makes it possible to look for trends in those topics, and analyze their changes and dynamics over time. Topic models are created upon several heuristics to determine the affiliations of words to topics. For the purposes of this thesis, Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) are utilized for topic modeling.

Indexing by Latent Semantic Analysis by Deerwester et al. (1990) and *Probabilistic Latent Semantic Analysis* by Hofmann (1999) explain the mechanics of Latent Semantic Analysis, a statistical technique for co-occurrence analysis, automated indexing (LSI), and the probabilistic extension PLSI. It is based on Singular Value Decomposition (SVD) of term-by-term co-occurrence document matrices resulting in a document representation composed of its factor weights. The PLSI extension introduces a so called *aspect model* and describes class association with a joint probability model. Furthermore, Hofmann suggested a maximum likelihood model to avoid overfitting.

A more sophisticated basis for topic modeling than LSI, which applies a probabilistic method instead of statistical co-occurrence, is presented in *Latent Dirichlet Allocation* by Blei, Ng, et al. (2003), *Topic Models* by Blei and Lafferty (2009) and in the recent survey *Introduction to Probabilistic Topic Models* by Blei (2012). The introduction of LDA by Blei, Ng, et al. (2003) is the main reference for the explanation that follows in chapter 5.2.3. In their publication Blei, Ng, et al. briefly defined the alternatives for information retrieval in text corpora, mentioned TF-IDF,

13 In January 2014 the size of the English Wikipedia had already reached a total of more than 4.4 million articles by it's own account at `https://wikipedia.org/`.

LSI and PLSI, and listed the disadvantages of these alternatives that the Latent Dirichlet Allocation aims to overcome. According to the authors, LDA is still a dimensionality reduction technique like LSI, but with proper probabilistic semantics, modularity and extensibility.

With *Probabilistic Topic Models* by Steyvers and Griffiths, there is another excellent work on topic models that provides a structured overview. Steyvers and Griffiths explained the models, variants of it, an algorithmic approach and similarity calculations.

Supervised Topic Models by Blei and McAuliffe is about an augmentation for the LDA topic modeling. Its application area covers document collections where documents are related to a response variable that is not contained in the words, for example movies with ratings or news articles related to sections in the news paper, so the data set is a corpus of document-response pairs. Using SLDA one can compute the topic structure of a document and predict the response variable. The previously known generative procedure of LDA is extended by a step where the response is drawn from a generalized linear model. A Poisson and a Gaussian model are presented as examples for generalized linear models for the response.

Tuulos and Tirri (2004) presented an application of topic modeling based on multinomial principal component analysis (PCA) in *Combining Topic Models and Social Networks for Chat Data Mining*. The authors wanted to uncover the topics of chat logs based on short snippets of text as they appear in chat rooms. After creating a model for this application, the authors successfully enhanced it further by providing a method to decrease noise using web-graph analysis of background information from a corresponding social network.

As mentioned before, the discovered topics will also be subject to a trend analysis. Methodically, the topics will be interpreted as separate documents themselves and form a corpus on which a modified Term Frequency - Inverse Document Frequency measure will mark and distinguish between short lived trends and long term key words of topics in subreddits. In *Term-Weighting Approaches in Automatic Text Retrieval* by Salton and Buckley (1988), typical term-weighting components and formulas for single-term-indexing models are presented and statistically evaluated and compared. The authors identified the best performing document and query weighting approaches, and gave advice on the

construction of components. The paper *An Information-Theoretic Perspective of TF–IDF Measures* by Aizawa (2003) turns the spotlight on some other variants and adaptations of the TF-IDF weighting in other theoretical studies. Motivated by these examples and the idea of finding a probability-oriented version of the vector-space-oriented TF-IDF, Aizawa also presented a mathematical definition of the probability-weighted amount of information (PWI).

4. Data Sets and Collection

This chapter describes the first collections of data with reddit's own Application Programming Interface (API) and the limitations of this API when it comes to scraping a complete record of the submissions to reddit back in time (4.1.1). The ambitious aim of this master thesis has been to produce a longitudinal study of reddit's evolution to its current state, as detailed as possible. The dataset, provided by Jason Baumgartner, consists almost 60 million submissions to reddit, the complete set of all submissions of five consecutive years, and their metadata as provided by the reddit API. The collection process is presented in chapter 4.1, and the contents of the data set and its extent are described in chapter 4.2 and summarized in table 4.1.

4.1. Data Collection

4.1.1. First Collections with Reddit API and PRAW

Reddit offers its own API[1] to allow controlled access to its data and functionality. It answers to Hypertext Transfer Protocol (HTTP) requests with JavaScript Object Notation (JSON)[2]. The simplified accessibility encourages the development of various bots, programs that automatically scrape submissions and comments, that post submissions themselves or answer to comments. Further ease of usage comes from several API wrappers that allow access to it within the domain of a certain programming language. For the purpose of this thesis, a first data set was created with PRAW[3], the Python Reddit API Wrapper. This package offers the

1 `http://reddit.com/dev/api`
2 A full description of the JSON data structures can be found at `https://github.com/reddit/reddit/wiki/JSON`
3 `https://praw.readthedocs.org`

reddit API calls as ready to use methods and ensures the adherence to the API rules. The use of the API is bound to several strictly monitored rules, which ensure that the provided options are not abused (e.g. by spammers) and the traffic caused by automations does not get out of hand. One of those rules is to send no more than 30 requests per minute to Reddit's servers, a rule that effectively slows down the scraping of submissions or comments and thwarts ambitious plans like crawling all submissions back to the beginning of Reddit on purpose[4].

Nonetheless, efforts where made to crawl Reddit backwards in time and a remarkable first set of the submissions posted in the timespan of the year 2012 has been collected by Dipl.Ing. Philipp Singer over the course of several months. This first data set was used as an experimental environment to test most of the methods presented in this thesis and to get an impression on what to expect from a statistical analysis of Reddit on larger scales.

4.1.2. Complete Data Set of Submissions

Jason Baumgartner collected submission data of Reddit using its interface from 2007 on and provided his extensive collection for the purposes of this thesis. He contributes regularly to *r/TheoryOfReddit*, a subreddit dedicated to provide an inquiring look on Reddit itself and to offer space for discussion on analytics, statistics, features and properties. Jason Baumgartner is also the owner and operator of the website `http://redditanalytics.com`, which is, as the name suggests, dedicated to support the analysis of reddit and where the collected data is visualized via a web interface.

Baumgartner collects the data on time by recording the stream of submissions at the moment they are submitted. A month later, when the scores have settled, the submission has been archived and frozen by the system so its values cannot change anymore, the recorded submissions are re-crawled again to deliver the fixed final result of the democratic voting process. That way he ensures to have the very latest submissions at disposal as well as the final scores they achieved. The limitations of reddit's API are not restraining this method, because the expensive

4 One can only request submissions sequenced backward in time.

crawling backward in time is not necessary. The handover of the data set took place in August 2013, containing all submissions from November 2007 to July 2013.

4.2. Description of the Data Set

As mentioned earlier, the data set includes, without limitations, all submissions to Reddit in the span from November 2007 to July 2013. For the purpose of this work the data set has been narrowed to the time span from January 2008 to December 2012, to support and display whole years. For calculation and analysis, each month is one unit of time, resulting in a timeline of 60 ticks as basis for longitudinal statistics and plots. Per submission, the following information is given:

- The Number of upvotes, downvotes, and the resulting score.
- The title of the submission, its author (account name of the poster), and the subreddit it was posted to.
- The author flair text, where moderators can put a tag to the author, and the author flair CSS for the (modified) styling of the author's flair text.
- The IDs of the subreddit and the submission itself.
- The link, if it was not a self-post, and the domain of the link.
- The self text, if it was a self-post and not a link submission.
- The link flair text, where moderators can put a tag to the submission, and the link flair CSS class for its representation.
- The number of comments posted to the submission.
- Timestamps for when the submission was created or edited.
- The number of reports (users can report a submission, to mark it for moderators to be reviewed, because of allegedly broken subreddit rules).
- The name of the moderator who removed the submission, if so.
- The name of the moderator who approved the submission, if so.
- A thumbnail.
- Flags that mark if the submission is hidden or contains adult (known as NSFW, an abbreviation for "Not Safe For Work") content.

The size of the dataset is remarkable, with 58,875,227 submissions in total from 4,910,850 distinguishable authors. Per month, roughly 981,237

submissions are posted to Reddit on average, or 31,653 submissions a day. The amount of submissions more than doubles each year, hinting at an exponential growth. In this data set, spanning the course of five years, there are 125,662 different subreddits in total, some of them still existent, others already closed and deleted. Each subreddit accounts for 469 submissions on average. A breakdown of the properties of the data set is condensed in table 4.1.

Table 4.1.: **Data set statistics**

Number of submissions					58,874,227
Average submissions per month					981,237.12
Number of subreddits					125,662
Average submissions per subreddit					469
Number of self-posts					14,979,707
Number of distinguishable domains					1,841,239
Average submissions per domain					31
Number of submissions with a top 100 domain					40,772,856
Proportion of submissions with a top 100 domain					69.25%
Proportion of self-posts					25.44%
Distinguishable users (authors)					4,910,850
Submissions per month					
142,916	147,713	168,723	167,700	177,275	191,698
218,336	213,050	257,497	283,500	274,430	284,894
333,392	330,553	364,660	359,054	356,846	386,147
428,885	437,748	444,146	462,776	456,700	493,376
555,779	511,861	613,267	626,481	527,416	490,386
511,372	549,113	610,269	641,228	687,952	739,761
837,995	822,302	976,817	971,371	1,081,578	1,153,048
1,264,991	1,448,347	1,482,575	1,590,673	1,634,431	1,772,219
1,981,577	1,961,817	2,158,965	2,279,491	2,293,901	2,393,973
2,663,529	2,782,752	2,595,238	2,797,808	2,726,056	2,755,873

5. Methodology

This work aims to describe the structure and content of reddit in its present state, as well as the dynamics it has undergone since its early days back in 2008. In order to do so, several features were selected and investigated using statistical methods and machine learning approaches. This chapter explains the chosen approaches, depicts the setup of the experiments using these approaches, specifies the assumptions that have been made, explains why assumptions were necessary in the first place, and outlines the possible outcomes. Size and growth of reddit, how it emerged to its current form, and if it can be fitted to established growth models, are the first aspects that are analyzed. The amount of submissions per month is taken into account for measuring the overall advance of reddit's extension (chapter 5.1.1). An ascending graph of these submissions over time suggests the study of its resemblance toward predefined mathematical models. This is executed by calculating parameters of the model's function so that its graph is as close as possible to the actual measured data (chapter 5.1.2).

Since reddit is a composition of subreddits which arise and disappear again, some more popular than others, their growth, development and popularity represents the growth of reddit as a whole. The subreddits are making the difference between reddit and other social link aggregators, therefore their comparison and advancement over time are essential to understand the structure (chapter 5.1.3).

The evaluation and visualization of subreddits leads to the second core of the analysis - the content. The key element of each submission to reddit is the link. A categorization of these links enables statistical measurements and visualizations to outline what is submitted to reddit in its entirety on one hand, and what is submitted to each channel on the other (chapter 5.2.1).

Term Frequency - Inverse Document Frequency (TF-IDF) and topic modeling are utilized for more sophisticated research of content. The goal is to identify trends and bursts within topics in subreddits and see how they change over time (chapter 5.2.3).

5.1. The Evolution of Reddit

5.1.1. Growth of Reddit

Growth is a rather vague term due to the many different ways it can be interpreted and measured. Reddit's growth can, for example, be seen as a growth in user-base, a rise in revenue, additional technological advancement, higher participation of users (a growth in acceptance of the system) or an increase in countries reached (where the users come from). But, as mentioned earlier, this work assumes that subreddits and submissions are the core features of reddit. Additionally the API rules do neither permit nor support the collection of a complete data set containing all accounts on reddit. Thus, the growth is seen in the context of submissions and subreddits. A growth in submissions per time interval provides a strong argument that other aspects of growth have also increased because one can assume that if reddit grows a lot in terms of extend, there must be more users using it and with a larger userbase, the systems must be popular in one way or the other. With more content and more users, potential income (even if it is still not yielding profits) and originating countries will increase eventually, and technological advancement is simply a necessity to continue to deliver the service in time. The focus in terms of the time interval is marked out by the data set. The beginning is set to January 2008 and the end of the observed period is set to December 2012.

5.1.2. Growth Models

The paper on the growth of Wikipedia by Suh et al. (2009) acts as a model for this analysis.The original idea that motivated Suh et al. for their

inquiries and the inspiration for their title came from *The Singularity is Near* by Kurzweil (2005). Kurzweil wrote that

> "technological evolution [is] an outgrowth of — and a continuation of — biological evolution."

He underlined this statement with various examples, from genetics and DNA sequencing to the Internet, nanotechnology and artificial intelligence, describing how technologies followed similar growth patterns to biological ones and what to expect from future research. According to Kurzweil, a paradigm grows exponentially until its potential is exhausted, at which point it is changed until it grows again.

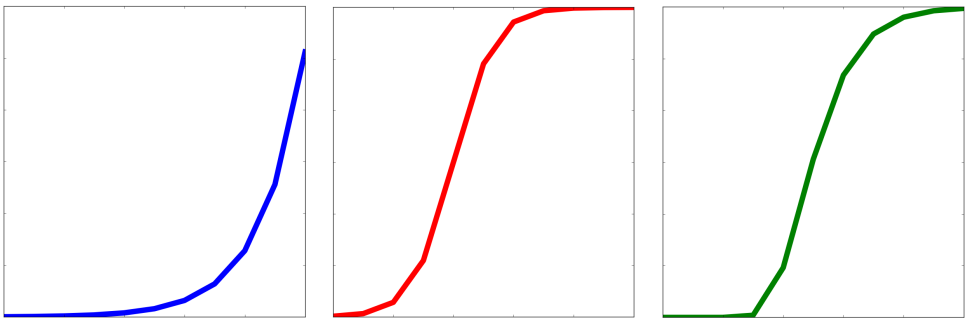
Thereupon, Suh et al. (2009) suggested that Wikipedia had grown exponentially as well. Comparing technology to biological principles, the authors interpreted Wikipedia as a habitat for articles and authors. That way the principles of population growth by Malthus (1826), for example, and other works on natural growth can serve as models, and their procedures can be adjusted and applied on Wikipedia data. Suh et al. proved that these assumptions were not too far-fetched and showed how well an exponential model depicts the actual growth. Wikipedia's growth curve flattens in the progress. Similar to other applications of these models, as described by Meadows et al. (1972), exponential growth, which is limitless by definition, can only be maintained until a certain point, where the growing rate begins to stagnate. Kurzweil (2005) called it *The Life Cycle of a Paradigm*, where each paradigm perambulates three stages, namely slow growth in the early phase, followed by an explosive growth as it is typical for exponential curves and a leveling off in the end. The resulting curve forms an *S*, which is typical for biological growth or any system with relatively fixed complexity that nourishes upon finite resources, according to Kurzweil.

Suh et al. suggested to interpret this deceleration of Wikipedia's growth with the theory that there is an upper boundary, a maximum knowledge that is available and can be described in a Wikipedia article, derived from the theory that there is an upper boundary in knowledge that can be gathered. The upper boundary can be defined statically, or with a function that describes a decreasing growth rate. To illustrate and establish these assumptions, Suh et al. selected and applied models that support midway exponential growth until a turning point, where the growth rate begins to decline until converging to a maximum value or

function, called two-phase exponential models. The logistic equations of Pierre-François Verhulst, also known as the law of population growth by Alfred Lotka, among others, are applied. These models produced better estimates of Wikipedia's growth than the exponential model.

Taking into account *The Life Cycle of a Paradigm* by Kurzweil (2005) and the quickly growing numbers from the data set statistics in table 4.1, reddit might have undergone a progression similar to Wikipedia, or might see itself confronted with a slowdown of submission growth in the future. The absolute numbers of submissions per month are a magnitude higher compared to Wikipedia's article counts. Aside from that, the data set statistics hint that the submission counts increased steadily: in 2007 the amount of submissions per month doubled (factor 199.3%, from 142,916 to 284,894 (a difference of 142,058 submissions), and in 2010 the numbers more than doubled (factor 211.5%), from 837,995 to 1,772,219 (a difference of 934,224). However, in 2012 the gain slowed down to a factor of 139.1%, from 1,981,577 to 2,755,873 (a difference of 774,296). Hence, in search for a description and model of reddit's growth in submissions, the approach of Suh et al. (2009) is adopted. Interpreting reddit as a habitat where submissions are its population, a similar exponential growth model could fit to its population counts. Following the assumption of an upper boundary to knowledge, which states that at some point the newly generated knowledge is constant, the content submitted to reddit could be limited as well. If reddit's submissions are moulded to some part out of knowledge and recent events, and we assume the growth in knowledge slows down, then it seems natural also for the growth of reddit to slow down in proportion to the slower generation of knowledge. This ultimate maximum is hardly imaginable for both Wikipedia and reddit, as there will always be new realizations and events to be described, new creations to be distributed. But in early stages the maximum value simply reforms the exponential curve, giving an idea of when the growth will begin to stagger. In regards to the large submission counts in the data set that are doubling after almost every year, this maximum, if a function that employs an upper bound fits to the sample data at all, will likely be of negligible effect.

Using the resulting data from chapter 5.1.1 as pattern to be matched, three models are tried to meet the expansion of reddit: exponential, logistic and the Gompertz growth model. The mathematical fitting



(a)Exponential Growth (b)Logistic Growth (c)Gompertz Growth

Figure 5.1.: **The growth models in their characteristic forms.** The exponential curve is presented by the blue line in figure (a), with a slow start and a very quick ascension in the end. The logistic curve in (b) has a similar growing phase but only until a breaking point, from which on it approaches a maximum value. The Gompertz curve in (c) grows steeper than the logistic one in the first phase, while the third phase, where it approaches the upper limit, is more dragged out to the right.

of the models, in terms of finding the parameters that result in the least square error, towards the actual data is calculated with SciPy[1], a compilation of scientific computing tools for the programming language Python, distributed as open source software under a Berkeley Software Distribution (BSD) license. The optimization algorithm works best with smaller numbers, so the growth data will be lowered by a magnitude of 10^5 for the fitting process, and solutions scaled back up. The Kullback-Leibler divergence (Kullback and Leibler, 1951) from the fitted curve to the data serves as a measure on which model suits better. It is calculated on the downscaled values to keep them readily comprehensible.

Exponential Model

The exponential growth appears when the time dependent value, e.g. population size in the context of the origins of growth models, increases by a proportional factor at regular intervals (Meadows et al., 1972). Over time, the value increases by a multiple thereof. Following its definition, this model depicts a limitless growth. The characteristic forms of the

1 http://scipy.org

growth models can be seen in figure 5.1a, typically featuring a long and slowly increasing forerun, and a steep growth in the later segment.

The basic formula of this model is defined in the following way:

$$\begin{aligned} x_t &= x_0 * (1+r)^t \\ \text{where } x_t &= x \text{ as a function of } t \\ x_0 &= x \text{ at point of time } 0 \\ r &= \text{growth rate} \\ t &= \text{time} \end{aligned} \tag{5.1}$$

Logistic Model

Exponential growth is without limits, which is an unlikely assumption for any system in the real world. At some point, the growth rate will rise slower and break down eventually, which leads us to a logistic model. The logistic model describes an exponential growth at first, until a turning point, and after that point a convergence to a maximum value or boundary condition[2](or *carrying capacity* in ecology). The resulting curve resembles an S form. In figure 5.1b, the red line depicts a typical logistic growth curve.

$$\begin{aligned} x_t &= \frac{x_{max}}{1+e^{-r*(t-t_{half}}} \\ \text{with } \lim_{t\to\infty} x_t &= x_{max} \\ \text{where } x_t &= x \text{ as a function of } t \\ x_{max} &= \text{upper bound of } x \\ t_{half} &= \text{symmetric inflection point} \\ r &= \text{growth rate} \\ t &= \text{time} \end{aligned} \tag{5.2}$$

x_{max} limits the value of x_t. If a curve resulting from this equation can be fitted to the data of reddit and it fits better than the one from the exponential formula, it would predict a maximum number of submissions

2 http://mathworld.wolfram.com/SigmoidFunction.html

to reddit. This might not be an absolute limit, which would be the end of reddit basically, but rather a forecast of when reddit's growth will probably noticeably stagnate.

Gompertz Model

Benjamin Gompertz defined a mathematical growth model that, similar to the logistic function, grows towards an upper limit, but in contrast to the logistic one it is not symmetric. The logistic curve has three phases, where the first features a slow growth in the beginning, which rises strongly like an exponential function in the second phase and again slows down in the third and last phase, slowing at the same rate as it grows in the first phase. The Gompertz function, however, features an asymmetry of the first and third phase as it approaches the upper asymptote more gradually. The green line in figure 5.1c represents a typical Gompertz curve. The formula with adjusted variable names for comparison to the other models is[3]

$$x_t = x_{max} * e^{k*e^{r*t}} \tag{5.3}$$

$$\text{with } \lim_{t\to\infty} x_t = x_{max}$$

$$\begin{aligned} \text{where } x_t &= x \text{ as a function of } t & x_{max} &= \text{upper bound of } x \\ r &= \text{growth rate} & t &= \text{time} \\ k &= \text{y displacement} & e &= \text{Euler's Number} \end{aligned}$$

x_{max} again describes the upper asymptote where one day reddit's growth will end, described as $x_{max} * e^{k*e^{-\infty}} = x_{max} * e^0 = x_{max}$. The constant k reshapes the curve towards left or right and both k and the growth rate r in this model are negative.

5.1.3. Dynamics of Subreddits

The many channels of reddit are themselves extending in size and count over time and are responsible for reddit's success. The previous chapter

3 http://mathworld.wolfram.com/GompertzCurve.html

already showed how reddit as a whole grew, so changes of subreddits will be represented proportional to the size at the moment. The resulting shares are then printed as a stackplot, where the development and progression of the various channels are well comparable. The time span again covers the whole data set from 2008 to 2012.

A measure for statistical dispersion called Gini coefficient (Gini, 1912) is applied to envelop the gathered information in a single number. The Gini coefficient measures inequality based on a relative mean difference and the Lorenz Curve. It results in a value between 0 and 1. If expressed as percentage (multiplied by 100) it is called *Gini Index*.

A Gini coefficient of 0 corresponds to perfect equality, while a Gini coefficient of 1 means perfect inequality. Perfect equality in this case means that every subreddit gains submissions equally, and perfect inequality stands for a setup where one subreddit gets all the submissions and all the others get none. The advantages are that it is independent of sample sizes and stays comparable, but adequately simple to apply and interpret.

The Gini coefficient is defined, if the data is sorted, as follows

$$G = \frac{\sum_{i=1}^{n}(2i - n - 1)X_i}{2\hat{X}n(n-1)} \tag{5.4}$$

$$\begin{aligned} \text{where } G &= \text{The resulting Gini coefficient of the sample set} \\ n &= \text{the size of the sample set} \\ i &= \text{the index in the sample set} \\ X_i &= \text{the value at index } i \end{aligned}$$

(Dixon et al., 1987)

5.2. Analysis of Content

The content of reddit of about 59 million submissions, is very diversified. Each subreddit is dedicated to its own theme, containing various topics that are transported via various media. Guided by the objectives of this thesis, categorization of normalized links and statistics form the device

to grasp what are the predominant means of content media, be it image or video or text.

Next, a method is introduced to excavate deleted submissions from political subreddits which enables a review of the extent of moderation in these communities and maybe notice some inequalities.

Topic modeling is then introduced as an approach for identification of the core topics in subreddits, what users are posting about, and how this changed over the course of a year. These topics are subject to a trend analysis to spot short time event-like topics and differentiate from persistent core topics of a community.

5.2.1. Categorization

In order to identify what users submit to reddit, the submissions must be categorized. The key element of each submission is the link. The link is the motivation of the submission, it is what the author of the submission wants to show and distribute, what users want to see and what is likely to kick off a discussion in the comments. The normalization of domains and the assignment of categories has been done in cooperation with Elias Zeitfogel.

Domains and Categories

Computers on the Internet are identified and accessed via an Internet Protocol (IP) address, a binary 32 bit (IPv4) or 128 bit (IPv6) number. Because IP addresses are not very memorable to humans, domain names were introduced. A domain name is a string that labels an IP, network or application specific resource. Domain names are hierarchically arranged in levels that are separated by a period, as it is depicted in figure 5.2. A domain name always contains a top-level domain at the end that is either generic (e.g. *com*, *net*, *org*, *info*) or country-coded (e.g. *at* for Austria, *de* for Germany). From right to left follow second-level domain (e.g. *youtube* in youtube.com or *co* in bbc.co.uk), optionally third-level domain (e.g. *bbc* in bbc.co.uk) and so on. Commonly, second-level or third-level subdomains name the purpose of the resource, e.g. the owner, company, service or product that is represented. A link typically starts

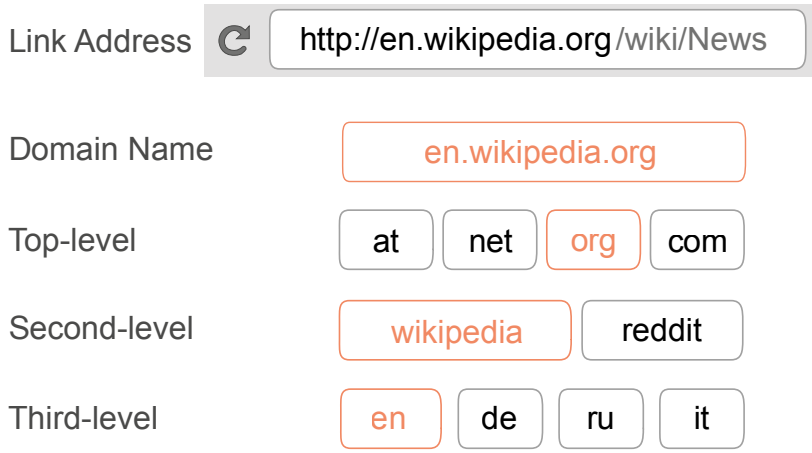


Figure 5.2.: **Typical hierarchical structure of a domain name.** A link address starts with the domain name of the resource. Every domain name has a top-level domain, such as *org*, *com* or *at* and a second-level domain, like *wikipedia*. Some domains also contain even further levels of subdomains, such as *en* in this example.
The domain normalization for this thesis would extract *wikipedia.org* as the identifying domain from the original link address. If another link address would, for example, point to a German Wikipedia article (`http://de.wikipedia.org/wiki/News`), the same identifying domain, *wikipedia.org*, is extracted, because it originates from the same organization and website.

with a domain name followed by further host dependent addressing of the particular resource.

The links of the submissions have to be generalized to a common basis to be able to deploy statistics about their origins. The reddit API and consequently the data set already offers the field *domain*, which contains the whole domain name of the link (for example *en.wikipedia.org* in figure 5.2). Often multiple different domains names belong to the same host, or a host has many subdomains. For accurate results, a common and minimal second-level or third-level domain that unambiguously identifies the source must be extracted. The minimal domain in this context is usually the address of the index page of the particular website itself.

For example, if the link is leading to a video hosted on the large video platform YouTube[4] it would look like this:

```
http://www.youtube.com/watch?v=tlI022aUWQQ
```

Thus the second-level domain that needs to be extracted is:

```
youtube.com
```

Some websites have multiple domains that redirect the user to the main one. For statistical purposes these domains are joined together to the same domain. The YouTube example can be consulted to explain this further. YouTube offers a second domain to shorten links to it if the user utilizes the *share* function. Now the link to the video above looks like this:

```
http://youtu.be/tlI022aUWQQ
```

Consequently the automatically extracted domain is different from YouTube's main one:

```
http://www.youtu.be/
```

The occurrences of multiple domain names of the same website are merged within the same name, which is in this case again:

```
youtube.com
```

4 http://youtube.com/

Many links and their domain names originate from server farms and cloud stores where content is hosted. This is often the case for images hosted on social networks and image platforms. Deviantart[5], a website popular among artists to host and sell their drawings, generates links for images according to the server, each with a different third-level domain and ".net" instead of ".com" as top-level domain:

```
fc00.deviantart.net
fc01.deviantart.net
fc02.deviantart.net
fc03.deviantart.net
...
```

Images hosted on Facebook come from even less recognizable domains, for example:

```
fbcdn-sphotos-c-a.akamaihd.net
```

This domain belongs to the Facebook content delivery network, where static data that is posted on Facebook is hosted. The origin that is interesting for the statistic and that is certainly where the user got it from in the first place, however, is *facebook.com*.

