technische universität
dortmund

Bachelorarbeit

Entdecken von Themen und Communitys in Online-Foren

Andreas Sitta
30.06.2017

Gutachter:
Prof. Dr. Katharina Morik
Lukas Pfahler (M.Sc)

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für Künstliche Intelligenz (LS 8)
http://www-ai.cs.uni-dortmund.de

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Motivation und Hintergrund 1

1.2 Aufbau der Arbeit 3

2 Begriffserklärungen und Definitionen **5**

2.1 Probabilistisches Modell 5

2.2 Generatives Modell 6

2.3 Latente/Beobachtbare Variablen 6

2.4 Community, Community Detection 7

2.4.1 Cliquenbildung 7

2.4.2 Hierarchische Clusteranalyse 8

2.5 Topic Models 8

3 Mathematisches Hintergrundwissen **11**

3.1 A-priori-Wahrscheinlichkeit 11

3.2 A-posteriori-Wahrscheinlichkeit 12

3.3 Satz von Bayes 12

3.4 Binomialverteilung 13

3.5 Multinomialverteilung 13

3.6 Dirichlet-Verteilung 13

3.7 Poisson-Verteilung 14

4 Latent Dirichlet Allocation **15**

4.1 Latent Dirichlet Allocation 15

4.1.1 Notation 15

4.1.2 Generativer Prozess 16

4.1.3 Inferenz 17

5 Link-Content Modell **19**

5.1 Notation 19

5.2 Generative Semantik und Plate Notation 21

5.3 Inferenzprozess . . . 22
5.4 Gibbs Sampling für das Link-Content Modell . . . 22
5.5 Algorithmus . . . 24
5.6 Ermitteln latenter Variablen . . . 24

6 Theorie der Datenvorverarbeitung 27
6.1 Rohdatenbereitstellung . . . 27
6.2 Lexikalische Vorverarbeitung . . . 28
6.3 Linguistische Datenvorverarbeitung . . . 28
6.3.1 Eliminierung von Wörtern mit geringer inhaltlichen Relevanz . . . 29
6.4 Morphologischer Ansatz zur Wortreduzierung . . . 31
6.4.1 Stammformreduzierung (Stemmer) . . . 31
6.4.2 Reduzierung auf die grammatikalische Grundform (Lemmatization) und Anwendung eines Thesauruses . . . 32

7 Datenvorverarbeitung des Link-Content Modells 35
7.1 relNet Datensatz . . . 35
7.2 Datenvorverarbeitung . . . 36
7.2.1 Vorverarbeitung . . . 37
7.2.2 Erstellung des Inputs . . . 38

8 Versuchsdurchführung 41
8.1 Themenerkennung . . . 41
8.1.1 Qualitative Auswertung . . . 41
8.1.2 Quantitative Auswertung: Perplexitätsberechnung . . . 42
8.1.3 Auswertung . . . 43
8.1.3.1 Vergleich Link-Content Modell zu LDA . . . 44
8.1.3.2 Vergleich LDA und Link-Content mit mehr Themen . . . 45
8.1.3.3 Variation der Communities . . . 45
8.1.3.4 Keine themenbezogene relativen Häufigkeiten . . . 50
8.1.3.5 Variation des Wortfilters . . . 50
8.1.3.6 Variation der Freundschaftsdefinition . . . 56
8.2 Community Detection . . . 61
8.2.1 Vergleich zwischen Link-Content und Louvain Algorithmus . . . 61

9 Zusammenfassung und Ausblick 67

Abbildungsverzeichnis 69

Literaturverzeichnis 75

Kapitel 1

Einleitung

Diese Bachelorarbeit befasst sich mit dem Entdecken von Themen und Communitys in Online-Foren. Dabei sollen verschiedene Methoden des Data-Minings angewendet und miteinander verglichen werden. Hierbei wird das Hauptaugenmerk auf die Anwendung eines relativ neuen Modelles zur simultanen Modellierung auf Topic- und Communityebene gelegt.

1.1 Motivation und Hintergrund

In der jüngsten Zeit haben sich verschiedene Aspekte des gesellschaftlichen Lebens immer mehr in das Internet verlagert. Wie schon Papst Johannes Paul II. im Jahr 2002 zum 36. Welttag der sozialen Kommunikationsmittel sagte, geschieht dies auch mit dem religiösen Leben.[1]

Während sich früher Informationen und so auch Religionen ausschließlich über primäre Medien (menschlicher Kontakt ohne technische Hilfsmittel, z.B. Wörter oder Gesten) und sekundäre Medien (technische Hilfsmittel werden nur beim Nachrichtensender benötigt, z.B. Bilder, Schriftrollen oder Bücher) verbreitet haben, wurde dies durch die tertiären Medien (technische Hilfsmittel beim Sender und Empfänger, z.B. Radio, Fernsehen oder Internet) enorm beschleunigt.[2] So liefert die Suchmaschine Google (Stand Januar 2017) für den Begriff „religion“ ca. 600 Millionen und für den Begriff „god“ ca. 1,6 Milliarden Ergebnisse. Hier hat sich eine große Anzahl von Internetforen gebildet, in denen über ein breites Spektrum an Themen diskutiert wird.

Dieses Phänomen ist für Religionswissenschaftler aus mehreren Gründen von großer Bedeutung. Zum einen unterscheidet sich das Internet von den anderen Medien dadurch, dass viel mehr Menschen und diese viel einfacher erreicht werden können. Es wird kein großes Knowhow benötigt, um z.B. einen Youtube-Kanal zu erstellen oder sich in einem Internetfo-

[1] Siehe auch: Botschaft von Papst Johannes Paul II. zum 36. Welttag der sozialen Kommunikationsmittel. Thema: Internet: Ein neues Forum zur Verkündigung des Evangeliums (2002).

[2] Kategorisierung von Medien nach Pross. [1]

rum anzumelden und dort Einträge zu verfassen. Damit einher geht auch die Entwicklung, dass nun jeder zur „religiösen Autorität“ in einer speziellen Community (Gruppe) oder bezogen auf ein spezielles Thema (eng. topic) aufsteigen kann. Zum anderen sind Benutzer durch die Anonymität und die daraus resultierende (zumindest suggerierte) Rechtsfreiheit weniger eingeschränkt. Sie können Meinungen, Ideen oder Kritik offen vertreten, welche in der Gesellschaft kontrovers oder tabu sind und müssen dadurch, wenn sie anonym bleiben, keine Konsequenzen in ihrem Alltagsleben erfahren. Durch diese Anonymität finden auch solche Meinungen und Ansichten den Weg an die Öffentlichkeit. Dies betrifft vor allem auch Religionen. Es können Menschen aus Ländern ohne Religionsfreiheit mit freien Menschen aus anderen Ländern kommunizieren und sich austauschen.
Ein weiterer Vorteil der großen Reichweite ist, dass sich Menschen mit seltenen Religionsansichten mit Hilfe von Suchmaschinen finden und austauschen können.
Im Fokus dieser Arbeit sollen allerdings die Online-Foren stehen, in denen sich User aus der ganzen Welt zusammenfinden, um Kontakte zu knüpfen oder über interessierende Themen zu diskutieren.
Es entstehen dabei riesige Mengen an Daten, welche nicht mit menschlichen Fähigkeiten sortiert, bearbeitet oder kategorisiert werden können. Zum einen ist dies aufgrund der enormen Datenmenge nicht möglich, zum anderen gelangen verschiedene Menschen durch ihre subjektive Wahrnehmung, z.B. bei einer Kategorisierung, zu unterschiedlichen Ergebnissen.
Um diese Entwicklung datengetrieben, also mit Methoden des Data-Minings [2, 3, 4, 5], zu untersuchen, können vor allem zwei Informationsarten aus den Daten extrahiert werden: die inhaltlichen Themen und die Communitys.
Dafür können sowohl Algorithmen des Topic Modelings, als auch Algorithmen der Community Detection verwendet werden. Ebenso kommen Algorithmen in Frage, die beides kombinieren [6].
Sowohl die Themen, als auch die Communitys sind also wichtig für die Untersuchung von Online-Foren und des Verhaltens der Nutzer. Neben dem thematischem Inhalt spielt auch die soziale Struktur für Religionswissenschaftler eine große Rolle, z.B. bei der Suche nach Autoritäten. Dementsprechend ist es naheliegend, beide Themenfelder simultan zu bearbeiten.
Ein weiterer Grund dafür ist z.B., dass Mitglieder innerhalb einer Community oftmals auch ähnliche Interessen aufweisen und sich über ähnliche Themen austauschen und daher die Präzision des Algorithmus verbessert werden kann. Die von den Algorithmen bearbeiteten Daten können anschließend weiter analysiert werden und in verschiedenen Anwendungsbereichen Nutzen finden. Diese Arbeit verwendet Datensätze aus dem relNet Datensatz, einem Abbild mehrerer religiösen Foren. Zum einen sind diese so aufbereiteten Daten für Religionsforscher interessant, zum anderen aber auch für andere Bereiche, wie z.B. der personenbezogen Werbung oder der Kontrolle und dem Verständnis von Prozessen der Radikalisierung, die zu Terrorismus führen.

1.2 Aufbau der Arbeit

Als erstes werden die benötigten Begriffe und Definitionen in Kapitel 2 vorgestellt. Dann wird in Kapitel 3 das mathematische Hintergrundwissen erläutert.
Die zu untersuchenden Algorithmen zum LDA- bzw Link-Content Model werden in den Kapiteln 4-5 vorgestellt. Für die Versuche wurde der LDA-Algorithmus dankenswerterweise vom Lehrstuhl für künstliche Intelligenz der Technischen Universität Dortmund zur Verfügug gestellt. Der Algorithmus für das Link-Content Modell wurde vom Ersteller dieser Arbeit selbst implementiert, da allgemein kein passender Quellcode verfügbar war. Die Implementierung erfolgte in der Programmiersprache Java. In Vorbereitung auf die Darstellung der Ergebisse wird in Kapitel 6 auf die Theorie der Datenvorverarbeitung und in Kapitel 7 auf die tatsächliche Umsetzung und den in dieser Arbeit verwendeten Datensatz eingegangen.
Die Versuchsdurchführung und die Diskussion der Ergebnisse folgen in Kapitel 8. Zum Abschluss wird die Arbeit in Kapitel 9 zusammengefasst und ein Ausblick auf mögliche Erweiterungen gegeben. Ich möchte mich an dieser Stelle bei meinen Lukas Pfahler (M.Sc) und dem Lehrstuhl für künstliche Inteligenz (LS8) für die Unterstützung bedanken.

Kapitel 2

Begriffserklärungen und Definitionen

In diesem Kapitel werden die für den weiteren Verlauf der Arbeit wichtigen Begriffe und Definitionen erläutert. Die Begriffe entsprechen den gängigen Definitionen in den Bereichen Community Detection und Topic Modeling. Zusätzlich werden hier auch kurz die Themenbereiche „Topic-Modelling“ und „Community-Detection“ vorgestellt.

2.1 Probabilistisches Modell

Probabilistische Modelle modellieren gegebene Daten durch einen stochastischen Prozess und treffen anschließend anhand eines bestimmten Sachverhaltes Wahrscheinlichkeitsaussagen über wiederkehrende Ereignisse. Sie stehen damit im Gegensatz zu deterministischen Modellen, in denen der Ausgang eindeutig ist. Die meisten realen Abläufe sind deterministisch. Kennt man beispielsweise die aktuelle Geschwindigkeit und Position eines Balles, lässt sich durch Gleichungen der Mechanik seine Flugbahn eindeutig bestimmen - seine Flugbahn ist somit deterministisch.
Ausnahmen realer Abläufe, in denen selbst das exakte Ergebnis nicht eindeutig ist, sondern nur Wahrscheinlichkeitsaussagen möglich sind, finden sich beispielsweise in der Quantenmechanik. Hier können in der Regel über Geschwindigkeiten, Positionen und Energien nur Wahrscheinlichkeitsaussagen getroffen werden, wodurch probabilistische Modelle zwingend notwendig sind.
Auch wenn die meisten Prozesse in Wirklichkeit deterministisch sind, sind probabilistische Modelle oft trotzdem sinnvoll. Eine Anwendungsmöglichkeit für probabilistische Modelle ergibt sich, wenn reale Prozesse approximiert werden sollen, um einen ansonsten zu komplexen Zusammenhang zu berechnen.
Probabilistische Modelle werden vor allem eingesetzt, wenn ein Prozess zu viele zu berücksichtigende Variablen und Parameter besitzt oder manche davon gänzlich unbekannt sind. Somit kann in der konkreten Anwendung, je nach probabilistischem Modell, die Datenmenge und die Berechnungszeit auf ein akzeptables Maß gesenkt werden.

