LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Department "Institut für Informatik"
Professur für Computational Social Science and Big Data
Prof. Jürgen Pfeffer

Masterarbeit

Not all those who wander are lost

Dynamiken bei der Interessensentwicklung in Online Communities

Oliver Baumann
<baumanno@cip.ifi.lmu.de>

Bearbeitungszeitraum:	30.04.2018 bis 29.10.2018
Betreuer:	Dr. Mirco Schönfeld
Verantw. Hochschullehrer:	Prof. Jürgen Pfeffer

Zusammenfassung

Die vorliegende Arbeit reiht sich in Forschungsliteratur zu interaktiven Tischen, interaktiven Arbeitsumgebungen, gekrümmten Multitouch-Displays und indirekten Multitouch-Mappings ein. Anhand einer Nutzerstudie wird die Wirkung zweier indirekter Eingabemodi auf den Nutzer untersucht. Dazu wurde für *Curve*, ein interaktiver Tisch mit gebogenem Display, eine prototypische Anwendung entwickelt, die entweder mit einer Maus oder über Multitouch-Gesten bedient werden kann. Im Gegensatz zu isolierten Tasks ermöglicht die Anwendung den von einer Desktopumgebung gewohnten Arbeitsablauf. Das System bietet für den Anwendungsfall "'Audio-Bearbeitung"' die Möglichkeit, in einem Audio-Sample zu navigieren und dieses zu modifizieren. Die beiden Interface-Varianten wurden auf ihre Wirkung auf das Nutzererlebnis und ihre Eignung zum Einsatz in interaktiven Arbeitsplätzen hin untersucht. Es wurde festgestellt, dass keine der beiden Varianten dabei übermäßig gut oder schlecht abschneidet. Beide Eingabetechniken sind dabei ähnlich gut für den speziellen Anwendungsfall geeignet. Ein Transfer zu anderen Einsatzmöglichkeiten schließt die Arbeit ab. Es sei darauf hingewiesen, dass die in dieser Studie präsentierten Ergebnisse anhand einer kleinen Stichprobe ermittelt wurden und möglicherweise nicht vollends generalisierbar sind.

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

. . . . . .

München, 26. November 2018

Inhaltsverzeichnis

1 Einleitung

> The world is a thing of utter inordinate complexity and richness and strangeness that is absolutely awesome.
>
> *Douglas Adams*

2 Grundlagen und verwandte Forschung

2.1 Topic-Modelle

2.1.1 LDA

2.1.2 Verwandte Arbeiten

2.2 Soziale Netzwerkanalyse

2.2.1 Graphen und Netzwerke

2.2.2 Ego-Netzwerke

2.2.3 Verwandte Arbeiten

2.3 Reddit

2.3.1 Begriffsklärung

2.3.2 Verwandte Arbeiten

Tabelle 3.1: wichtige Schlüssel-Wert-Paare des Datensatzes

Schlüssel	Wert
author	Nutzername des Kommentar-Erstellers
id	eindeutige ID des Kommentars
parent_id	eindeutige ID des Elements, auf das sich der Kommentar bezieht
subreddit	Name des Subreddits, in dem der Kommentar erstellt wurde

3 Datenanalyse

Dieses Kapitel liefert einen Überblick über die Methodik sowie die Ergebnisse der Datenanalyse. Zunächst wird der verwendete Datensatz präsentiert und Kritik daran erörtert. Weiterhin wird dargelegt, wie die betrachteten Topic-Modelle erzeugt werden und welche Methoden der sozialen Netzwerkanalyse Anwendung finden, sowie welche Software-Komponenten jeweils zum Einsatz kommen. Im zweiten Teil des Kapitels werden dann die Ergebnisse vorgestellt, ohne dabei jedoch einer Interpretation zu weit vorzugreifen.

3.1 Methodik

3.1.1 Datensatz

Die Grundlage der Analyse bildet ein frei zugänglicher Datensatz mit Reddit-Kommentaren. Jason Baumgartner, der unter dem Pseudonym *stuck_in_the_matrix*[1] selbst auf Reddit aktiv ist, stellt monatliche Zusammenfassungen aller erstellten Kommentare zum Download bereit [3]. Diese reichen zum gegenwärtigen Zeitpunkt von Oktober 2018 zurück bis Dezember 2005.

Struktur Die monatlichen Datensätze liegen in Form von Textdateien vor, in denen jede Zeile einen Kommentar sowie Metadaten enthält. Das maschinenlesbare JSON-Format, in dem die Daten abgelegt sind, ermöglicht dabei eine effiziente computergestützte Auswertung. Tabelle 3.1 führt die für diese Arbeit relevanten Schlüssel-Wert-Paare des Datensatzes auf. Der Schlüssel *parent_id* bezeichnet dabei das Element, auf welches sich der Kommentar bezieht. Dies können Beiträge oberster Ordnung sein, sog. „Links“, oder selbst Kommentare [8]. Zu beachten ist hier insbesondere, dass der eigentliche Textinhalt des Kommentars für diese Auswertung nicht genutzt wird.