The extraction of the correct domains and problem of concealed, multiple and server-dependent domain names is solved by both using the public suffix list[6], a project by Mozilla Foundation, and applying a manually compiled dictionary to translate and conflate described entities.

Reddit features not only link submissions, but also so called self-posts, which instead of a link only have a text entered by the author of the submissions. This kind of submission is often used for posing questions to the vast user base (e.g. the r/AskReddit subreddit), offering services (e.g. *ask me anything* in case of the r/IAmA subreddit) or directly engaging a discussion (e.g. the r/SRSDiscussion subreddit for progressive-oriented discussions of social justice).

If every link is generalized to the domain it originates from, a statistical evaluation is possible. Tracking the occurrences of each condensed domain over time yields which websites are especially popular and

5 http://deviantart.com
6 http://publicsuffix.org

Table 5.1.: **The six categories** for domains.

Category	Content
self	self-posts
text	News, Blogs, Articles, Papers, everything with text as focus
image	Images and frame-based Animations in the Graphics Interchange Format (GIF)
video	Video platforms like YouTube or Vimeo
audio	Audio platforms like SoundCloud
misc	Miscellaneous, e.g. link shorteners like TinyURL or Hosting Services like Amazon Web Services (AWS)

often submitted to reddit and how their popularity evolved over time. A further interesting aspect is the development of the self-posts in contrast to the links, if they can even compete in terms of quantity against prestigious external websites.

The vast diversity of domains makes it difficult to recognize what the contents are. A minimalistic and simple categorization clarifies this. There is data available online, like the DMOZ open directory project[7] that categorizes domains. However, databases like DMOZ come with the downside of far too sophisticated and enormous sets of categories to find a common basis, often with ambiguous or outdated entries or missing newer domains. As an alternative solution, this work applies a limited set of six self defined categories to describe the content behind a link: self, text, image, video, audio and miscellaneous (table 5.1). The classification of the subject matter provided by a link by means of the specified categories is performed manually. Used as input is a compiled list of the top 100 domains (by amount of links submitted to reddit from 2007 to 2012). Each domain is visited and categorized by its content and the services stated by the website. In case of ambiguity, the most common usage of the domain on reddit is determined to find the right category. The most common usage is elicited by reviewing the functionality provided via the locator */domain/*[8]. Consequently, the categorization is biased on the usage of the links on reddit, which part of the content is shared and submitted to it. It also implicates that this

7 `http://dmoz.org`
8 E.g. `http://reddit.com/domain/youtube.com`

categorization is not universally applicable, because it is tightly tailored to the focus and needs of this thesis. Both Elias Zeitfogel and the author of this work applied the categorisation individually. After consultation with advisor Philipp Singer, the categorizations have been compared and the few differences equalized.

Dynamics of Domains

The cataloguing and categorization of the domains provides basis for many statistics in this work. First of all, the dynamics of domains themselves can be looked upon. A progressional depiction of the 20 most frequently posted domains shows where the content of reddit originates from most of the time and how this changed since 2008. The Gini coefficient as described in chapter 5.1.3 serves again as a comparative measurement.

Categorization of Submissions

Using the created catalogue that maps domains to categories, the submissions can now be categorized. The resulting segmentation gives a quick answer to whether reddit is a social news aggregation website or rather an image board. Further on, the development of the distribution of the categories over time gives insights over the dynamics of the content, in which direction reddit is evolving and what it will consist of in the future.

Content Composition and Development within Subreddits

While the categorization of submissions looks at reddit as a whole, the same methods used upon subreddits might provide a deeper, more detailed understanding of it. Again, the submissions are categorized. Then the categorization is added up for each subreddit. All subreddits have their own rules and topics, which limit the ways users can contribute to them. These limitations will be clearly outlined when categorized. Simultaneously the results show what one can expect from reddit and the particular subreddit. It gives a more detailed view on the developments

and dynamics that are noticeable with the categorization of submissions alone.

5.2.2. Moderation

Although reddit is not professionally edited as News networks or papers are, it is still moderated and manipulated by users. Moderators, as already described in chapter 2.2.2, are users empowered with options to intervene in the workings of a subreddit. These users monitor subreddits, submissions and comments within the subreddits, whether the rules of the subcommunity are complied with, the title is not misleading, the content is suitable, and other criteria set by subreddit rules and the *reddiquette*[9], an etiquette for redditors. The moderators can mark submissions approved or delete them, and even ban users from the subreddit. Now, with so many users on reddit and channels like r/politics or r/worldnews, some moderators might have critical power in controlling which information comes through, or how it is presented. Reddit itself and the rules of general political subreddits that are not themed towards a particular political bearing like r/politics state that they are unbiased.

There is a subreddit called r/POLITIC, not to be confused with r/politics, which has the mantra "Politics without Suppression". In r/POLITIC there are no moderators, only the rules of reddit are enforced. Here a bot called PoliticBot mirrors all submissions from all political subreddits, listing r/news, r/politics/, r/conspiracy, r/Conservative, r/socialism, r/worldnews, r/MensRights, r/progressive, and many more, at the moment they are submitted to the respective subreddit - before a moderator can review or delete it. The original subreddit and author of the submission are saved in the link-flair-text property of the mirrored submission.

The r/POLITIC subreddit and its PoliticBot are a very recent development, as it started only in 2012. The existence of such a mirror may be enough to discourage judgemental or controlling abuse of the options a moderator of critical subreddits has.

9 http://reddit.com/wiki/reddiquette

In this experiment, each submission to r/POLITIC is traced back to its origin to see if it has been deleted. All deleted submissions are collected and statistically investigated, which words and which domains are most frequently removed. Although the data available is very limited, since the bot started working in spring of 2012 and only a couple of months have been observed. The resulting statistics have to be taken with a grain of salt, because submissions can also be deleted by its own author, not only by moderators. Yet, this method would probably spot bans of domains or words, as it happened to the Russian news website RT[10] in r/news in August 2013 (Alfonso, 2013), or the major term blocking of about 50 critical words (such as *NSA*, *CIA* or *net neutrality*[11]) in r/technology unveiled early 2014, where usage of a blacklisted word in the submission title caused its removal (Collier, 2014).

5.2.3. Topics and Trends

Each subreddit is dedicated to an overall subject described by its title. The submissions to the subreddit are expected to fit into this subject. Consequently, the titles of the submissions should depict aspects of that matter, and a relevant constant as well as temporal vocabulary are expected to be perceptible.

Automated identification of such aspects and vocabulary needs linguistic preprocessing, where relevant terms of the text corpus get extracted using methods like tokenization, stop word removal, semantic enrichment and monolingual dimensionality reduction. Generalized vector space models (VSM) are a common method for semantic enrichment, where related terms are found by co-occurrence measures. Dimensionality reduction is then achieved by performing latent semantic analysis (LSA) or a probabilistic variant of it (Sammut and Webb, 2010). An application of a non-parametric Bayesian model called *topic modeling* combines and applies these techniques. Topic models based on LSI and LDA, as they are described by Hofmann (1999) and Blei, Ng, et al. (2003), are utilized to automatically find out what vocabulary is of relevance in subreddits (chapter 5.2.3) and which combinations of terms merge to core topics within their subreddit.

10 http://rt.com
11 http://redd.it/22yewf

For this experimental setup, only one year of submissions is analyzed, the time frame of 2012 which is the most recent complete year and the most voluminous one in the data set. The limitation is necessary, because the subjects in most subreddits vary greatly and quickly, and with a larger time frame, the holistic topics involve higher amounts of verbalisms and phrases. Furthermore, this study aims to be able to trace the calculated topics back, and identify potential real world events and situations that triggered the intensive communication on the subject.

In the first approach, topics are extracted from all submissions in 2012 at once, which means that all submission titles of a subreddit are collected and interpreted as a single text corpus. The results should identify the persistent or often reoccurring topics and vocabulary on reddit. In order to distinguish between the permanent topics and short-lived trends, the topic extraction is then applied on the monthly segments of 2012. A modified version of TF-IDF identifies which terms in the monthly topics are trending.

Topic Models

A popular and simple approach to retrieve the core words in a collection of documents is TF-IDF. The method sets the count of a word in a document i in contrast to the number of documents that contain the word at least once (Salton and Buckley, 1988). Even so, Blei, Ng, et al. (2003) states that TF-IDF does not provide enough information about *"[...]inter- or intradocument statistical structure"*. LSI and probabilistic LSI were introduced to overcome some of these limitations, but both still relied on the assumption that the order of words is not important and both words and documents are exchangeable (*bag-of-words assumption*). An advancement in this area is provided by LDA based topic modeling, which aims to find intra-document statistical structure using mixed distribution.

First of all, the terminology that is used in the context of these models has to be clarified to describe the methodology:

- *word*: The smallest unit in the data, in this case a substring of a submission title containing unicode characters that are delimited by whitespaces or punctuations.

- *document*: A document is a sequence of words. For this application of topic models a document is a title of a submission to reddit.
- *corpus*: The corpus is the whole collection of documents used. In the context of this analysis, the corpus contains the titles of all submissions to reddit in the time span from January 1st 2012 to December 31st 2012 without exceptions.

Topic models try to assign probabilities from one document to other similar ones and to find sets of probabilities of co-occurring words that define these similarities. The methods rely on decomposition of large text corpora and automatic information retrieval within them based on the vector space model. Both topic model approaches that are chosen for this purpose, LSI as well as LDA, start with a set of documents, and assume that given documents have common latent topics. These topics are arbitrarily shuffled over those documents.

For example, the topic model could be set up to find four topics in a set of documents that all somehow relate to the Greek classical elements. Words like *hot*, *bright* or *dry* will generate high co-occurrence probabilities and get labelled as a topic that is related to *fire*. The topic does not get the literal name *fire*. It is merely a vector with the length of the vocabulary from all documents containing the probabilities if the words that fit into the topic. Sorted by value and rematched with the vocabulary, the words of the topic can be listed, but the theme of the topic has to be derived as the quintessence of all those words.

The same, of course, applies for a *water* related topic with co-occurring words like *wet* or *cold*, *air* related for *wind* or *gust*, and *earth* related for *dust* or *dirt*. Words without a dedicated meaning or relevance like articles and pronouns will be evenly distributed and not result in a topic.

The LSI approach, which is the simpler, more straightforward one to topic modeling, handles the topics as a form of co-occurrences of words represented in a weighted matrix. LDA topic modeling, on the other hand, utilizes multinomial variables, which represent the probability distributions over sets of words, as topics and assumes that the topic distributions contain a Dirichlet prior.

The first step for LSI is building a weighted term-document matrix to find all the unique terms in the set of documents. In the term-document matrix, a row represents a term and a column represents a document,

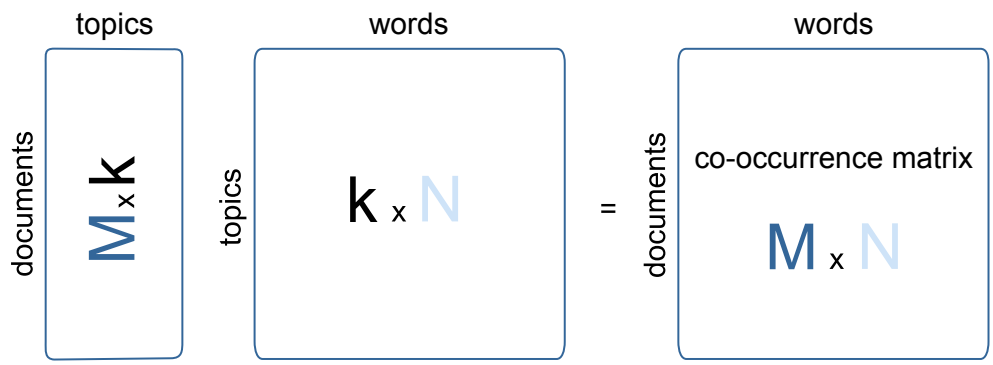


Figure 5.3.: **The term-document matrix**, as described by Steyvers and Griffiths (2005). The variable M stands for the number of documents, N for the number of words and k for the number of topics.

resulting in an $M * N$-sized matrix, where M is the count of documents and N the number of terms (figure 5.3). Optionally the cells in the matrix are then weighted with

- a local term weight, representing the relative rate of the term in the respective document,
- and with a global weight, representing the relative rate of the term in all documents.

For these weights, TF-IDF can be used, TF for the local weight and IDF for the global one. Other options for local weighting are logarithmic weight and augmented normalized TF. Entropy, global frequency-IDF and normal weighting are alternatives for the global weight (Berry and Browne, 2005).

On the weighted matrix, a rank-reduced singular value decomposition is calculated, which finds the connections between terms and concepts in the corpus (Hofmann, 1999). Rank reduction is the key idea of LSA and accomplished by sorting the singular values by size, keeping the largest k values and replacing the remaining smaller ones with zero. A multiplication of the resulting matrices then approximately results again in the matrix before SVD calculation, but with a rank of $k(\leq N)$.

If interpreted geometrically and the rows of the reduced matrices of singular value decomposition are seen as coordinates of points, these

points display documents and terms in a k-dimensional space. An Inner Product calculated between coordinates of points offers comparability (Deerwester et al., 1990).

LDA is a more sophisticated approach that is expected to yield far better results than LSI. Blei, Ng, et al. (2003) describes the basic LDA with a plate notation (figure 5.4), where

> D is the text corpus containing all documents,
> M is the number of documents,
> N is the number of words in a document,
> V is the number of words in the vocabulary of all documents,
> θ is the topic mixture,
> z is a N-sized vector holding the topic for each word, identified by its index between 1 and k,
> k is the dimension of the topic variable z,
> α is a k-sized vector with components $\alpha_i > 0$, representing the Dirichlet prior weights of the topics per document.,
> β is a V-sized vector, representing the Dirichlet prior weights of the words per topic, which, if calculated over all topics, results in a $k * V$-sized matrix, where $\beta_{ij} = p(w^j = 1|z^i = 1)$, and
> w is the set of N words given in a document.

Furthermore, Blei, Ng, et al. (2003) defines that LDA requires a probabilistic generative process, which also exhibits the interaction between documents and the latent documents. This generative process relies on the Dirichlet distribution for the topic mixtures, and assumes that document lengths N_i follow a Poisson distribution.

For each document

1. choose $N \sim Poisson(\xi)$ ($N_i \forall i \in \{1, ..., M\}$),
2. choose $\theta \sim Dir(\alpha)$ ($\theta_i \forall i \in \{1, ..., M\}$).
3. For each of the N words in the document ($\forall i \in \{1, ..., M\}$ and $\forall j \in \{j, ..., N_i\}$)
   a) choose a topic $z_{ij} \sim Multinomial(\theta_i)$,
   b) choose a word w_{ij} from $p(w_{ij}|z_{ij}, \beta)$ a multinomial probability conditioned on the topic z_{ij}.

In contrast to a Dirichlet-multinomial clustering model, which has two levels where only a multinomial clustering variable is sampled once for

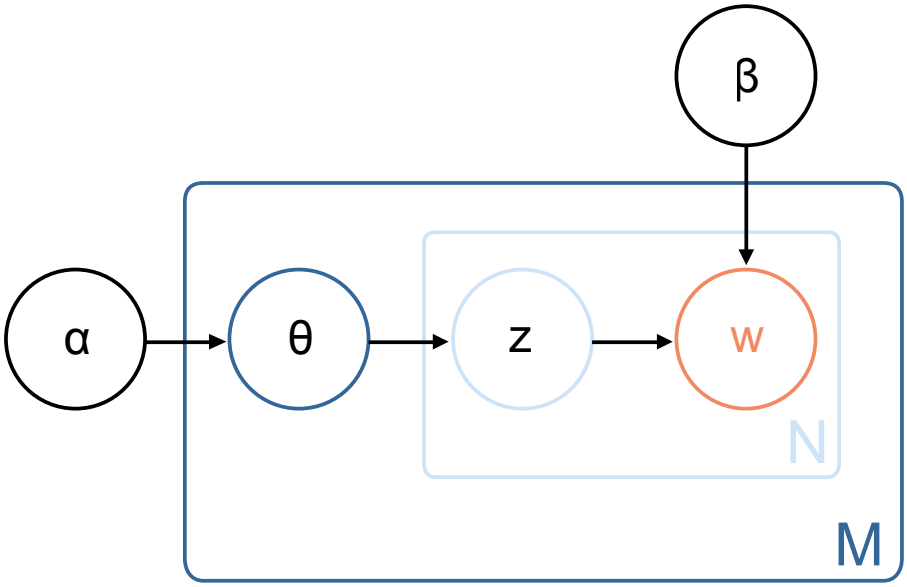


Figure 5.4.: **The plate notation describing the Latent Dirichlet Allocation** by Blei, Ng, et al. (2003). The outer box marked with the M in the bottom right corner represents the set of documents called *corpus* and contains all the variables related to the document i. M stands for the number of documents and also the number of times the process is repeated - once for each document. The inner box with N in its corner, where N stands for the number of words in the document i, carries the variables related to each of the words in the document. The circles (or plates, thus the name) represent the variables, the edges mark dependencies between the variables. The variable w symbolizes the words from the document, z the topics of the current document, and β is a V sized vector, where V is the size of the vocabulary in the documents, that is holding the prior weights of words in a topic. Collecting β for each topic results in a $k * V$ matrix, where k is the dimensionality of the topic variable z. θ is a k dimensioned vector of the topic distribution for the current document. α is a k dimensioned vector of the prior weights of each topic in a document.

each document in the corpus, the LDA model contains three levels with a topic variable (the counterpart to the multinomial clustering variable) that is sampled multiple times per document. This way, it is possible to associate documents with multiple topics (Blei, Ng, et al., 2003).

The plate model in figure 5.4 shows the three environments for its variables. The outermost area is without a plate, depicting α and β as variables that are calculated once when the model is set up. The first plate, marked out by the box with the M in the bottom right corner, describes the variable θ that is calculated once for each of the M documents. The innermost plate with N in its bottom right corner shows the variables z and w that are calculated for each word in each document.

Mathematically, Blei, Ng, et al. defined the marginal distribution of a document $p(w|\alpha, \beta)$ (5.5) and subsequently the probability of a corpus $p(D|\alpha, \beta)$ (5.6) from the product of the marginal probabilities of the documents as follows:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)) d\theta \tag{5.5}$$

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) d\theta_d \tag{5.6}$$

The formula for $p(w|\alpha, \beta)$ describes the joint probability distribution of w with the input weights of α and β, consisting of the integration of the summed up joint distributions of all topics z.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)) \tag{5.7}$$

A joint distribution of a topic mixture θ (5.7) results from the joint probability of θ given α multiplied by the product of all N probabilities of topic z_n given θ, and w_n with topic z_n including the prior weight β.

When summing up, $p(\theta|\alpha)$ can be factored out to yield the marginal distribution for a document after integration over θ (5.5). A multiplication of the marginal probabilities of all M documents in the corpus yields the marginal probability distribution of the corpus D given the input weights α and β (5.6).

This model can be further enhanced in several ways, e.g. to extract interlingua components for cross-lingual text mining if extended to perform clustering on a set of parallel corpora. (Sammut and Webb, 2010)

The Python library *gensim*[12] by Řehůřek and Sojka delivers the functionality for this analysis. The decomposition algorithms used by *gensim* follow the guiding presented by Halko, Martinsson, and Tropp, 2011. The library, however, utilizes an estimation algorithm that is based on the online variational Bayes algorithm for LDA, introduced and developed by Hoffman, Blei, and Bach, 2010, to execute Latent Dirichlet Allocation for reasons of computational and especially memory efficiency. Consequently, the algorithm is very fast and allows streaming of huge document collections, which facilitates the application on the large textual corpora of this study with available computational power.

The setup for both LSI and LDA topic modeling is the following: The text corpus consists of the collection of submission titles from the 20 largest subreddits in 2012. A list of stopwords and punctuation characters (. ? , ; : ! ″ ( ) - [ ] _ { } < > — ′ \ ˜ ˆ * + = /) are removed beforehand. For each of the 20 largest subreddits in 2012, the number of topics to be found is set to 20. Each topic displays the top 10 terms in it. This results in two times $20 * 20$ topics, each with 10 terms, for 2012. Because of the length of this list it is not featured in its entirety in the results review in chapter 6.2.4, but the LDA results can be found en bloc in the appendix C. Additionally to the topics of the whole year of submissions the monthly ones are calculated. The resulting $4,800$ topics are used for the trend analysis, described in chapter 5.2.3.

12 http://radimrehurek.com/gensim

TF-IDF and its Variation for finding Short Term Trends

TF-IDF assumes that terms that are important and specific for a document are mentioned frequently in the document while simultaneously mentioned rarely in most other documents of the corpus (Sammut and Webb, 2010). By straightforward counting of the occurrences of a term t in a document d (term frequency), one recognizes how common a word is to a document. This basically forms the local weight of the term in this specific document.

If this term is common not only in document d but in all documents in the corpus D, it is nothing special and should have a low TF-IDF weight. Therefore, the term frequency needs to be divided by the document frequency, the number of documents where term t is mentioned at least once. This is called the Inverse document frequency, or global weight, since it defines the uncommonness and distinctiveness of a term in relation to all documents. If the term is not in the corpus at all, this would result in a division by zero, which is why the denominator is extended by $+1.0$.

This way, however, larger documents would have an advantage over shorter ones, simply because the more words are contained in a single document, the higher are the term frequencies in it, no matter how special and distinctly descriptive the terms are for the document. Additionally, more of these words might not be used in shorter documents at all. In order to compensate or prevent the bias towards larger documents, the frequencies are usually logarithmically scaled and normalized.

For the purposes of this work, where the documents will consist of the monthly topics of subreddits, all of them of the same length, this bias is of no concern and the basic formula for TF-IDF suffices for a template:

$$
\begin{aligned}
tfidf(t,d,D) &= \frac{f(t,d)}{1.0 + f(t,D)} \qquad (5.8)\\
\text{where } D &= \text{The set of documents (corpus)}\\
d &= \text{A document of the corpus } D\\
t &= \text{A term in document } d\\
f(t,d) &= \text{Frequency of term } t \text{ in document } d\\
f(t,d,D) &= \text{Frequency of documents } d \text{ containing term } t \text{ in } D\\
& \quad (|\{d \in D : t \in d\}|)
\end{aligned}
$$

In order to find the trending terms within the topics found with LDA topic modeling, the TF-IDF formula is modified to represent timely changes. In the case of TF, all topics found in the subreddit s in the month m are interpreted as one single document. Each subreddit forms its own corpus where the list of topics form a document every month.

The topic models have been set to 20 topics with 10 terms each for the top 20 subreddits over 12 months. This results in 12 documents, each with 200 words, for every subreddit. A term can occur 20 times in each month at most, in case it is featured in every single topic.

The IDF part in the denominator on the other hand is used to represent the previous month. In this perspective, each topic is a document by itself, containing 20 words. The corpus D is the set of topics in the month $m-1$ from subreddit s. Counting the document frequency of a term is equal to counting the topics that contained the term in the previous month.

$$
\begin{aligned}
tfidf_{trend}(t,d,D) &= \frac{f(t_i,d_j)}{1.0 + f(t_i, D_{s,m-1})} \forall t_i \in d, \forall d_j \in D_{s,m} \qquad (5.9)\\
\text{where } D_{s,m} &= \text{Set of topics in subreddit } s \text{ in month } m\\
d_j &= \text{Topic } j \text{ of set of topics } D_{s,m}\\
t_i &= \text{Term } i \text{ of topic } d \text{ from set } D_{s,m}\\
f(t_i,d_j) &= \text{Frequency of term } t_i \text{ in document } d_j\\
f(t_i, D_{s,m-1}) &= \text{Frequency of term } t_i \text{ in all documents } D \text{ from month}\\
& \quad m-1 \text{ and subreddit } s \ (|\{d \in D_{s,m-1} : t_i \in d\}|)
\end{aligned}
$$

In short, the approach is to set the number of topics that contained the term in the current month m in proportion to the number of topics that contained the term in the previous month $m - 1$, resulting in a ratio of trendiness between those two months. If the resulting value is high, it means that the observed term is contained in many topics of month m, but is not featured that often in the previous month. On the other hand, a low value indicates that the denominator is high, hence the term was already featured in many topics in the $m - 1^{th}$ month and either stayed influential or decreased in importance.

6. Results

In this chapter, the results of the previously introduced methodologies are presented and described. Most methods are composed of statistical approaches and visualizations that are subject to interpretation. Commencing with the growth measurements and modeling in chapter 6.1, insights into reddit's evolution are granted. Chapter 6.2 relates to the content of reddit, the results and implications of the categorization, the statistics of categorized submissions and subreddits. Furthermore, the extent of moderation is evaluated statistically by reviewing deletion rates of domains and terms. Finally, the topic modeling results are presented in chapter 6.2.4, and short term trends are identified and described.

6.1. The Evolution of Reddit

This first part of the inquiries aims to show the progress of reddit from a rather mediocre website back in 2008 to its state and size at the end of 2012 (chapter 6.1.1). The course of this advance is reminiscent of certain models, especially the one of exponential growth. Section 6.1.2 compares the growth of reddit with said models, and investigates whether their functions and parameters are optimized to approximate the curve. With the introduction of subreddits, one can observe how these subchannels progressively overtake the structure of reddit, arranging the content (chapter 6.1.3). In the course of these statistics, the 20 largest subreddits are introduced, which are subject to further investigation in section 6.2.

6.1.1. The Growth of Reddit

There are diverse indicators for growth of a website like reddit. The quantity of users, comments, or page hits could be investigated and would yield interesting insights. This work, however, concentrates on the growth in terms of its primary service and content feature, the submissions. Furthermore, reddit does not provide exhaustive page hit statistics or user data at the moment.

As the data set already hinted, reddit grew extraordinarily over the past five years. Figure 6.1 displays the progress of monthly submissions to reddit. The first two years show a rather steady gain in submissions, almost linear until 2010, where two pronounced disruptions are noticeable.

Explanations for these spontaneous declines of submissions are hardly verifiable in hindsight, but an investigation of reddit's blog archive gives an idea of the problems the website was facing in 2010. During this time, reddit struggled repeatedly with technical issues, mostly high server loads, server failures and consequently slow responsiveness.

In May 2010, reddit (2010) reported significant increases of storage requirements and subsequent problems with the distributed database system Cassandra. Data integrity problems occurred with broken listings in the system. Nevertheless, the traffic reddit experienced grew further, while reddit was still running on a tight budget and not able to provide additional hardware. In July, Schiraldi (2010b) from the reddit team asked for financial help from the community, because the website's performance deteriorated constantly. Money should not come from donations, but from a new product: reddit Gold, a premium membership, was introduced and soon turned out to be a success (Slowe, 2010b). Schiraldi (2010a) also stated and proved that external traffic analysis providers, namely Compete, Quantcast and Alexa, displayed wrong numbers and underestimated reddit's traffic by a long way.