Im Beispiel der Latent Dirichlet Allocation (Kapitel 4) werden Daten (hier Dokumente mit Wörtern) als Beobachtungen angesehen, die durch einen generativen, probabilistischen Prozess entstanden sind. Anschließend kann durch Inferenz (z.B. mittels Gibbs Sampling) der verborgene Erzeugungsprozess dieser Daten invertiert werden, d.h. es wird basierend auf den Dokumenten mit Wörtern eine Themenzugehörigkeit jedes Dokumentes approximiert. Um die Qualität eines approximierten probabilistischen Modelles zu prüfen, können anschließend weitere Dokumente betrachtet und daraufhin geprüft werden, wie gut das Dokument zu dem approximierten Modell und den Ursprungsdaten passt.

2.2 Generatives Modell

Generative Modelle beschäftigen sich mit der Frage, wie bestimmte Daten entstanden sind. Diese Modelle können sowohl deterministisch als auch probabilistisch sein.
Konkrete Beispiele im für diese Arbeit relevanten Zusammenhang zur Dokumentenerzeugung wären Folgende:

- Ein Beispiel für ein generatives deterministisches Modell bei einer Dokumentenerzeugung wäre, ein ganzzahliges N zu definieren und anschließend jedes N-te Wort des Dudens in das Dokument einzusetzen. Das erzeugte Dokument wäre für ein konstantes N immer identisch.
- Ein Beispiel für ein probabilistisches generatives Modell liegt der Latent Dirichlet Allocation (Kapitel 4) zugrunde. Vereinfacht dargestellt wird dort angenommen, dass jedes Wort mit einer bestimmten Wahrscheinlichkeit auftritt und die Wörter bei der Erzeugung des Dokumentes aus einem Wörterpool anhand dieser Wahrscheinlichkeiten gezogen werden.

2.3 Latente/Beobachtbare Variablen

Beobachtbare Daten sind jene Daten, welche direkt gesehen und gemessen werden können. In einem Dokument wären dies die einzelnen Wörter, die Satzstruktur und weitere Informationen, wie Autor oder Veröffentlichungsdatum (sofern diese am Dokument angehängt sind). In einem sozialen Netzwerk (hier ist ein allgemeines soziales Netzwerk gemeint, nicht speziell der Online-Dienst) wären dies die Freundschaften, falls diese bekannt oder offensichtlich sind.
Latente Variablen sind im Gegensatz zu beobachtbaren Variablen nicht direkt messbar. Es sind Informationen, welche verborgen sind, aber durch beobachtbare Variablen hergeleitet werden können. Im Beispiel eines Dokumentes sind dies z.B. die Themen, mit welchen sich das Dokument beschäftigt. Im Beispiel eines soziales Netzwerkes sind dies Freundeskreise, dessen Mitglieder untereinander eine hohe Freundschaftsdichte haben.

2.4 Community, Community Detection

Allgemein gesagt sind Communitys Gruppen aus einer Vielzahl von Individuen, die im Vergleich zu anderen Individuen eine oder mehrere gemeinsame oder änliche Eigenschaften besitzen. Allerdings hat sich für den Begriff Communtity keine allgemein anerkannte exakte Definition durchgesetzt. Für unterschiedliche Anwendungsgebiete sind daher auch unterschiedliche Definitionen erforderlich. Gängige Definitionen wurden von Fortunato zusammengefasst [7]. In dieser Arbeit zeichnen sich Mitglieder einer Community durch eine hohe Dichte von Freundschaften untereinander und durch ein Interesse an ähnlichen Themen aus. Um Communitys zu identifizieren, wird also nicht nur der soziale Graph verwendet. Es wird davon ausgegangen, dass Individuen zu mehreren Communitys gehören können, z.B. Familie, Sportgemeinschaften, Arbeitsgruppen. Der soziale Graph soll also in überlappende Communitys bzw. überlappende Cluster, genannt Cover, aufgeteilt werden. Hier sind Communitys das Ergebnis eines komplexeren Algorithmus und sind somit durch den verwendeten Algorithmus definiert und daher das bloße Ergebnis des Algorithmus, ohne eine genaue A-priori Definition. Die Community Detection befasst sich dementsprechend mit der Erkennung von Communitys innerhalb einer größeren Menge von Individuen.
Hier gibt es viele verschiedene Vorgehensweisen, deren Wahl von mehreren Faktoren abhängig ist. Zum einen hängt die Wahl von der Definition des Begriffes „Community“ ab. Zum anderen hängt es von der Struktur der vorhandenen Daten ab. Es macht schließlich einen Unterschied, ob z.B. ein Graph mit ungewichteten und ungerichteten Kanten vorliegt oder ob vielleicht noch gar keine direkt verwendbaren Daten vorhanden sind. Dies ist bei Online-Foren normalerweise der Fall, da von außen keine Informationen zu Freundschaften, Abonnements oder Blockierungen zwischen Usern vorliegen. Man muss also die vorhandenen Daten erst umwandeln, um diese einem Algorithmus übergeben zu können. Daher ist man bei der Wahl des Algorithmus nicht auf eine feste Struktur des sozialen Graphen angewiesen.
Um die Verschiedenheit der Vorgehensweisen zu zeigen, werden nun zwei einfache, dafür aber in Bezug auf Online-Foren eher ungeeignete Beispiele für Community Detection Verfahren vorgestellt. Einen detailierteren Überblick über das Thema Community Detection gibt Fortunato in seinen Werken [7].

2.4.1 Cliquenbildung

Eine Menge von Knoten wird Clique genannt, in welchen jeder Knoten mit jedem direkt verbunden ist. Bei der Cliquenbildung gibt es verschiedene Vorgehensweisen. Eine wäre das Suchen der größten Cliquen, d.h. das Suchen von Cliquen, die nicht schon in einer anderen, größeren Clique enthalten sind. Die Kanten sind hierbei ungerichtet und ungewichtet. Dadurch, dass ein Knoten zu mehreren Cliquen und somit auch zu mehreren Communitys gehören kann, würden User in Online-Foren mit verschiedenen Interessen

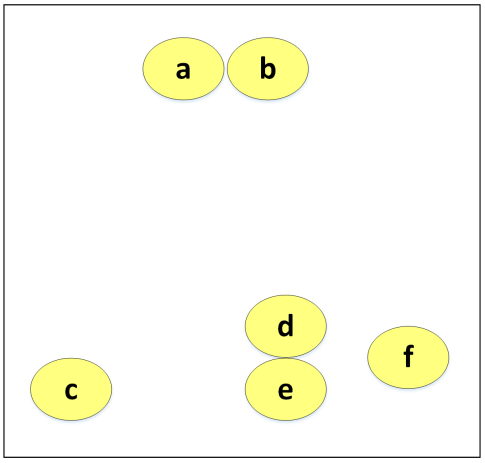

Abbildung 2.1: Datensatz: Knoten mit relativen Entfernungen

besser repräsentiert werden. Ein Nachteil bezogen auf Foren ist, dass je nach Definition einer Kante nur sehr viele kleine Cliquen auftreten.

2.4.2 Hierarchische Clusteranalyse

Eine andere Vorgehensweise wäre die Hierarchische Clusteranalyse. Hierbei besitzen die Knoten gewichtete Kanten. Diese Gewichtung kann je nach Zusammenhang als Distanz oder als Ähnlichkeit verschiedener Knoten interpretiert werden. Auch hierbei gibt es verschiedene Verfahren. So wird bei agglomerativen Clusterverfahren (auch „Bottom-up Verfahren“ genannt) wie im Beispiel in Abbildung 2.2 zuerst davon ausgegangen, dass alle Knoten jeweils einem Cluster der Größe 1 angehören. Anschließend werden diese Cluster anhand der Distanz der Knoten zusammengefasst. Dies wird für alle iterativ so lange durchgeführt, bis nur noch ein Cluster vorhanden ist, welches aus allen Knoten besteht. Ein Vorteil hierbei ist, dass die Kanten gewichtet sein können und keine scharfe Unterscheidung zwischen einer vollwertigen Kante und keiner Kante erforderlich ist. In Foren müsste also nicht entschieden werden, ob sich zwei User, die sich gerade erst kennenlernen nun „befreundet“ sind oder nicht. Ein weiterer Vorteil ist, dass die Clustergröße und somit die Größe von Communitys beliebig klein oder groß gewählt werden kann.

2.5 Topic Models

Topic Models basieren auf der Annahme, dass jedes Dokument verschiedene Themen beinhaltet, die jeweils unterschiedlich stark gewichtet sind. Man nimmt also die Existenz einer Wahrscheinlichkeitsverteilung über die Themen innerhalb des Dokuments an. Ebenso wird für jedes dieser Themen die Existenz einer weiteren Wahrscheinlichkeitsverteilung angenommen, welche das Vorkommen der Wörter innerhalb des Themas beschreibt.
Topic Models sind generative Modelle, in welchen davon ausgegangen wird, dass ein zu un-

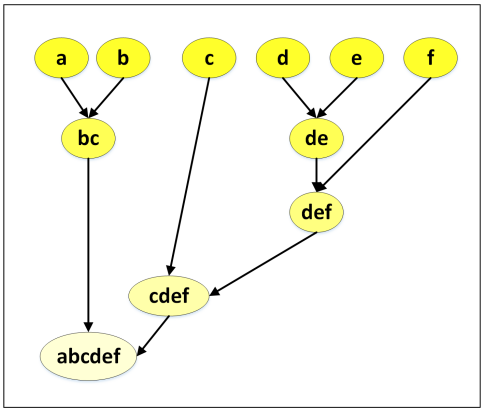

Abbildung 2.2: Baumdiagramm des Bottom-up Clusterings

tersuchendes Dokument aus einer Menge von Dokumenten folgendermaßen entstanden ist: Für das Dokument wurde zuerst eine Wahrscheinlichkeitsverteilung über Themen gewählt und die Anzahl in dem Dokument vorkommender Wörter festgelegt. Jedes Wort wurde erzeugt, indem aufgrund dieser Verteilung jeweils ein Thema gewählt wurde und aus der zugehörigen Wahrscheinlichkeitsverteilung jeweils ein Wort gezogen wurde. Ziel eines Topic Models ist es nun, für ein fertiges Dokument aufgrund dieser Annahmen Rückschlüsse auf die verwendeten Verteilungen und Themen zu ziehen. In anderen Worten soll der Entstehungsprozess mittels Inferenzstatistik invertiert werden.

Kapitel 3

Mathematisches Hintergrundwissen

In diesem Kapitel wird das für den weiteren Verlauf der Arbeit wichtige mathematische Hintergrundwissen erläutert.