Kohärenz des Datensatzes Im März 2018 haben Gaffney und Matias eine Analyse des Baumgartner-Korpus vorgelegt [4]. Der vollständige Korpus enthält neben Kommentaren auch Datensätze mit allen monatlich erstellten Beiträgen, im folgenden auch „Submissions“ genannt. Gaffney und Matias kommen zu dem Schluss, dass die Erfassung sowohl der Submissions als auch der Kommentare Lücken aufweist, also Elemente gänzlich nicht im Datensatz vorhanden sind. Für den Gegenstand der vorliegenden Arbeit ist dieser Umstand insofern von Bedeutung, als dass

[1] https://www.reddit.com/user/stuck_in_the_matrix

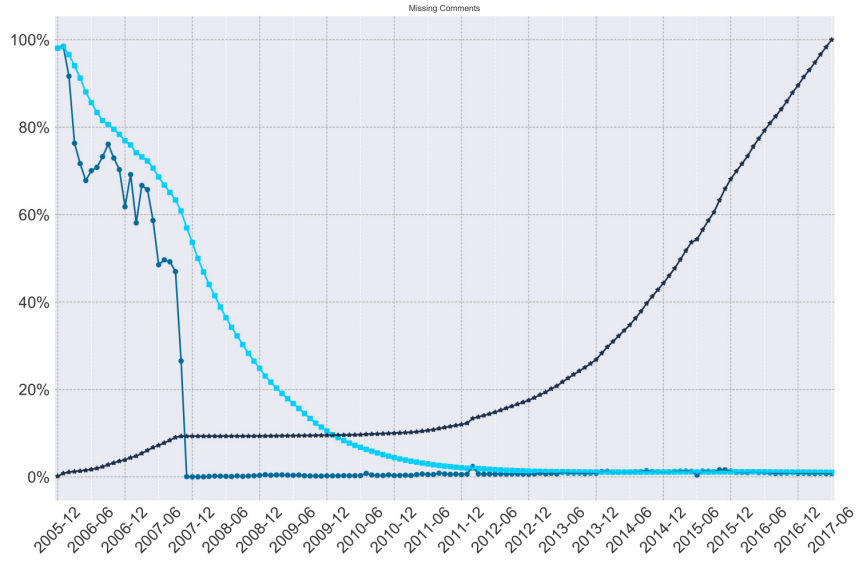

Abbildung 3.1: Anteil fehlender Kommentare. Die hellblauen Quadrate (obere Linie) stellen den gleitenden Mittelwert fehlender Kommentare in Prozent dar, die mittelblauen Punkte (mittlere Linie) den prozentualen Anteil fehlender Kommentare, und die dunkelblauen Kreuze (untere Linie) die kumulierte Gesamtzahl fehlender Kommentare [4]

fehlende Kommentare die Topic-Affinität von Nutzern verzerren können, etwa wenn ein Topic, in dem der Nutzer durchaus aktiv war, aufgrund fehlender Daten unterrepräsentiert ist. Auch Gaffney und Matias stellen fest, dass Studien, welche auf die vollständige Historie von Nutzern zugreifen, dem höchsten Risiko ausgesetzt sind, lückenhafte Daten zu betrachten [4].

Reddit weist jedem Kommentar eine eindeutige numerische ID zu. Das von Baumgartner eingesetzte System nimmt zusammenhängende Blöcke von jeweils 100 solcher IDs und versucht, die zugehörigen Kommentare über die Reddit-API[2] aufzulösen [2]. Da Reddit auch Anfragen nach gelöschten Elementen sinnvoll beantwortet, also nicht etwa mit einer Fehlermeldung, sollte ein Bereich von 100 sequentiellen IDs auch vollständig im Datensatz abgebildet sein, inklusive als gelöscht markierter Elemente. Gaffney und Matias stellen jedoch für den Zeitraum Dezember 2005 bis Februar 2016 fest, dass 943.755 Kommentar- und 1.539.583 Beitrags-IDs nicht in den Datensätzen enthalten sind. Als mögliche Gründe für das Fehlen nennen Gaffney und Matias dreierlei: sog. „dangling references“, also Verweise, bei denen das Element, auf das verwiesen wird, nicht auffindbar ist; öffentlich zugängliche Daten, die aus unbekanntem Grund nicht von Reddit an Baumgartners System übertragen wurden; oder Daten aus als privat eingestuften Communities, die nicht öffentlich sondern nur von Mitgliedern mit Zugangsberechtigung einsehbar sind [4].

In Abbildung 3.1 stellen die mittelblauen Punkte den Anteil fehlender Kommentare in Prozent dar. Ab etwa April 2006 beginnt dieser Anteil zu sinken, fällt ab etwa August 2007 stark ab und stabilisiert sich ab etwa November 2007 im niedrigen einstelligen Bereich. Um den Einfluss fehlender Kommentare so gering wie möglich zu halten, wurden daher für die vorliegende Arbeit die Datensätze beginnend mit November 2007 bis einschließlich Februar 2018 ausgewertet.
Jason Baumgartner hat als Folge der Veröffentlichung von Gaffney und Matias angekündigt,

[2] https://api.reddit.com/

Tabelle 3.2: Kennzahlen der Altersverteilung.