After several extensions of the server architecture, Slowe (2010a) reported a tremendous raise of traffic by the end of August that deemed an ordeal for the efforts done previously. In this article, a takeover of the

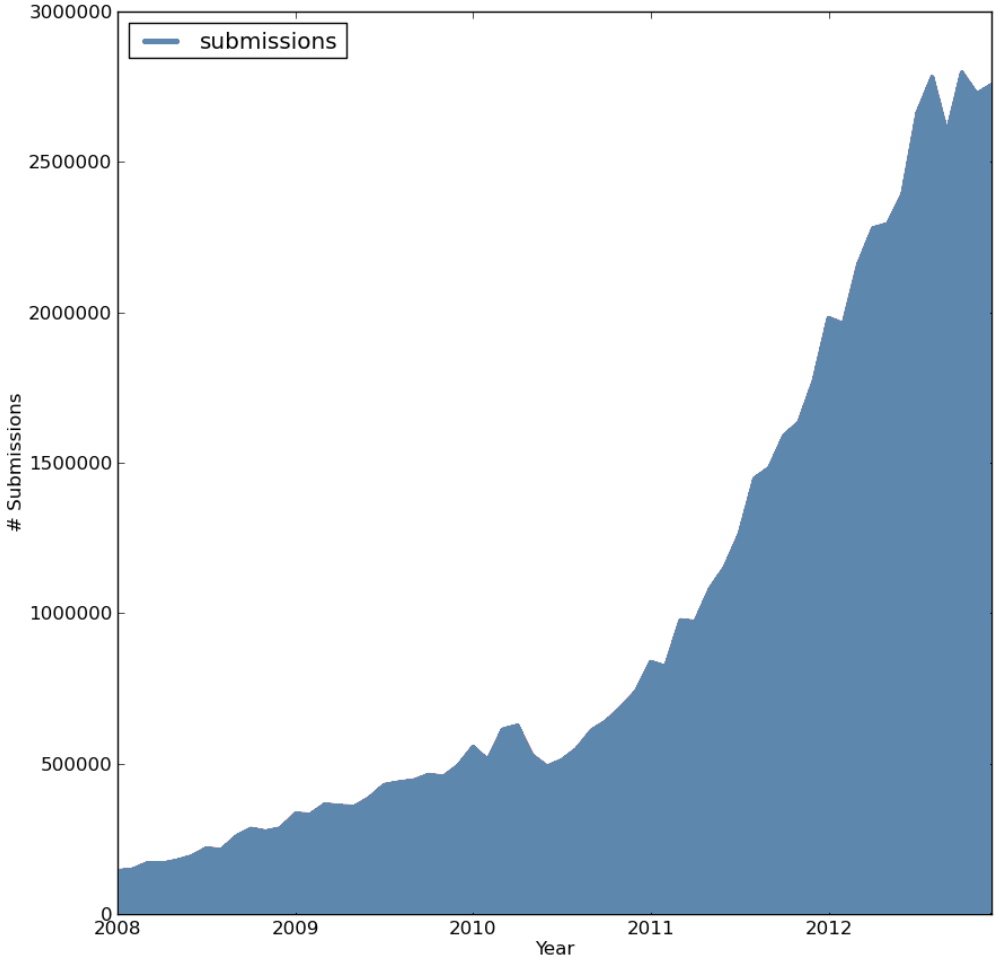


Figure 6.1.: **The growth of reddit** in submissions over time.

community of Digg[1] is mentioned to be the cause for the high increase in traffic.

In August 2010, Digg, a reddit competitor utilizing a very similar system and service, released its fourth version. But the new release of Digg was afflicted with many problems regarding bugs, glitches and an unpopular overall change of the business model. Its community was alienated, felt exploited and made pleas to return the former Digg version 3 (Finn, 2010).

The Digg situation even drew the attention of news media. Friedman (2010) of TIME summarized the new features and letdowns of Digg, and explains the consequences. One of these new features allowed media networks to automatically submit all their publications and articles without users posting them, which led to congestions of News articles and simultaneously to posts of mediocre popularity that still stayed on Digg's front page for far too long. That way, large websites easily suppressed smaller ones. As a consequence, users switched to reddit, and an automatic article feed to Digg using aforementioned feature was implemented that streamed reddit links to it as a provocation. The Los Angeles Times wrote especially about reddit's profit from Digg's problems (Milian, 2010), and featured an unconcerned statement of Digg founder Kevin Rose.

The events of the Digg affair as well as the technical reports on reddit's blog correlate perfectly with the process of the submission growth curve in figure 6.1, both with the interruption in the first half of 2010 and the enormous gain in the second half. A further hint on the impact of Digg's mistakes and reddit's expansion is given by the Google Trends[2] service in a record of the frequencies in which the names of both websites were searched via Google in 2010, as it is depicted in figure 6.2. One can clearly recognize the peaks of Digg and reddit in August and September 2010. Digg, however, lost this interest soon thereafter, while reddit managed to conserve the push.

In 2012, Digg was split up and sold, which again might have had beneficial influence on reddit's further growth. However, Digg relaunched

1 http://digg.com
2 `http://google.com/trends`

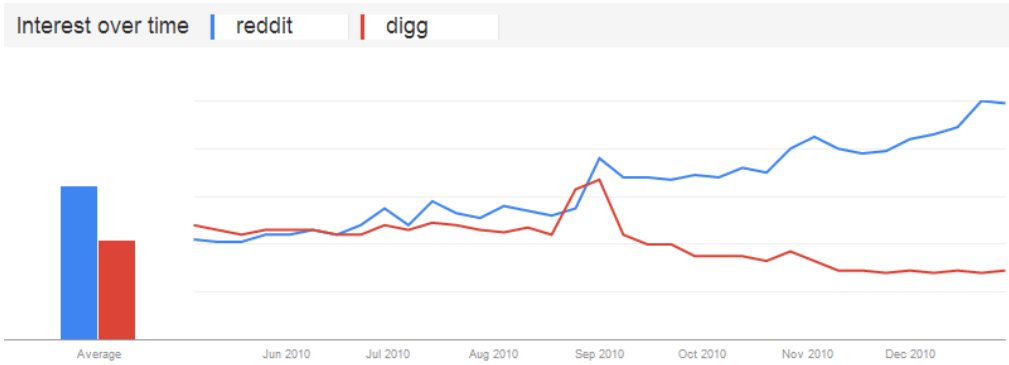


Figure 6.2.: **The Google Trends analysis** of reddit and Digg in 2010. The blue line stands for the interest in the search term *reddit*, the red one for the interest in the search term *digg*. Starting with a close lead for Digg, the first half of 2010 reflects the competition between both websites. At the end of August, first Digg rises strongly due to the relaunch. Reddit follows shortly thereafter, but preserves the gained attention, while Digg loses a lot of interest between September and October.

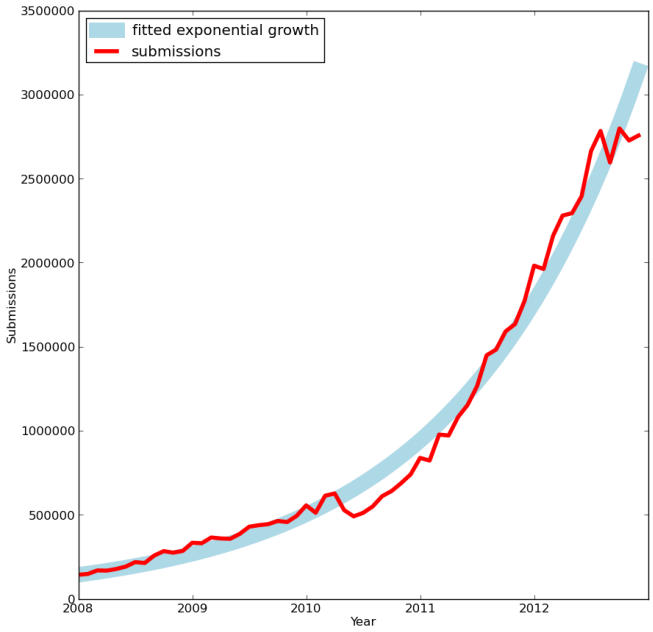
on the 31st of July, which probably accounts for the disturbance in the mid-2012 upward trend in the figure.

6.1.2. Growth Models

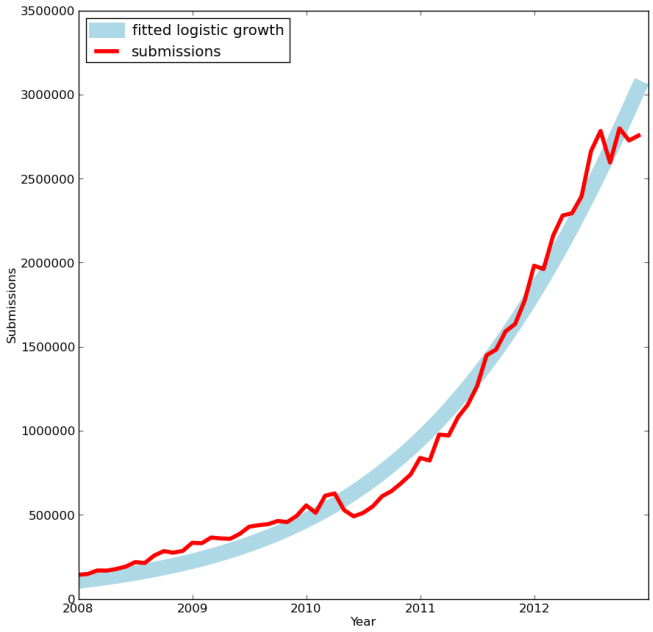
In this chapter, three models are applied to fit towards the data. The curve depicted in figure 6.1 has, as presumed previously in chapter 5.1.2, a distinct form, similar to an exponential curve to some degree. But in contrast to that, the final months display a slowdown in growth, resembling the logistic model or the Gompertz one.

The Exponential Growth Model

The exponential growth model (as described in chapter 5.1.2) fits the curve of reddit's submissions really well. In figure 6.3a, the target empirical data is plotted as a red line, and the light blue line marks the fitted function. The calculated optimal value for the growth rate of the model to fit to the data, after minimizing the summed up squared error,



(a)The fitted exponential function.



(b)The fitted logistic function.

Figure 6.3.: These plots depict the results of the **mathematical fitting of models**. The blue line stands for the fitted function, the red line for the growth data in submissions per month.

was 0.054, at a starting value of 142,872, which is pretty close to the first submission count of 142,916 of January 2008. The KL-Divergence between the fit and the downscaled growth data resulted in 3.28.

The Logistic Growth Model

The logistic model (as described in chapter 5.1.2) comes pretty close to the real growth curve. At the end of 2012, almost three million submissions were posted each month, as one can see in the figure. The predicted x_{max} of approximately 9 million submissions (8,882,424.44 to be exact) each month is triple the growth reddit has right now. If this model is correct in any form, reddit will become much larger than it already is. The parameter for the growth rate was optimized to 0.064.

Figure 6.3b displays the fitted function again with a light blue line, and the red line represents the actual growth data. The predicted maximum is far higher than the number of submissions in late 2012, so the specific trait of the model, which is the flattening curve in the third phase, is not noticeable yet. Only the first phase and the early part of the second phase are visible, where the model advances similarly to the exponential one.

The KL-Divergence gives indication of whether the exponential growth model or the logistic one corresponds better to the data. Its calculation for the logistic model to the downscaled growth data equals 12.73, which is considerably larger, and therefore worse, than the KL-Divergence of 3.28 of the exponential model. Thus, the exponential model is a better representation of reddit's submission growth over the course of five years.

The Gompertz Growth Model

The characteristic difference between logistic and Gompertz model is that the Gompertz curve is not symmetric, because it flattens more slowly in the right half. The curve of the test data does not display such a feature, and so it is hardly surprising that it is not possible to find fitting parameters describing the curve in a Gompertz formula.

Summarizing, the exponential model seems to offer the best description for the evolution of reddit[3]. The logistic model is not far off either. Future development might continue in either way, because of the irregularity, that marks the second half of 2012, and that is not educible by any of the models.

6.1.3. Growth of Subreddits

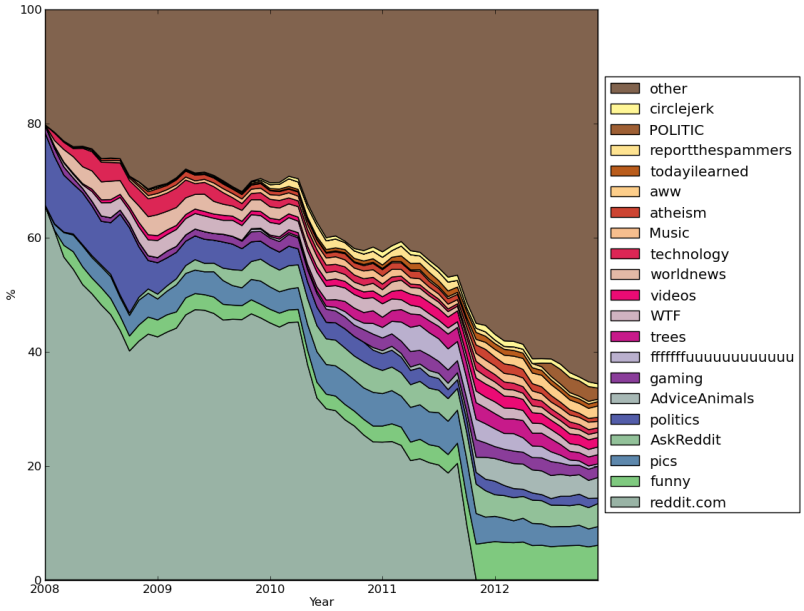
When observing the growth of subreddits in comparison to each other, some very salient features are detectable. Size and growth are again measured in submissions to the subreddit at the time. All in all, there are 125,662 different subreddits. Only 504 of them are of considerable size, with more than 10,000 submissions posted to them. Summed up, these 504 subreddits contain 48,191,547 submissions, 82% of all submissions in the data set. The distribution of submissions to subreddits is far from uniform.

In figure 6.4a, the development of all subreddits is depicted, with their size relative in percent to the overall size of reddit at that moment. The 20 largest subreddits are each represented by an individual color, and submissions to all the other subreddits are summarized in the color brown. In relation to reddit as a whole, the fragmentation into more and more different small subreddits is increasing, as the growing brown part in the figure suggests. At the end of 2012, the smaller subreddits combined contain the majority of submissions to reddit, with more than 60% of all postings.

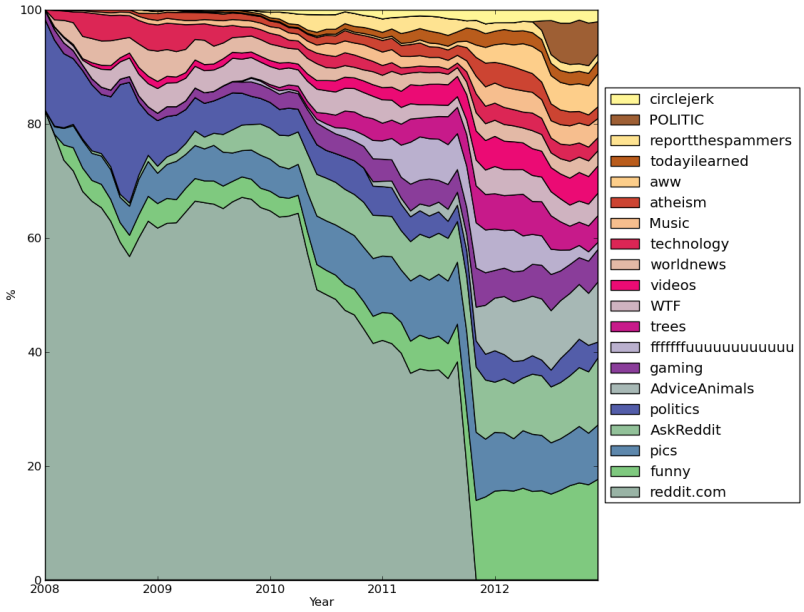
While the largest 20 subreddits contained close to 80% of all submissions in the years of 2008 to 2010, their relative share declined rapidly, and in the end of 2012 they accounted for less than 40% of all submissions. Reddit seems to increase its diversity in terms of subreddits and thematization.

This development is based on the 20 overall largest subreddits. It is even more pronounced if the subreddits are partitioned by the submissions to the 20 largest subreddits per month and opposed by all the other

3 Note that more sophisticated statistical methods for comparing the fits of several candidate distributions exist (e.g. likelihood ratio test). For this application, however, KL-Divergence has been deemed appropriate and sufficient.



(a)The evolution of all subreddits.



(b)A detailed depiction of the evolution of the top 20 subreddits.

Figure 6.4.: **The evolution** of all subreddits in a), where *other* contains all the subreddits not ranked 1 to 20 combined, and a detailed view of the 20 largest subreddits in b) from 2008 to 2012.

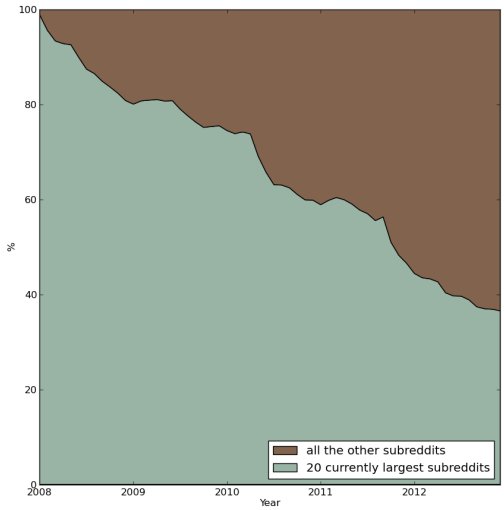


Figure 6.5.: **The unequal distribution of submissions to subreddits.** The green section of the plot contains submissions to the 20 largest subreddits of the respective month instead of the overall top 20.

subreddits, as it is depicted in figure 6.5, which underlines the unequal distribution in every month. Almost every submission has been posted to one of the (at that moment) largest subreddits in the first quarter of 2008. The first two years in this figure show a higher share of the top 20 subreddits, because some of the overall largest had just been founded or were about to be. The second half of the figure, almost exactly from the beginning of 2010 onwards, resembles the shape of figure 6.4a.

The Gini coefficient, a measure for statistical dispersion, results in 0.97 in mid 2008, 0.95 in mid 2010 and 0.94 by the end of 2012. These numbers show that the inequality in the sizes of the subreddits declines slowly and steadily, albeit it is still strongly pronounced.

Another very prominent feature of this plot is the decreasing and in 2011 vanishing part of r/reddit.com. In the beginning of reddit, there were no subreddits, only an all-embracing one: r/reddit.com. All submissions were posted into this single subreddit, and thus the content on reddit was entirely uncategorized.

Even when subreddits were introduced, a great deal of submissions were still posted to r/reddit.com, mostly because back then, there were

no default subreddits except r/reddit.com that a visiting, unregistered user would see initially. With more and more subreddits founded, r/reddit.com shrunk, until in October 2011 a set of 21 default subreddits was introduced, which made r/reddit.com obsolete and it was closed (Martin, 2011b).

Figure 6.4b now excludes the smaller subreddits and concentrates on the largest 20. Most of them are default subreddits. The subreddit r/POLITIC stands out, because it starts as recently as 2012, but escalates immediately to one of the 20 largest subreddits of all time. As mentioned in chapter 5.2.2, r/POLITIC is a subreddit for mirroring submissions from other subreddits, and this is done automatically, thus its spontaneous growth is no surprise. Proportional to the other subreddits, r/politics decreases to a fraction of its earlier relative size. The subreddit r/worldnews, which focuses on news as well, perseveres its allotted share, and r/technology registers noticeable accession only in the last three years. On the other hand, r/funny, r/pics, r/AdviceAnimals and r/fffffffuuuuuuuuuuuu grow enormously. When r/reddit.com ends, these subreddits record the largest surges. Furthermore, the subreddit r/AskReddit emerges to be among the top five, with content that completely consists self-posts, as it is shown later in the content related section of this work (chapter 6.2.2). This suggests that news, informative and text-based content loses more and more ground to image-based, entertaining content and self-posts.

6.2. Analysis of Content

The core of this work aims to shed light on the contents of reddit. What type of content is most prominent? What are the topics? How did it change over time? A quick impression to this subject is given by a word cloud, as featured in figure 6.6. This figure presents the most frequent terms in all submission titles posted to the 20 largest subreddits in 2012. The size of the word depends on the frequency, the larger it is the more often it appears in submission titles. Despite the simplicity of the generation of a word cloud, it gives a clue on several aspects of reddit, that are reviewed using more sophisticated methods in this chapter. For example, the term *reddit* is the most frequently used one

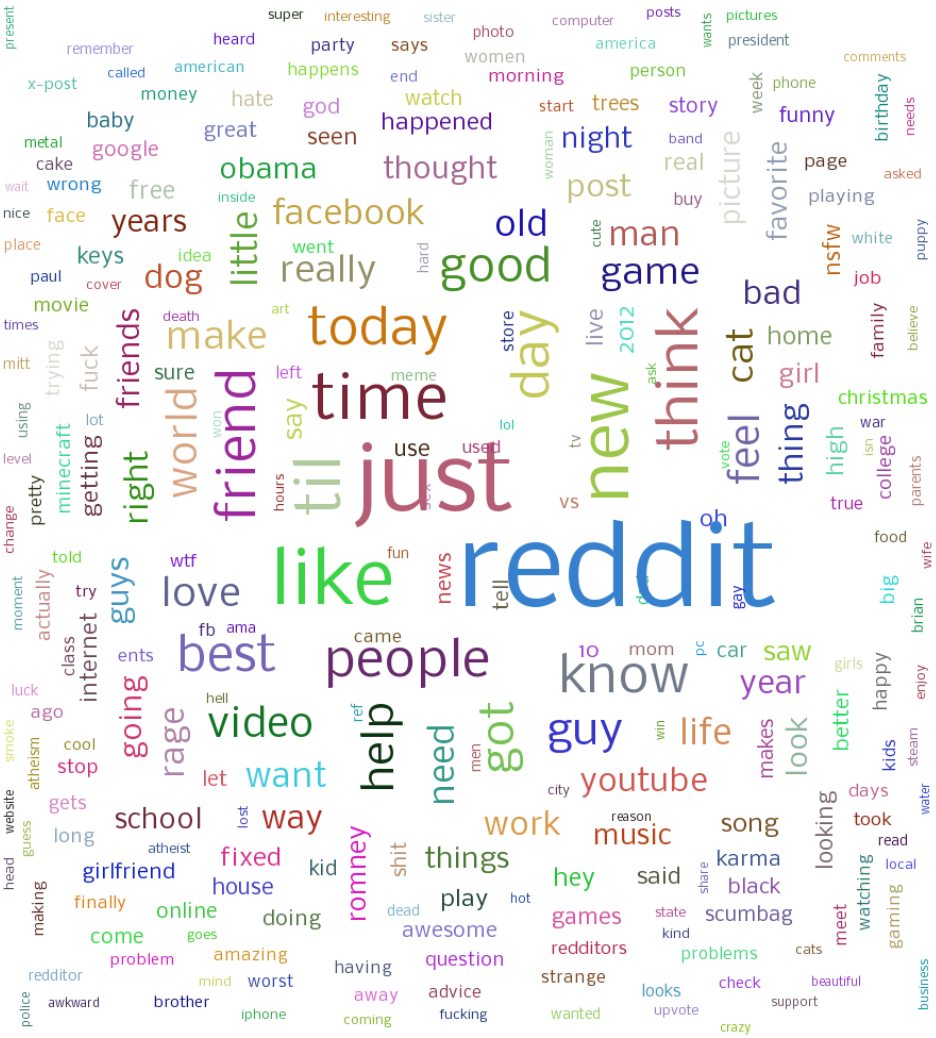


Figure 6.6.: **A wordcloud** of the 280 most frequent words in submission titles in the 20 largest subreddits of 2012. It gives a small impression on the content, vocabulary and attitudes on reddit. The term *reddit* is most frequently used, probably because reddit users address the community with this term and self-referentiality. Titles often express the actuality of submissions using *just* (as in *I just flew 800 miles...*), or downplay the contribution (*Just a painting of my son*). Several terms listed hint that personal stories are of relevance, such as *girlfriend*, *love*, *mom*, *friend*, *family* or *kids*. Reddit is fond of animals, especially *cats* and *dogs*. The term *til* refers to the abbreviation of *today I learned*, used as a prefix in the subreddit r/todayilearned. Users of r/trees call themselves *ents*. The swearword *scumbag* is not so frequently used because of unmannerly behavior of reddit users, but because of a popular internet meme in r/AdviceAnimals

in titles of submissions, thus the largest one in the word cloud, hinting at certain degree of self-referentiality in submissions. The impact of some of the subreddits is also visible, for example is *til*, the abbreviation for *today I learned* that is used in the subreddit r/todayilearned, one of the most common terms, more so than *Obama*, *college* or *American*. The wordcloud demonstrates, that there is background knowledge necessary to understand, why certain terms are so frequently used on reddit, and how they are interconnected. The results presented in this chapter, especially the topic analysis in section 6.2.4, uncover some of these coherences and provide a better understanding.

6.2.1. Domains

All in all, there are 1,841,239 distinguishable domains in the data set. On average, there are 31 submissions posted per domain. The compiled ranking list of the top 100 most submitted domains, which is the basis for the content categorization described in chapter 5.2.1, covers the domains of links in 40,772,856 submissions, which accounts for about 69.25% of all submissions done to reddit. 14,979,707 of these submissions are self-posts, accounting alone for 25.44%.

In figure 6.7, the evolution of the relative proportion of domains in submissions to each other is displayed. The 20 most frequent domains are drawn individually, the domains ranked 21 to 100 are condensed in the brown part, and all the other domains are summarized in grey.

The relative share of self-posts is not very high initially, but receives a boost in 2009 and grows until it covers almost consistently about 30% of the monthly submissions from 2011 onwards. Although the image hosting service Imgur was founded only in 2009, it quickly rose to become the origin of the largest proportion of external submission links. In July 2010, 7.32% of all links were from Imgur, and at the end of 2012 26.6%.

Alan Schaaf, the founder of Imgur, stated that it was his intention to develop an image hosting service specialized to meet reddit's needs[4].

4 http://redd.it/7zlyd

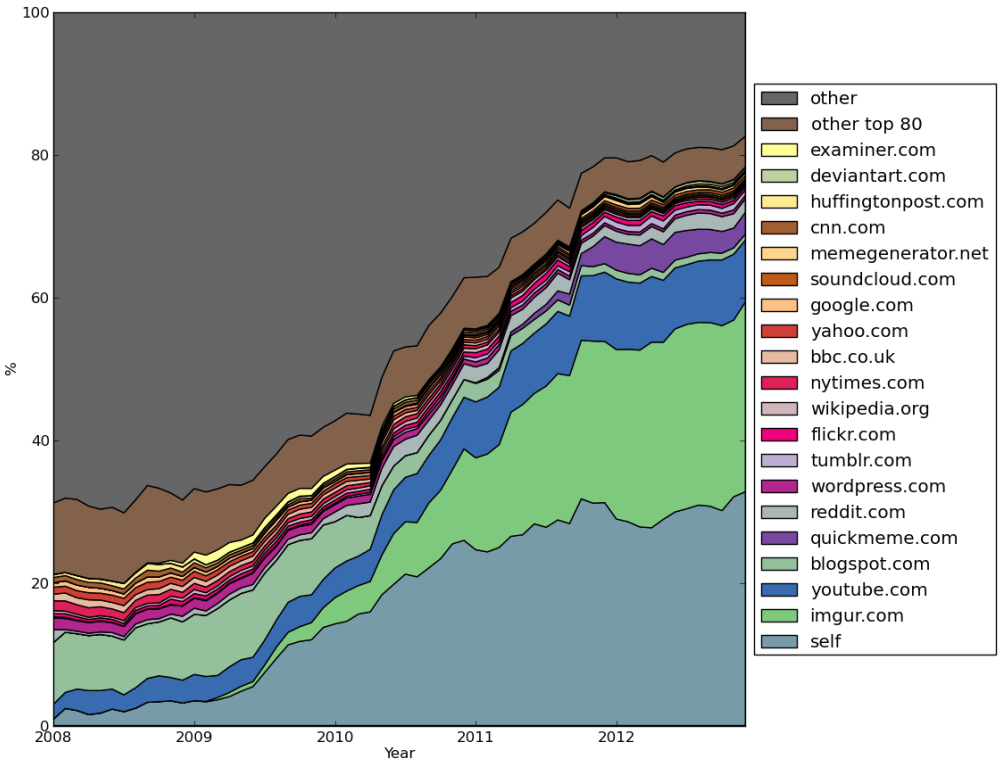


Figure 6.7.: **The distribution and evolution of the domains** of links submitted to reddit. The self-posts form their own domain category to differentiate them from reddit internal links. The section called *other top 80* contains domains ranked 21 to 100, and *other* all remaining domains together. In 2012 self-posts, Imgur and YouTube serve more than 60% of all submissions.

Thus, Imgur forms a third party extension to reddit, and it is well received as such by the community.

YouTube is the second largest external link target by number of submissions. In contrast to Imgur, links from YouTube feature an almost steady increase, with 2.37% in July 2008, 6.24% in 2010 and 8.68% in December 2012. Blogspot (or later Blogger[5]), however, loses much of its significant share of 7.68% in 2008 over 3.03% in 2010 to 0.83% in 2012. Next in the ranking is the image captioning website Quickmeme[6], which was founded in 2011. It enjoys great popularity almost immediately, raising its share of submissions to about 3.05% until December 2012.