3.1 A-priori-Wahrscheinlichkeit

Eine A-priori-Wahrscheinlichkeit, auch Ursprungswahrscheinlichkeit genannt, ist eine Wahrscheinlichkeitsverteilung, die in der Bayesschen Statistik einem Ereignis zugeordnet wird, bevor Daten erhoben wurden. Mögliches Vorwissen geht dabei schon in die A-priori-Wahrscheinlichkeit ein. Dabei wird versucht, die A-priori-Wahrscheinlichkeit so vorurteilsfrei wie möglich zu wählen, um kein falsches Vorwissen einzubauen. Ist überhaupt kein Vorwissen vorhanden, sollte jedem möglichen Ausgang die gleiche Wahrscheinlichkeit zugeordnet werden.
Will man beispielsweise den Ausgang eines Münzwurfes vorhersagen und hat kein Vorwissen über die Münze, würde man als A-priori-Wahrscheinlichkeit für beide Ausgänge einen Wert von 0.5 wählen. Wüsste man zusätzlich, dass die Münze unfair ist, also eine der beiden Seiten nur mit 20% auftritt, allerdings nicht welche von beiden, wäre 0.5 als A-priori-Wahrscheinlichkeit für Kopf weiterhin eine vernünftige Wahl. Denn hier müsste man sowohl das 20 als auch das 80 prozentige Auftreten von Kopf noch als gleichwahrscheinlich gewichten, sodass Kopf und Zahl weiterhin gleichwahrscheinlich wären.
Allerdings ist bei vorhandenem Vorwissen eine Gleichverteilung nicht immer die bevorzugte A-priori-Wahrscheinlichkeit. Folgendes Beispiel kann dies veranschaulichen: möchte man eine Wahrscheinlichkeitsverteilung aufstellen, wie viele von 6 Spielen bei einem Turnier torlos geblieben sind, so ist es bei dieser Fragestellung noch vertretbar, eine Gleichverteilung zwischen 0 und 6 anzunehmen, da kein weiteres Wissen, nicht einmal die Sportart, bekannt ist. Erhält man nun Vorwissen in Form der Information, dass insgesamt 8 Tore gefallen sind, so ist eine Gleichverteilung der torlosen Spiele zwischen 0 und 6 als A-priori-Wahrscheinlichkeit alleine deshalb schon nicht mehr sinnvoll, weil durch das zusätzliche

Vorwissen 6 torlose Spiele ausgeschlossen werden können. Vorwissen wie dieses wird für die A-priori-Wahrscheinlichkeit nach dem Prinzip des maximalen Unwissens (Entropie) eingearbeitet. Ein weiteres Beispiel findet sich in 3.3.

3.2 A-posteriori-Wahrscheinlichkeit

Die A-posteriori-Wahrscheinlichkeit unterscheidet sich von der A-priori-Wahrscheinlichkeit dadurch, dass hier nun Informationen aus einem Datensatz, also einer Beobachtung, eingegangen sind. Als Beispiel kann die in 3.1 genannte in unbekannter Richtung unfaire Münze herangezogen werden. Weiß man nicht, welche der Seiten zu 80% auftritt, ist eine Vorhersage von 50% für Kopf A-priori vernünftig. Hat man die Münze jedoch 6 mal geworfen und dabei 5 mal Kopf gesehen, würde unter Einbeziehung dieser Daten die A-posteriori-Wahrscheinlichkeit für Kopf beim folgenden Wurf entsprechend höher ausfallen.
Ein weiteres Beispiel findet sich im folgenden Abschnitt 3.3.

3.3 Satz von Bayes

Der Satz von Bayes ist einer der wichtigsten Sätze aus der Wahrscheinlichkeitstheorie. Dieser ermöglicht es, aus einer bedingten Wahrscheinlichkeit $P(B|A)$, welche die Wahrscheinlichkeit für das Ereignis B unter der Bedingung bezeichnet, dass A eingetreten ist, die Wahrscheinlichkeit für das Ereignis $P(A|B)$ zu ermitteln. Voraussetzungen dafür sind, dass die A-priori-Wahrscheinlichkeiten (Abschnitt 3.1) $P(A)$ und $P(B)$ bekannt sind. Ist dies der Fall, ergibt sich nach der Bayes-Formel folgende Umformung:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.1}$$

Ein Beispiel: man gehe davon aus, dass insgesamt eine Kugel aus zwei Behältern A_1 und A_2 mit je 10 Kugeln gezogen werden soll. Behälter A_1 besitzt 9 schwarze und 1 weiße Kugel, Behälter A_2 9 weiße und 1 schwarze. Die Wahrscheinlichkeit, dass aus Behälter A bzw. Behälter B gezogen wird, beträgt $\frac{1}{2}$. Diese Wahrscheinlichkeiten, einen der beiden Behälter zu wählen, werden A-priori-Wahrscheinlichkeit genannt. Wenn nun aber bekannt ist, dass eine weiße Kugel gezogen wurde und die Wahrscheinlichkeit gefragt ist, mit welcher sie aus Behälter A_1 gezogen wurde, dann wird diese Wahrscheinlichkeit „A-posteriori“ genannt. Diese kann anhand der Bayes-Formel berechnet werden:

- $P(A_1), P(A_2)$: Wahrscheinlichkeit aus Behälter A_1 bzw. Behälter A_2 eine Kugel zu ziehen.
- $P(B_1), P(B_2)$: Wahrscheinlichkeit eine weiße bzw. schwarze Kugel zu ziehen.
- $P(A_1|B_1)$, ... : Wahrscheinlichkeit, dass aus Behälter A_1 gezogen wurde, wenn die Kugel weiß ist.

- $P(B_1|A_1)$, ... : Wahrscheinlichkeit, dass die Kugel weiß ist, wenn aus Behälter A_1 gezogen wurde.

Nach der Bayes-Formel erhält man nun für vorhergehende Fragestellung:

$$P(A_1|B_1) = \frac{P(B_1|A_1) \cdot P(A_1)}{P(B_!)} = \frac{1/10 \cdot 1/2}{(0.5 \cdot 1/10) + (0.5 \cdot 9/10)} = 0.1 \tag{3.2}$$

3.4 Binomialverteilung

Die Binomialverteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die das Ergebnis einer Serie von identischen Versuchen beschreibt, die für sich genommen jeweils nur 2 Ergebnisse mit bekannten Wahrscheinlichkeiten haben können. Ein Beispiel für eine Binomialverteilung wäre die Anzahl an gewürfelten Zweien bei insgesamt 10 Würfen. Die Wahrscheinlichkeit für eine zwei bei einem einzelnen Wurf ist bekannt. Die Wahrscheinlichkeit für das Auftreten von 0, 1, 2, ..., 10 Zweien sind binomialverteilt.

3.5 Multinomialverteilung

Die Multinomialverteilung ist eine Verallgemeinerung der Binomialverteilung und damit ebenfalls eine diskrete Wahrscheinlichkeitsverteilung. Der wesentliche Unterschied ist lediglich, dass die Anzahl verschiedener Ergebnisse der Einzelversuche nicht mehr auf 2 beschränkt ist.

3.6 Dirichlet-Verteilung

Die Dirichlet-Verteilung [8, 9] gibt an, wie wahrscheinlich eine Multinomialverteilung ist. Die Wahrscheinlichkeitsdichtefunktion der Dirichlet-Verteilung mit einer k-dimensionalen Zufallsvariable und einem k-dimensionalem Parametervektor α lautet:

$$Dir(\alpha_1, ..., \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{3.3}$$

Die Dirichlet-Verteilung ist eine multivariate generalisierte Beta-Verteilung und ein konjugierter Prior der Multinomialverteilung [9]. Da bei einem konjugierten Prior der Prior und der Posterior zur gleichen Familie der Wahrscheinlichkeitsverteilungen gehören, vereinfacht die Verwendung eines konjugierten Priors die Berechnung des Posteriors deutlich und macht sie in vielen komplexen Fragestellungen, wie in dieser Arbeit, überhaupt erst möglich.

3.7 Poisson-Verteilung

Die Poisson-Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, mit welcher die Häufigkeit des Auftretens von Ereignissen beschrieben werden kann. Anhand des Parameters λ können den natürlichen Zahlen $k = 0, 1, 2, ...$ die Wahrscheinlichkeiten

$$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{3.4}$$

mit der eulerschen Zahl e zugewiesen werden. λ ist dabei zugleich Erwartungswert und Varianz dieser Verteilung.

Kapitel 4

Latent Dirichlet Allocation

In diesem Kapitel wird die Latent Dirichlet Allocation (LDA) vorgestellt, welches später mit dem Link Content Model verglichen werden soll.

4.1 Latent Dirichlet Allocation

Das wahrscheinlich bekannteste Topic Model ist die 2003 von Blei et al. [10] vorgestellte Latent Dirichlet Allocation, welches im folgenden erläutert wird. Latent Dirichlet Allocation ist ein generatives probabilistisches Modell, welches anhand von beobachtbaren Daten nicht sichtbare Zusammenhänge entdeckt. Die primäre Anwendung findet im Topic Modeling statt, wobei eine Sammlung von Dokumenten mit Wörtern untersucht wird und anhand dieser die inhaltlichen Themen/Topics ermittelt werden. Es existieren allerdings auch andere Anwendungsgebiete. So wird dieses Modell auch in der Bilderkennung angewendet, um z.B. in einem Bild zu erkennen, ob ein Wald oder ein geschlossener Raum dargestellt ist. Dabei stellt ein Bild ein Dokument dar und kleine Ausschnitte des Bildes werden als die Wörter angesehen[11].

4.1.1 Notation

Im LDA gibt es folgende drei Einheiten:

- Wörter eines Vokabulars {1,...,V} stellen die kleinste Einheit dar.
- Ein Dokument $\mathbf{w} = w_1, w_2, ..., w_N$ ist aus einer Anzahl N Wörtern zusammengesetzt. Hier bezeichnet w_n das n-te Wort in dem Dokument.
- Ein Korpus $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$ ist die Menge aller M relevanten Dokumente. $\mathbf{w}_m$ ist das m-te Dokument im Korpus.

Weiterhin ist:

- θ die Themenverteilung für alle Dokumente und θ_m die Verteilung für Dokument m

- ϕ die Verteilung der Wörter über alle Themen und speziell ϕ_z die Verteilung für der Wörter im Thema z

- β eine eine $K \times V$ Matrix, wobei K die Anzahl aller Themen ist. Diese Matrix gibt mit β_{ij} die Wahrscheinlichkeit an, dass aus einem Thema i das Wort j gezogen wird.