N	arithm. Mittel	SD	Min	Q1	Median	Q3	Max
28.029.716	6,67	12,41	1	1	2	6	124

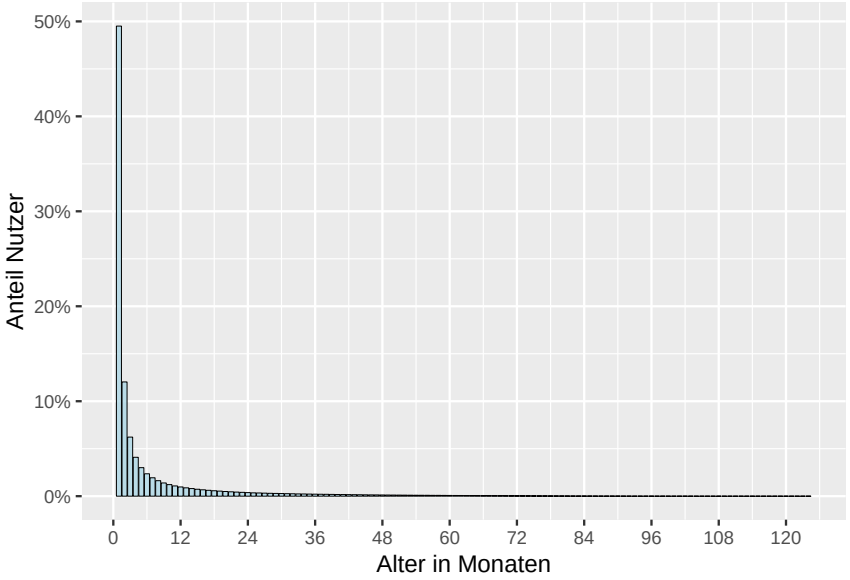

Abbildung 3.2: Verteilung der Alterswerte.

fehlende Kommentare und Beiträge nachträglich zu erfassen [1].

3.1.2 Stichprobe von Nutzern

Nachdem der Datensatz einer zeitlichen Einschränkung unterworfen wurde, müssen Kriterien für die Auswahl von Nutzern herangezogen werden. Dieser Abschnitt gibt Aufschluss über die Altersverteilung von Nutzern im Datensatz und beschreibt die Ziehung einer Stichprobe für die Datenanalyse.

Altersverteilung Nach der zeitlichen Eingrenzung der Daten liegen für den betrachteten Zeitraum von 124 Monaten etwa 3,6 Milliarden Kommentare vor, verfasst von ca. 28 Millionen Nutzern. Für jeden dieser Nutzer wurde bestimmt, in wie vielen Monaten er insgesamt im Datensatz enthalten ist; nachfolgend wird dies als „virtuelles Alter“ oder schlicht „Alter“ des Nutzers bezeichnet.

Tabelle 3.2 bietet eine Übersicht über die Verteilung der Alterswerte. 50% der Nutzer sind zwischen einem und sechs Monaten auf Reddit aktiv (unteres resp. oberes Quartil), und 50% sind in zwei Monaten oder weniger enthalten (Median). Der arithmetische Altersdurchschnitt liegt bei etwa sieben Monaten. Da ein Nutzer mindestens einen Kommentar verfasst haben muss, um gezählt zu werden, liegt das minimale Alter bei einem Monat.

Das Histogramm in Abbildung 3.2 veranschaulicht das hohe Vorkommen eher kleiner Werte. Die ALtersverteilung weist einen sog. „Long Tail“ auf: sehr viele Nutzer sind eher kurz aktiv,

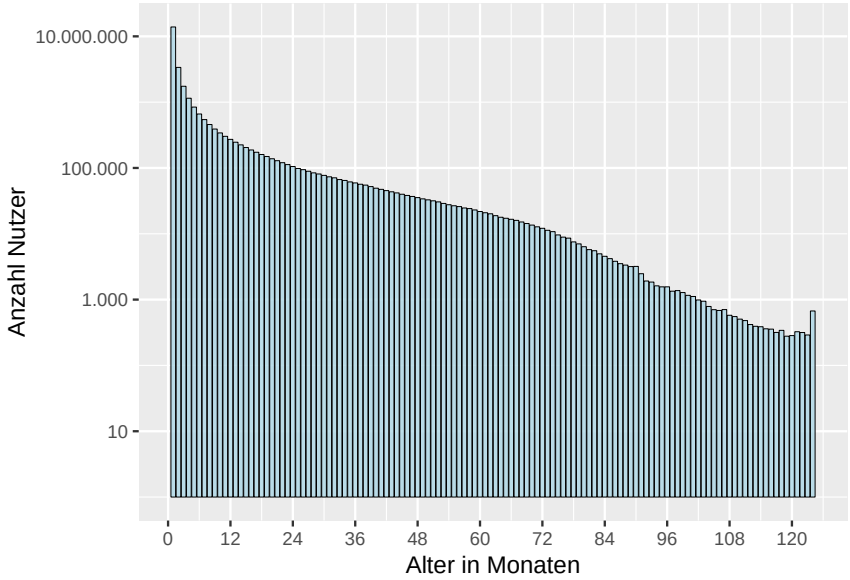

Abbildung 3.3: Verteilung der Alterswerte, logarithmische Darstellung.

während Nutzer mit eher langer Aktivität nur einen gerinen Anteil ausmachen. Die lineare Skala des Histograms macht es schwierig, die Werte des Long Tail sinnvoll darzustellen. Abbildung 3.3 nutzt daher eine logarithmische Darstellung.