To sum up, the most remarkable feature of figure 6.7 is the takeover of the majority of submissions from few distinct websites and consequently the strong decreasing part of the *other*-section of the plot, which outlines the share of all the websites not within the top 100 domains. The absolute number of recorded domains is growing from 34,082 in mid-2008, over 68,577 in mid-2010 to 103,660 at the end of 2012. The relative shares in submissions of the domains not within the top 100, however, declines from 70.19% in July 2008 to 46.95% in July 2010, and finally descents to only 17.42% in December 2012. Thus, the diversity of domains increases in terms of absolute number of domains, but the diversity in submissions originating from them (and therefore also the user's perceived diversity) declines decidedly for the benefit of a handful of domains. This phenomenon is reflected by the Gini coefficient, which increases significantly from 0.78 in July 2008, over 0.83 in July 2010 to 0.95 in December 2012.

6.2.2. Categorization

The manual categorization concentrates on self-posts and the top 100 domains, which represent almost 70% of all submissions to reddit. As described in chapter 5.2.1, these domains are segmented into the categories

self, containing solely self-posts,

5 https://blogger.com
6 http://quickmeme.com

text, for domains that primarily deliver textual content such as news, blogs, articles, papers,
image, for domains that are mainly used to link images (e.g. Imgur),
video, for video streaming services such as YouTube or Vimeo,
audio, for audio platforms like SoundCloud, and
misc, containing miscellaneous domains of link shorteners, website toolkit and hosting services.

Submissions are consequently categorized by reference to their domain, and statistics on the composition of content on reddit as a whole are presented, followed by an investigation of the development over time and a comparison of the compositions of several subreddits.

Categorization of Submissions

A breakdown of all submissions to reddit by categories of their domains delivers an overview of the content submitted to reddit. Out of all submissions that contain a categorized domain or self-post, there are 37.1% image submissions, 36.7% self-posts, 12.8% are text based submissions, 11.8% are categorized as video and the minorities are miscellaneous with 1.1% and audio with only 0.5% (figure 6.8). In this static point of view, images and self-posts balance each other out. Less than half as many submissions are categorized to be text, which includes news websites, or video. For this reason alone, it is easily comprehensible that the label *social news aggregator* is no longer a very accurate one for reddit.

Adding time as a further dimension, the progression, as displayed in figure 6.9, highlights that the static statistic of figure 6.8 is not very representative for reddit, because the composition of content on the web portal is highly dependent on the time period of observation or consumption.

Self-posts have not always been the predominant type of submissions. From 2008 to mid 2009, the majority of submissions are text-based. Consulting the growth analysis of subreddits in figure 6.4, one can see that for the same period of time r/politics, a subreddit where news articles are very common, has its highest relative share of reddit as well.

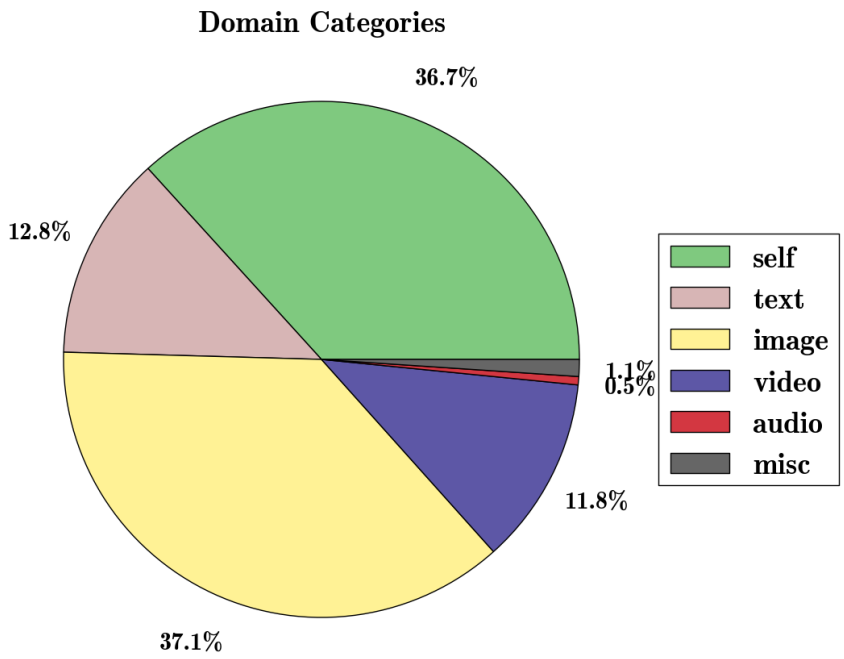


Figure 6.8.: This pie-chart displays the **ratio of the six categories of domains**: self, text, image, video, audio and misc. It is based on all submissions with categorized domains. In each case, the shares of two categories almost balance each other out. Images and self-posts are most prominent, both with shares of about 37%. Textual and video content include $12 - 13\%$ respectively. The categories of miscellaneous and audio content come in last with noticeable small shares.

The definition *social news aggregator* had been definitely eligible at that time.

The subreddit growth analysis in figure 6.4 also reveals that the large self-post-oriented subreddit r/AskReddit initially appears in 2009, but gains participation quickly. This is reflected by the boost of the self category in 2009 in figure 6.9. From 2010 onward, the share of self-posts stabilized between 35% and 45%, only interrupted by a short decline at the turn of the year 2011 to 2012.

Image-based submissions are always rather popular. For the first year, about a third of all submissions are images. In 2009, there is a considerable decline clearly recognizable, probably because self-posts come into fashion at this time. The image category recovers itself in 2010, rises well above the 40% mark and settles at about 37% in 2012.

In contrast to the ups and downs of self, text and image, video-based submissions deliver an almost constant share of close to 10% throughout the data set, with a slightly growing trend. Although video media is more modern, more sophisticated and offers comprehensive and multidimensional information transport, and despite the immense success and popularity of video streaming websites, such as YouTube and Vimeo, images are still far more widespread and utilized on reddit.

Miscellaneous domains, containing template website hosting services and link shorteners, have paled into insignificance over time. There are only two domains in the top 100 that are categorized as audio: SoundCloud[7], founded in 2007, and Bandcamp[8], founded in 2008. The audio dedicated websites grow to significance by 2010, and slowly acquired about 1% until December 2012. There are many music-oriented subreddits, and one of them, r/Music, is one of the largest 20 subreddits. Still, audio-based submissions are few and far in between.

Content Composition and Development within Subreddits

Further insights can be gathered if another relevant dimension is added: subreddits. Reddit is a fractal, and a user compiles a personal front

7 https://soundcloud.com
8 http://bandcamp.com

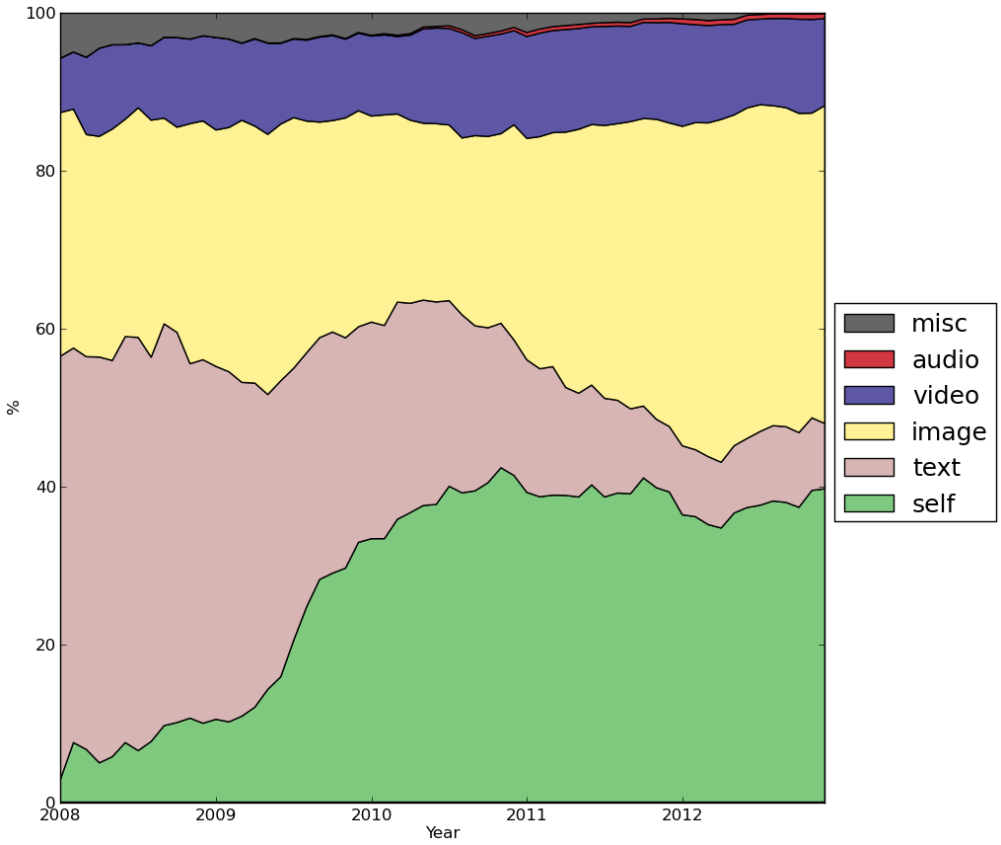


Figure 6.9.: **The development of the six categories** of submissions self, text, image, video, audio and misc over time clearly differentiates itself from the static plot in figure 6.8. Self-posts have not always been predominant. In 2008, the majority of submissions are categorized as text. The image section displays volatility and a decline in 2009, before it rises its share again in 2011 and 2012. Video, on the other hand, features a slow and steady gain, and miscellaneous a continuous decline.

page by subscribing to or unsubscribing from subreddits. So, while the overall evolution of reddit's content might be interesting and revealing, it is not what a user consumes and sees when visiting the website.

There are too many possible combinations of subreddits to analyze them all, but to understand how different subreddits and their content can appear, the same approach is applied to the largest 20 of them.

Figure 6.10a contains the structure of r/reddit.com. It resembles the composition of reddit as a whole in figure 6.9, featuring similar increase in self-posts, decrease in text submissions, while the shares of image, video and misc remain almost constant. Until its shutdown in 2011, this subreddit formed a cross section of reddit, because of its size and the lack of thematic scope.

Subreddit r/funny impersonates the triumphal march of the image submissions in figure 6.10b. While its content was almost balanced out between text, image and video (and even self for some time) in the years of 2008 to 2010, image submissions became rampant from 2011 on.

The subreddit r/pics (figure 6.10c) delivers what one would expect. Although in the first three years there is still a considerable amount of text submissions, probably before the rules were adapted and prohibited non-image submissions.

Where r/pics is a paragon for image focused subreddits, r/AskReddit is one for self-posts. Figure 6.10d proves that there is nothing but self categorized submissions, which is also ensured by the rules of the subreddit.

Since all news media and network websites are categorized as text, one would assume that the subreddit r/politics is a text-only channel (figure 6.10e). However, this is not entirely the case. The majority of submissions is text based indeed, but there are considerable shares of video, image and self submissions still in place. The sharp drop of self-posts in 2011 indicates a temporary change of rules (or entire deactivation of self-posts), which has been revoked shortly after.

In figure 6.10f the content composition of r/gaming is depicted. A few trends are noticeable, such as the steady increase in video submissions and the continuous decrease in text submissions. Image submissions have had a turbulent development in the first three years. Towards the

end of the observed time frame image submissions become more and more prominent. The share of self-posts peaks at more than 50% in 2009, but declines after 2010 for the benefit of image submissions.

Video streaming services, such as YouTube and Vimeo, are unsurprisingly the most common category in r/videos (figure 6.10g). News networks not only feature articles on their websites, but often news reports and documentary video clips as well. For that reason, the category text has a moderate, but shrinking, share of submissions in r/videos.

An interesting case is r/music in figure 6.10h. Although music is the obvious theme of the subreddit, the domains are mainly categorized as video. Only few submissions originate from audio domains. It appears that music is shared in this subreddit via YouTube links most of the time, containing music videos or simply showing the album cover or images as video during the song. Sometimes, articles about music (release dates, news and gossip) are also posted, which form the small share of the text category. More frequent than that are self-posts, often to ask the community for opinions or recommendations.

In r/technology posts are expected to contain updates and news on technology, preferably from appropriate technology focused news websites. The figure 6.10i brings out that, while the majority are text submissions indeed, the focus has not been enforced until the end of 2011. Moderators have changed the rules, disabled self-posts and forbid URL shorteners. Later, in 2013, video and image submissions have been banned as well.

The remaining 11 subreddit categorizations of the 20 largest subreddits can be found in appendix B.

Recapitulating, one can see that by adding dimensions, and concentrating on one subreddit at a time, reddit's content structure and development look differently. In an analogous manner, users perceive reddit's content differently, since registered users modify their front page with their choice of subreddits. A feature that most subreddits have in common and that is slightly visible in the overall plot in figure 6.9, is that the proportion of self-posts has risen slowly before reaching its zenith of popularity from 2009 to 2011, and decreasing again in 2012. This is noticeable with varying degrees of distinctness even in subreddits specifically dedicated to other categories, such as r/videos, r/pics or r/Music.

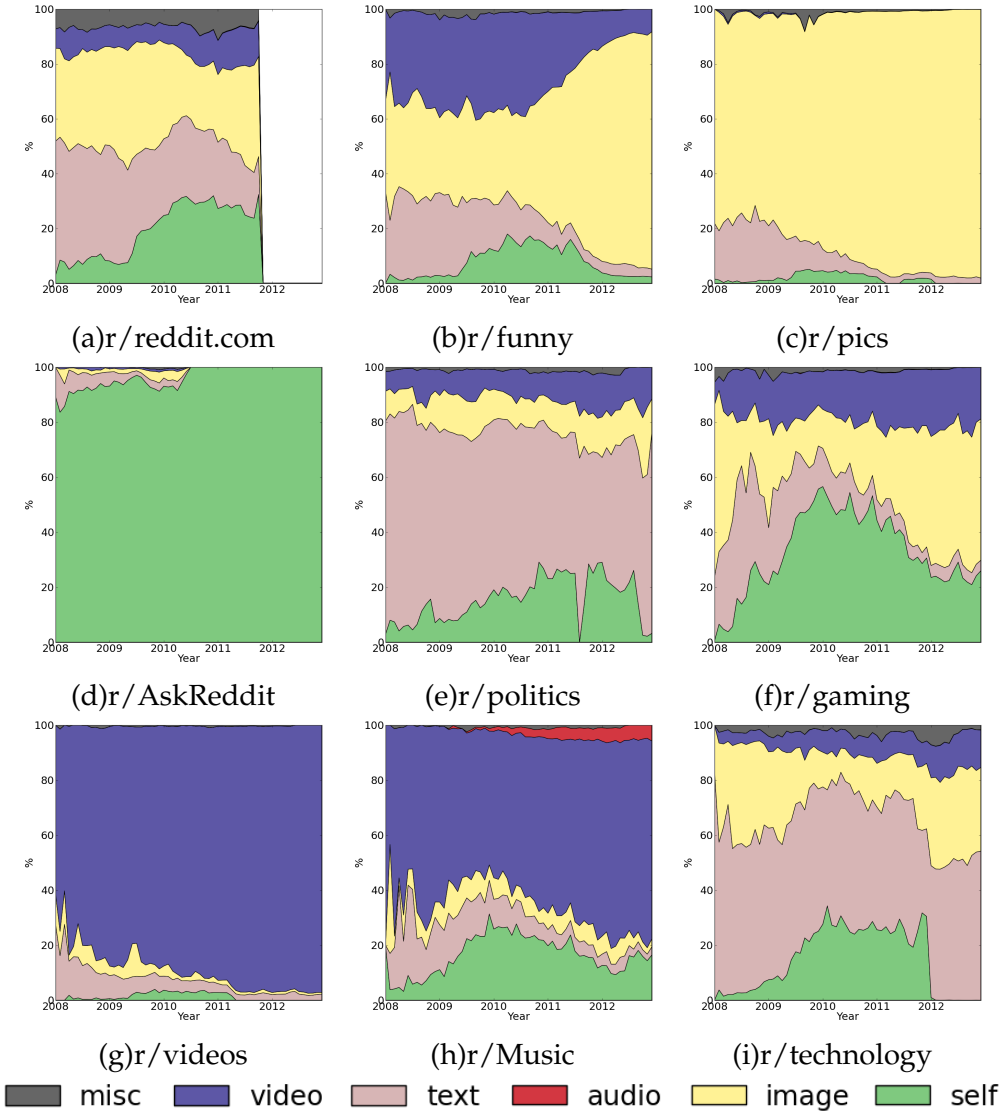


Figure 6.10.: **The development of the six categories** self, text, image, video, audio and misc over time in the scope of several subreddits. All these subreddits evolved differently, because of different thematic focuses and different sets of rules.

6.2.3. Moderation

The extent of moderation on reddit is difficult to measure. Each subreddit has different rules, different moderators, different numbers of them, and they interfere in different ways with the submissions. Moderators control what stays on reddit. Their tools are user banning and submission deletion. One would have to record all submissions the moment they are posted, and try to load them again later to see if they have been removed. In this chapter, an investigation of the extent of moderation is presented using the example of r/POLITIC.

The subreddit r/POLITIC offers this look into the past, at least for some political subreddits. A computer program reposts all submissions from those subreddits immediately to r/POLITIC. The submissions in r/POLITIC can now inversely be tracked down to their origin. If a submission no longer exists there, it has been deleted. A comparison of the relative deletion rates of domains and terms outlines whether certain items are noticeably deleted more often than others, and therefore indicate a certain bias of the subreddits mirrored in r/POLITIC.

Since r/POLITIC is a very recent development on reddit, and has taken up its work in a consistent manner only in June 2012[9], the statistics in this chapter are limited to submissions from June 2012 to December 2012.

Deleted Domains

Table 6.1 lists the most frequently deleted domains, their number of deletions and the ratio of deleted submissions to all submissions with the given domain, in the time where r/POLITIC and the u/PoliticBot have been active (which is about seven months in the data set). The domain *reddit.com* also contains self-posts, because from r/POLITIC's point of view these are reddit internal links, since the u/PoliticBot submits a link to the original self-post rather than copying its body and reconstructing it as a self-post in r/POLITIC.

In absolute numbers, *reddit.com* is in the lead with 13,975 deleted submissions. However, the percentage of deleted submissions in relation to

9 See figure 6.4 and, in more detail, figure B.1j.

all mirrored submissions from this domain suits better for comparison. In this perspective, about 26.1% of *reddit.com* submissions have been deleted. The highest ratios of deletions have *nationalmemo.com* with remarkable 44.5%, *quickmeme.com* with 31.7% and *imgur.com* with 30.3%. Submissions from websites which are not dedicated to news (either networks or newspapers), such as Imgur, Facebook, YouTube or reddit itself, aggregate higher rates of deletions. The domains *thinkprogress.org*, *bloomberg.com* and *alternet.com* are, in this context, least likely to be deleted.

Deleted Words

While news dedicated websites might have a bias towards one political orientation or the other, a clearer insight in the question, if political subreddits suppress certain topics, can be given when investigating the titles. Similar to the approach with the domains, table 6.2 lists a statistic of 75 terms in deleted submissions, ranked by the absolute number of occurrences.

In 2012, the United States of America held presidential elections between the officiating president Barack Obama and his contender Mitt Romney. With presidential campaigns running, various debates and electoral analyses in the forefront and results and discussions in the aftermath of the election, this was a major political topic in 2012. Both Obama and Romney are mentioned very often, and also frequently deleted, leading the table by far with 6,938 and 6,587 deletions. Compared to the submissions mentioning both candidates, they are removed at an almost equal ratio, with 18.3% for *obama* and 18.2% for *romney*, and close to the ones of *election* (18.5%), *president* (17.9%) or *presidential* (19.2%). A similar percentage of 17.4% (at 1,613 deletions) is held by another presidential candidate, Ron *Paul*.

It is worth mentioning that the terms *republican* and *republicans*, naming the *Republican Party* of the United States, are among the 75 most frequently deleted terms, yet their opposing party, the *Democratic Party* of the United States, is not. It is exceptionally unpopular to mention *facebook* (with 26.5% removed), *youtube* (deleted in 27.4% of all times), *shooting* (with 27.5% submissions gone missing), or *reddit* (which is removed in 31.1% of all occurrences).

Table 6.1.: **The most frequently deleted domains** of the subreddits mirrored in r/POLITIC. The absolute number of deletions of a domain (#) is given as well as the percentage of deletions in relation to all submissions with the domain (%). The domain *reddit.com* contains all internal links, including self-posts.

Domain	**#**	**%**	**Domain**	**#**	**%**
reddit.com	13975	26.1	youtube.com	7790	25.8
imgur.com	5680	30.3	nytimes.com	1703	14.9
huffingtonpost.com	1590	16.4	cnn.com	1439	20.2
bbc.co.uk	1117	21.7	yahoo.com	1093	16.7
washingtonpost.com	994	14.4	reuters.com	846	15.8
guardian.co.uk	816	16.2	rawstory.com	601	17.1
blogspot.com	553	14.0	go.com	528	20.4
dailymail.co.uk	505	27.2	foxnews.com	498	19.4
quickmeme.com	495	31.7	google.com	435	24.4
politico.com	433	14.9	cbc.ca	401	19.3
dailykos.com	391	12.5	rt.com	377	20.4
msn.com	368	19.2	wordpress.com	365	12.4
aljazeera.com	361	16.2	nbcnews.com	351	17.4
telegraph.co.uk	335	16.9	latimes.com	330	14.2
cbsnews.com	330	15.7	npr.org	327	15.1
businessinsider.com	323	17.1	wsj.com	320	16.7
theglobeandmail.com	307	20.7	alternet.org	306	10.9
usatoday.com	292	19.4	wikipedia.org	284	26.8
nationalpost.com	281	21.2	nationalmemo.com	260	44.5
thinkprogress.org	258	7.1	nydailynews.com	257	26.6
facebook.com	255	27.6	twitter.com	245	26.0
ap.org	243	15.3	salon.com	236	12.0
talkingpointsmemo.com	221	10.2	washingtontimes.com	213	16.2
thestar.com	208	24.2	time.com	203	20.0
thehill.com	183	12.9	bloomberg.com	180	9.4

Table 6.2.: **The most frequently deleted words** of the subreddits mirrored in r/POLITIC. The absolute number of deletions of a word (#) is given as well as the percentage of deletions in relation to all submissions containing the word in its title (%).

Word	**#**	**%**	**Word**	**#**	**%**	**Word**	**#**	**%**
obama	6938	18.3	romney	6587	18.2	news	2267	20.9
mitt	2261	18.4	people	2246	20.9	president	1879	17.9
election	1719	18.5	paul	1613	17.4	vote	1587	20.2
video	1448	20.0	state	1410	16.9	debate	1406	20.2
government	1372	17.2	police	1319	17.7	post	1312	21.6
2012	1276	19.4	tax	1270	14.8	year	1265	18.9
party	1253	17.5	america	1252	19.5	right	1244	19.4
youtube	1243	27.4	world	1218	18.5	time	1214	19.2
reddit	1162	31.1	years	1159	19.0	ryan	1137	17.2
republican	1122	16.7	politics	1108	23.8	american	1102	18.4
political	1069	19.5	know	1057	24.1	war	1053	16.8
campaign	1037	16.5	canada	1021	24.4	gop	1019	14.0
help	1011	23.5	shooting	975	27.5	day	935	19.5
presidential	916	19.2	israel	913	17.4	republicans	910	15.7
anti	898	17.5	women	897	19.0	states	832	18.0
house	822	15.4	gun	821	18.4	school	802	22.6
attack	760	17.7	court	753	14.9	money	744	17.8
men	740	23.8	woman	739	21.5	support	725	20.5
voting	725	18.5	law	718	15.8	iran	715	16.5
speech	708	20.3	facebook	704	26.5	fox	698	18.8
rights	689	18.4	white	683	17.7	national	675	17.9
ron	671	18.1	media	668	16.9	dead	662	23.1
killed	661	20.6	country	656	19.0	million	655	16.1
free	630	17.1	americans	627	16.3	china	623	15.1
gay	622	21.1	way	620	19.8	real	616	19.5

6.2.4. Topics and Short Term Trends

Due to the simplicity of LSI based topic models, their results are functionally adequate, yet, compared to those of LDA based topic modeling, not very satisfying, because they overrated irrelevant terms (such as verbs with little specific meaning, e.g. *come*) as a symptom of the bag-of-words basis of this method. For this reason, the following sections concentrate on the results yielded by the LDA calculations.

The following chapter presents the results of two applications of the LDA topic model: First, topics, that have been discovered in a text corpus containing all submissions in 2012, separated per subreddit, are demonstrated and described in chapter 6.2.4. The limitation of the time frame, as mentioned before, is applied for reasons of clarity and comprehensibility, and it facilitates the traceability of events. Second, topics from monthly text corpora of submissions in 2012 are identified, and the yields are processed to find salient differences in the terms the topics are made of. These differences are investigated in chapter 6.2.4, and outstanding ones are investigated and explained as trending terms at the time.

Topics of Reddit

The topics of the 20 largest subreddits in 2012 are listed in its entirety in appendix C.

The subreddit r/AdviceAnimals is about image macros called memes, pictures with exchangeable captions. These memes typically have names that are mentioned in the titles of the submissions. There is also a website called Know Your Meme[10], which identifies and lists memes, their history and various appearances. This gives r/AdviceAnimals ideal preconditions for extraction and verification of topics. Table C.1 includes many memes, but only two of them are almost in every topic: *Scumbag Steve*[11], addressing unethical behavior, and *Good Guy Greg*[12] (or *GGG*), which is basically the opposite of *Scumbag Steve*. Both memes date back

10 http://knowyourmeme.com
11 http://knowyourmeme.com/memes/scumbag-steve
12 http://knowyourmeme.com/memes/good-guy-greg

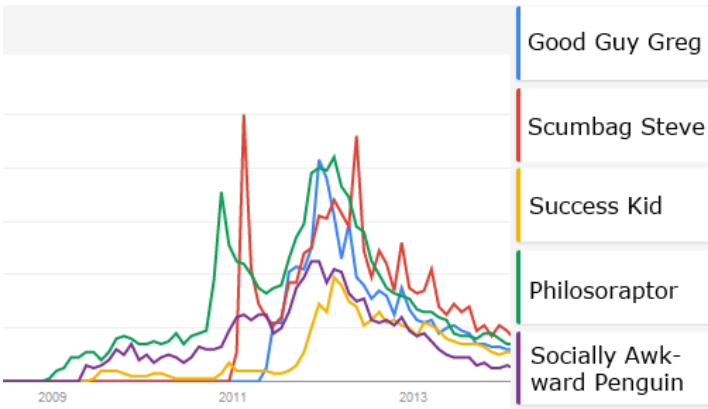


Figure 6.11.: **The Google Trends analysis** of the mentioned memes in the topics of r/AdviceAnimals, namely *Good Guy Greg*, *Scumbag Steve*, *Success Kid*, *Philosoraptor* and *Socially Awkward Penguin*.

to 2011 and accumulated great popularity. Other memes among the 20 topics of r/AdviceAnimals are, for example,

- Socially Awkward Penguin[13],
- Success Kid[14],
- Philosoraptor[15],

among many others. Consulting the relative search interest by Google Trends, as depicted in figure 6.11, one can see that the attention these memes received was not limited to reddit.

The subreddit r/AskReddit (table C.3) is self-post focused, aiming to establish discussions with the community. Titles usually address the community, which is why every single topic involves the term *reddit*. Most topics revolve around the term *help*, often combined with a term that hints at what the users asked help for, such as *life*, *sex* or *SOPA*, the Stop Online Piracy Act, which many websites rallied against in 2012, including Wikipedia and Google (Wortham, 2012).

The topics of r/aww in table C.4 point out what reddit users find adorable. The most common terms are *cat*, *dog*, minimizations of said species, and *baby*.

13 http://knowyourmeme.com/memes/socially-awkward-penguin
14 http://knowyourmeme.com/memes/success-kid-i-hate-sandcastles
15 http://knowyourmeme.com/memes/philosoraptor

With exception of the discontinued r/reddit.com, the subreddit r/funny is the largest in number of submissions. Humorous submissions in r/funny revolve primarily around *facebook posts*, adult content (*nsfw*, *sex*, *porn*), *girls* and *girlfriends*. Also appearing are memes with terms like *good*, *guy*, *scumbag*, *world* and *problems* (from the meme *First World Problems*[16]), although memes have been banned from r/funny later. Screenshots from facebook or other social networks have been banned as well in 2013.