4.1.2 Generativer Prozess

Die LDA geht davon aus, dass alle Dokumente einer Dokumentsammlung aus einem generativen Prozess erzeugt wurden. Zuerst wird eine Anzahl von zu erzeugenden Dokumenten bestimmt. Anschließend ist die Vorgehensweise für die Erstellung jedes Dokumentes gleich und wird nun für ein einzelnes Dokument erläutert. Aus einer Poisson-Verteilung wird die Anzahl der Wörter bestimmt, die in dem Dokument vorkommen sollen. Aus einer Dirichlet-Verteilung wird eine Verteilung über Themen bestimmt. Das heißt, dass ein Dokument für jedes mögliche Thema eine Wahrscheinlichkeit aufweist. Diese Wahrscheinlichkeiten summieren sich zu eins auf und geben an, welche Themen wie stark in diesem Dokument vertreten sind. Hier entstehen für jedes Dokument andere Themengewichtungen. Es wird davon ausgegangen, dass in einem Dokument nur sehr wenige Themen vorkommen. Dies wird über die Dirichlet-Verteilung $Dir(\alpha)$ (siehe Kapitel 3.6, Formel (3.3)) realisiert. Alle Komponenten des Vektors α werden im Normalfall mit dem gleichen Wert belegt. Ein kleiner Wert für α bedeutet, dass die Dokumente eine starke Tendenz zu einem Thema aufweisen und nur selten weitere Themen im Dokument vorhanden sind. Ein hoher Wert bedeutet, dass die Dokumente jeweils viele Themen enthalten, welche im Normalfall gleich stark gewichtet sind. Jedes Thema besitzt eine Wahrscheinlichkeitsverteilung über Wörter. Diese Verteilungen geben für das jeweilige Thema an, wie häufig jedes Wort relativ gesehen vorkommen sollte. Wenn man z.B. ein Buch über das Thema „Meer“ schreibt, könnten 2% aller Wörter „Meer“, 1% „Wasser“ und 0.1% der Wörter im Buch „Strand“ sein. Diese Verteilungen ϕ_z können je nach Thema z unterschiedlich sein und werden aus der Dirichlet-Verteilung $Dir(\beta)$ gezogen. Auch hier sind einzelne Wörter für ein Thema relevanter als andere. Daher ist auch β normalerweise klein (α, $\beta < 1$). Nun wird das Dokument mit Wörtern gefüllt, indem wiederholt zuerst ein Thema aus der Themenverteilung des Dokumentes gezogen wird und anschließend aus diesem Thema ein Wort gezogen wird. Hierbei ist es wichtig zu erwähnen, dass die Reihenfolge der Wörter eines Dokumentes als nicht relevant angenommen wird. Die Vorgehensweise lässt sich anhand des Plattenmodells (Abbildung 4.1) zusammenfassend wie folgt beschreiben:

1. Ziehe dokumentenübergreifend $\phi \sim Dir(\beta)$, eine Verteilung von Wörtern über alle Themen

Wiederhole für jedes zu erzeugende Dokument:

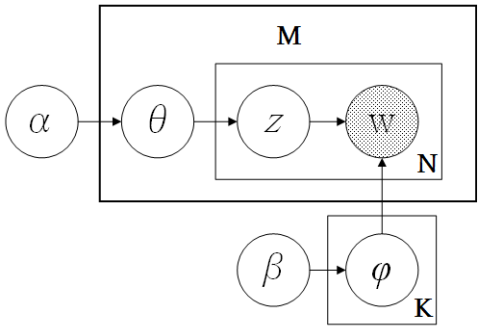

Abbildung 4.1: Plate Modell für LDA

2. Ziehe $N \sim Poisson(\xi)$, die Anzahl der Wörter des Dokumentes (siehe Kapitel 3.7, Formel (3.4))
3. Ziehe $\theta \sim Dir(\alpha)$, eine gewichtete Verteilung von Themen des Dokumentes
4. Für jedes w_n der N Wörter:
   (a) Ziehe ein Thema $z_n \sim Multinomial(\theta)$
   (b) Ziehe ein Wort w_n anhand von ϕ_z aus diesem Thema

4.1.3 Inferenz

Üblicherweise liegen die erzeugten Wörter als Daten vor, man ist jedoch an den verborgenen Wahrscheinlichkeiten im Entstehungsprozess interessiert. Der obige generative Prozess wird nun in der LDA invertiert. Man verwendet Bayes-Inferenz, um aus den erzeugten und sichtbaren Wörtern die posterior Verteilung über die Themenverteilung θ für jedes Dokument und die Wortverteilungen jedes Thema zu berechnen. An dieser Stelle ist in Kapitel 3.6 erwähnte Tatsache von Bedeutung, dass die Dirichlet-Verteilung der zur Multinomialverteilung konjugierte Prior ist. Dadurch lässt sich der Ausdruck für den Posterior und seine Bestimmung deutlich vereinfachen. Im folgenden Kapitel 5 wird dies anhand des konkreten Algorithmus des Link-Contet-Modells veranschaulicht und erneut aufgegriffen. Dennoch ist der posterior in der Regel nicht analytisch bestimmbar, sodass man sich anderweitig behelfen muss. Hierbei gibt es drei übliche Ansätze: Expectation Propagation [12], Variational Bayes [10] und Collapsed Gibbs Sampling [13]. Die letzteren beiden Methoden sind die Methoden, die am häufigsten verwendet werden [14]. In dieser Arbeit wird Gibbs Sampling verwendet. Vereinfacht dargestellt ist Gibbs Sampling eine Markov Kette, die am Ende nach mehrfacher Iteration gegen die gesuchte A-Posteriori-Wahrscheinlichkeitsverteilung konvergiert, siehe auch [15, 13, 16]. Auf die Vorgehensweise wird etwas detaillierter in dem Kapitel 5 „Link-Content Modell“ anhand des Algorithmus nochmal eingegangen.

Kapitel 5

Link-Content Modell

Das Paper „Community Detection in Content-Sharing Networks“ von Natarajan et al., 2013, [6] befasst sich am Beispiel Twitter mit der Themen- und Communityerkennung in sozialen Netzwerken. Dabei wird ein probabilistisches, generatives Modell vorgestellt, welches beide Probleme simultan bearbeitet. Die Idee basiert auf der Annahme, dass die soziale Struktur und der thematische Inhalt voneinander abhängen. So haben befreundete User wahrscheinlich auch gemeinsame Interessen und anders herum ist die Wahrscheinlichkeit, dass User mit ähnlichen Interessen befreundet sind, höher als bei Usern, die völlig unterschiedliche Interessen haben. Im Gegensatz zu anderen Modellen, wie der LDA, wird davon ausgegangen, dass die Themen der Dokumente von den Communitys gezogen werden und nicht von den Usern. Im Folgendem wird der Inhalt des Modelles und Papers vorgestellt. Dabei werden einige Hinweise zum Paper gegeben und es wird auf einige kleinere Fehler im Paper und im Algorithmus hingewiesen. Dies ist nicht als Kritik an die Autoren aufzufassen, sondern dient lediglich dem Verständnis und der Vollständigkeit.

5.1 Notation

$\mathbf{U}$ bezeichnet eine Menge von Usern. $S \subseteq \mathbf{U} \times \mathbf{U}$ ist eine Menge von gerichteten Kanten zwischen jeweils zwei Usern und stellt das soziale Geflecht dar. Im Folgenden wird eine gerichtete Kante als (gerichtete) Freundschaft definiert. Dabei ist zu beachten, dass es sich hierbei nicht nur um eine Freundschaft handeln muss, sondern auch eine andere soziale Beziehung darstellen kann. Im Link Content Modell ist im Anwendungsfall Twitter User v ein Freund von u, wenn u dem User v folgt oder anders gesagt u ein Follower von v ist. Dies wird dargestellt als $u \to v$. Die Menge aller Freunde von u ist $L_u = \{v | u \to v \in S\}$. Eine Besonderheit bei diesem Modell ist, dass Freundschaften gerichtet sind. So kann z.B. User u ein Interesse an den Posts von User v haben, aber anders herum v User u gar nicht kennen oder die Posts von u entsprechen nicht den Interessen von v. Die Freundschaften sind vorher bekannt und werden dem Algorithmus vorgegeben. Die Parameter K

Observed quantities	
S	Social network graph
$\mathbf{D}$	Docs. in the social network
$\mathbf{U}$	Users in the social network
D_u	Content uploaded by user u
L_u	Friends of u
V	Vocabulary size

Input parameters	
K	Number of communities
T	Number of topics

Hyper-parameters	
ν	Hyper-parameter for δ_u
μ	Hyper-parameter for ψ_k
α	Hyper-parameter for θ_k
β	Hyper-parameter for ϕ_z

Counts	
M_{uk}	#docs. $\in D_u$ community k generates
M_{kz}	#docs. community k & topic z generate
F_{uk}	#friends $\in L_u$ community k generates
F_{kv}	#times community k generates v
N_{zw}	#times topic z generates w

Latent variables	
δ_u	User u's preference over communities
ψ_k	Community k's distribution over users
θ_k	Topic distribution of community k
ϕ_z	Word distribution of topic z

Abbildung 5.1: Zusammenfassende Legende der Notation, übernommen aus [6].

und T bezeichnen die Anzahl an zu findenden Communitys bzw. Topics und werden ebenso vorgegeben. $\mathbf{D}$ bezeichnet eine Menge von Dokumenten. Diese sind im Falle Twitter einzelne Posts, die von einem User erstellt wurden. Ein soziales Netzwerk besteht aus dem Triple $\langle \mathbf{U}, S, \mathbf{D} \rangle$.

Die Modellierung von Communitys und Topics geschieht, wie auch beim LDA, über multinomiale Wahrscheinlichkeitsverteilungen. Communitys werden als multinomiale Verteilung über die User U dargestellt. ψ_k ist die Verteilung der k-ten Community. $\psi_k(u)$, im Folgenden mit ψ_{ku} abgekürzt, ist der Anteil bzw. Stellenwert eines Users u an der Community k. Andersherum existiert jeweils eine Verteilung von Communitys über einen User δ_u, welche jeweils den Einfluss einer Community k auf den User u mit $\delta_u(k)$, oder kurz δ_{uk}, beschreibt. Es wird davon ausgegangen, dass jede Community Themenschwerpunkte hat. Die Verteilung der Themen in einer Community k ist θ_k und der Anteil eines bestimmten Themas z in dieser Community dementsprechend θ_{kz}. Die Verteilung aller Wörter des Vokabulars über ein Thema z ist ϕ_z, bzw. der Anteil eines Wortes w am Thema z ist ϕ_{zw}. Wenn man dies aus Sicht des generativen Prozesses betrachtet, ist ϕ_{zw} die Wahrscheinlichkeit, dass aus einem gezogenem Thema z das Wort w gezogen wird. Weiterhin werden verschiedene Counts verwendet. M_{uk} zählt die Anzahl der Dokumente von User u, die durch Community k generiert werden. M_{kz} zählt, wie viele Dokumente genau aus dem Community-Themen-Paar $\langle k, z \rangle$ entstanden ist. F_{uk} gibt an, wie viele Freunde von u aus der Community k stammen und F_{kv} dementsprechend wie viele Freunde insgesamt die Community k generiert hat. Zuletzt zählt N_{zw} in allen Dokumenten, wie oft ein Wort w aus einem Thema z gezogen wurde. Da es schwer fallen wird sich direkt alle Variablen zu merken gibt Abbildung 5.1 einen schnellen Überblick.

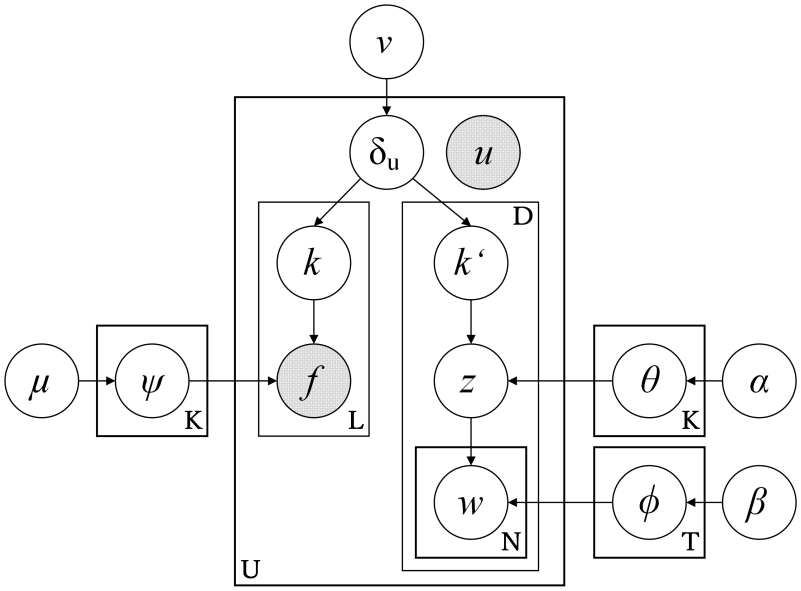

Abbildung 5.2: Platte Notation des Link-Content Modells, übernommen aus [6].