Markant ist bei dieser Visualisierung der Ausschlag am äußersten rechten Rand. Offenbar entfallen auf die Altersgruppe „124 Monate“ noch einmal deutlich mehr Nutzer als noch auf die vorhergehenden Gruppen. In der Tat beträgt der Unterschied zwischen den beiden letzten Altersgruppen 382 Nutzer Obwohl an dieser Stelle keine Erklärung für diese Beobachtung geliefert werden kann, ist es denkbar, dass es einen „harten Kern“ von Nutzern gibt, die Reddit seit langer Zeit, möglicherweise sogar von Anfang an nutzen, und regelmäßig aktiv sind.

Kriterien für die Stichprobe Um Nutzer für die weitere Datenanalyse auszuwählen, wurden zwei Kriterien festgelegt. Zum einen sollte ein Nutzer über einen ausreichend langen Zeitraum hinweg aktiv sein. Hierdurch wird sichergestellt, dass zeitliche Verläufe möglichst keine Lücken enthalten und genügend Daten vorliegen, um Trends zu identifizieren. Zum anderen sollte der Nutzer ein Mindestmaß an Interaktionen pro Monat aufweisen, damit Aussagen sowohl über seine thematische Affinität als auch die sozialen Kreise möglich sind, in denen er sich bewegt.

Um dem zeitlichen Kriterium zu genügen wird eine zufällige Auswahl aus den ältesten 10.000 Nutzern getroffen. Indem nur Nutzer einbezogen werden, die über die gesamten 124 Monate hinweg mindestens 50 Kommentare pro Monat erstellt haben, wird der Datensatz weiter gefiltert und dem Volumen-Kriterium entsprochen. Das Histogramm in Abbildung 3.4 zeigt die zugehörige Verteilung der Altersgruppen nachdem die beiden Einschränkungen vorgenommen wurden.

Wegen der Auswahl der 10.000 ältesten Nutzer ist die erste, „jüngste“ Säule nicht vollständig gefüllt. Das Mindestalter in dieser neuen Verteilung liegt bei 103 Monaten. Dies entpricht einer Überschneidung zu 83.06% mit dem gesamten Untersuchungszeitraum von 124 Monaten.

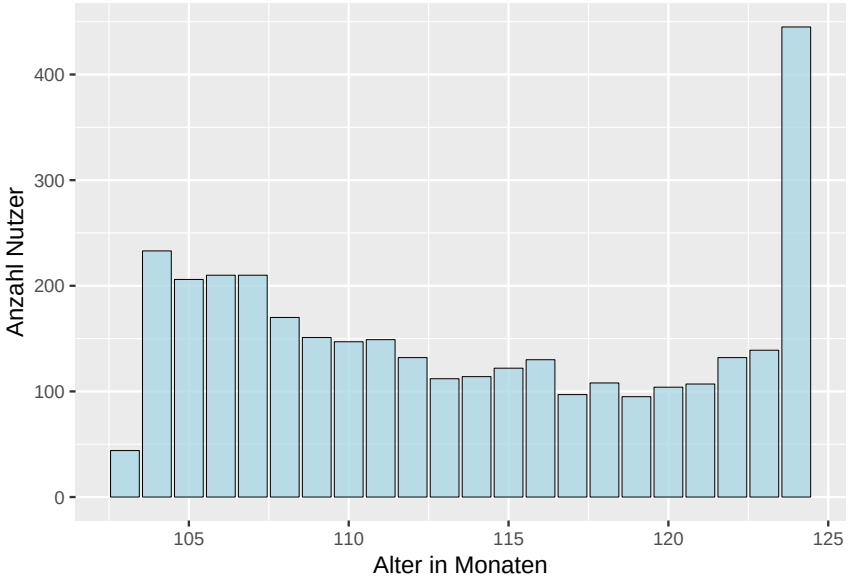

Abbildung 3.4: Altersverteilung nach Einschränkungen.

3.1.3 Topic-Modelle

Für alle Nutzer, die über dem Durchschnittsalter von 6.67 Monaten liegen, wurde festgehalten, in welchen Subreddits sie kommentiert hatten. Durch diese Altersbeschränkung wird verhindert, dass Subreddits in das Topic-Modell einbezogen werden, die ausschließlich von Nutzern aufgesucht werden, die die Plattform nach kurzer Aktivität verlassen.

Für die so erhaltenen 419.616 Subreddits wurden über die Reddit-API jeweils maximal 50 Beiträge aus dem Listing „Top“ abgerufen. Diese Sortierung liefert Beiträge mit dem besten Score, also der Differenz aus Up- und Downvotes [7]. In der Folge erhält man so diejenigen Beiträge, die von der Community am besten bewertet wurden. In dieser Arbeit wird davon ausgegangen, dass ein Beitrag mit hohem Score auch repräsentativ für die Inhalte der Community ist.