It seems natural that r/gaming discusses new, popular and successful computer games and video game consoles. However, as presented in the topics in table C.8, only the games *The Elder Scrolls - Skyrim*, *Super Mario*, *Minecraft*, *Star Wars - The Old Republic* (*swtor*) and *online poker* are identified in topics. Especially *Skyrim* seems to be exceedingly popular on reddit, because it is featured in 14 out of 20 topics. The digital distribution software *Steam*[17] experiences a lot of attention as well, since it is part of 12 topics. It bears mentioning that, from all the various video game consoles on the market in 2012 (Sony Playtation, Nintendo Wii,...), only Microsoft's *Xbox (360)* is among the topics of r/gaming. Also, the Stop Online Piracy Act (*SOPA*) turns up again in this subreddit. Manual investigation has shown that submissions to r/gaming only need to be related to games in any way, thus memes, jokes and entertaining content are frequent. More serious submissions and content about computer games with the purpose of informing or initiating discussions can be found in r/games.

League of Legends, developed and published by the company *Riot*, is a very competitive online multiplayer game that, according to Gaudiosi (2012) from Forbes, is

> "officially the most played PC game in North America and Europe."

Unsurprisingly, the corresponding subreddit r/leagueoflegends is one of the largest subreddits in 2012. The topics of this subreddit, listed in table C.9, revolve around the aliases of famous players and teams (*Dyrus*, *Saintvicious*, *TSM*, *M5*, *CLG*) and game specific terms (such as *elo*, *champion* or *solo queue*). There are topics where users are discussing

16 http://knowyourmeme.com/memes/first-world-problems
17 http://store.steampowered.com

tournaments (*kiev*, *IEM*) in combination with players or teams, the *streaming* of matches, game mechanics (*jungle* or *support champion*) and the ranking system.

Another subreddit that is dedicated to a single computer game is r/Minecraft. Minecraft is an independent computer game that has been created by Markus Persson. In contrast to League of Legends, this game's mechanics are oriented towards creativity, constructiveness and building. The users of Minecraft write about *new servers*, modifications of the game (*mod*) and game specific terms (*redstone*, *word*, *map*, *mob*,...). When overlooking the topics in table C.10, it is interesting that, although the theme of the subreddit is clear due to its title and address, users seem to consistently refer to the name of the game (and the subreddit) over and over again. The term *Minecraft* is in every single topic, and in most topics even ranked first. In r/leagueoflegends, the game and subreddit name is mentioned quite frequently too using the abbreviation *lol*, yet not to that extent.

The music oriented subreddit r/Music usually links to music on video streaming websites, as depicted in figure 6.10g. The titles mirror this convention, and *video* is in almost every topic in table C.11. Otherwise, the topics of r/Music meet typical expectations, because all of them revolve around *live music*, *bands*, *guitars*, *covers*, *remixes*, *lyrics* and expressions of fondness.

Due to the unfettered theme of r/pics, the topics in table C.12 are not very informative. The topics mainly contain personal and mundane terms, and only few drop a clue on the possible motives of the images that have been titled with them. The terms *birthday*, *new year* and *cake* hint that special days are the cause of many submissions. Again, *cat* and *kitty* are appearing, and also *sopa* is in a topic, although the other terms do not fit to it (*topic 14: like, just, cat, reddit, today, picture, sopa, oh, awesome, year*).

The American political scene of the year 2012, according to r/politics, must have been quite monotonous (table C.13). The presidential election, evaluations, debates and the campaigns of the candidates dominated this subreddit. Many key personalities are repeatedly featured in the topics, namely Barack Obama, the officiating president of the United States of America, and the challengers Mitt Romney, Rick Santorum, Newt Gingrich and Ron Paul from the Republican Party (or *GOP*, Grand

Old Party). Along these names are the controversial subjects, that have been debated and important for reddit. The most prominent issue, which even radiated into other subreddits, is the Stop Online Piracy Act, and a similar bill called *PIPA*, the Preventing Real Online Threats to Economic Creativity and Theft of Intellectual Property Act. These bills aim to regulate internet usage, which is why they are of certain concern to the reddit community.

SOPA and *PIPA* are in the limelight of r/technology as well (table C.14). Aside from the political turmoil, this subreddit features submissions about *mobiles*, in various combinations with *Google* and *Apple*, *Android* and *IPhone*, about *news*, (*web*) *design*, *Facebook*, *Megaupload* and *Wikipedia*. Recent development in software and hardware are both discussed, services and trade shows (such as the *CES*, the Consumer Electrics Show) are presented. Technical or engineering details are not subject of this subreddit.

It seems like the submissions to r/videos are almost always *awesome* or *amazing* (or similar wooing descriptions), and contain the things reddit is fond of in other subreddits as well, such as *cats* and *dogs*, *girls* and *guitars* (table C.18). Videos might be *NSFW*, *commercials* or *trailers*. Although there is a subreddit for music (which also posts primarily videos), r/videos has music related submissions. The vocabulary in the topics suggest that titles often request the community to watch it. Following the trend of other subreddits, *SOPA* is also part of a topic in this one.

The focus of r/worldnews lies at major news from outside of the United States of America. However, *SOPA* and *PIPA* turn up in the topics list of this subreddit, which can be found in table C.19 in the appendix, as well, along with *ACTA*, the Anti-Counterfeiting Trade Agreement, which expands some of the ideas of the other two American acts to a global scale as a multinational treaty. Another large story on r/worldnews is *Iran*'s *nuclear* program, which has developed the dimensions of an international crisis in 2012 (CNN, 2012). The *Syrian* civil *war*, which escalated in 2012, forms multiple topics as well. The reason why *Megaupload*, a file hosting service, is part of topics in r/technology and r/worldnews, is that the website has been shut down in the United States of America in 2012, and legal charges against its founder have been laid (Kravets, 2012).

All in all, the calculated topics are well replicable and unveil the most frequent, and probably most important, topics for reddit users in 2012, such as the repetitively, subreddit-independently reoccurring *SOPA* or *PIPA*. Reddit seems to be highly interested in politics, at least when it comes to Internet regulation bills. Furthermore, each of the subreddits' themes produces distinct topics, which in turn reflect the theme and perfectly give away the essence of the subreddit. Every topic is closely related to either the theme of the subreddit, as determined by the subreddit's title, or to the characteristic of the subreddit in terms of customary dealings, behavior and verbalisms. Consequently, natural language processing with topic models proves to be a suitable and effective method to capture the essential ongoings in subreddits and to deepen one's knowledge about thematic content on reddit.

The topic modeling approach in such a long time frame serves the purpose of description and explanation of content in subreddits very well, and enhances the comprehension of reddit, which is the goal of this work. Besides the descriptive factor, the topics hardly support a contention for or against reddit being a suitable social news aggregation website, because the extent of the time frame concedes no assertion upon relevance to current events. However, some events, which dominated the headlines of newspapers in 2012, are prominent in the topics as well and provide at least a point for news aggregation with regard to thematization.

Short Term Trends of Topics

In this chapter, short term trends within topics are observed by generating 20 topics with 20 terms each per subreddit on a monthly basis in the year 2012, and calculating the differences to the previous month. In the following figures, the trend scores of the terms are plotted as lines. The 19 most prominent terms are presented by name, while the others are summarized in thin lines, which are labeled as **other* in the legend.

Figure 6.12a represents the monthly trending terms of r/AdviceAnimals. The higher the peak is, the larger is the difference of occurrences of the term between the current and the previous month. This figure highlights that, for example, in March 2012 the term *kony* has trended.

Kony 2012[18] is an online campaign, triggered by a YouTube video from a humanitarian organization, to raise awareness about Joseph Kony, leader of an Ugandan guerilla group. In June, a parody song became viral on YouTube and gave birth to the meme *Overly Attached Girlfriend*[19], which became popular on reddit as well. With the presidential campaign running and the re-election in November 2012, a photograph of President Barack Obama summoned the *Upvoting Obama*[20] meme, which causes a peak in August. Other notable, identifiable memes in this figure are *Bad Luck Brian*, *Scumbag Steve* and *Actual/Malicious Advice Mallard*.

The topic trends of r/AskReddit in figure 6.12b are very balanced, the terms underwhelming. Users ask advice for *college* in August, since the U.S. college year usually starts in late August or early September, and about *Halloween* combined with *costume* in October, in preparation to 31 October.

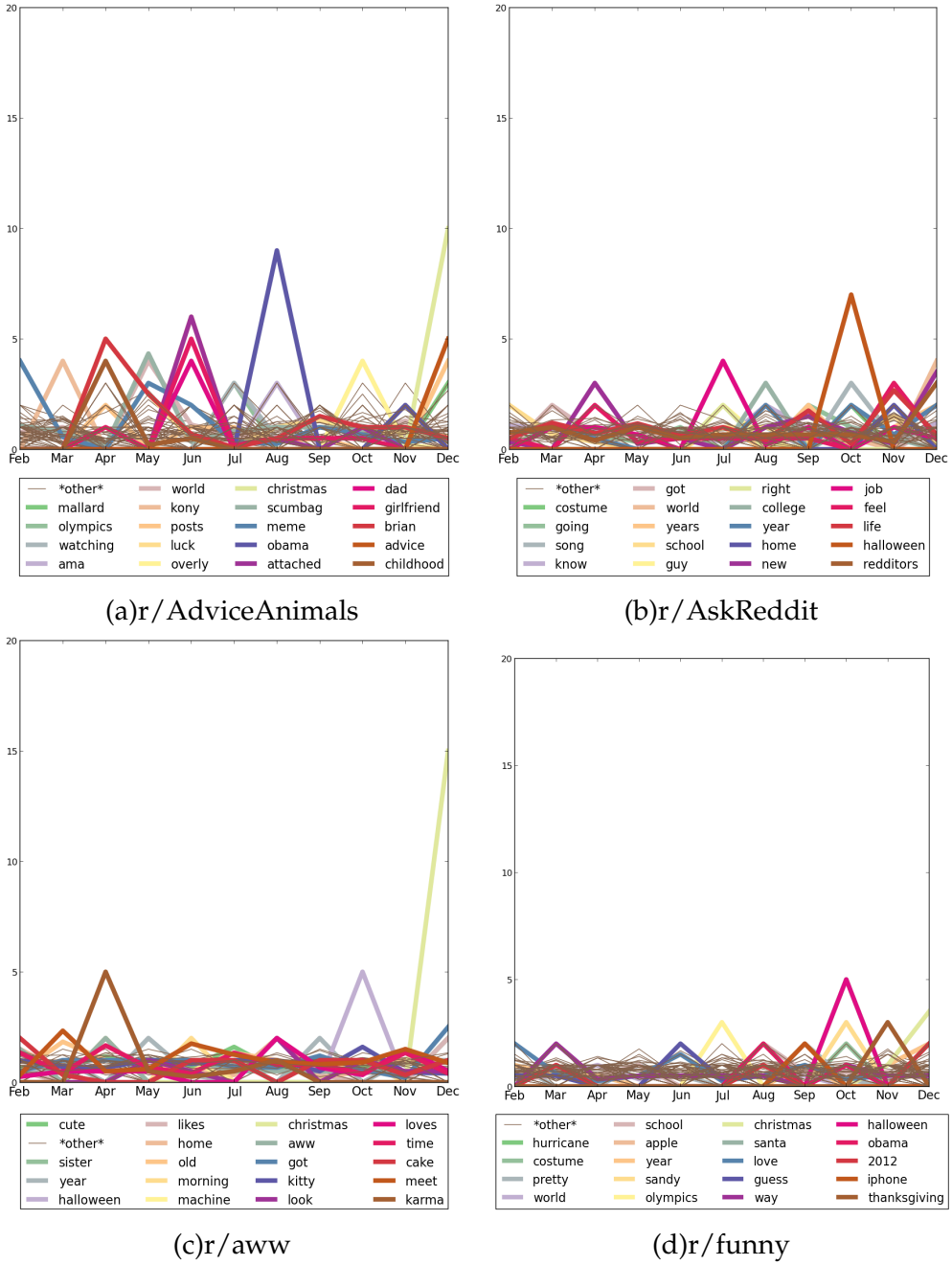
The subreddit r/aww in figure 6.12c features similar steadiness, which seems natural, because endearing images of cats and dogs are seldom exposed to trends. Yet, there are three striking peaks:

- *Karma* in April, due to the description *karma machine* for subjects in images on r/aww,
- *Halloween* in October, and
- *Christmas*, enormously trending, in December.

Figure 6.12d displays that the trends in r/funny are equally related to seasons, holidays and events. The 2012 Summer *Olympics* in London in July are reflected as well as *Halloween*, the *Hurricane Sandy*, that ravaged great parts of America in October, *Thanksgiving* in November and *Christmas* in December.

The computer games related subreddits r/gaming, r/leagueoflegends and r/Minecraft are far more volatile. Figure 6.13a displays enormous spikes, and each of these spikes stands for a recent development or announcement in the industry. The first one is *Mass Effect 3*, a game released in March, followed by a smaller trend from the *Assassin's Creed* series, which surfaces again in October, when the third part was released. *Diablo III*, developed by Blizzard Entertainment and released in May,

18 http://knowyourmeme.com/memes/events/kony-2012
19 http://knowyourmeme.com/memes/overly-attached-girlfriend
20 http://knowyourmeme.com/memes/upvoting-obama



(a)r/AdviceAnimals (b)r/AskReddit

(c)r/aww (d)r/funny

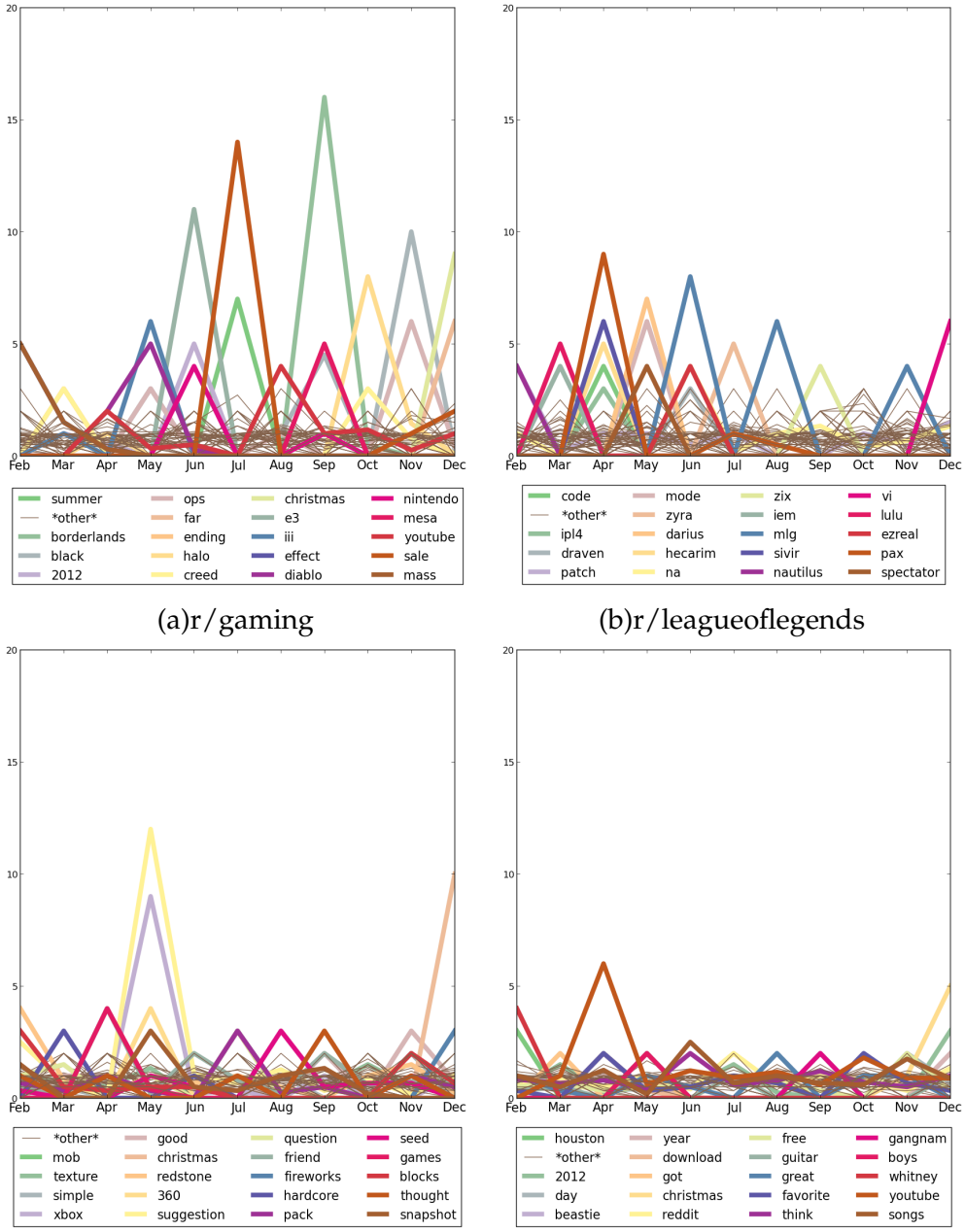
Figure 6.12.: The trends of r/AdviceAnimals, r/AskReddit, r/aww and r/funny

causes a similar hype. The Electronic Entertainment Expo, abbreviated as *E3*, is an exclusive, industry-only event, where upcoming games and gaming related technologies are presented. News from the *E3* are understandably enough very relevant to and welcome at r/gaming. However, even this event is surpassed by the *Steam Summer Sales*, an annual sale event on the digital distribution platform Steam, and most of all the release of the game *Borderlands 2*. Further notable trends are fuelled by the game-releases of *Halo: Combat Evolved Anniversary*, *Call of Duty: Black Ops II* and *Far Cry 3*.

A trait of League of Legends is that new playable characters are added to the game on a regular basis. These characters are always promoted and discussed prior to or in the first few weeks of their introduction to the game. This is reflected in the trends of r/leagueoflegends in figure 6.13b, where *Draven*, *Zyra*, *Darius*, *Hecarim*, *Zix* (full name is *Kha'Zix*), *Nautilus*, *Vi* and *Lulu* are newly released characters, and *Sivir* and *Ezreal* are characters that have been changed. League of Legends is a competitive game with several annual tournaments that are enthusiastically watched and discussed by the community, such as the *IPL4* (IGN ProLeague tournament), *IEM* (Intel Extreme Masters tournament), *MLG* (Major League Gaming tournament) and the *PAX* (Penny Arcade Expo tournament). Larger changes to the game cause comparable trends, so when Riot introduced the *Spectator Mode* in April, it was a central conversational topic in r/leagueoflegends.

Minecraft's trends, depicted in figure 6.13c, involve only two larger ones: the release of the game on the *Xbox* 360 video game console in May, and again the *Christmas* spike in December that is in so many subreddits. Minecraft players and users of r/Minecraft like to come up with new ideas for the game, which are tagged with *[Suggestion]*, and discuss them, hoping that the developers of the game would see them and consider an implementation. The extension of the player base with the release on Microsoft's Xbox 360 set off an avalanche of new ideas and suggestions.

Figure 6.13d shows that the trends of r/Music are mostly balanced. The first peak is *Whitney Houston*, an American singer who died in February 2012. The second one is the start of a longer trend: starting in April 2012, users of r/Music mention *YouTube* very often in their titles. The reason might be changes of the subreddit rules, but this could not be verified.



(a)r/gaming (b)r/leagueoflegends

(c)r/Minecraft (d)r/Music

Figure 6.13.: The trends of r/gaming, r/leagueoflegends, r/Minecraft and r/Music

In July 2012, the Korean musician Psy released the single *Gangnam Style*, which earned world fame for becoming the first YouTube video to reach a billion views in December 2012 (Gruger, 2012). While it is trending in r/Music in September, one would assume that this song would have had a larger impact on this subreddit, considering its focus on music and YouTube links.

Due to the lack of a dedicated theme other than the definition of the preferred medium, r/pics is a conglomeration of everything interesting enough to be the subject for a photograph. As one can see in figure 6.14a, this subreddit has again seasonal trends, such as *Valentine's Day*, *Halloween* and *Christmas*. The *Mars* Rover Curiosity landed in August 2012[21] and its first images attracted the attention of this subreddit. *Hurricane Sandy* is of equal significance in September.

The topic modeling results of r/politics in 2012 already revealed that the subreddit is basically defined by the presidential election campaigns of several candidates on the one hand, and Internet regulation bills on the other. The monthly trends, depicted in figure 6.14b, now show the short term subjects of presidential campaigns or the rise of new hot topics. The figure as a whole displays the erratic nature of politics and news, where one headline hunts the other, topics are sky-rocketing on one day and as good as forgotten on the next. An important subject of the campaigns revolved around the support of insurance coverage for *birth control*, which has been discussed heavily in February. This topic escalated, when the conservative radio talk show host *Rush Limbaugh* gave provocative and offensive comments about a female law student, whose statement was denied at a hearing to the topic (Fung, 2012). Another incident, an anti-gay advertising campaign on buses in London that has been prevented by the mayor (Booth, Mulholland, and Strudwick, 2012), forms the titles in the subreddit in April. In contrast to SOPA and PIPA, which are core topics in r/politics and other subreddits throughout the year, the Cyber Intelligence Sharing and Protection Act (*CISPA*), again a similar Internet *control* act only bubbles up in April, when the bill was passed in the House of Representatives. Debates on the *health care* peak in June, when the American *Supreme Court* upheld the contested overhaul of the system introduced by President Barack Obama (Liptak, 2012). In August, the presidential candidate Mitt Romney publicly an-

21 http://solarsystem.nasa.gov/news/msl_landing.cfm

nounced *Paul Ryan* as his vice-presidential running mate (Zeleny, 2012). Another reoccurring important topic in American politics is *gun control*, coming up in December due to President Obama's announcement of new policies (Altman, 2012), a subject matter that brings the National Rifle Association of America (*NRA*) to the scene.

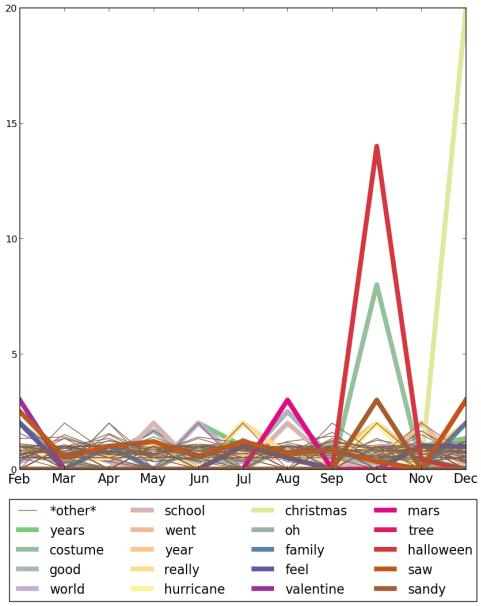
In contrast to r/politics, the trends of r/technology in figure 6.14c are almost unimpressive. The trends identify some of the, according to r/technology, most important technological news in 2012:

- Apple launches its third generation *Ipad* tablet in March, and announces the fourth generation and a *mini* version of the *Ipad* in October.
- *CISPA* has been passed in April.
- *Microsoft* announces Microsoft Surface, its first tablet PC, in June and releases it in October, simultaneously with Windows 8 and Windows Phone 8.
- The *Mars* Rover Curiosity lands successfully in August.
- *Samsung* loses patent infringement charges against Apple in August, and is ruled to pay $1.05 billion in damages (Vascellaro, 2012).

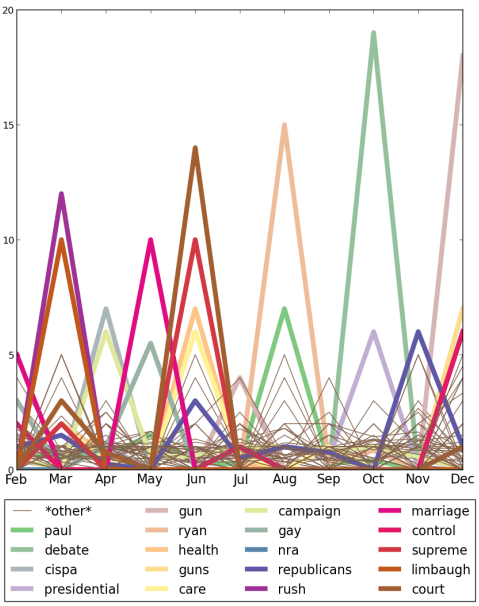
Although r/Music missed out on the success of the song *Gangnam Style* by the musician Psy, it is the only trend worth mentioning in r/videos besides *Christmas*, presented as a significantly high spike in September in figure 6.14d.

Figure 6.15 shows that r/worldnews features similar characteristics to r/politics. Some of the news that are identified as trends have also been short term trends or lasting topics in other subreddits. Whitney Houston's Death is the first distinguished one, along with an offensive of the Syrian army (Borger and Mahmood, 2012) in the Syrian civil war in February.

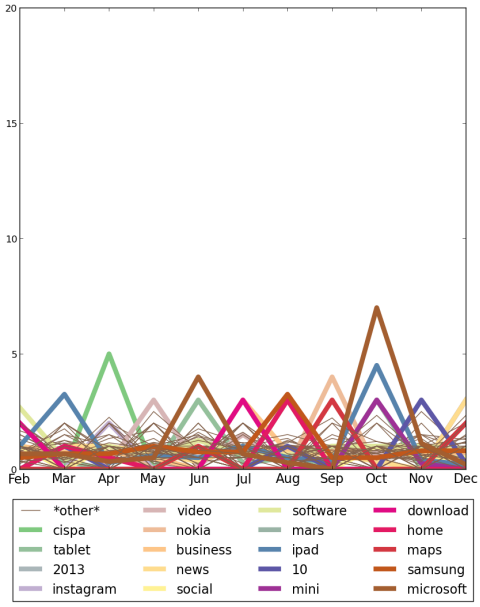
North Korean Rocket launches and the suspension of food aid to North Korea by the United States of America are major subjects in March (*Korea*, *government*). However, these events are even eclipsed by the media frenzy caused by the aforementioned *Kony* campaign in March, and the critics of the humanitarian organization behind it, the *Invisible Children*, in April. In June, the Syrian civil war escalated again after a ceasefire attempt from April onwards. The 2012 Summer *Olympics* in



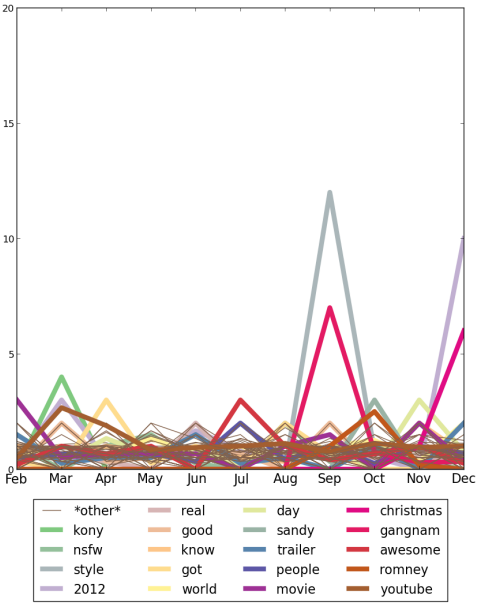
(a)r/pics



(b)r/politics



(c)r/technology



(d)r/videos

Figure 6.14.: The trends of r/pics, r/politics, r/technology and r/videos

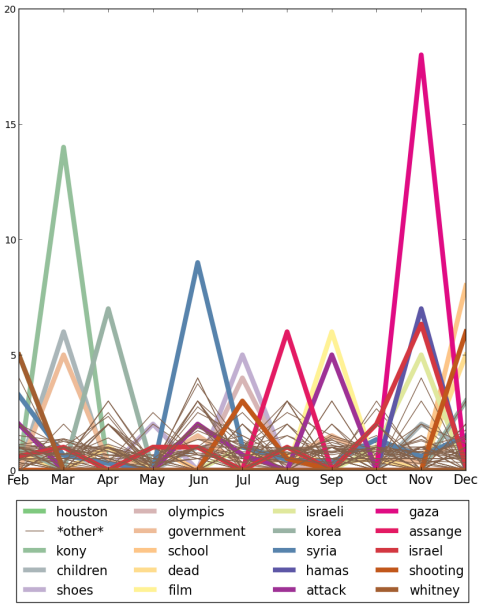


Figure 6.15.: r/worldnews

London form a peak in July, followed by the news of political asylum for Julian *Assange*, the founder of WikiLeaks, in August (Ferran and Bruner, 2012). The Gaza-Israel conflict burst into severe fighting in October, and *Gaza*, *Israel*, and *Hamas* dominated the titles of the r/worldnews. The last event marked in the figure is the tragedy of the *Sandy Hook Elementary School shooting* in December, a mass murder in Sandy Hook, Connecticut (Barron, 2012).