5.2 Generative Semantik und Plate Notation

Das Link-Content Modell bildet soziale Netzwerke mit ihren Usern, Links und Dokumenten/Einträgen ab. Da Twitterposts oft nur kurz und die Anzahl der Zeichen auf 140 begrenzt sind, wird davon ausgegangen, dass ein Post auch nur ein Thema besitzt. Der generative Prozess, auf dem der Algorithmus basiert, wird nun mithilfe der Plate Notation (Abb. 5.2) vorgestellt. Hierbei stehen die Kreise/Knoten für Zufallsvariablen und die Rechtecke/Plates für Wiederholungen. Graue Knoten stehen für beobachtbare Zufallsvariablen und die weißen Knoten werden mittels Inferenz bestimmt. Im Folgenden wird der generative Prozess beschrieben: für jedes Dokument, dass User u erzeugen soll, wählt das Modell anhand der Communityverteilungen δ_u von u eine Community aus, für die das Dokument geschaffen wird. Anschließend wird für das Dokument ein Thema z aus k's Themenverteilung θ_k und Wörter aus diesem Thema anhand von ϕ_z. Auf der sozialen Ebene wird für jeden Freund $v \in L_u$ erst eine Community aus δ_u gewählt und dann ein Freund mittels ψ_k generiert. So basiert jede Freundschaft auf einer Community. Da die Freundschaften gerichtet sind, gibt es zwei Verbindungen zwischen zwei Freunden, falls die Freundschaft beidseitig ist. Weiterhin existieren Hyperparameter ν, μ, α, β für die latenten Variablen δ_u, ψ_k, θ_k, ϕ_z. Diese sind vorgegebene Parameter, durch deren Variation beeinflusst werden kann, ob davon ausgegangen wird, dass z.B. Themen eines Dokumentes mit höherer Wahrscheinlichkeit gleich verteilt (größere Hyperparameter) sind oder einzelne Themen mit höherer Wahrscheinlichkeit das Dokument dominieren (kleinere Hyperparameter).

5.3 Inferenzprozess

Ein sehr wichtiger Aspekt von generativen Modellen ist die Bestimmung der latenten Variablen, auf denen das Modell basiert. Dafür ist die Inferenz zuständig. Es wird ein Gibbs-Sampling Algorithmus angewendet um Samples zu ziehen.

Im Folgenden stehen die großen Buchstaben der latenten Variablen Θ, Φ, etc. für die Gesamtheit der jeweiligen Wahrscheinlichkeitsverteilungen. Die Wahrscheinlichkeit, dass ein Dokument d von User u erzeugt wird, ist $P(d|\delta_u, \Theta, \Phi)$. Verallgemeinert gesagt wird die Wahrscheinlichkeit, dass genau das Tupel $\langle d, k, z\rangle$ auftritt, über alle Themen und Communitys summiert:

$$P(d|\delta_u, \Theta, \Phi) = \sum_z \sum_k P(d, k, z|\delta_u, \Theta, \Phi) \tag{5.1}$$

mit

$$P(d, k, z|\delta_u, \Theta, \Phi) = P(k|\delta_u)P(z|\theta_k) \prod_{w \in d} P(w|\phi_z) = \delta_{uk}\theta kz \prod_{w \in d} \phi_{zw} \tag{5.2}$$

Die Wahrscheinlichkeit, dass v ein Freund von u ist, kann gleichermaßen umgeformt werden:

$$P(v|\delta_u, \Psi) = \sum_k P(v, k|\delta_u, \Psi) mit P(v, k|\delta_u, \Psi) = P(k|\delta_u)P(v|\psi_k) = \delta_{uk}\psi_{kv} \tag{5.3}$$

Mit diesen Wahrscheinlichkeiten und unter Berücksichtigung der Hyperparameter kann das komplette soziale Netzwerk $\langle \mathbf{U}, S, \mathbf{D}\rangle$ dargestellt werden als

$$P(\mathbf{U}, S, \mathbf{D}, \mathbf{K}, \mathbf{Z}, \Theta, \Phi, \Delta, \Psi; \alpha, \beta, \nu, \mu) \propto$$
$$\prod_{\substack{u=1\\k=1}}^{U,K} \delta_{uk}^{M_{uk}+F_{uk}+\nu_k-1} \prod_{\substack{k=1\\u=1}}^{K,U} \psi_{ku}^{F_{ku}+\mu_u-1} \prod_{\substack{k=1\\z=1}}^{K,T} \theta_{kz}^{M_{kz}+\alpha_z-1} \prod_{\substack{z=1\\w=1}}^{T,V} \phi_{zw}^{N_{zw}+\beta_w-1} \tag{5.4}$$

Diese Gleichung kann durch Marginalisierung (also Integration über $\Theta, \Phi, \Delta, \Psi$) vereinfacht werden zu:

$$P(\mathbf{U}, S, \mathbf{D}, \mathbf{K}, \mathbf{Z}; \alpha, \beta, \nu, \mu) =$$
$$\prod_{u=1}^{U} \frac{\prod_k \Gamma(M_{uk} + F_{uk} + \nu_k)}{\Gamma(M_{u\cdot} + F_{u\cdot} + \sum_k \nu_k)} \prod_{k=1}^{K} \frac{\prod_u \Gamma(F_{ku} + \mu_u)}{\Gamma(F_{k\cdot} + \sum_u \mu_u)} \prod_{k=1}^{K} \frac{\prod_z \Gamma(M_{kz} + \alpha_z)}{\Gamma(M_{k\cdot} + \sum_z \alpha_z)} \prod_{z=1}^{T} \frac{\prod_w \Gamma(N_{zw} + \beta_w)}{\Gamma(N_{z\cdot} + \sum_w \beta_w)} \tag{5.5}$$

Hierbei steht ein im Subscript stehender Punkt "$\cdot$" für die Summe über alle entsprechenden Indizes, also $M_{u\cdot} = \sum_{k=1}^{K} M_{uk\cdot}$.

5.4 Gibbs Sampling für das Link-Content Modell

Für Topic-Models können verschiedene Inferenzverfahren verwendet werden. Aufgrund der Genauigkeit, relativen Effizienz und der Einfachheit wurde in [6] Gibbs Sampling [17] verwendet. Um Gibbs Sampling verwenden zu können sind für alle Zuordnungen des Modells

bedingte Wahrscheinlichkeiten erforderlich. Diese geben an, wie eine Zuordnung von allen anderen abhängig ist. Auf die detaillierte Herleitung aus Gleichung 5.5 wird hier verzichtet, sie entspricht dem Standardverfahren [13]. Bei jeder Samplingiteration (Zeilen 14-21 des Algorithmus in Abb. 5.3 unten) wird jeder Freundschaft eine Community zugewiesen und jedem Dokument wird ein Community-Topic-Paar zugewiesen. Dies geschieht durch ziehen aus folgenden bedingten Wahrscheinlichkeitsverteilungen: Die Wahrscheinlichkeit, dass die User u und v durch die Community k verbunden sind (Z. 16 des Algorithmus), lautet:

$$P(k|\mathbf{U}, S, \mathbf{D}, \mathbf{K}, \mathbf{Z}; \alpha, \beta, \nu, \mu) \propto \frac{M^{-}_{uk} + F^{-}_{uk} + \nu_k}{M^{-}_{u\cdot} + F^{-}_{u\cdot} + \sum_{k'} \nu_{k'}} \frac{F^{-}_{kv} + \mu_v}{F^{-}_{k\cdot} + \sum_{u'} \mu_{u'}} \tag{5.6}$$

Zu beachten hierbei ist, dass ein "$-$" im Superskript bedeutet, dass bei den Counts die Entität, über die gesamplet wird, nicht berücksichtigt wird. D.h. in Zeile 16 bezeichnet F^{-}_{uk} die Häufigkeit, wie oft Community k einen Freund von u generiert hat, ohne den Freund v über den gesamplet wird zu berücksichtigen. Im Paper wurden alle Counts mit einem Minus im Superskript dargestellt. Dies ist unter Umständen etwas missverständlich. So wird in Zeile 16 über die Freunde gesamplet und nicht über Dokumente. Das heißt, dass die Superskripte in M^{-}_{uk} und $M^{-}_{u\cdot}$ wirkungslos sind. In Gleichung (5.6) misst der erste Term, wie oft k verwendet wurde, um Dokumente von u zu erzeugen, oder um andere Freunde von u (außer v) zu erzeugen. Der zweite Term gibt an, wie oft User v von Community k als ein Freund anderer User erzeugt wurde. Die Wahrscheinlichkeit, dass ein Dokument d von u das Community-Topic-Paar $\langle k, z\rangle$ erhält (Z. 19-20 des Algorithmus), lautet:

$$\begin{aligned} &P(k, z|\mathbf{U}, S, \mathbf{D}, \mathbf{K}, \mathbf{Z}; \alpha, \beta, \nu, \mu) \propto \\ &\frac{M^{-}_{uk} + F^{-}_{uk} + \nu_k}{M^{-}_{u\cdot} + F^{-}_{u\cdot} + \sum_{k'} \nu_{k'}} \frac{M^{-}_{kz} + \alpha_z}{M^{-}_{k\cdot} + \sum_{z'} \alpha_{z'}} \prod_{w \in d} \prod_{i=1}^{n_{dw}} \frac{N^{-}_{zw} + i - 1 + \beta_w}{N^{-}_{z\cdot} + \sum_{w' < w} n_{dw'} + i - 1 + \sum_{w'} \beta_{w'}} \end{aligned} \tag{5.7}$$

Da hier über Dokumente gesamplet wird, sind die Superskripte in F^{-}_{uk} und $F^{-}_{u\cdot}$ wirkungslos, da in diesen beiden Counts die Freunde und nicht die Dokumente eine Rolle spielen. In Zeile 18 ist anzumerken, dass nicht über alle Dokumente D iteriert wird, sondern nur die Dokumente des Users u, also „**for** $d \in D_u$ ". Weiterhin iteriert im Gegensatz zu Gleichung 5.1 das Produkt $\prod_{w \in d}$ im dritten Term nicht mehrfach über ein mehrfach vorkommendes Wort im Dokument. Dafür wird ein weiteres Produkt mit der auftretenden Häufigkeit des Wortes verwendet. Der erste Term misst, wie oft k verwendet wurde, um andere Dokumente (das Dokument über welches gerade iteriert wird, wird nicht berücksichtigt) von u zu erzeugen, oder um dessen Freunde zu erzeugen. Der zweite Term misst, wie oft Thema z von Community k erzeugt wurde. Dies geschieht ebenso ohne Berücksichtigung des aktuellen Dokumentes. Der letzte Term misst die Wahrscheinlichkeit, dass die Wörter in d von Thema z erzeugt wurden. Beim Lesen des Originalpapers ist hierbei zu beachten, dass sich anscheinend auf Seite 4 (unten rechts) ein Fehler eingeschlichen hat. Es soll wohl nicht gemeint sein: „... generate other documents posted by u or *by* friends of u, ...", sondern „...

generate other documents posted by u or *generate* friends of u, ...". Denn der erste Term von Zeile 19 im Algorithmus ist identisch mit dem ersten Term von Zeile 16 (außer natürlich dem Einfluss des „−" im Superskript). Die Beschreibung der beiden identischen Terme unterscheidet sich im Text von Natarajan et al., 2013, allerdings inhaltlich signifikant.

5.5 Algorithmus

Im Folgenden wird der Algorithmus des Gibbs Sampling im Link-ContentModells vorge stellt und erläutert. Der Algorithmus ist in drei Phasen unterteilt:
Zuerst wird die Initialisierung durchgeführt. Dabei wird allen Freundschaften jedes Users eine zufällige Community zugewiesen. Zusätzlich erhalten alle Dokumente jedes Users ein zufälliges Community-Topic-Paar. Anschließend kommt die Burn-in Phase, in dem sich die Markov-Kette stabilisieren soll. Die Counts sind von den Zuweisungen der Initialisierung abhängig. Da die Initialisierung komplett zufällig geschieht, macht es wenig Sinn, direkt Samples zu sammeln. Wegen der anfänglich rein zufälligen Counts würden die ersten Samples das Endergebnis verfälschen. Der Burn-in Prozess sorgt dafür, dass die Counts Werte besitzen, die der Realität eher entsprechen. Nach dem Burn-in wird weiter gesamplet. Allerdings werden nun die Samples gesammelt, um daraus die latenten Variablen zu ermitteln (siehe folgendes Kapitel). Da unmittelbar aufeinanderfolgende Samples autokorreliert sind, werden regelmäßig Samples verworfen, um möglichst voneinander unabhängige Samples zu sammeln.