Beim Abrufen der Top-Listings über die Reddit-API traten zum Teil zwei Arten von HTTP-Fehlern auf: „403 Forbidden“ sowie „404 Not found“. Da keine weiteren Informationen zu den Fehlern übermittelt wurden, können nur Vermutungen über deren Ursache angestellt werden. Da Listings auf Subreddit- und nicht auf Beitragsebene abgerufen werden, ist ein Fehler immer im Kontext des Zugriffs auf eine Community zu verstehen. In diesem Fall könnte der Grund für einen 403-Fehler auf mangelnde Zugriffsrechte auf das Subreddit zurückzuführen sein, sprich: nur Mitglieder dürfen Beiträge lesen und schreiben, die Community ist privat. Ähnlich ist ein 404-Fehler zu interpretieren: obwohl das Subreddit im Datensatz enthalten ist, kann es zum gegenwärtigen Zeitpunkt nicht abgerufen werden; aller Voraussicht nach wurde es gesperrt oder gelöscht.

Die Titel der 207.056 erfolgreich abgerufenen Beiträge wurden konkateniert und zusammen mit dem Namen des Subreddits gespeichert. Jedes Subreddit entspricht damit einem Dokument, dessen Inhalt die Titel der Top 50 Beiträge bilden. Die Inhalte wurden in Kleinschreibung umgewandelt und um Satzzeichen und Ziffern bereinigt; redundante Leerzeichen wurden entfernt. Zudem wurden Dokumente verworfen, deren Inhalt weniger als 500 Zeichen umfasste. Die so erhaltene Sammlung

Tabelle 3.3: Startparameter des LDA-Algorithmus

Parameter	Wert
α	0.195
β	0.1
k	256
niter	2000

Tabelle 3.4: Kennzahlen der Topic-Verteilung

N	arithm. Mittel	SD	Min	Q1	Median	Q3	Max
256	808,81	4.040,09	1	6	9	97,5	45.577

an Dokumenten diente dem LDA-Algorithmus als Eingabe.

Wegen der hohen Anzahl an Dokumenten wurde wurde eine effiziente Implementierung des LDA-Algorithmus benötigt. Die Wahl fiel dabei auf „GLDA“, eine „verbesserte Version“ [5] der „GibbsLDA++“-Implementierung [6], welche sich die hohe Rechenleistung moderner Grafikkarten zunutze macht. Die Start-Parameter des Algorithmus führt Tabelle 3.3 auf.

Da sich eine Evaluation verschiedener k-Parameter wegen der hohen Zahl an Dokumenten als schwierig herausstellte, wurde die Zahl der zu bestimmenden Topics auf 256 festgelegt. Unklar ist, ob diese Zahl in der Nähe eines Optimums liegt; dies festzustellen wird jedoch nicht Gegenstand dieser Arbeit sein.

Der LDA-Algorithmus beruht auf der Annahme, dass sich jedes Dokument aus verschiedenen (latenten) Themen[3] zusammen setzt. Das Ergebnis liefert für jedes Dokument eine Wahrscheinlichlichkeitsverteilung über die zu bestimmenden Topics. Für die weitere Analyse in dieser Arbeit wird aus dieser Verteilung von Topic-Wahrscheinlichkeiten eine 1:1-Zuordnung abgeleitet, indem für jedes Dokument das Topic als maßgebend angesehen wird, dem der Algorithmus die höchste Wahrscheinlichkeit zuweist.

Abbildung 3.5 stellt die Zuordnung von Subreddits zu Topics als Histogramm dar. Jede Klasse entlang der x-Achse entspricht einem der 256 Topics, die der LDA-Algorithmus identifizieren sollte. Die y-Achse ist logarithmisch skaliert und notiert die Anzahl an Subreddits, die dem jeweiligen Topic zugewiesen wurden. Wie bereits zuvor beim Alter der Nutzer zeigt diese Darstellung eine Verteilung mit einem Long Tail: auf einen großen Teil der Topics entfallen vergleichsweise wenige Subreddits; 50% der Topics sind weniger als neun Subreddits zugeordnet (siehe auch Tabelle 3.4).