The short term trends analysis results of the subreddits

r/atheism,
r/circlejerk,
r/fffffffuuuuuuuuuuuu,
r/tf2trade,
r/todayilearned,
r/trees, and
r/WTF

can be found in appendix D.

In conclusion, the short term trends identified with reddit channel topic

modeling verifiably enable automated detection of events in the subreddit's specific category. Subreddits, which are not aligned to a certain topic, such as r/pics or r/videos, tend to feature no characteristic trends other than holidays, namely *Valentine's Day*, *Halloween* and *Christmas*. Other subreddits, however, produce distinct short time trends, which are nearly always related to real world events in a timely manner. The paragons are r/politics and r/worldnews, where every short term trend is pronounced and revolves around an event that caused newspaper headlines and sensation. In this regard, reddit seems to fulfill the necessity of a social news aggregation website, at least in some subreddits.

The method of combining topic modeling over short time frames with a term-weighting-measure turns out to be a suitable approach to identify and emphasize red-hot topics. Thus, the approach could find uses in many different domains as well. In some subreddits, the trend generation can be used for market investigation and marketing purposes. Reddit itself could use it to profile subreddits and attract potentially interested advertisers. In r/politics, the subreddit for American political topics, the presidential campaigns, discussions on programs and several political affairs exhibit distinct trends. Analyses of these could be useful for opinion research and future campaign planning. Automatic trend detection could reveal perfect timing for an *Ask me Anything* submission of a politician, which could bring voters over to his or her side, comparable to a media conference or a public forum. Another possible application for short term trends would be to support recommender systems, either for reddit-internal uses (e.g. subreddit recommendations, advertising recommendation) or external ones.

These applications would probably need to further shorten the time frame of the topic modeling process, refine the trend calculations to receive more detailed trend data, and experiment with extensions. This means that it is a domain open to future work as well. It would be interesting to combine the topics and short term trends with automated comparison to Google Trends[22], headline feeds of newspapers and networks, or to research user participation in trends by using comment and score data, similar to the approaches in *"What is Twitter, a Social Network or a News Media?"* by Kwak et al. (2010).

22 http://google.com/trends

7. Discussion of Results

Recapitulating the results, it is clear that reddit has had a turbulent history. In only two years, from 2011 until the end of 2012, after solving problems with server capacity and profiting from a struggling competitor, it nearly quadrupled its monthly number of submissions. The diversity in terms of subreddits has grown as well. Over time, more and more subreddits have been created and used, approaching a more uniform distribution of submissions to subreddits. The discontinuation of the general, all-encompassing subreddit r/reddit.com marks an important step towards clarity of content arrangement and thus accessibility and usability for new users.

However, the development of the domains of links submitted to reddit draws a completely different picture than the evolution of subreddits. In this context, the contrary is the case. Since mid-2009, the diversity of domains declines steadily. Self-post, Imgur, YouTube and Quickmeme submissions are posted with increasing frequency, at the expense of everything else, especially Blogspot (or Blogger) and Wordpress. Link shorteners and hosting service websites are frowned upon on reddit, and disappear eventually. On the other hand, websites that specialize their service intentionally for usage on reddit rise to great success[1].

Relating the content on reddit, one can say that there is a pronounced tendency for image and self submissions. Both categories decidedly predominate from a static point of view at all submissions in the data set. Yet, when adding the aspect of time, it is evident that the predominance of image and self submissions has not always been the case. Before 2010 and the growth burst, the major part of submissions had originated from text based websites. From 2010 to the third quarter of 2011, self-posts had become the most prominent submission type, to such extent

1 Imgur is designed to match reddit's needs, stated by its founder at `http://redd.it/7zlyd`

that even most of the 20 largest subreddits with a clear focus on other, non-self categories, still have their highest ratio of self-posts in this time span. Only from 2011 onwards, image submissions have caught up. The relative shares of Quickmeme and the r/AdviceAnimals subreddit grow continuously, a symptom of the advancing memes and mind candy in form of images. The large success of image content is reasonable, considering the amount of time required to consume it and decide to vote on the submission, in contrast to the time required to do the same on text or video, even though these categories carry more information content. Less effort and less time is required to look at a picture, which essentially increases the throughput of consuming users, which in turn increases the potential score.

Reddit's honorable intentions are to be the *front page of the Internet* and *a source for what's new and popular on the web*[2]. These might have been true in its earlier days, when content was more balanced out and originated from more diverse sources. Until mid-2009, every second submission had originated from a text-centered domain. Today, reddit as a whole is better described as a source for images and self-posts. This shift to self-referentiality and quick consumable content, which is particularized in the paper by Singer et al. (2014), is evident by

- the outright majority of image and self submissions,
- the merging of image hosting almost exclusively on Imgur established by a consensus of the reddit community, and
- the success of respective subreddits, such as the image exclusive r/funny, r/AdviceAnimals, r/ffffffuuuuuuuuuuuu and r/aww (all of which feature creations by reddit's own users), or the self-post exclusive r/AskReddit, r/todayilearned, r/IAmA and the self-ironic r/circlejerk.

Over time, all submissions with external links other than images have declined continuously and given way to Imgur and self-posts. In Singer et al. (2014), the notion of a shifting composition of content and the changes of interests are further pursued and expanded by an analysis of attention and perception, by popularity measures and a user survey.

Many submissions in these subreddits require inside knowledge of the culture and demeanor that is fostered on reddit in order to understand

2 http://reddit.com/wiki/faq

them, which is interpretable as a further reference to the concentration on the community and reddit itself. The extent of the necessity of inside knowledge is comprehensible when investigating the identified topic models and short time trends of some of the mentioned subreddits, where abbreviations, verbalisms, phrases and memes are omnipresent. Other subreddits are almost untouched by inside knowledge, trends or ongoings in the world, and revolve around little more than pets and holidays, such as r/pics and r/aww.

On the other hand, there are r/politics, r/worldnews, r/technology and other, even more specialized subreddits, that deliver relevant content in a timely manner, and considering the most frequent domains they do so from a variety of sources. Especially subreddits about news and politics still contain large shares of text categorized submissions from the various news networks and papers. The topic models and short term trends point out that these subreddits have their finger on the pulse of the time. A prime example for this is an Ask Me Anything thread by user u/jammastajayt in 2007[3], who kept answering questions on reddit from inside the Virginia Polytechnic Institute and State University in Blacksburg, Virginia, United States, while a shooting took place, claiming 33 victims[4]). Zafar (2011) from TIME subtly worded it in the following way:

> "Reddit is quickly challenging Twitter's turf as a place for real-time updates and citizen journalism."

Summarizing, one can say that reddit has many facets and can be whatever a user wants it to be - entertaining or distracting, informational or educational, stimulating or deterrent, personal or impersonal, fast or slow - as long as this user finds and subscribes to the respective subreddits. In its entirety, however, reddit clearly turns its attention towards entertainment and community features.

3 http://redd.it/n56uf
4 http://nytimes.com/2007/04/16/us/16cnd-shooting.html

8. Conclusion

The results presented in this work draw an amply descriptive picture of reddit. A model has been fitted to reddit's growth curve, and shifts in structure and content have been discovered. On this basis, answers to the research questions prompted in chapter 1.2 can be given.

1. What model fits reddit's growth best?

It has been demonstrated that the number of monthly submissions to reddit grows at an ever more increasing rate, and its growth curve is mathematically very similar to the model of exponential growth. A significant interruption is identified in 2010, caused by competition, technical and financial problems, and a second one in late-2012, where the causes could not be retraced.

2. What kind of content is submitted to reddit and what is the dominant media?

Reddit features a wide range of diverse content on the Web. Many submissions originate from news websites, image hosting, video or audio streaming services, or reddit itself. Over time, the diversity of sources declines in favor of self-posts, Imgur or Quickmeme images and YouTube videos. The categorization concept in this work divides these sources into six types, namely text, image, video, self, audio and miscellaneous. From a static point of view, self and image are the most dominant media.

3. Is reddit a social news aggregator, or rather an image board?

The answer to the fundamental question, whether reddit truly is a social news aggregator, as most frequently described, or rather an image board, depends on the time and the scope in which reddit is seen. Reddit in its entirety has been thoroughly focused on news aggregation

in its earlier days. Until mid-2009, social news aggregation has been a valid description of the web portal, because in this time submissions from news websites have been prevalent, although already seconded by a significant share of images. This composition switched almost completely, at first to a majority of self and later of image submissions. At this point hardly a tenth of all submissions per month contain external textual content and news. In 2012, reddit resembles far more an image community board than a news aggregation website. Reddit still serves the purpose of news aggregation in many subreddits, but it is no longer the most prominent and defining service.

In these subreddits, the extent of moderation has been examined as well. While a decent quantity of submissions are deleted over a year, an overall bias could not be identified by any of the available means. The measuring approach of moderation in this work is simplified, and further investigation might yield more informative results.

The topics of subreddits, calculated using LDA topic modeling, are satisfying for the following reasons:

The identified latent topics describe the character of each subreddit well. Depending on the subreddit, the topics either contain core subjects of the subreddit (e.g. in r/politics or r/technology), or the buzz and gossip in less serious and thematized ones (e.g. r/pics or r/AskReddit). Additionally, the topics point out how much prior knowledge is necessary to fully understand a subreddit and be a part of it, since many subreddits and their users foster their own terms, abbreviations and phrases in common parlance.

The short term trends in topics are of equivalent success. In subreddits, where submissions revolve around a single, simple topic, such as r/aww, or around nothing at all besides special days, such as r/pics, the trends are unsurprising and provide little information. The trends of r/AdviceAnimals reveal insights in the development of memes and their Internet subculture very accurately. News focused subreddits, however, mirror important real world events consistently and timely, although they are biased towards the interests of American users, naturally due to the fact that the majority of the users come from the USA.

8.1. Limitations

This thesis offers an insight in reddit's development in the years of 2008 to 2012, and makes assertions on its state in terms of size in submissions, origins of submissions and content. It is revealed, that reddit changed its setup in terms of content several times, and it will clearly continue to do so in the future. Thus, the results and conclusions are to be seen in their historical context, and are not necessarily valid to the state of reddit afterwards.

Limitations of this study and the concluding answers to the research questions are constituted by the choice of methods and statistical techniques, as well as their potential inappropriate use.

1. There could be a growth model that fits the monthly growth of submissions better than the three preselected ones in chapter 5.1.1 do.
2. The content analyses are based on the domain categorization, which are possibly biased due to the manual execution of the categorization, or simply wrong. The intersubjectivity could be low - a limitation that was slightly mitigated by the merging of multiple categorizations. More detailed, extensive or otherwise different categorizations result in accordingly different statistics.
3. The definition of reddit depends on the content analysis for justification, and is equally limited. The topic models are an unsupervised machine learning approach, which means that they derive topics from clustering and probability features without additional external input. Because of this, it is possible that topic modeling reports patterns in the text corpora that are present, but uninteresting. Furthermore, the generated topics are vectors of terms, but they are not labeled. Thus, the theme of each topic had to be interpreted. The number of extracted topics and displayed terms were limited to meet the requirements of this study. After multiple executions and experiments with different configurations, the best setting in terms of performance, processability and presentability was selected.

Technical changes to measures and methods influence the results as well, and might pose a limitation as well.

8.2. Outlook

This work provides a basic understanding for reddit, and, accompanied with the findings of the paper (Singer et al., 2014), should motivate and inspire further research of it. As mentioned before, reddit has (in contrast to, for example, Twitter) drawn little scientific attention so far, despite its success and size. Much of the research conducted on Twitter could be conducted on reddit as well, for example:

- Studies on the influence of certain users, as shown by Bakshy et al. (2011) on Twitter users, could reveal the impact that some of the users have that use reddit to communicate their opinions or interests.
- Rowe, Angeletou, and Alani (2011) identified tweets as discussion starters, which could also be done for reddit's comment threads. Since the design of the comment section encourages back-and-forth discussions, and there are dedicated subreddits for discussions, it would be fascinating if the indicators for discussion starting tweets also apply to reddit comments or submission titles.
- Reddit can be used as corpus for sentiment analysis and opinion mining, similar to Pak and Paroubek (2010).

Reddit could be used to review the platform independent validity of some insights that were gathered on Twitter, or as environment for social experiments, opinion mining for various purposes (economical, political or social ones, to name a few), or to analyze the interconnectivity between users in the context of anonymity. Comparisons to other similar or related systems such as Hacker News, Digg, Imgur or Slashdot, would outline similarities and differences, and reveal the status that reddit has among them.

The results and conclusions presented in this thesis are based on a data set of submissions to reddit. While submissions are the core functionality of this web portal, the aspects of comments, comment-threads and registered users are of great interest, especially with the noticeable trend of self-referentiality and community focus in mind. It would be compelling to investigate the topics and the behavior within comments, or the social network traits among reddit's users, or to perform use case studies and surveys to understand the motivations of the community. Combined results would yield an even better understanding of reddit

and its evolution. With submission-, comment- and user-data it would be possible to apply network analysis on the response behavior between users, similar to retweet-analyses on Twitter.

At the time of the analyses of this work, reddit's API does not support the crawling of its users, and the collection and storage of comments over a considerable span of time (let alone to crawl them backwards in time until, for example, 2008) requires vast technical efforts, simply because of the huge amount of data that is generated in comments every day and the traffic restrictions of the API.

The bias of moderation on reddit as well as a study on a potential underlying political orientation of the community would shed light on the effects of reddit on public opinion and the possibility of manipulation and propaganda.

Topic modeling and short term trend analysis on reddit could find many applications, but need further development and refinement to do so. As mentioned earlier in chapter 6.2.4, with a higher frequented short term trend extraction from topics on suitable subreddits, it would be very interesting to evaluate and compare the results to a similar application on news network feeds. Studies could also investigate the dimension of user participation in trends, and how they are affected by it. While future work in terms of the bias of moderation would investigate the extent of opinion forming already in progress on reddit, topic models could be used to gather insights into the public opinion of reddit users and probably how to amplify or mitigate certain notions in the long run.

Appendix

Appendix A.

A short description on Subreddits in this Thesis

There are many Subreddits mentioned, many are targets of evaluations. The subjects of analysis are chosen by its size in submissions over the timespan of interest. For most experiments, the whole data set is considered so the timespan is from the 1st of January 2008 to the 31st of December 2012. In this section, the Subreddits that are mentioned throughout this work are described for a better understanding.

Table A.1.: All subreddits mentioned throughout this work explained. [A-E]

Subreddit	Description
r/AdviceAnimals	A subreddit dedicated to internet memes, popular images with short texts used as macros. Only direct links to these images are allowed.
r/announcements	An official subreddit that summarizes the most important items from reddit's blog.
r/AskReddit	Users can post questions to the reddit community, thus only self-posts are allowed, containing the question. It is dedicated to open-ended discussion, polls and surveys are not allowed.
r/askscience	Similar to r/AskReddit, but dedicated to scientific questions.
r/atheism	This subreddit contains all manner of content related to atheism and agnosticism.
r/aww	The theme is "things that make you go AWW!", meaning images, videos and stories of endearing things, mostly pets (cats and dogs).
r/bestof	A collection of the best submissions or comments to reddit. Therefore this subreddit features reddit internal links only.
r/birdswitharms	Images of photoshopped birds with human arms attached to them.
r/blog	An official subreddit listing all the items from reddit's blog that are about the community or otherwise not directly related to functionality changes.
r/books	Discussions about and presentations of books, authors and genres can be found in this subreddit.
r/circlejerk	On reddit, *circlejerk* is the term for running gags and common opinions between like-minded users in (short phrased) comments, often satirizing reddit, where each participant upvotes each other participant. The positive feedback loop, powered by the Karma system, rewards popular opinion and catchphrases, while punishing unpopular or poorly worded ones.
r/Conservative	A community dedicated to conservatism.
r/conspiracy	A list of conspiracy theories on all topics and coincidences imaginable by its subscribers.
r/earthporn	EarthPorn collects images of very beautiful natural landscapes.
r/explainlikeimfive	Similar to r/AskReddit users can ask the community for explanations in very basic English.

Table A.2.: All subreddits mentioned throughout this work explained. [F-P]

Subreddit	Description
r/fffffffuuuuuuuuuuuu	The term fffffffuuuuuuuuuuuu describes a special kind of comics, called *rage comic*. Stylistic devices are very crudely drawn macros of faces and a very short storyline, often describing real life events or experiences that end with a humorous punchline.
r/funny	Submissions are expected to be funny.
r/Games	Serious discussions and content about computer games.
r/gaming	A subreddit for everything related to computer games.
r/gifs	A subreddit containing only images in GIF featuring animations.
r/IAmA	The title of this subreddit means "I Am A" and "Ask Me Anything". Users state their profession or something special about them, and offer to answer questions of the community.
r/KarmaConspiracy	This subreddit assumes that all successful submissions on reddit only aim to gather as much karma as possible and that there is a conspiracy behind them. It has a humorous intent.
r/KarmaCourt	A community that plays a meta game on reddit, where violations against the reddiquette[1] with some own modifications are collected and court hearings are reenacted.
r/leagueoflegends	This subreddit is dedicated to the exceedingly successful and popular online computer game League of Legends.
r/MensRights	Men's rights and infringements of those are discussed here.
r/Minecraft	Minecraft is the name of an independent computer game developed by Markus Persson, that became very popular and successful.
r/movies	Everything related to movies can be found here.
r/Music	Users submit songs and albums and discuss some aspects of music. Its intent is to provide a platform where people can discover music they like.
r/newreddits	In this subreddit, users can promote their newly created subreddits to get attention and early subscribers.
r/news	A subreddit about news, that is not primarily political.
r/pics	Photographs and pictures are submitted here, as long as the image is not a meme, screenshot, animated format or contains nudity or gore.
r/POLITIC	A non-moderated subreddit to counter censorship on reddit, where all submissions to political subreddits are automatically mirrored.

Table A.3.: All subreddits mentioned throughout this work explained. [P-Z]

Subreddit	Description
r/politics	A place for current matters of politics in the United States of America.
r/progressive	A community for the political Modern Progressive Movement.
r/random	This is not a real subreddit, but rather a redirecting link where the address looks like a subreddit. Following it, a user is forwarded to a random subreddit.
r/reddit.com	The default subreddit until October 2011, when it has been closed, archived and replaced by a set of subreddits as default subscriptions.
r/redditdev	A subreddit about developing programs with the reddit API.
r/reportthespammers	Here spammers can be reported, and confirmed reports are forwarded to reddit administrators.
r/socialism	A subreddit for the socialist ideology.
r/technology	Articles related to technology; images and videos are not allowed.
r/television	Conversations, articles and videos about television programming are featured here.
r/tf2trade	The computer game "Team Fortress 2" by Valve offers dealing with digital items within the game. This subreddit acts like a market place for these items.
r/TheoryOfReddit	This is a self-critical and self-questioning subreddit dedicated to inquiring into what makes reddit what it is, how and why it works the way it does. Here questions and theories about reddit arise, studies are submitted and discussed.
r/todayilearned	Interesting and specific facts can be posted here, which are supposedly learned by the user the day the fact got posted. Thus all submission titles all start with TIL, *Today I Learned*.
r/trees	A subreddit for posts related to cannabis.
r/videos	Video content of all kinds can be found here, except politics, pornography or gore.
r/worldnews	A subreddit about news from outside of the United States of America.
r/WTF	This subreddit features shocking or surprising content, often containing gore.

Appendix B.

Categorization of Subreddits

This appendix lists the categorization results of the following subreddits:

r/AdviceAnimals
r/fffffffuuuuuuuuuuuu (abbreviated as r/f7u12)
r/trees
r/WTF
r/worldnews
r/atheism
r/aww
r/todayilearned
r/reportthespammers
r/POLITIC
r/circlejerk

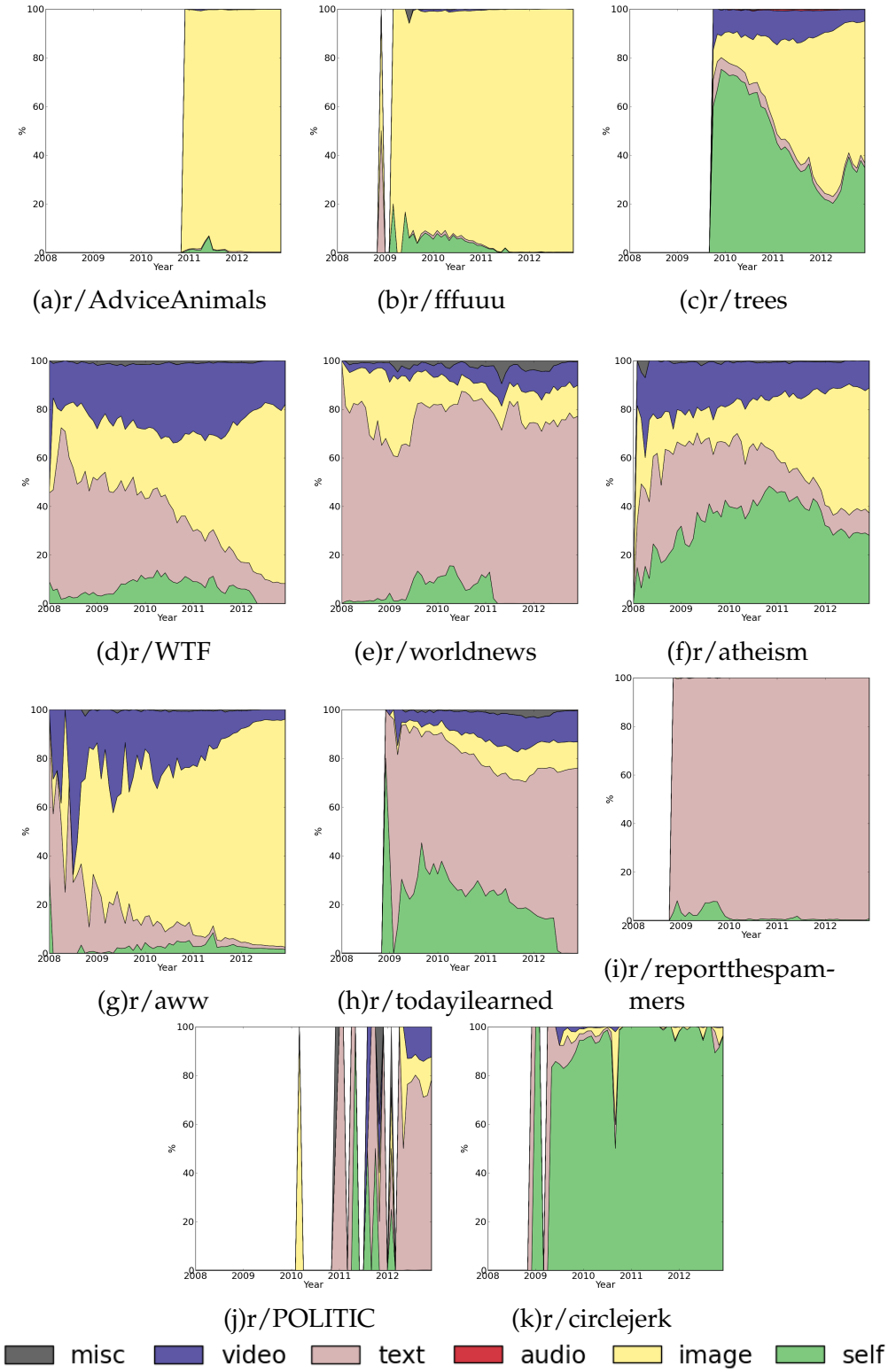


Figure B.1.: **The development of the six categories** in the scope of several subreddits.

Appendix C.

LDA Topics in 2012

These are the results of topic modeling based on LDA upon a corpus containing all submissions from 2012 in the respective subreddits. This model has been calculated for the 20 largest subreddits by submissions in 2012, namely

1. r/AdviceAnimals
2. r/AskReddit
3. r/atheism
4. r/aww
5. r/circlejerk
6. r/fffffffuuuuuuuuuuuu
7. r/funny
8. r/gaming
9. r/leagueoflegends
10. r/Minecraft
11. r/Music
12. r/pics
13. r/politics
14. r/technology
15. r/tf2trade
16. r/todayilearned
17. r/trees
18. r/videos
19. r/worldnews
20. r/WTF

A short explanation on the subreddits can be found in table A.1. For each subreddit, 20 topics are generated, and of each topic, the 10 highest weighted terms are listed.