5.6 Ermitteln latenter Variablen

Nachdem die Samples gesammelt wurden, können anhand folgender Formeln die latenten Variablen bestimmt werden:

$$\begin{aligned}
\delta_{uk} &= \frac{M_{uk} + F_{uk} + \nu_k}{M_{u\cdot} + F_{u\cdot} + \sum_{k'} \nu_{k'}}\,, \\
\theta_{kz} &= \frac{M_{kz} + \alpha_z}{M_{k\cdot} + \sum_{z'} \alpha_{z'}}\,, \\
\phi_{zw} &= \frac{N_{zw} + \beta_w}{N_{z\cdot} + \sum_{w'} \beta_{w'}}\,, \\
\psi_{ku} &= \frac{F_{ku} + \mu_u}{F^{-}_{k\cdot} + \sum_{u'} \mu_{u'}}\,.
\end{aligned} \tag{5.8}$$

Da das Sampling auch vom Zufall abhängt, kann es sein, dass im unglücklichen Fall einzelne Samples nicht der Realität entsprechen. Daher werden, wie Ende des vorherigen Kapitels 5.5 beschrieben, viele möglichst unabhängige Samples gesammelt, um einen Durchschnittswert zu bilden, welcher eine genauere Schätzung der latenten Variablen ermöglicht.

Algorithm 1: Gibbs sampling for Link-Content

/*Initialize*/
for $u \in \mathbf{U}$ **do**
  for $v \in L_u$ **do**
    $k \sim$ uniform$[1 \dots K]$
    Assign community k to link $u \rightarrow v$
  for $d \in D_u$ **do**
    $k \sim$ uniform$[1 \dots K]$
    $z \sim$ uniform$[1 \dots T]$
    Assign community-topic pair $\langle k, z \rangle$ to d
/* Burn-in */
$I \leftarrow$ number of burn-in iterations
$i \leftarrow 0$
while $i < I$ **do**
  for $u \in \mathbf{U}$ **do**
    for $v \in L_u$ **do**
      $k \sim \frac{M^{-}_{uk}+F^{-}_{uk}+\nu_k}{M^{-}_{u\cdot}+F^{-}_{u\cdot}+\sum_{k'}\nu_{k'}} \frac{F^{-}_{kv}+\mu_v}{F^{-}_{k\cdot}+\sum_{u'}\mu_{u'}}$
      Assign community k to link $u \rightarrow v$
    for $d \in D_u$ **do**
      $\langle k, z \rangle \sim \frac{M^{-}_{uk}+F^{-}_{uk}+\nu_k}{M^{-}_{u\cdot}+F^{-}_{u\cdot}+\sum_{k'}\nu_{k'}} \frac{M^{-}_{kz}+\alpha_z}{M^{-}_{k\cdot}+\sum_{z'}\alpha_{z'}}$
      $\prod_{w \in d} \prod_{i=1}^{n_{dw}} \frac{N^{-}_{zw}+i-1+\beta_w}{N^{-}_{z\cdot}+\sum_{w'<w} n_{dw'}+i-1+\sum_{w'}\beta_{w'}}$
      Assign community-topic pair $\langle k, z \rangle$ to d
  $i = i + 1$
/* Collect samples */
$A \leftarrow$ number of samples to be collected
$B \leftarrow$ number of iterations to be skipped
for $i \in 1 \dots A$ **do**
  for $j \in 1 \dots B$ **do**
    Sample $\langle k, z \rangle, \forall d \in \mathbf{D}$ and $k, \forall v \in L_u, u \in \mathbf{U}$
  Collect sample

Abbildung 5.3: Gibbs Sampling für das Link-Content Modell, ergänzte Abb. aus [6]. Für das Verständnis und dem Vergleich zum Paper wurde in grün ein fehlender Subskript ergänzt und in blau die Counts markiert, deren Superskripts ohne Bedeutung sind.

Kapitel 6

Theorie der Datenvorverarbeitung

Im Laufe der Bearbeitung dieser Arbeit wurde deutlich, dass gerade die Aufbereitung der zur Verfügung stehenden Daten einen wichtigen Aspekt darstellt und für eine erfolgreiche Datenanalyse maßgeblich ist. Daher wird in diesem Kapitel genauer auf theoretische Methoden der Datenvorverarbeitung eingegangen. Dazu wurden einige Techniken des sogenannten „Information Retrieval“ (IR) [18, 19, 20, 21] angewendet und der Effekt der Wortfilterung auf das Ergebnis untersucht.

Der Begriff Information Retrieval beschreibt den Prozess der (Wieder-)Gewinnung von Informationen aus einem bestehenden Datenbestand zur Lösung von Fragestellungen in einer konkreten Situation [21, S. 18 ff]. Er steht somit in engem Zusammenhang mit Suchmaschienen. In der Literatur wird dieser Begriff aber oft weiter gefasst und umfasst auch das Speichern, den Zugriff und die Auswertung der Daten. Der Prozess wird hierbei meist mit dem Suchen von Informationen in großen Datenbeständen in Verbindung gebracht [22, S. 7] [23, S. 7]. Auch wird allgemein betont, dass die Qualität der Datenvorverarbeitung entscheidenden Einfluss auf die Qualität der Ergebnisse der weiteren Auswertung hat.
Hierbei ist jedoch zu beachten, dass jede Anwendung unterschiedlich auf die verschiedenen Ansätze reagiert [24]. Es werden verschiedene Ansätze verwendet, die zum einen auf eine Normalisierung der Daten und zum anderen auf eine Reduktion der Daten auf Teile mit hoher Informationswertigkeit abzielen. In den folgenden Abschnitten werden die wichtigsten Methoden beim Information Retrieval kurz erläutert.

6.1 Rohdatenbereitstellung

Wenn, wie in dieser Arbeit, die Datenbestände außerhalb der Anwendung eingesetzt werden sollen, müssen diese zunächst extrahiert werden. Im Folgenden wird die Vorgehensweise bei der Verarbeitung von Textdaten betrachtet.
Bei dem eigentlichen Extrahieren werden oft die Zeichensätze bestimmt und gegebenenfalls konvertiert. Anschließend erfolgt eine Analyse der Rohdaten hinsichtlich ihrer Struktur,

was eine Gewinnung von Metainformationen (Informationen über die Daten) einschließt. So können aus einem Datenextrakt eines Forums auch Detailinformationen wie Autor, Veröffentlichungsdatum, Threadname sowie der eigentliche Text des Threads bzw. eines Posts gewonnen und getrennt gespeichert werden. Gerade diese Metainformationen waren für diese Arbeit wichtig und standen im relNet Projekt bereits zur Verfügung.

6.2 Lexikalische Vorverarbeitung

In diesem Zusammenhang erfolgt die Normalisierung der Wörter. Dabei erfolgt im Allgemeinen eine Umwandlung von Großbuchstaben in Kleinbuchstaben sowie eine Behandlung der Sonderzeichen. So werden alle diakritischen Zeichen, also Akzente über oder unter Buchstaben wie è, é, ê, entweder entfernt oder wie bei den deutschen Umlauten ¨ durch den Buchstaben „e“ ersetzt, beispielsweise ä $\rightarrow$ ae. Vorkommende Zahlen werden häufig ebenso entfernt. Wenn die Zahlen jedoch für einen Kontext von Bedeutung sind, kann auch die Bildung einer Phrase hilfreich sein. Beispielsweise würde bei einer Textstelle wie „schwerer Unfall im Jahre 1991“ ein Wegfallen des Datums einen signifikanten Informationsverlust bedeuten. Im Topic Modeling kann im besten Fall alleine durch ein einziges Datum einem Dokument ein sehr wahrscheinlich vorkommendes Thema zuordnen. So weist z.B. das Datums „09112001“ mit hoher Wahrscheinlichkeit auf die Terroranschläge des 11. Septembers hin und dementsprechend auf das Thema „Terror“. Gerade bei Suchmaschienen lassen sich bei Berücksichtigung der Jahreszahl bessere Suchergebnisse erzielen. Auch kann eine Erweiterung von Abkürzungen und die Anwendung eines einheitlichen Datenformates erfolgen.
In den folgenden Schritten erfolgt, je nach Zielsetzung, eine Zerlegung auf Absatz-, Satz- und schließlich Wortebene. Letztere, im englischen auch „Tokenizing“ genannt, ist die eigentlichen Aufgliederung des Textes in Wörter. Wenn keine Analyse auf Satzebene erfolgen soll, werden die Satzzeichen entfernt und damit die Satzstruktur verworfen. Für diese Arbeit war lediglich eine Zerlegung auf Wortebene erforderlich.

6.3 Linguistische Datenvorverarbeitung

Der Übergang zur linguistischen Datenvorverarbeitung wird in der Literatur nicht einheitlich betrachtet. In manchen Fällen wird die Normalisierung in diesem Bereich mit eingeschlossen. Eigentliches Ziel der linguistischen Datenvorverarbeitung ist es, den Datensatz zu reduzieren, um die weitere Verarbeitung zu beschleunigen, dabei aber die relevanten Informationen zu behalten. In den folgenden Abschnitten werden die wichtigsten Schritte der typischen Vorgehensweise kurz beschrieben.

6.3.1 Eliminierung von Wörtern mit geringer inhaltlichen Relevanz

Die im englischen „stop words“ genannten Wörter (Füllwörter mit geringer Unterscheidungskraft) liefern nur einen geringen Beitrag zum Informationsgehalt. [4]. Hierzu zählt man Artikel, Konjunktionen, Pronomen und Präpositionen. Im Folgenden werden verschiedene Vorgehensweisen zur Eliminierung von stop words vorgestellt [25, S. 7].

- Vorgegebene Stopwortlisten:
  Die einfachste und klassische Form ist die Nutzung von festen, „manuell“ erstellten Wortlisten. In der Literatur findet man unter anderem Vorschläge von van Rijsbergen [26] mit 250 und Christopfer Fox [27] mit 421 Wörtern für den englischen Sprachraum.
- Das Zipfsche Gesetz als Basis der empirischen Ermittlung von Wörtern geringer inhaltlicher Bedeutung:
  Sortiert man Wörter in einem Dokument nach der Häufigkeit ihres Auftretens, so besagt das Zipfsche Gesetz aus den 1930er Jahren [28], dass die Häufigkeit eines Wortes $f(w)$ umgekehrt proportional zu seiner Position $r(w)$ in der Liste ist [29, 30]. Die Position des Wortes in einer nach absteigender Häufigkeit sortierten Liste wird auch als Rang bezeichnet. Anders ausgedrückt besagt das Zipfsche Gesetz also, dass das Produkt aus der Häufigkeit und dem Rang konstant ist [21]:

$$r(w) \cdot f(w) = c \tag{6.1}$$

  Die Konstante c ist dabei abhängig vom betrachteten Wortschatz. Das Zipfsche Gesetz lässt sich aber auch auf andere Bereiche anwenden. Als Beispiel stellt Abbildung 6.1 die Abhängigkeit der Anzahl der Einwohnerzahl einer Stadt von ihrem Rang dar.