LDA liefert nicht nur für jedes Dokument eine Verteilung von Topics, sondern auch zu jedem Topic eine Wahrscheinlichkeitsverteilung von Wörtern. Sortiert man die Auftretenswahrscheinlichkeiten der Wörter eines Topics, erhält man die für dieses Thema charakteristischen Begriffe. Tabelle A.1 im Anhang enthält für alle Topics, denen mindestens 500 Subreddits zugeordnet wurden, die 25 häufigsten Wörter. Dabei fällt auf, dass die Topics 235, 122, 69 und 194 beinahe ausschließlich

[3] In dieser Arbeit wird hauptsächlich die englische Wortform „topic(s)“ gebraucht, um explizit die algorithmisch bestimmten Themenkomplexe zu bezeichnen

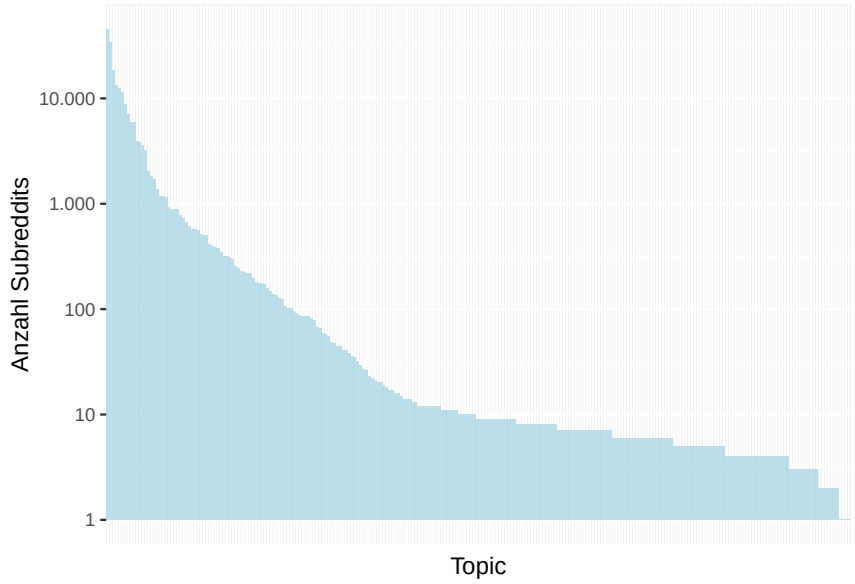

Abbildung 3.5: Anzahl der Zuordnungen von Subreddits zu Topics. Die Gesamtzahl aller kategorisierten Topics beträgt 207.056, die Zahl der Topics 256. Aus Gründen der besseren Darstellung wurde auf eine Auszeichnung der x-Achse verzichtet; die Sortierung erfolgt absteigend nach Anzahl der Subreddits.

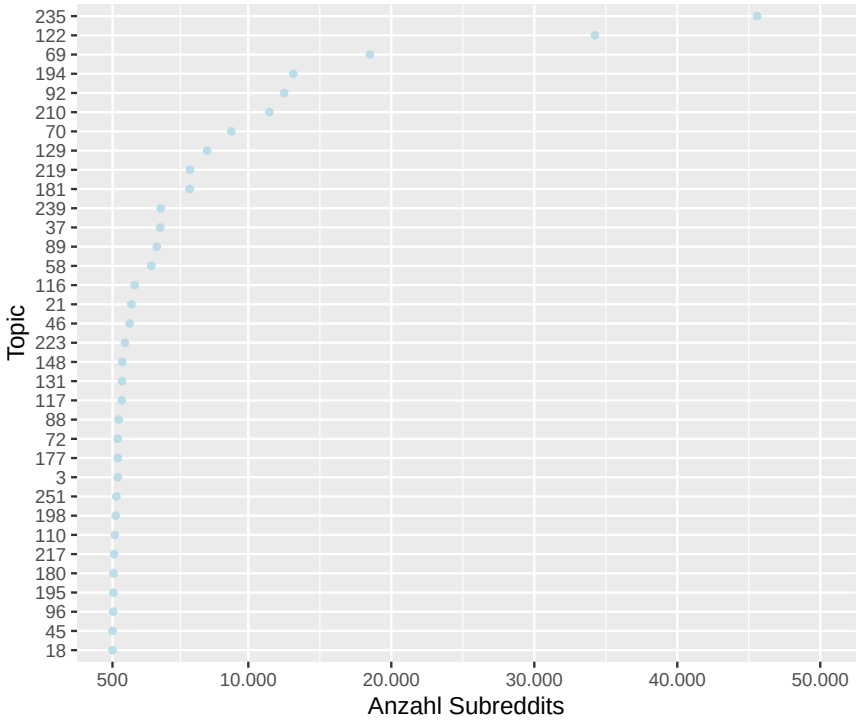

Abbildung 3.6: Anzahl der Zuordnungen von Subreddits zu Topics. Dargestellt sind alle Topics, denen 500 oder mehr Subreddits zugeordnet wurden. Die Sortierung erfolgt analog zu Abbildung 3.5 nach Anzahl Zuordnungen in absteigender Folge.

aus Stoppwörtern bestehen. Da dies keine sinnvollen Rückschlüsse auf Nutzerinteressen zulassen, werden bei sie der weiteren Analyse zwar aufgeführt, jedoch nicht näher berücksichtigt.