Table C.1.: The 20 discovered topics with LDA in r/AdviceAnimals in the time of 2012

Topic 1	guy reddit	good today	greg fixed	friend advice	sap scumbag
Topic 2	world know	problems steve	scumbag cat	just new	ggg teacher
Topic 3	happens facebook	success guy	scumbag know	good dog	time class
Topic 4	world steve	problems new	scumbag night	guy meme	good year
Topic 5	kid bad	just scumbag	insanity success	wolf happened	day internet
Topic 6	college freshman	day good	joke awkward	bad socially	eel penguin
Topic 7	time world	forever good	scumbag greg	problems college	guy reddit
Topic 8	success reddit	scumbag kid	happened futurama	fry good	girl bachelor
Topic 9	scumbag school	college freshman	feel new	guy just	good time
Topic 10	scumbag introducing	fixed inappropriate	success kid	feel world	radio dj
Topic 11	man feel	world scumbag	interesting parrot	paranoid today	redditor reddit
Topic 12	just wolf	happens problem	world insanity	philosoraptor success	time right
Topic 13	scumbag new	happened black	man just	successful lazy	college feel
Topic 14	scumbag college	awkward reddit	socially sap	penguin world	today bad
Topic 15	good reddit	guy man	greg problems	scumbag coon	world lame
Topic 16	scumbag guy	happened good	man facebook	successful today	people black
Topic 17	world kid	problems scumbag	good friend	happened guy	success just
Topic 18	scumbag world	keanu freshman	conspiracy steve	brain success	college kid
Topic 19	reddit black	like man	night feel	problems successful	meme boromir
Topic 20	good philosoraptor	guy greg	scumbag day	reddit world	time gets

Table C.2.: The 20 discovered topics with LDA in r/AskReddit in the time of 2012

Topic 1	reddit	think	help	like	good
	time	question	sopa	thing	just
Topic 2	reddit	people	need	think	help
	does	make	thing	time	best
Topic 3	reddit	thing	like	just	did
	does	world	time	best	good
Topic 4	reddit	does	time	did	year
	really	getting	new	help	just
Topic 5	reddit	people	just	know	help
	does	best	did	like	question
Topic 6	reddit	best	know	good	does
	thing	make	did	people	help
Topic 7	reddit	like	best	help	work
	sopa	people	just	does	need
Topic 8	reddit	time	just	does	help
	hey	good	tell	life	best
Topic 9	reddit	help	way	got	best
	just	need	know	sex	does
Topic 10	reddit	help	does	know	need
	best	way	story	people	things
Topic 11	reddit	hey	best	need	think
	just	life	want	did	know
Topic 12	reddit	think	did	people	like
	does	best	help	hey	need
Topic 13	reddit	redditors	thing	help	did
	need	just	best	say	question
Topic 14	reddit	help	does	thing	make
	favorite	day	like	good	time
Topic 15	reddit	make	does	people	going
	help	life	just	like	use
Topic 16	reddit	new	need	help	think
	thing	know	internet	did	best
Topic 17	help	life	need	reddit	time
	best	want	people	just	think
Topic 18	reddit	like	want	does	time
	life	help	know	think	people
Topic 19	reddit	just	know	help	need
	think	did	website	does	like
Topic 20	reddit	does	people	help	know
	life	like	just	thing	work

Table C.3.: The 20 discovered topics with LDA in r/atheism in the time of 2012

Topic 1	religion	church	atheism	like	god
	atheist	people	facebook	jesus	catholic
Topic 2	god	atheist	like	atheism	christians
	church	atheists	help	does	christian
Topic 3	god	atheist	post	bible	facebook
	just	friend	think	guys	thought
Topic 4	god	atheist	just	christian	atheism
	like	scumbag	religion	does	need
Topic 5	christian	atheism	atheist	school	religion
	today	facebook	right	thought	god
Topic 6	god	christianity	atheism	evolution	need
	help	think	atheists	new	friend
Topic 7	atheist	atheism	just	really	people
	atheists	love	christians	christianity	fuck
Topic 8	atheism	good	atheist	like	atheists
	thought	religion	christian	guy	question
Topic 9	atheism	atheist	like	god	new
	jesus	christian	christians	say	love
Topic 10	god	religion	atheist	christian	atheism
	like	just	jesus	tebow	facebook
Topic 11	just	atheist	atheism	people	faith
	god	atheists	know	think	christian
Topic 12	atheists	atheist	atheism	religion	god
	just	think	christian	tebow	help
Topic 13	jesus	religion	atheists	god	atheism
	video	bible	hate	reason	old
Topic 14	religion	god	think	tebow	jesus
	people	religious	tim	atheist	bible
Topic 15	atheism	religious	atheist	logic	church
	good	right	christian	want	little
Topic 16	christian	jesus	religious	people	religion
	god	bible	atheist	say	know
Topic 17	god	atheist	just	atheism	people
	believe	christian	think	new	like
Topic 18	atheism	just	christian	atheist	fixed
	religion	ahlquist	religious	think	new
Topic 19	religious	atheists	god	jesus	religion
	atheist	atheism	believe	hate	good
Topic 20	religion	god	atheists	christians	right
	bible	time	atheist	jesus	atheism

Table C.4.: The 20 discovered topics with LDA in r/aww in the time of 2012

Topic 1	old	baby	dog	love	new
	like	cat	cute	year	friend
Topic 2	puppy	cat	dog	new	friend
	just	like	baby	home	think
Topic 3	dog	time	just	kitty	old
	little	year	today	day	friend
Topic 4	dog	kitty	cat	just	guy
	little	new	picture	puppy	reddit
Topic 5	like	cat	kitten	friend	cute
	puppy	reddit	today	think	time
Topic 6	cat	just	dog	baby	love
	friend	little	best	cute	sleep
Topic 7	love	dog	just	day	kitty
	time	cat	got	right	cute
Topic 8	just	cat	dog	reddit	like
	likes	friend	old	cats	dogs
Topic 9	reddit	just	dog	got	kitten
	little	kitty	home	new	best
Topic 10	just	love	cat	aww	year
	new	kitten	dog	reddit	baby
Topic 11	cat	dog	puppy	reddit	adorable
	cute	like	just	kitten	new
Topic 12	puppy	little	got	dog	just
	new	old	like	best	baby
Topic 13	cat	little	love	day	just
	dog	picture	cute	baby	cake
Topic 14	new	best	dog	cats	puppy
	love	just	think	loves	little
Topic 15	dog	like	just	cat	little
	guy	know	love	cute	loves
Topic 16	cat	just	love	kitten	little
	time	loves	baby	reddit	cute
Topic 17	meet	little	kitty	cute	dog
	new	reddit	just	guy	cat
Topic 18	day	cat	love	little	baby
	sleeping	post	oh	dog	cake
Topic 19	puppy	new	love	meet	day
	cute	dogs	just	little	dog
Topic 20	cat	friend	reddit	dog	love
	cute	likes	baby	post	day

Table C.5.: The 20 discovered topics with LDA in r/circlejerk in the time of 2012

Topic 1	upvote literally	reddit circlejerk	upvotes iama	til hitler	ama left
Topic 2	iama til	circlejerk upvotes	fixed karma	upvote sopa	ama make
Topic 3	ama upvotes	post guys	til request	upvote just	frothy new
Topic 4	paul literally	ron vote	til reddit	tebow arrow	upvote just
Topic 5	iama karma	ama paul	day ron	cake know	upvote reddit
Topic 6	upvote like	til pussy	friend upvotes	lol just	zone ron
Topic 7	upvote ron	circlejerk did	reddit dae	paul think	til ama
Topic 8	upvote reddit	circlejerk til	ron upvotes	ama like	paul 1984
Topic 9	reddit circlejerk	dae soap	just iama	ama think	like upvotes
Topic 10	reddit like	upvote fap	day upvotes	paul know	ron dae
Topic 11	upvote hey	karma acta	just iama	reddit like	ama guys
Topic 12	upvote cunt	paul upvotes	post iama	ron gets	ama literally
Topic 13	le reddit	karma say	just iama	upvote hey	ama circlejerk
Topic 14	paul reddit	sopa til	ron upvotes	blackout pipa	ama upvote
Topic 15	upvote iama	ama circlejerk	brave know	reddit paul	santorum atheist
Topic 16	paul upvote	ron upvotes	reddit circlejerk	dae fuck	just sopa
Topic 17	reddit nope	upvotes post	paul left	ron new	upvote karma
Topic 18	sopa rick	upvote 2012	reddit best	literally circlejerk	santorum post
Topic 19	karma help	ama paul	upvote sopa	reddit ron	circlejerk page
Topic 20	ama make	nigger page	reddit upvote	hitler literally	request iama

Table C.6.: The 20 discovered topics with LDA in r/ffffffuuuuuuuuuuuu in the time of 2012

Topic 1	time	know	rage	like	day
	going	reddit	fixed	feel	comic
Topic 2	rage	comic	stupid	friend	just
	hate	troll	happened	game	worst
Topic 3	rage	time	think	reddit	really
	like	life	troll	good	hate
Topic 4	rage	reddit	fixed	just	mom
	time	best	troll	hate	school
Topic 5	rage	reddit	just	fuck	happens
	shower	time	dog	good	forever
Topic 6	rage	reddit	comic	page	45
	class	old	today	fixed	time
Topic 7	rage	did	reddit	fixed	just
	mom	time	troll	know	girl
Topic 8	rage	happened	troll	just	people
	time	oh	story	really	day
Topic 9	rage	just	little	happened	time
	fixed	day	reddit	way	best
Topic 10	rage	know	day	time	new
	fixed	reddit	troll	shit	think
Topic 11	rage	time	happens	new	mom
	today	make	love	school	comic
Topic 12	rage	just	new	comic	year
	time	facebook	life	morning	happened
Topic 13	rage	know	true	fixed	new
	school	hate	story	girl	try
Topic 14	rage	fixed	damn	best	day
	just	school	use	high	reddit
Topic 15	rage	story	just	happens	true
	new	bad	happened	like	troll
Topic 16	rage	fixed	troll	like	comic
	friend	45	time	page	post
Topic 17	rage	reddit	friend	shower	people
	feel	like	need	new	time
Topic 18	rage	reddit	life	fixed	thought
	night	close	time	troll	sad
Topic 19	rage	comic	day	reddit	time
	story	true	damn	brain	girlfriend
Topic 20	just	rage	love	internet	comic
	life	oh	fuck	reddit	time

Table C.7.: The 20 discovered topics with LDA in r/funny in the time of 2012

Topic 1	today	like	new	hey	think
	use	hear	happened	looks	forever
Topic 2	just	new	oh	god	sure
	fucking	does	dog	time	seriously
Topic 3	reddit	day	thanks	comments	life
	think	going	got	best	really
Topic 4	fuck	love	think	fixed	right
	facebook	men	little	youtube	music
Topic 5	like	just	reddit	true	year
	feel	funny	story	old	girlfriend
Topic 6	did	know	good	hate	just
	cat	damn	moment	reddit	fuck
Topic 7	oh	just	level	day	reddit
	horse	fail	friends	real	kid
Topic 8	legit	need	cat	got	think
	shit	funny	kid	world	job
Topic 9	think	fixed	want	did	sfw
	game	people	new	know	cat
Topic 10	got	better	man	like	friend
	way	right	facebook	badass	reddit
Topic 11	make	day	guy	good	people
	feel	nice	google	said	just
Topic 12	facebook	saw	post	cat	thought
	friend	feel	makes	sure	today
Topic 13	reddit	like	work	friends	thing
	did	redditor	saw	guys	think
Topic 14	right	doing	friend	shit	look
	guy	people	say	video	girls
Topic 15	fixed	love	best	world	like
	time	friend	things	wait	funny
Topic 16	say	got	movie	face	new
	guys	old	true	guy	man
Topic 17	world	problems	did	fixed	lol
	porn	new	college	page	sex
Topic 18	nsfw	know	like	oh	dog
	reddit	thank	best	girl	today
Topic 19	time	just	bad	scumbag	cat
	sorry	x-post	joke	best	favorite
Topic 20	reddit	just	friend	time	tell
	facebook	feel	like	people	got

Table C.8.: The 20 discovered topics with LDA in r/gaming in the time of 2012

Topic 1	game	play	gaming	best	games
	just	help	steam	need	skyrim
Topic 2	gaming	just	good	video	time
	know	final	games	game	like
Topic 3	skyrim	game	xbox	gaming	love
	steam	live	does	help	new
Topic 4	game	new	games	pc	gaming
	skyrim	trailer	free	time	steam
Topic 5	game	gaming	think	just	time
	play	new	games	video	friend
Topic 6	game	steam	gaming	best	video
	online	think	like	new	just
Topic 7	game	help	just	gaming	pc
	reddit	time	skyrim	free	steam
Topic 8	game	steam	skyrim	gaming	video
	2012	best	like	got	games
Topic 9	games	video	minecraft	skyrim	new
	swtor	game	steam	play	2012
Topic 10	games	game	gaming	xbox	minecraft
	just	old	playing	skyrim	like
Topic 11	games	game	just	gaming	got
	video	time	help	think	skyrim
Topic 12	games	game	super	steam	online
	know	mario	pc	world	poker
Topic 13	gaming	play	games	new	playing
	game	dead	friend	just	best
Topic 14	games	game	gaming	xbox	video
	like	steam	skyrim	online	playing
Topic 15	gaming	like	just	game	skyrim
	video	play	games	used	good
Topic 16	game	games	new	gaming	best
	steam	skyrim	does	online	time
Topic 17	game	play	just	best	day
	games	new	skyrim	like	online
Topic 18	game	fixed	playing	like	video
	did	just	nintendo	right	skyrim
Topic 19	game	play	video	playing	games
	new	like	remember	steam	skyrim
Topic 20	game	steam	gaming	video	help
	like	games	just	thought	sopa

Table C.9.: The 20 discovered topics with LDA in r/leagueoflegends in the time of 2012

Topic 1	elo kiev	game people	ranked hell	stream did	iem play
Topic 2	game like	stream champion	ranked just	team lol	tsm idea
Topic 3	game need	streaming stream	ranked play	league jungle	team vs
Topic 4	league lol	vs champion	legends riot	sejuani tsm	new think
Topic 5	just vs	elo game	ranked tsm	hell saintvicious	lol people
Topic 6	new league	game skin	stream just	play looking	games champion
Topic 7	game think	lol ap	vs mid	new league	just friend
Topic 8	league lee	game stream	legends support	vs sin	does riot
Topic 9	lol dyrus	game kiev	games like	play know	league good
Topic 10	playing like	lol legends	league best	vs elo	ranked game
Topic 11	lol iem	elo stream	kiev support	league eu	clg legends
Topic 12	lol league	new iem	like kiev	play team	player post
Topic 13	elo queue	riot solo	stream streaming	ranked time	league team
Topic 14	new streaming	help ranked	support champion	solo lol	legends league
Topic 15	just solo	elo queue	game new	support playing	league lol
Topic 16	league play	legends best	team new	streaming m5	lol vs
Topic 17	league lol	legends vs	elo question	hell just	game help
Topic 18	lol players	game riot	help legends	league stream	new sejuani
Topic 19	riot games	lol streaming	champion game	does league	like playing
Topic 20	new support	champion video	time jungle	lol games	skin come

Table C.10.: The 20 discovered topics with LDA in r/Minecraft in the time of 2012

Topic 1	minecraft	new	server	play	build
	world	let	problem	lets	smp
Topic 2	minecraft	mod	play	build	new
	help	need	let	world	just
Topic 3	minecraft	build	help	need	mod
	server	reddit	think	map	episode
Topic 4	minecraft	like	know	guys	make
	play	think	help	jungle	time
Topic 5	minecraft	help	just	world	spawn
	reddit	make	blocks	bug	creative
Topic 6	minecraft	server	got	mod	just
	make	map	looking	build	mojang
Topic 7	minecraft	server	build	pack	texture
	new	just	tree	want	make
Topic 8	minecraft	build	map	mob	survival
	did	new	play	help	think
Topic 9	mod	minecraft	redstone	help	new
	idea	request	making	friend	server
Topic 10	minecraft	just	server	mod	new
	build	play	suggestion	video	survival
Topic 11	minecraft	think	new	server	like
	play	good	let	episode	just
Topic 12	minecraft	server	like	mod	world
	build	guys	just	skin	new
Topic 13	minecraft	server	reddit	survival	new
	think	does	fun	creative	jungle
Topic 14	minecraft	world	creative	make	like
	know	build	need	help	mod
Topic 15	minecraft	new	survival	world	know
	jeb	like	good	caves	just
Topic 16	server	minecraft	idea	skin	new
	build	make	episode	seed	little
Topic 17	minecraft	new	build	jungle	world
	just	think	server	suggestion	biome
Topic 18	minecraft	new	server	help	map
	world	redstone	village	spawn	need
Topic 19	minecraft	server	way	map	idea
	caves	spellbound	mod	help	episode
Topic 20	minecraft	server	build	new	just
	help	house	map	play	world

Table C.11.: The 20 discovered topics with LDA in r/Music in the time of 2012

Topic 1	song	cover	music	best	band
	love	know	rock	world	good
Topic 2	music	love	new	song	cover
	live	free	time	rock	album
Topic 3	music	video	new	cover	like
	band	love	amazing	think	song
Topic 4	live	new	good	band	songs
	great	just	music	video	love
Topic 5	song	video	reddit	live	new
	help	rock	band	know	like
Topic 6	song	music	new	think	just
	best	listen	world	guys	friend
Topic 7	music	band	reddit	video	friend
	time	think	song	best	2011
Topic 8	music	band	new	like	video
	just	songs	reddit	album	best
Topic 9	song	music	video	love	best
	remix	feel	reddit	amazing	live
Topic 10	music	song	new	video	band
	album	just	reddit	friend	live
Topic 11	music	new	song	cover	best
	like	album	remix	need	think
Topic 12	music	new	band	video	love
	know	song	cover	great	reddit
Topic 13	love	new	album	music	like
	video	james	help	black	time
Topic 14	new	band	reddit	music	like
	album	2012	song	live	cover
Topic 15	music	song	cover	like	know
	video	help	new	years	best
Topic 16	music	like	songs	video	rock
	song	love	listen	guitar	new
Topic 17	music	cover	just	album	video
	like	live	song	new	guitar
Topic 18	music	song	new	like	cover
	rock	video	live	band	amazing
Topic 19	love	music	song	band	video
	man	best	live	heard	just
Topic 20	music	video	cover	band	lyrics
	live	good	love	remix	free

Table C.12.: The 20 discovered topics with LDA in r/pics in the time of 2012

Topic 1	photo	reddit	day	picture	like
	just	year	old	work	cake
Topic 2	just	reddit	like	got	guy
	right	think	live	awesome	time
Topic 3	time	reddit	think	today	like
	little	new	feel	car	really
Topic 4	picture	new	cat	got	guys
	guy	like	time	just	life
Topic 5	reddit	right	think	just	new
	friend	doing	like	need	fun
Topic 6	just	picture	today	great	reddit
	saw	right	art	think	facebook
Topic 7	just	did	day	love	world
	know	think	cake	real	people
Topic 8	just	new	think	like	face
	picture	year	reddit	fuck	cat
Topic 9	know	new	true	day	best
	just	art	today	reddit	right
Topic 10	got	just	reddit	today	friend
	love	life	took	make	forever
Topic 11	like	just	reddit	day	new
	birthday	friend	make	look	night
Topic 12	friend	picture	just	like	reddit
	day	want	best	old	got
Topic 13	friend	reddit	think	work	make
	best	just	picture	cat	feel
Topic 14	like	just	cat	reddit	today
	picture	sopa	oh	awesome	year
Topic 15	reddit	friend	just	like	got
	today	think	nice	kitty	world
Topic 16	reddit	look	today	like	little
	year	thought	just	picture	know
Topic 17	think	just	today	does	day
	reddit	friend	cake	life	photo
Topic 18	work	doing	good	right	reddit
	morning	time	new	friend	just
Topic 19	reddit	friend	night	day	new
	best	saw	fixed	happy	just
Topic 20	got	really	day	like	new
	reddit	friend	facebook	cake	love

Table C.13.: The 20 discovered topics with LDA in r/politics in the time of 2012

Topic 1	sopa	santorum	rick	romney	ndaa
	republican	obama	people	new	support
Topic 2	sopa	obama	gingrich	paul	president
	romney	ron	2012	mitt	campaign
Topic 3	sopa	paul	santorum	ron	obama
	state	new	people	did	internet
Topic 4	obama	newt	paul	ron	sopa
	gingrich	2012	romney	like	mitt
Topic 5	paul	ron	romney	santorum	colbert
	america	mitt	gop	party	government
Topic 6	romney	paul	obama	ron	gingrich
	sopa	mitt	republican	newt	santorum
Topic 7	obama	sopa	paul	ron	pipa
	gop	new	vote	political	santorum
Topic 8	romney	paul	gingrich	new	ron
	state	obama	president	iowa	union
Topic 9	sopa	pipa	paul	president	ron
	government	new	obama	reddit	war
Topic 10	paul	ron	romney	mitt	sopa
	just	reddit	santorum	new	gop
Topic 11	obama	romney	sopa	reddit	newt
	gingrich	new	santorum	like	paul
Topic 12	obama	paul	new	ron	santorum
	tax	congress	rick	president	gingrich
Topic 13	paul	ron	sopa	romney	santorum
	new	rick	debate	support	gingrich
Topic 14	romney	paul	ron	mitt	obama
	sopa	pipa	people	gingrich	right
Topic 15	state	romney	paul	ron	gingrich
	santorum	says	obama	occupy	iowa
Topic 16	sopa	romney	obama	paul	gingrich
	santorum	ron	pipa	republican	newt
Topic 17	paul	santorum	sopa	ron	rick
	obama	people	romney	internet	think
Topic 18	paul	ron	romney	people	sopa
	republican	santorum	obama	mitt	vote
Topic 19	obama	romney	iowa	gingrich	america
	mitt	government	newt	santorum	caucus
Topic 20	obama	paul	ron	need	people
	romney	gop	states	just	court

Table C.14.: The 20 discovered topics with LDA in r/technology in the time of 2012

Topic 1	google android	phone apple	website sopa	new software	free development
Topic 2	iphone blog	technology business	blackout facebook	online free	seo windows
Topic 3	new sopa	review online	android tablet	2012 facebook	tech wikipedia
Topic 4	sopa business	design apple	video new	social pc	computer pipa
Topic 5	sopa megaupload	google online	company services	seo pipa	new apple
Topic 6	new technology	2012 design	best ces	web social	future phone
Topic 7	google twitter	mobile world	2012 internet	sopa vs	online free
Topic 8	web hosting	india mobile	computer services	new tablet	use price
Topic 9	development video	web company	software best	facebook website	apple ipad
Topic 10	mobile search	iphone ipad	social 2012	data software	media google
Topic 11	sopa facebook	internet car	apple india	services service	google reddit
Topic 12	free new	android sopa	app web	megaupload facebook	google phone
Topic 13	ipad megaupload	mobile home	solar download	iphone services	video mac
Topic 14	design google	data free	services online	india software	web new
Topic 15	new company	web software	design google	windows free	phone review
Topic 16	best video	iphone tablet	design real	apple news	new mobile
Topic 17	google 2011	sopa training	news web	design facebook	new computer
Topic 18	google iphone	2012 free	online make	facebook phone	development internet
Topic 19	sopa best	new facebook	internet 2012	google apple	pipa apps
Topic 20	sopa using	internet company	best business	pipa hosting	iphone marketing

Table C.15.: The 20 discovered topics with LDA in r/tf2trade in the time of 2012

Topic 1	ref refined	keys metal	strange festive	pc store	33 reclaimed
Topic 2	strange festive	ref rec	keys weapons	launcher scrap	metal crates
Topic 3	strange pc	ref scrap	vintage stranges	metal store	weapons paint
Topic 4	ref pc	metal buds	strange festive	keys hats	66 hat
Topic 5	metal keys	ref stranges	buds items	offers strange	hats weapons
Topic 6	refined pc	keys festive	strange unusual	genuine ref	metal valve
Topic 7	refined unusual	metal 33	ref strange	pc vintage	keys buds
Topic 8	ref scrap	keys offers	hat 33	strange metal	pc clean
Topic 9	ref offers	keys 33	metal hats	strange launcher	66 pc
Topic 10	metal launcher	keys festive	strange hats	offers inside	buds scrap
Topic 11	metal 66	ref unusual	keys 33	offers crates	strange launcher
Topic 12	ref buds	pc offers	strange hats	metal 33	unusual coal
Topic 13	ref hat	keys strange	scrap bills	weapons festive	metal refined
Topic 14	keys buds	pc offers	ref team	metal hats	unusual bills
Topic 15	keys hat	metal pc	strange ref	hats bills	refined weapons
Topic 16	keys buds	metal offers	strange hat	bills pc	ref 10
Topic 17	strange ref	offers buds	keys pc	refined rec	33 price
Topic 18	keys ref	strange crates	festive refined	metal holiday	weapons pc
Topic 19	keys pc	ref festive	strange rec	unusual $1	key paypal
Topic 20	buds rec	offers refined	ref metal	strange festive	key 33

Table C.16.: The 20 discovered topics with LDA in r/todayilearned in the time of 2012

Topic 1	til school	world google	make just	like actually	people money
Topic 2	til world	movie new	called people	like american	sex paul
Topic 3	til 10	used world	years called	man actually	people american
Topic 4	til actually	reddit free	like know	united money	states life
Topic 5	til years	used called	people new	world man	year named
Topic 6	til man	used learned	time make	today called	buffalo google
Topic 7	til people	new man	years wrote	year make	house old
Topic 8	til time	reddit make	world day	work home	called does
Topic 9	til people	make year	like years	free called	actually word
Topic 10	til years	just world	named life	year song	reddit states
Topic 11	til video	people new	called used	years known	make actually
Topic 12	til like	today day	make old	people black	learned fish
Topic 13	til women	black work	men named	movie used	people reddit
Topic 14	til world	people called	use car	word voice	new day
Topic 15	til air	people just	actually used	know movie	bacon line
Topic 16	til use	used years	new make	word reddit	today actually
Topic 17	til just	called book	new company	day war	world years
Topic 18	til people	called just	like new	used film	white man
Topic 19	til years	learned actually	called money	world year	today google
Topic 20	til english	world used	new man	people million	website free

Table C.17.: The 20 discovered topics with LDA in r/trees in the time of 2012

Topic 1	trees	ents	feel	just	know
	fellow	love	like	dealer	smoke
Topic 2	guy	ent	just	right	10
	like	munchies	guys	want	mflb
Topic 3	trees	new	ents	love	ent
	thought	just	guys	today	know
Topic 4	ents	new	smoke	trees	guys
	10	hey	spot	favorite	like
Topic 5	just	ents	think	world	ent
	good	marijuana	need	guy	got
Topic 6	ents	friend	trees	thought	good
	new	guy	got	feel	just
Topic 7	time	good	feel	trees	night
	post	smoke	know	use	bong
Topic 8	high	help	new	time	just
	day	happened	10	best	trees
Topic 9	like	trees	scumbag	ents	smoke
	time	10	best	smoked	really
Topic 10	weed	ent	ents	10	did
	trees	need	best	got	frient
Topic 11	trees	new	just	10	smoke
	guy	need	time	ents	ent
Topic 12	got	just	think	time	new
	piece	smoking	best	trees	today
Topic 13	ents	trees	just	smoke	guys
	time	got	like	enjoy	smoking
Topic 14	smoking	friend	know	love	trees
	high	think	10	bowl	marijuana
Topic 15	ents	night	trees	love	just
	song	got	good	friend	enjoy
Topic 16	just	trees	high	weed	10
	think	ents	new	like	post
Topic 17	just	like	trees	smoking	ents
	thing	ent	time	10	high
Topic 18	trees	like	ents	guys	thought
	time	help	just	need	new
Topic 19	smoke	like	day	guy	feel
	true	high	trees	pickle	friend
Topic 20	ents	just	like	high	trees
	good	post	guy	love	fuck

Table C.18.: The 20 discovered topics with LDA in r/videos in the time of 2012

Topic 1	new	awesome	girl	video	reddit
	funny	just	makes	love	guy
Topic 2	video	say	2012	shit	real
	awesome	old	did	reddit	guy
Topic 3	video	music	just	guy	song
	people	friend	think	love	like
Topic 4	think	reddit	youtube	video	man
	make	like	just	guy	shit
Topic 5	world	watch	video	reddit	baby
	best	little	think	nsfw	girl
Topic 6	youtube	new	just	reddit	year
	video	people	friend	trailer	kid
Topic 7	say	shit	video	man	new
	reddit	think	people	time	year
Topic 8	guy	best	video	watch	dog
	just	dubstep	know	right	good
Topic 9	video	just	think	reddit	like
	school	stop	sopa	help	friend
Topic 10	video	like	guy	new	friend
	man	little	thing	awesome	guys
Topic 11	amazing	level	kid	2012	love
	reddit	man	just	video	girl
Topic 12	like	video	reddit	song	nsfw
	know	really	just	friend	film
Topic 13	video	amazing	just	song	best
	cat	reddit	year	shit	commercial
Topic 14	video	man	just	asian	day
	best	school	love	cat	know
Topic 15	video	say	shit	girls	year
	new	friend	like	song	reddit
Topic 16	vs	like	song	funny	video
	guy	cat	girl	time	guys
Topic 17	video	just	like	watch	guy
	cat	reddit	check	wait	time
Topic 18	video	say	love	shit	amazing
	best	dog	guy	friend	new
Topic 19	dog	people	good	car	seen
	video	guitar	awesome	new	cat
Topic 20	video	man	best	life	shit
	know	gets	reddit	say	free

Table C.19.: The 20 discovered topics with LDA in r/worldnews in the time of 2012

Topic 1	2012	best	online	world	north
	year	news	new	says	facebook
Topic 2	people	new	iran	india	china
	business	ship	iranian	nuclear	2012
Topic 3	news	iran	new	2012	oil
	real	10	jobs	says	israel
Topic 4	india	life	new	insurance	online
	government	home	auto	good	man
Topic 5	iran	2012	new	year	american
	israel	ship	site	life	cruise
Topic 6	iran	new	india	oil	online
	video	china	2012	jobs	world
Topic 7	news	buy	nuclear	police	new
	day	online	iran	israel	car
Topic 8	new	syria	iran	news	india
	year	bbc	police	day	china
Topic 9	new	2012	dead	year	world
	syria	death	sopa	syrian	arab
Topic 10	online	video	2012	protest	iran
	china	world	internet	man	israel
Topic 11	online	iran	time	sopa	free
	new	news	2012	million	internet
Topic 12	news	online	new	vs	live
	2012	city	government	world	time
Topic 13	uk	news	new	year	marketing
	iran	killed	real	design	syria
Topic 14	online	world	new	2012	best
	news	help	iran	live	arab
Topic 15	iran	news	2012	online	new
	world	years	man	police	video
Topic 16	iran	new	video	live	china
	megaupload	attack	business	world	hotel
Topic 17	iranian	iran	online	oil	syria
	eu	2012	company	blog	south
Topic 18	iran	seo	says	new	services
	government	india	israel	acta	world
Topic 19	iran	nuclear	war	sopa	anonymous
	ship	dead	new	news	says
Topic 20	new	says	war	china	india
	sopa	iran	business	military	video