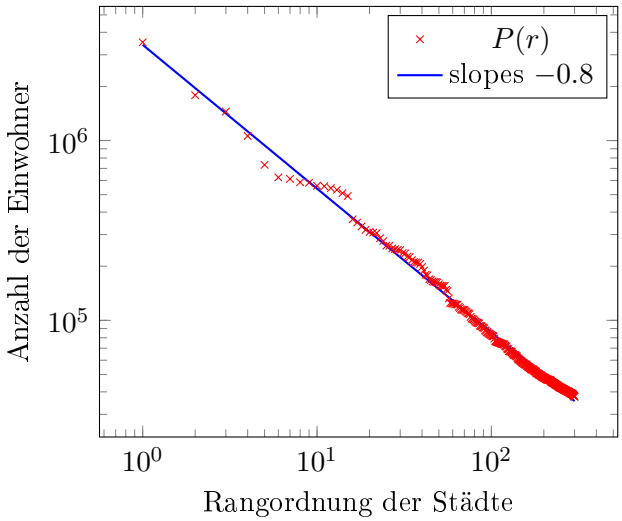

Abbildung 6.1: Zipfsches Gesetz am Beispiel der Größe und Rangfolge der 300 größten Städte in Deutschland [2015] [31]

Luhn [32] nahm 1957 diese Gesetzmäßigkeit auf, um zwei Begrenzungen des Vokabulars am oberen und unteren Rand der Verteilung festzulegen. Während Wörter mit sehr hohen Häufigkeiten als „zu allgemein“ angesehen werden, werden Wörter mit sehr geringer Häufigkeit als „zu selten“ angesehen, um signifikant zum Verständnis des Inhaltes beizutragen [33, S.18 ff]. Abbildung 6.2 zeigt die von Luhn graphisch dargestellte Beziehung.

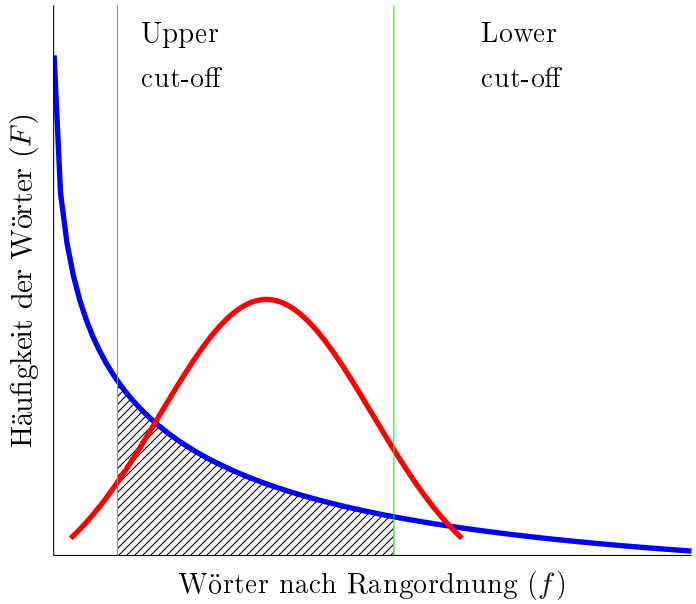

Abbildung 6.2: Schematische Darstellung der Häufigkeit eines Wortes und seines Ranges unter gleichzeitiger Betrachtung ihrer Signifikanz, wie von Luhn [32] beschrieben.

- Absolute Häufigkeit von Wörtern:
  Auf Basis der absoluten Häufigkeit wären erste Ansätze ein einfaches Abschneiden einer festen Anzahl von Wörtern im oberen bzw unteren Häufigkeitsbereich. Im oberen Bereich stellt dies ein einfaches Mittel dar, um Füllwörter zu eliminieren. Aber auch Wörter, die nur einmal in einem Dokument oder gar im gesamten Datenbestand vorkommen, liefern so gut wie keinen Beitrag zum Verständnis des Textes und können durch eine untere Grenze entfernt werden. Bei der Wahl der Grenzen spielt allerdings die absolute Textlänge eine Rolle. Daher ist eine weiterer sinnvoller Schritt, nach dem Abschneiden bei einer festen absoluten oberen Häufigkeit im oberen Bereich, die nächsten x% der Wörter für die Analyse zu verwenden (siehe auch Pareto 80/20 Regel [34]).

- Inverse Dokumentenhäufigkeit (IDF):
  Eine weitere Möglichkeit der Filterung besteht darin, die Wörter nach ihrer Relevanz zu gewichten. Dabei sollen Wörter, die nur in wenigen Dokumenten vorkommen, im Vergleich zu Wörtern, die in allen Dokumenten vorkommen, bei gleicher Gesamthäufigkeit höher gewichtet werden. Spärk Jones [35] schlug 1972 die inverse

Dokumentenhäufigkeit als Quotient aus der Gesamtzahl der Dokumente N und der Anzahl der Dokumente n_i in denen ein bestimmtes Wort vorkommt, als eine geeignete Gewichtung vor. Diese auch „inverse document frequency“ (IDF) genannte Gewichtung wurde von Salton und Yang experimentell untersucht und ihre Wichtigkeit bestätigt [36]. Da auch diese Größe dem Zipf'schen Gesetz folgt, bietet sich eine logarithmische Skalierung für die Gewichtung an:

$$idf_i = \log_2(N/n_i) \tag{6.2}$$

- TF-IDF Gewichtungsschema:
  Das TF-IDF Gewichtungsschema ist eine heute weit verbreitete Methode [37, 38]. Es kombiniert dabei sowohl obige IDF, als auch eine auf Luhn, 1957, [32] zurückgehende Gewichtung zu [39]

$$w_{i,j} = \begin{cases} (1 + \log_2 f_{i,j}) \cdot \log_2(N/n_i) & \text{für } f_{i,j} < 0\,, \\ 0 & \text{für } f_{i,j} \geq 0\,, \end{cases} \tag{6.3}$$

  wobei die „term frequency“ ein Maß für die Häufigkeit des Terms i in Dokument j ist. Gerade für den Bereich niedriger Häufigkeiten wurden mit dem TF-IDF Bewertungsschema gute Ergebnisse erzielt [39].

6.4 Morphologischer Ansatz zur Wortreduzierung

Morphologischer Ansätze zur Wortreduzierung fallen unter das sogenannte Natural Language Processing (NLP) [40, 41], welches jedoch noch weitergehende Analysen z.B. zur Satzanalysen beinhalten.
Schließlich soll z.B. zwischen den Wörtern „schwimmen“„schwimmst“ und „schwamm“ in der weiteren Verarbeitung nicht unterschieden werden. Es geht hierbei also um die Eliminierung der Flexion (Beugung) von Wörtern. Beugungsformen sind abhängig von der Wortart und werden bestimmt durch Genus, Kasus, Person, Numerus, Tempus und Modus [42, S. 60 ff] [43].

6.4.1 Stammformreduzierung (Stemmer)

Reduktion eines Wortes auf eine Stammform (Stemming):
Die einfachste Form der Stammformreduktion ist das Stemming. Hier wird in Abhängigkeit der Sprache und ihren typischen Formen der Wortbeugung vom Ende des Wortes ein Teil abgeschnitten. Dadurch entstehen jedoch auch künstliche Stammformen. Regelbasierte Verfahren sind am sinnvollsten, wenn die Sprache überwiegend aus regelmäßigen Wortbildung verfügt. In diesen Fällen ist die Zahl der zu implementierenden Regeln und Ausnahmen von den Regeln überschaubar. Neben festen Regeln wurden auch Algorithmen erprobt, die

auf statistischen Methoden beruhen. Ein erster Algorithmus zum Stemming wurde bereits 1968 von Lovins entwickelt [44]. In der Literatur finden sich auch Vergleiche der Leistungsfähigkeit verschiedener Algorithmen [45]. Da diese Methoden auch für Topic-Modelling interessant sind, folgt hier ein kurzer Überblick, der sich an [45] orientiert:

- Abschneidende Stemmer:
  Diese Stemmer sind die ersten Versuche einer sinnvollen Stammformreduzierung. Beispiele sind die Stemmer von Lovins [44], Porter [46], Dawson [47] und Paice/Husk [48]. Der Lovins Stemmer arbeitet mit einer 249 Endungen umfassenden Tabelle, unterstützt von 29 Bedingungen und 35 Umformungsregeln. Der von Dawson erweiterte Stemmer verfügt über 1200 Endungen, arbeitet aber auch als einstufiger Algorithmus. 1980 wurde die erste Version des Porter Stemmers vorgestellt. Hier wird davon ausgegangen, dass man die Vielzahl langer Endungen in mehrere kleine zusammengesetzte Endungen aufteilen kann. Diese werden dann nach einem fünfstufigem Schema abgearbeitet. Die Vorgehensweise beim Stemmer von Paice/Husk beruht auf 120 Regeln, welche auf dem letzten Buchstaben des Wortes aufsetzen.
- Statistische Methoden der Stammbildung:
  Während die abscheidenden Stemmer sprachabhängig sind und besonders gut mit Sprachen wie Englisch arbeiten, können auf statistischen Methoden basierende Algorithmen sprachunabhängig eingesetzt werden. Beispiele hierfür sind der N-Gram [49], Hmm- [50] und der Yass Stemmer [51], siehe auch [52] .
- Weitere Ansätze sind Korpus basierende und Kontext sensitive Stemmer [25].

6.4.2 Reduzierung auf die grammatikalische Grundform (Lemmatization) und Anwendung eines Thesauruses

Unter dem englischen Begriff „Lemmatization“ versteht man die Umwandlung eines Wortes in seine grammatikalische Grundform (Lemma). Dabei erfolgt der Abgleich der Wörter mit einem Wörterbuch, indem zu einem Lemma alle gebeugten Formen enthalten sind. Der Umfang eines solchen Wörterbuches hängt von der jeweiligen Sprache ab. Der große Vorteil liegt hier darin, dass keine künstlichen Wortformen entstehen. Im Gegensatz zum Englischen, wo viele Vergangenheitsformen mit dem gleichen Wortstamm nur durch Anhängen von ed"gebildet werden, sind viele Vergangenheitsformen im im Deutschen (und auch Französischen) mit Wortstammänderungen verbunden. Daher ist diese Methode gerade im Deutschen von Vorteil, auch wenn gleichzeitig die Erstellung des Wörterbuches zunächst sehr aufwendig und umfangreich ist. Das ist auch der Grund dafür, dass entsprechene Wörterbücher nicht frei und kostenlos verfügbar sind, während es frei verfügbare Stoplisten in fast allen Sprachen gibt.
Durch eine Ergänzung dieser Wörterbücher um einen Thesaurus kann diese Form der

Wortoptimierung noch weiter verbessert werden. Bei einem Thesaurus handelt es sich um ein Wortnetz, um z.B. bedeutungsgleiche Wörter, unterschiedliche Schreibweisen und Abkürzungen zusammenzufassen. Allerdings sind solche Verfahren sehr ressourcenintensiv. Daher wurden gerade beim Thesaurus Versuche unternommen, gute Algorithmen zu erstellen, siehe auch [52, 24].

Kapitel 7

Datenvorverarbeitung des Link-Content Modells

In diesem Kapitel wird der verwendete Datensatz vorgestellt. Anschließend wird darauf eingegangen, wie dieser vorverarbeitet wurde, um als sinnvolle Eingabe für die zu untersuchenden Algorithmen zu dienen.

7.1 relNet Datensatz

Der relNet Datensatz ist eine Sammlung von verschiedenen Datensätzen. Diese Datensätze sind Kopien von verschiedenen religiösen Online-Foren, einschließlich der Threads und Beiträge des jeweiligen Forums. Diese Datensätze beinhalten allerdings nur die Daten, welche öffentlich zugänglich sind. Es sind keine internen Informationen, auf welche, wenn überhaupt, nur der Forenadministrator Zugriff hat, d.h. es sind beispielsweise keine Informationen über private Nachrichten zwischen Usern, Freundschaften, Kontaktlisten oder die Herkunft der User über die IP-Adressen vorhanden.
Einer der Datensätze stammt aus dem Forum „Ahlu-Sunnah.com“ mit dem Untertitel „Islam nach Quran & Sunnah“, welches von dem Besitzer am 28.07.2014 gesperrt und anschließend gelöscht wurde. Gründe dafür seien gewesen: „Unqualifizierte Meinungsbeiträge, blinder Fanatismus im Anhängen an Gruppen und Personen, seifenopernartige Ergüsse zum eigenen Privatleben, ins Perverse gesteigerter Takfirismus und dergleichen mehr“. [1] Das Forum soll laut dem Administrator auf dem „Konzept eines öffentlichen, anonymen, von der Teilnehmerschaft her unbeschränkten islamischen Internetforums“[1] beruhen.
Der Datensatz besitzt folgende Eckdaten:

1. Größe: ca. 450MB

2. Anzahl der User: 3787

[1] Aus dem Kommentar und Ankündigung des Administrators zu der Schließung des Forums.