3.1.4 Interaktionsgraphen aus Kommentaren

Die Kommentarverläufe von Reddit lassen sich als Interaktionsgraphen modellieren. Jeder Knoten in einem solchen Graph stellt einen Akteur in einem sozialen Netzwerk dar. Zwischen Akteuren manifestieren sich Kanten, wenn sie miteinander interagieren, in diesem Fall in Form von Kommentaren auf Reddit. Die Richtung der Kanten gibt dabei an, welcher Nutzer den Kommentar verfasst hat (Quelle) bzw. an welchen Nutzer der Kommentar gerichtet ist (Senke). Im Datensatz sind Kanten über die Beziehung zwischen den *id*- bzw. *parent_id*-Attributen realisiert. Seien dazu U, V Akteure im sozialen Netzwerk und K_U, K_V von U resp. V verfasste Reddit-Kommentare. Zwischen U und V wird eine gerichtete Kante (u, v) eingefügt, wenn gilt: $K_U.parent\_id = K_V.id$.

Für jeden wie in Abschnitt 3.1.2 beschrieben ausgewählten Nutzer werden monatliche Interaktionsgraphen bestimmt. Da diese einen zeitlich abgegrenzten Ausschnitt aus dem gesamten sozialen Netzwerk eines Nutzers darstellen, werden sie im folgenden auch als Snapshot-Graphen oder einfach Snapshots bezeichnet. Da ausgehend von einem Nutzer dessen unmittelbare Kontakte erfasst werden, spricht man hier zudem von egozentrischen Netzwerken. Dabei ist zu beachten, dass abweichend von gängigen Definitionen des Begriffs (etwa [9, S. 42], [10]) in dieser Arbeit keine Strukturen zwischen den Alteri erfasst werden, sondern nur zwischen Ego und Alteri.

3.2 Ergebnisse

3.2.1 Verteilung von Topics

3.2.2 Fallstudie

4 Diskussion

5 Zusammenfassung und Ausblick

Anhang

A Häufigste Wörter je Subreddit

Tabelle A.1: Charakteristische Wörter der größten Topics. Aufgeführt sind jeweils die Topic-ID, die Anzahl an zugeordneten Subreddits (n, mit der Einschränkung $n \geq 500$) sowie die 25 häufigsten Wörter in dem jeweiligen Topic.

Topic	n	Wörter
235	45.577	the, a, to, i, is, you, this, of, and, in, it, that, on, for, when, my, are, what, me, not, be, your, have, like, just
122	34.240	to, a, for, the, and, i, you, is, of, in, how, on, what, with, this, your, do, my, it, help, can, or, are, have, anyone
69	18.504	my, a, i, the, this, of, and, to, in, for, on, it, from, is, with, me, first, just, you, was, at, new, so, got, made
194	13.146	the, to, for, of, on, and, new, is, in, a, now, we, at, with, this, be, will, update, from, our, has, are, up, first, out
92	12.511	in, her, and, a, the, with, on, girl, xpost, ass, hot, sexy, big, from, she, pussy, cock, gif, tits, black, teen, sex, blonde, girls, porn
210	11.476	the, to, of, in, and, a, for, on, is, with, trump, by, us, from, as, that, are, after, at, be, not, his, who, about, has
70	8.818	by, book, the, online, download, movie, p, free, full, read, how, link, without, of, watch, no, ipad, pc, mp, english, iphone, format, android, tablet, pdf
129	7.118	in, the, at, of, a, to, from, on, for, and, city, park, new, this, area, looking, xpost, local, san, lake, near, st, night, th, with
219	5.922	the, of, and, a, in, to, is, on, for, by, from, how, an, that, with, are, as, about, why, what, world, science, life, be, new
181	5.894	the, music, song, video, album, by, new, of, cover, live, remix, on, band, official, ft, rock, songs, feat, mix, love, guitar, tour, show, full, in
239	3.885	to, with, for, a, in, and, on, how, using, google, windows, data, app, web, from, code, released, tutorial, linux, use, free, is, software, an, source
37	3.839	the, of, a, and, part, in, story, by, from, world, war, dark, wp, book, death, an, king, one, history, battle, man, books, life, first, ii
89	3.608	the, episode, season, and, of, show, on, movie, podcast, film, with, se, in, series, spoilers, trailer, tv, from, discussion, john, his, review, s, movies, watch
58	3.223	vs, the, game, league, team, season, to, in, round, football, match, win, cup, week, for, draft, thread, fc, at, highlights, player, nfl, sports, goal, players
116	2.052	game, games, play, video, ps, youtube, pc, lets, part, gameplay, super, xbox, gaming, mario, gta, steam, nintendo, trailer, channel, funny, new, v, fallout, best, switch