Table C.20.: The 20 discovered topics with LDA in r/WTF in the time of 2012

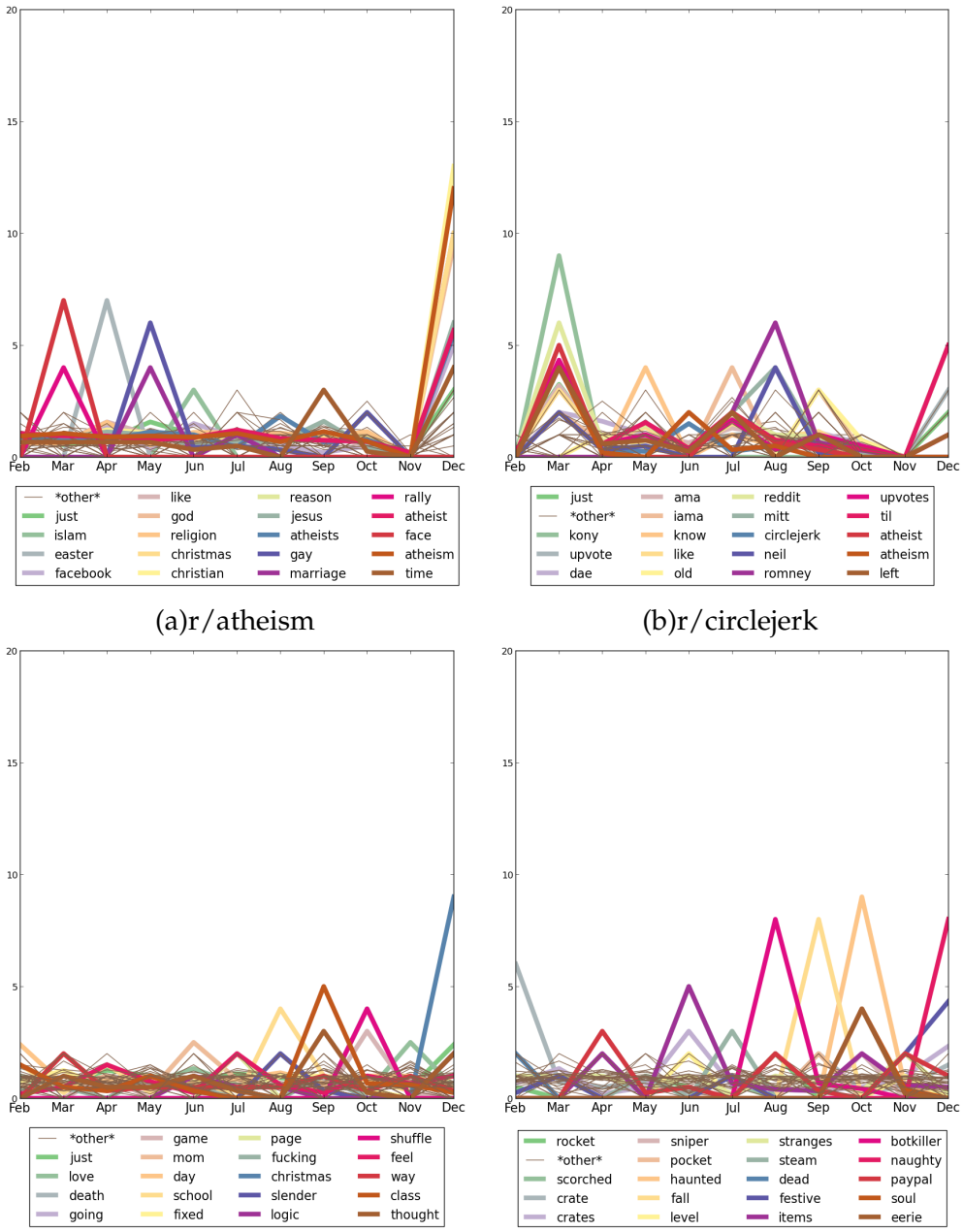
Topic 1	nsfw	wtf	people	want	nope
	reddit	like	think	today	time
Topic 2	wait	like	really	wtf	guy
	nsfw	shit	world	best	video
Topic 3	video	nsfw	wtf	sure	new
	say	happened	fuck	let	looking
Topic 4	nsfw	real	today	like	oh
	wtf	shit	feel	internet	love
Topic 5	wtf	good	day	friend	old
	best	man	words	nsfw	today
Topic 6	just	people	think	shit	day
	really	man	like	friend	look
Topic 7	just	like	fuck	nsfw	guy
	know	year	google	time	really
Topic 8	wtf	nsfw	just	love	nope
	man	nsfl	reddit	like	people
Topic 9	just	guy	new	wtf	right
	think	did	thing	facebook	dog
Topic 10	wtf	facebook	just	know	reddit
	seriously	time	like	fuck	got
Topic 11	wtf	like	man	got	just
	want	oh	fuck	know	sex
Topic 12	just	wtf	like	woman	thing
	looking	real	day	facebook	life
Topic 13	wtf	video	man	look	girl
	news	sure	nsfw	seen	google
Topic 14	wtf	nsfw	think	did	just
	man	youtube	reddit	know	like
Topic 15	like	just	man	nsfw	wtf
	going	did	guy	say	need
Topic 16	nsfw	just	youtube	new	wtf
	know	did	years	big	man
Topic 17	wtf	fuck	nsfw	girl	man
	world	thought	came	words	woman
Topic 18	nsfw	really	guy	people	video
	school	oh	wtf	new	know
Topic 19	wtf	just	new	facebook	wat
	oh	reddit	video	nsfw	think
Topic 20	just	people	wtf	like	man
	japan	facebook	asian	know	friends

Appendix D.

LDA Short Term Trends 2012

In this appendix chapter, the short term trend analysis figures of the remaining 20 largest subreddits of 2012 are listed, namely:

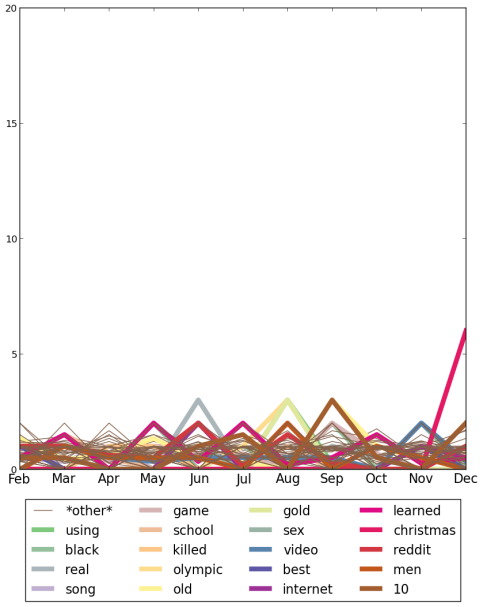
1. r/atheism
2. r/circlejerk
3. r/fffffffuuuuuuuuuuuu
4. r/tf2trade
5. r/todayilearned
6. r/trees
7. r/WTF



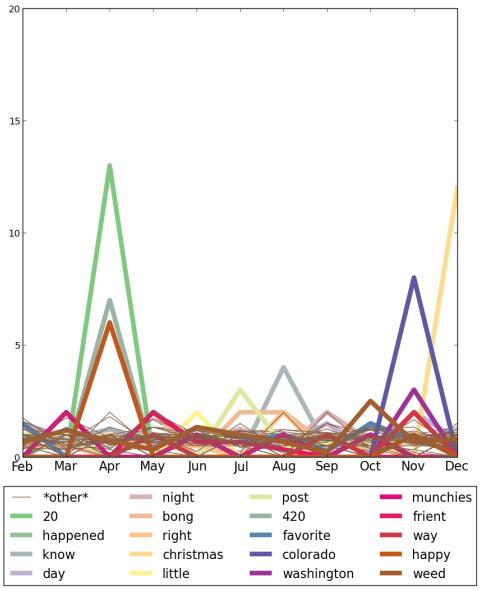
(a)r/atheism (b)r/circlejerk

(c)r/ffffffffuuuuuuuuuuuu (d)r/tf2trade

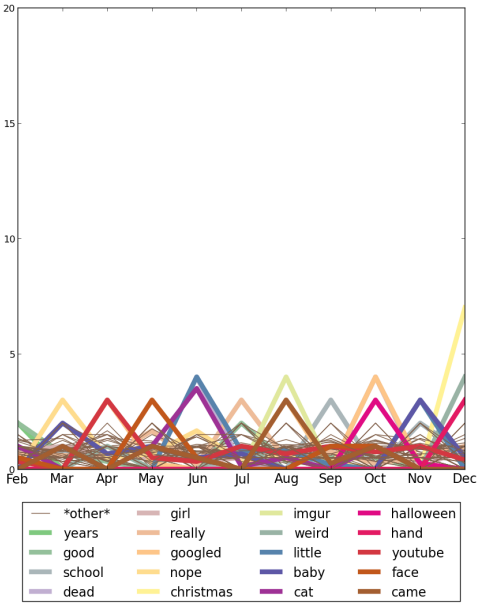
Figure D.1.: The trends of r/atheism, r/circlejerk, r/f7u12 and r/tf2trade



(a)r/todayilearned



(b)r/trees



(c)r/WTF

Figure D.2.: The trends of r/todayilearned, r/trees and r/WTF

Bibliography

Aizawa, Akiko (2003). "An information-theoretic perspective of tf-idf measures." In: *Information Processing & Management* 39.1, pp. 45–65. DOI: 10.1016/S0306-4573(02)00021-3 (cit. on p. 34).

Alfonso, Fernando (2013). *Reddit's r/news bans RT.com for alleged spamming*. URL: http://dailydot.com/news/rt-russia-today-banned-reddit-r-news/ (visited on 04/22/2014) (cit. on p. 54).

Altman, Alex (2012). *Obama Takes a First Step on Gun Control After Sandy Hook*. Time. URL: http://swampland.time.com/2012/12/19/obama-takes-a-first-step-on-gun-control-after-sandy-hook/ (visited on 02/15/2012) (cit. on p. 102).

ATD (2012). *How Reddit Began*. YouTube. Video interview. URL: http://youtu.be/5U-NCG1zZds (visited on 10/28/2013) (cit. on p. 7).

Bakshy, Eytan et al. (2011). "Everyone's an Influencer: Quantifying Influence on Twitter." In: *Proceedings of the fourth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, pp. 65–74. ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935845 (cit. on pp. 20, 21, 114).

Barron, James (2012). *Children Were All Shot Multiple Times With a Semi-automatic, Officials Say*. The New York Times. URL: http://nytimes.com/2012/12/16/nyregion/gunman-kills-20-children-at-school-in-connecticut-28-dead-in-all.html (visited on 02/15/2012) (cit. on p. 104).

Benevenuto, Fabrício et al. (2009). "Characterizing User Behavior in Online Social Networks." In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. IMC '09. New York, NY, USA: ACM, pp. 49–62. ISBN: 978-1-60558-771-4. DOI: 10.1145/1644893.1644900 (cit. on p. 26).

Bernstein, Michael S et al. (2011). "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community." In: *Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press. ISBN: 978-1-57735-505-2. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2873 (cit. on pp. 5, 23, 24).

Berry, Michael W. and Murray Browne (2005). *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. 2nd ed. SIAM, Society for Industrial and Applied Mathematics. ISBN: 0-89871-581-4 (cit. on p. 57).

Blei, David M. (2012). "Introduction to Probabilistic Topic Models." In: *Communications of the ACM* 55, pp. 77–84. DOI: 10.1145/2133806.2133826 (cit. on p. 32).

Blei, David M. and John Lafferty (2009). "Topic Models." In: *Text Mining: Classification, Clustering, and Applications*. Ed. by Ashok N. Srivastava and Mehran Sahami. London, England: Taylor & Francis. ISBN: 978-1420059403. URL: http://dl.acm.org/citation.cfm?id=1571651 (cit. on p. 32).

Blei, David M. and Jon D. McAuliffe (2007). "Supervised Topic Models." In: *Advances in Neural Information Processing Systems*. Ed. by John C. Platt et al. Vol. 20. Curran Associates, Inc. URL: http://papers.nips.cc/paper/3328-supervised-topic-models (cit. on pp. 27, 33).

Blei, David M., Andrew Y. Ng, et al. (2003). "Latent Dirichlet Allocation." In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: http://jmlr.org/papers/volume3/blei03a/blei03a.pdf (cit. on pp. 32, 54, 55, 58–60).

Booth, Robert, Helene Mulholland, and Patrick Strudwick (2012). *Anti-gay adverts pulled from bus campaign by Boris Johnson*. The Guardian. URL: http://theguardian.com/world/2012/apr/12/anti-gay-adverts-boris-johnson (visited on 02/15/2012) (cit. on p. 101).

Borger, Julian and Mona Mahmood (2012). *Syrian troops bombard sealed-off suburb of Homs*. The Guardian. URL: http://theguardian.com/world/2012/feb/09/syrian-troops-suburb-homs (visited on 02/15/2012) (cit. on p. 102).

CGPGrey (2013). *What is Reddit?* YouTube. video. URL: http://youtu.be/tlI022aUWQQ (visited on 10/28/2013) (cit. on pp. 7, 8).

CNN (2012). *Timeline of Iran's controversial nuclear program*. CNN. URL: http://edition.cnn.com/2012/03/06/world/meast/iran-timeline/ (visited on 02/15/2012) (cit. on p. 95).

Collier, Kevin (2014). *Reddit mods are censoring dozens of words from r/technology posts*. URL: http://dailydot.com/news/reddit-technology-banned-words/ (visited on 04/22/2014) (cit. on p. 54).

Deerwester, Scott et al. (1990). "Indexing by Latent Semantic Analysis." In: *Journal of the American Society for Information Science* 41.6, pp. 391–407. URL: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.8490 (cit. on pp. 32, 58).

Dieberger, A. et al. (2000). "Social Navigation - Techniques for Building More Usable Systems." In: *Interactions* 6, pp. 36–45. DOI: 10.1145/352580.352587 (cit. on p. 31).

Dixon, Philip M. et al. (1987). "Bootstrapping the Gini Coefficient of Inequality." In: *Ecology* 68, pp. 1548–1551. DOI: 10.2307/1939238 (cit. on p. 46).

Dourish, Paul and Matthew Chalmers (1994). "Running out of space: Models of Information Navigation." In: *Proceedings of HCI'94*. URL: http://www.dourish.com/publications/1994/hci94-navigation.pdf (cit. on pp. 9, 30).

Duan, Yajuan et al. (2010). "An Empirical Study on Learning to Rank of Tweets." In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Association for Computational Linguistics, pp. 295–303. URL: http://dl.acm.org/citation.cfm?id=1873781.1873815 (cit. on p. 20).

Duggan, Maeve and Aaron Smith (2013). *6% of Online Adults are reddit Users*. Pew Research Center's Internet & American Life Project. URL: http://pewinternet.org/~/media//Files/Reports/2013/PIP_reddit_usage_2013.pdf (visited on 10/28/2013) (cit. on p. 1).

Edberg, Jeremy (2009). *Moving to the Cloud*. reddit. URL: http://redditblog.com/2009/11/moving-to-cloud.html (visited on 10/28/2013) (cit. on p. 8).

Ferran, Lee and Raisa Bruner (2012). *Ecuador Grants WikiLeaks Founder Julian Assange Political Asylum*. ABC News. URL: http://abcnews.go.com/Blotter/ecuador-grants-wikileaks-founder-assange-political-asylum/story?id=17018133 (visited on 02/15/2012) (cit. on p. 104).

Finn, Greg (2010). *Digg v4: How To Successfully Kill A Community*. Search Engine Land. URL: http://searchengineland.com/digg-v4-how-to-successfully-kill-a-community-50450 (visited on 10/28/2013) (cit. on p. 68).

Friedman, Megan (2010). *Digg Users Lash Out At New Format, Join Forces with Reddit*. Time. URL: http://newsfeed.time.com/2010/08/30/digg-users-lash-out-at-new-format-join-forces-with-reddit/ (visited on 10/28/2013) (cit. on p. 68).

Fung, Katherine (2012). *Rush Limbaugh On Sandra Fluke, Obama Call: Having 'So Much Sex'; Parents Should Be 'Embarrassed'*. The Huffington Post. URL: http://huffingtonpost.com/2012/03/02/rush-limbaugh-sandra-fluke-sex-slut_n_1316625.html (visited on 02/15/2012) (cit. on p. 101).

Gaudiosi, John (2012). *Riot Games' League Of Legends Officially Becomes Most Played PC Game In The World*. URL: http://forbes.com/sites/johngaudiosi/2012/07/11/riot-games-league-of-legends-officially-becomes-most-played-pc-game-in-the-world/ (visited on 02/20/2014) (cit. on p. 93).

Gilbert, Eric (2013). "Widespread Underprovision on Reddit." In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. New York, NY, USA: ACM, pp. 803–808. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441866 (cit. on pp. 28, 29).

Gini, Corrado (1912). "Variabilita e mutabilita contributo allo studio della distribuzioni." In: *Studie Economico-Guiridici della R. Universita di Cagliari* (cit. on p. 46).

Gómez, Vicenç, Andreas Kaltenbrunner, and Vicente López (2008). "Statistical Analysis of the Social Network and Discussion Threads in Slashdot." In: *WWW '08 Proceedings of the 17th international conference on World Wide Web*. Ed. by Jinpeng Huai et al. ACM, pp. 645–654. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367585 (cit. on p. 25).

Gruger, William (2012). *PSY's 'Gangnam Style' Video Hits 1 Billion Views, Unprecedented Milestone*. Billboard biz. URL: http://billboard.com/biz/articles/news/1483733/psys-gangnam-style-video-hits-1-billion-views-unprecedented-milestone (visited on 02/15/2012) (cit. on p. 101).

Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp (2011). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions." In: *SIAM Review* 53.2, pp. 217–288. DOI: 10.1137/090771806 (cit. on p. 61).

Hardin, Garreth (1968). "The Tragedy of the Commons." In: *Science* 162. DOI: 10.1126/science.162.3859.1243 (cit. on p. 28).

Hoffman, Matthew D., David M. Blei, and Francis R. Bach (2010). "Online Learning for Latent Dirichlet Allocation." In: *Advances in Neural Information Processing Systems*. Vol. 23. Curran Associates, Inc., pp. 856–864. URL: http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation (cit. on p. 61).

Hofmann, Thomas (1999). "Probabilistic Latent Semantic Analysis." In: *Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*. San Francisco, California, USA: Morgan Kaufmann, pp. 289–296. ISBN: 1-55860-614-9. URL: http://uai.sis.pitt.edu/papers/99/p289-hofmann.pdf (cit. on pp. 32, 54, 57).

Höök, Kristina, David Benyon, and Alan J. Munro (2003). *Designing Information Spaces: The Social Navigation Approach*. Springer. ISBN: 978-1-4471-0035-5 (cit. on p. 31).

Huffman, Steve (2005). *On Lisp*. reddit. URL: http://redditblog.com/2005/12/on-lisp.html (visited on 10/28/2013) (cit. on p. 8).

Huffman, Steve (2008). *Reddit Goes Open Source*. reddit. URL: http://redditblog.com/2008/06/reddit-goes-open-source.html (visited on 10/28/2013) (cit. on p. 8).

Jakić, Bruno (2012). "Predicting Sentiment of Comments to News on Reddit." MA thesis. University of Amsterdam, Intelligent Systems Lab Amsterdam. URL: http://scribd.com/doc/103712449/Predicting-Sentiment-of-Comments-to-News-on-Reddit (visited on 07/30/2013) (cit. on pp. 2, 29).

Johnston, Casey (2012). *Reddit founders made hundreds of fake profiles so site looked popular*. URL: http://arstechnica.com/business/2012/06/reddit-founders-made-hundreds-of-fake-profiles-so-site-looked-popular/ (visited on 04/18/2014) (cit. on p. 3).

Kravets, David (2012). *Feds Shutter Megaupload, Arrest Executives*. Wired. URL: http://wired.com/threatlevel/2012/01/megaupload-indicted-shuttered (visited on 02/15/2012) (cit. on p. 95).

Kullback, Solomon and Richard A. Leibler (1951). "On Information and Sufficiency." In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694 (cit. on p. 43).

Kurzweil, Ray (2005). *The Singularity is Near. When Humans Transcent Biology*. Viking Penguin, a member of Penguin Group (USA) Inc. ISBN: 0-670-03384-7 (cit. on pp. 31, 41, 42).

Kwak, Haewoon et al. (2010). "What is Twitter, a Social Network or a News Media?" In: *WWW'10: Proceedings of the 19th International*

World Wide Web Conference. Raleigh, North Carolina, USA: ACM, pp. 591–600. DOI: 10.1145/1772690.1772751 (cit. on pp. 3, 5, 19, 20, 105).

Lakkaraju, Himabindu, Julian McAuley, and Jure Leskovec (2013). "What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media." In: *Seventh International AAAI Conference on Weblogs and Social Media*. The AAAI Press. ISBN: 978-1-57735-610-3. URL: http://aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6085 (cit. on pp. 27, 28).

Lerman, Kristina (2006). "Social Networks and Social Information Filtering on Digg." In: *The Computing Research Repository* abs/cs/0612046. URL: http://arxiv.org/abs/cs/0612046 (cit. on pp. 2, 24, 25).

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg (2009). "Meme-Tracking and the Dynamics of the News Cycle." In: *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris, France: ACM, pp. 497–506. ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557077 (cit. on p. 26).

Liptak, Adam (2012). *Supreme Court Upholds Health Care Law, 5-4, in Victory for Obama*. The New York Times. URL: http://nytimes.com/2012/06/29/us/supreme-court-lets-health-law-largely-stand.html (visited on 02/15/2012) (cit. on p. 101).

Malthus, Thomas Robert (1826). *An Essay on the Principle of Population*. Sixth edition. John Murray, London (cit. on p. 41).

Martin, Erik (2011a). *Independence*. reddit. URL: http://redditblog.com/2011/09/independence.html (visited on 10/28/2013) (cit. on p. 7).

Martin, Erik (2011b). *Saying Goodbye to an Old Friend and Revising the Default Subreddits*. reddit. URL: http://redditblog.com/2011/10/saying-goodbye-to-old-friend-and.html (visited on 11/04/2013) (cit. on pp. 17, 75).

Meadows, Donella H. et al. (1972). *The Limits to Growth*. New York: Universe Books. ISBN: 0-87663-165-0 (cit. on pp. 41, 43).

Mieghem, Piet Van (2011). "Human Psychology of Common Appraisal: The Reddit Score." In: *IEEE Transactions on Multimedia* 13.6, pp. 1404–1406. URL: http://dblp.uni-trier.de/db/journals/tmm/tmm13.html#Mieghem11 (cit. on pp. 2, 9, 18, 27).

Milian, Mark (2010). *Reddit considers itself a benefactor of Digg user revolt*. Los Angeles Times. URL: http://latimesblogs.latimes.com/technology/2010/08/reddit-digg.html (visited on 10/28/2013) (cit. on p. 68).

Nielsen, Jakob (2006). *Participation Inequality: Encouraging More Users to Contribute*. Nielsen Norman Group. URL: http://nngroup.com/articles/participation-inequality/ (visited on 10/28/2013) (cit. on p. 11).

Nonnecke, Blair and Jenny Preece (2000). "Lurker Demographics: Counting the Silent." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '00. The Hague, The Netherlands: ACM, pp. 73–80. ISBN: 1-58113-216-6. DOI: 10.1145/332040.332409 (cit. on p. 11).

Olson, Randal S. (2013). *Retracing the evolution of Reddit through post data*. URL: http://dx.doi.org/10.6084/m9.figshare.650851 (visited on 02/12/2014) (cit. on p. 30).

Pak, Alexander and Patrick Paroubek (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In: URL: http://www.lrec-conf.org/proceedings/lrec2010/summaries/385.html (cit. on pp. 21, 30, 114).

reddit (2010). *Reddit's May 2010 "State of the Servers" report*. reddit. URL: http://redditblog.com/2010/05/reddits-may-2010-state-of-servers.html (visited on 10/28/2013) (cit. on p. 66).

reddit (2013). *Reddit Myth Busters*. reddit. URL: http://redditblog.com/2013/08/reddit-myth-busters_6.html (visited on 10/28/2013) (cit. on p. 8).

Řehůřek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora." In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50. URL: http://is.muni.cz/publication/884893/en (cit. on p. 61).

Rowe, Matthew, Sofia Angeletou, and Harith Alani (2011). "Predicting Discussions on the Social Semantic Web." In: *ESWC'11 Proceedings of the 8th extended semantic web conference on The semanic web: research and applications*. Heraklion, Crete: Springer, pp. 405–420. ISBN: 978-3-642-21063-1. URL: http://dl.acm.org/citation.cfm?id=2017969 (cit. on pp. 21, 22, 114).

Salton, Gerard and Christopher Buckley (1988). "Term Weighting Approaches in Automatic Text Retrieval." In: *Information Processing and Management: an International Journal* 24, pp. 513–523. DOI: 10.1016/0306-4573(88)90021-0 (cit. on pp. 33, 55).

Sammut, Claude and Geoffrey I. Webb, eds. (2010). *Encyclopedia of Machine Learning*. Springer. ISBN: 978-0-387-30768-8. DOI: 10.1007/978-0-387-30164-8 (cit. on pp. 54, 61, 62).

Schiraldi, Mike (2010a). *"Experts" misunderestimate our traffic, and we don't know why*. reddit. URL: http://redditblog.com/2010/07/experts-misunderestimate-our-traffic.html (visited on 10/28/2013) (cit. on p. 66).

Schiraldi, Mike (2010b). *reddit needs help*. reddit. URL: http://redditblog.com/2010/07/reddit-needs-help.html (visited on 10/28/2013) (cit. on p. 66).

Schonfeld, Erick (2008). *Facebook Is Not Only The World's Largest Social Network, It Is Also The Fastest Growing*. TechCrunch. URL: http://techcrunch.com/2008/08/12/facebook-is-not-only-the-worlds-largest-social-network-it-is-also-the-fastest-growing/ (visited on 01/04/2014) (cit. on p. 26).

Singer, Philipp et al. (2014). "Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?" In: *WWW'14: Proceedings of the 19th International World Wide Web Conference*. Seoul, South Korea: ACM. URL: http://arxiv.org/abs/1402.1386 (cit. on pp. 6, 108, 114).

Slowe, Christopher (2010a). *Everything went better than expected*. reddit. URL: http://redditblog.com/2010/08/everything-went-better-than-expected.html (visited on 10/28/2013) (cit. on p. 66).

Slowe, Christopher (2010b). *This was a triumph.* reddit. URL: http://redditblog.com/2010/07/it-was-triumph.html (visited on 10/28/2013) (cit. on p. 66).

Spiliotopoulos, Tasos and Ian Oakley (2013). "Understanding Motivations for Facebook Use: Usage Metrics, Network Structure, and Privacy." In: pp. 3287–3296. DOI: 10.1145/2470654.2466449 (cit. on pp. 22, 23).

Steyvers, Mark and Tom Griffiths (2005). "Latent Semantic Analysis: A Road to Meaning." In: *Latent Semantic Analysis: A Road to Meaning.* URL: http://faculty.washington.edu/jwilker/559/SteyversGriffiths.pdf (cit. on pp. 33, 57).

Suh, Bongwon et al. (2009). "The Singularity is not near: Slowing Growth of Wikipedia." In: *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. New York, NY, USA: ACM, pp. 1–10. ISBN: 978-1-60558-730-1. DOI: 10.1145/1641309.1641322 (cit. on pp. 4, 31, 32, 40–42).

Tuulos, Ville H. and Henry Tirri (2004). "Combining Topic Models and Social Networks for Chat Data Mining." In: *Web Intelligence*. IEEE Computer Society, pp. 206–213. ISBN: 0-7695-2100-2. URL: http://dblp.uni-trier.de/db/conf/webi/webi2004.html#TuulosT04 (cit. on p. 33).

Vascellaro, Jessica E. (2012). *Apple Wins Big in Patent Case*. The Wall Street Journal. URL: http://on.wsj.com/SyBH8v (visited on 02/15/2012) (cit. on p. 102).

Weninger, Tim, Xihao Avi Zhu, and Jiawei Han (2013). "An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community." In: *Proceedings of the 2013 IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM 2013)*, pp. 579–583. ISBN: 978-1-4503-2240-9. DOI: 10.1145/2492517.2492646 (cit. on pp. 2, 28).

Wortham, Jenna (2012). *Public Outcry Over Antipiracy Bills Began as Grass-Roots Grumbling*. The New York Times. URL: http://nytimes.com/2012/01/20/technology/public-outcry-over-antipiracy-bills-began-as-grass-roots-grumbling.html (visited on 02/15/2012) (cit. on p. 92).

Zafar, Aylin (2011). *Student Posts Live Reddit Q&A During Virginia Tech Lockdown*. Time. URL: http://newsfeed.time.com/2011/12/10/student-posts-live-reddit-qa-during-virginia-tech-lockdown/ (visited on 03/01/2014) (cit. on p. 109).

Zeleny, Jeff (2012). *Romney Chooses Ryan, Pushing Fiscal Issues to the Forefront*. The New York Times. URL: http://nytimes.com/2012/08/12/us/politics/mitt-romney-names-paul-ryan-as-his-running-mate.html (visited on 02/15/2012) (cit. on p. 102).

Zuckerman, Ethan (2013). *Reddit: A Pre-Facebook Community in a Post-Facebook World*. The Atlantic. URL: http://theatlantic.com/technology/archive/2013/07/reddit-a-pre-facebook-community-in-a-post-facebook-world/277583/ (visited on 01/04/2014) (cit. on p. 1).