21	1.825	game, update, play, new, v, guide, patch, players, steam, beta, player, level, games, map, charac playing, build, bug, alpha, server, version, battle, mode, notes, event
46	1.701	de, la, a, en, el, que, y, para, o, del, los, no, e, do, un, por, da, con, se, em, las, es, como, una, al
223	1.378	bitcoin, blockchain, ico, crypto, coin, exchange, wallet, token, trading, mining, cryptocurrency network, ethereum, price, btc, platform, market, tokens, –, listed, coins, with, buy, —, decentraliz
148	1.181	free, for, w, code, h, sale, off, and, card, selling, amazon, trade, giveaway, buy, gift, or, cards, shipping, price, codes, get, k, gold, paypal, account
131	1.166	food, and, with, chicken, recipe, pizza, cheese, vegan, eat, chocolate, ice, coffee, cream, eating cake, tea, recipes, sauce, make, breakfast, milk, bread, bacon, meat, meal
117	1.152	in, best, business, your, for, online, company, top, services, marketing, money, service, market, h social, management, to, sales, india, tips, home, blog, credit, media, real
88	933	review, with, pro, k, x, gb, g, camera, pc, gaming, build, tv, laptop, setup, mm, v, best, usb, pow drone, led, mini, wireless, battery, pi
72	881	die, der, in, und, für, von, mit, das, ist, im, auf, ein, den, zu, ich, es, aus, des, was, nicht, am, an, dem, wie
177	879	water, home, with, diy, machine, glass, house, wood, table, make, design, steel, paper, wall, cleaning, hand, made, build, metal, knife, custom, set, kit, room, door
3	875	cat, dog, baby, a, xpost, dogs, cats, fish, his, little, bear, cute, animal, her, puppy, boy, pet, fishi he, meet, kitty, bird, animals, kitten, she
251	773	chapter, anime, no, cosplay, manga, ch, english, original, fanart, spoilers, hentai, japanese, naru volume, girl, episode, girls, art, chapters, japan, sakura, translation, disc, maid, wa
198	724	car, insurance, bike, race, ride, cars, truck, driver, ford, racing, auto, road, gt, drive, driving, spe s, motorcycle, honda, miles, engine, electric, r, bmw, crash
110	667	by, x, art, oc, artist, ×, drawing, wallpaper, painting, draw, deviantart, tattoo, photo, xpost, sket image, portrait, wallpapers, digital, concept, artists, canvas, photography, ink, artwork
217	611	de, la, le, les, et, du, à, des, en, un, pour, sur, au, une, dans, france, pas, est, par, avec, je, que, », qui
180	571	black, red, blue, dress, fashion, wedding, shoes, white, leather, boots, jacket, shirt, size, wear, st sale, tshirt, hat, vintage, mens, clothing, socks, bag, color, wearing
195	566	health, weight, cancer, treatment, pain, body, loss, workout, surgery, diet, fat, medical, and, thera care, fitness, disease, sleep, brain, drug, lbs, skin, depression, blood, after
96	553	f, me, looking, m, for, mf, fm, fun, kik, chat, daddy, snapchat, friends, fa, add, mm, play, yo, wa girl, sissy, message, snap, pics, seeking
45	507	looking, war, clan, guild, recruiting, join, raid, group, ps, players, server, for, members, destiny community, active, reddit, pvp, discord, th, rp, pc, base, crew, xbox
18	506	i, in, up, found, killed, a, levelled, xp, completed, treasure, trail, dragon, crystal, boss, elite, invention, hard, skills, over, events, all, fragment, quest, triskelion, recent

Literatur

[1] Jason Baumgartner. *"... anticipate that it will take between 4-6 weeks to fill in the largest gaps for missing comments. I will then rescan all missing ids in the sequential areas (ids over 27 billion for comments) and ingest the missing data there. Probably 1-2 months before complete." 6. April 2018, 20:13 Uhr*. URL: https://twitter.com/jasonbaumgartne/status/982456309726547968. Tweet.

[2] Jason Baumgartner. *Ingesting Data — Using high performance Python code to collect Data*. 7. Mai 2018. URL: https://pushshift.io/ingesting-data%E2%80%8A-%E2%80%8Ausing-high-performance-python-code-to-collect-data/ (besucht am 07. 05. 2018). Blog-Post.

[3] Jason Baumgartner. *pushshift.io*. URL: https://files.pushshift.io/reddit/comments/ (besucht am 23. 04. 2018).

[4] Devin Gaffney und J. Nathan Matias. „Caveat Emptor, Computational Social Science: Large-Scale Missing Data in a Widely-Published Reddit Corpus“. In: (13. März 2018). arXiv: 1803.05046v1 [cs.SI].

[5] Mian Lu u. a. „Accelerating Topic Model Training on a Single Machine“. In: *Asia-Pacific Web Conference*. 2013, S. 184–195. DOI: 10.1007/978-3-642-37401-2_20. URL: https://github.com/lumianph/glda.

[6] Xuan-Hieu Phan und Cam-Tu Nguyen. *GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA)*. 2007. URL: http://gibbslda.sourceforge.net/.

[7] Reddit. *reddit. historical code from reddit.com*. Source code. Reddit Inc. URL: https://github.com/reddit-archive/reddit (besucht am 22. 11. 2018).

[8] Reddit. *reddit.com: api documentation*. 20. Nov. 2018. URL: https://www.reddit.com/dev/api.

[9] Stanley Wasserman und Katherine Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994. ISBN: 9780521387071.

[10] Christof Wolf. „Egozentrierte Netzwerke: Datenerhebung und Datenanalyse“. In: *Handbuch Netzwerkforschung*. Hrsg. von Christian Stegbauer und Roger Häußling. Wiesbaden: VS Verlag für Sozialwissenschaften, 2010, S. 471–483. ISBN: 978-3-531-92575-2. DOI: 10.1007/978-3-531-92575-2_41. URL: https://doi.org/10.1007/978-3-531-92575-2_41.