3. Anzahl der Threads: 28282
4. Anzahl der Posts in den Threads: 168591 (einschließlich Erstellungsposts)
5. Anzahl aller Wörter insgesamt[2]: 22789200
6. Anzahl der Wörter im Vokabular, also Anzahl unterschiedlicher Wörter[2]: 312470

Der Datensatz besteht aus einer Liste von Posts, welche folgende Informationen enthalten:

1. author: Autor des Posts
2. html: Der HTML Code enthält neben dem reinem Text, auch weitere Informationen wie Smileys, Absätze, Kursiv- oder Fettschrift, farbige Schrift, Zitate.
3. id: ID des Posts
4. media: Smileys, etc.
5. no: Position eines Posts im Thread
6. parent: ID des vorherigen Posts
7. published: Erstellungsdatum (YYYY-MM-DD)
8. quoted_posts: Enthält bei einer Zitierung eines Posts dessen ID
9. text: Enthält den reinen Text, ohne Zusatzinfos (Kursivschrift, etc.) in Kleinbuchstaben
10. thread: Enthält sowohl die Thread-ID, als auch den Titel des Threads.

7.2 Datenvorverarbeitung

Nun wird beschrieben, wie die Daten vorverarbeitet wurden, um als Input für das Link-Content Modell dienen zu können. Weiterhin soll durch das Entfernen von redundanten, irrelevanten und negativ beeinflussende Daten der Speicherbedarf reduziert werden, die Geschwindigkeit verbessert und die Qualität der Daten erhöht werden. Dazu wurde die Vorverarbeitung in eine Reduktionsphase und eine Umwandlungsphase (für den Input) aufgeteilt.

[2]Hierbei bezeichnet ein „Wort“ eine Zeichenkette, bestehend aus Kleinbuchstaben, die durch ein Leerzeichen getrennt sind. Dabei ist zu beachten, dass bei der Zählung alle Buchstaben als Kleinbuchstaben betrachtet wurden und Umlaute durch entsprechende Buchstaben ersetzt wurden. Wörter die z.B. durch Tippfehler entstanden sind, sind noch in dem Vokabular enthalten.

7.2.1 Vorverarbeitung

In einem ersten Schritt wurden vorläufig irrelevante Informationen, wie der HTML Code, das Erstellungsdatum etc. entfernt. Da Strings relativ viel Speicher benötigen, wurde versucht diese möglichst durch Integer zu ersetzen. So wird der Titel des Threads gelöscht, die ID zur Identifizierung allerdings behalten. Da in Online-Foren häufig wenig Wert auf Rechtschreibung und Groß- und Kleinschreibung gelegt wird, kommt es entgegen, dass der Text ausschließlich aus Kleinbuchstaben besteht. Somit kann ein ursprünglich großgeschriebenes Wort und ein kleingeschriebenes Wort als selbes identifiziert werden. Umlaute „ä, ö, ü“ wurden zu „ae, oe, ue“ und „ß“ zu „ss“ geändert, da oftmals User eine andere Schreibweise verwenden. Es wurde eine deutsche Stop-Word Liste mit 596 Wörtern und eine englische mit 570 Wörtern verwendet, um thematisch irrelevante Wörter wie Artikel oder Präpositionen herauszufiltern.[3] Wörter mit einer Länge von weniger als drei Zeichen wurden ebenfalls entfernt. Anschließend wurde anhand aller restlichen vorkommenden Wörter ein Vokabular erstellt. Aus diesem wurden folgend die häufigsten 30 Wörter entfernt, siehe Kapitel 6.3.1. Da auffiel, dass die meisten Wörter im Vokabular keine echten Wörter sind, sondern teilweise zufällige Zeichenfolgen oder Wörter mit Tippfehlern, wurden Wörter, die seltener als drei mal vorkamen und anschließend, je nach Parameterwahl, auch die 80%, 90% oder 95% seltensten Wörter aus dem Vokabular entfernt. Dies geschah in Anlehnung an das Paretoprinzip[53], welches besagt, dass 80% der Ergebnisse mit 20% des Gesamtaufwandes erreicht werden.
Dies klingt nach einem zu hohem Wert, allerdings ist zu beachten, dass mit wachsendem Datensatz die Anzahl nicht existenter Wörter steigt. Daher ist ein absoluter Wert nicht für Datensätze unterschiedlicher Größe geeignet.
Mit dem endgültigem Vokabular wurden aus allen Posts die Wörter entfernt, welche nicht im Vokabular vorkamen. Da sehr kurze Posts meistens ohne relevantem Inhalt sind, wurden alle Posts, die weniger als fünf Wörter besitzen, als irrelevant angesehen und entfernt.
Die Wörter wurden anschließend anhand der entstandenen Vokabulartabelle als Integer codiert und können später zurückcodiert werden. Somit wird der Speicherbedarf reduziert und die Performance erhöht, da sich Integer leichter als Strings vergleichen und sortieren lassen.
Die Posts wurden weiterhin umgeformt, sodass sie nicht mehr die einzelnen Wörter beinhalten, sondern zu jedem vorkommenden Wort nur noch die ID und die Häufigkeit des Wortes in dem Dokument gespeichert wird. Dies hat neben einer Speicherreduktion den Vorteil, dass manche Berechnungen vereinfacht und die Laufzeiten optimiert werden können. So muss beispielsweise bei der Berechnung des Produktes $\prod_{w \in d} \phi_{zw}$ aus Gleichung (5.1) des Kapitels 5 nicht über jedes einzelne Wort iteriert werden, sondern für jedes Wort kann $\prod_{w' \in d} \phi^n_{zw'}$ berechnet werden, wobei w' für ein in ID und Häufigkeit zusammengefasstes

[3] übernommen am 14.05.2017 aus *https://github.com/6/stopwords-json*

Wort steht und n für dessen Häufigkeit. Die so bearbeiteten Posts wurden anschließend zur Weiterverarbeitung gespeichert.

7.2.2 Erstellung des Inputs

Nachdem die Daten optimiert wurden, müssen diese umgewandelt werden, um von dem Link-Content Modell verwendet werden zu können. Das Link-Content Modell erwartet als Eingabe eine Menge von Usern mit Dokumenten und eine Menge von gerichteten Kanten zwischen den Usern. Die Kanten werden „Freundschaften“ genannt. Dabei ist nur gemeint, dass diese User durch ähnliche Interessen miteinander verbunden sind. Sehr wichtig für die Qualität der Ausgabe des Algorithmus ist die Definition von Dokumenten und Freundschaften. Im Folgenden wird die Vorgehensweise erläutert: als Dokument wird hier ein ganzer Thread bezeichnet. Der Ersteller des Threads erhält das Dokument bestehend aus allen zusammgefassten Posts. Dies wurde so gewählt, da normalerweise der Threadersteller ein Thema vorgibt und entscheidet, worum es in dem Thread geht. Die User, die auf den Thread antworten, haben sich an das Thema zu halten. User, die Posts veröffentlichen, die nicht zum Thema passen, werden häufig von Administratoren verwarnt und der Post als „Off-Topic“ bezeichnet. Daher wird davon ausgegangen, dass alle Posts zum Thema des ersten Post gehören und zusammengefasst werden können. Das bedeutend größere Problem ist die Wahl der Freundschaften. Ursprünglich wurde das Link-Content Modell von Natarajan et al. [6] auf Twitter angewendet. Durch die Kenntnis der Follower eines Nutzers ist dort die Definition der für das Modell benötigten gerichteten Freundschaften einfach. Im Fall von Online-Foren, wie in dieser Arbeit, ist dies nicht mehr so einfach, da es viele mögliche Vorgehensweisen zur Freundschaftsdefinition gibt: Man könnte dies anhand von Zitierungen, Namensnennungen, Antworten in Threads eines Nutzers, etc. umsetzen. Da die Möglichkeiten hierzu vielzählig und teilweise sehr komplex sind, kann im Rahmen dieser Arbeit nicht auf alle eingegangen werden, auch wenn im Anschluss an die Arbeit weitere Experimente dazu interessant wären. Hier wird ein User v als Freund von u bezeichnet, wenn u eine bestimmte Anzahl an Antworten/Posts auf einen beliebigen Thread von v liefert. User, die weder Dokumente erstellt haben, noch Freunde in dem Forum besitzen (d.h. keinem User häufig genug geantwortet haben), wurden anschließend entfernt.
Die folgende Tabelle gibt anhand verschiedener Thresholds (Grenzen) an, wie viele gerichtete Freundschaften existieren und wie viele User keine Freunde besitzen. Dabei sei noch einmal erwähnt, dass die Freundschaften gerichtet sind, also nicht auf Gegenseitigkeit beruhen müssen.

Threshold Antworten	# User	# Freundschaften insgesamt	# User ohne Freunde
1	3600	29266	463
2	3289	12894	1751
3	3179	7110	2181
4	3130	4569	2386
5	3098	3220	2507
6	3084	2379	2598

Interessant ist, dass schon bei einem Threshold von nur zwei benötigten Antwort mehr als die Hälfte aller User überhaupt keine Freundschaften aufweisen. Das bedeutet, dass diese User nur in selbst erstellten Threads gepostet haben. Dagegen wies der von Natarajan et al. für Twitter verwendete Datensatz „SMALL“ mit 7646 Usern und 267637 Tweets ganze 335524 Kanten (Following-Beziehungen) auf [6]. Dieser Unterschied könnte sich dadurch erklären, dass bei Online-Foren möglicherweise der Inhalt eine größere Rolle spielt als das knüpfen sozialer Kontakte. Selten meldet man sich bei einem Forum nur an, um sich mit bestimmten Personen in Kontakt zu setzen oder über neueste Beiträge dieser informiert zu werden. Bei Twitter hingegen möchte man häufig das Neueste über prominente Personen oder Freunde herausfinden. Wie sich herausstellen wird, könnte die dadurch resultierende vergleichsweise winzige Kantenzahl ein Problem für den Algorithmus darstellen.

Kapitel 8

Versuchsdurchführung

In diesem Kapitel werden nun die Algorithmen an den Datennätzen des relNet Projektes getestet. Hierbei werden nun verschiedene Parameter der Algorithmen bzw. verschiedene Parameter der Datenvorverarbeitung variiert und verglichen. Im ersten Teil des Kapitelserfolgt eine Fokussierung auf die Themenerkennung und im zweiten Teil wird die Communityerkennung betrachtet. Dankenswerterweise hat der betreuende Lehrstuhl für Künstliche Intelligenz der Technischen Universität Dortmund den Algorithmus für das LDA Modell zur Verfügung gestellt. Da eine Impementierung des Link-Content Modelles nicht zur Verfügung stand war eine eigenständige Implementierung wesentlicher Bestand dieser Arbeit.

8.1 Themenerkennung

In diesem Abschnitt wird das Link-Content Modell zum einen mit sich selber - unter einer anderen Parameterwahl - als auch mit der LDA verglichen. Die Bewertung geschieht anhand von qualitativen und quantitativen Kriterien. Es werden bei der qualitativen Analyse die entstehenden Wortlisten der Themen mit den Wortlisten einer anderen Parameterwahl bzw der LDA verglichen. Bei der quantitativen Analyse werden wird jeweils Perplexität als Qualitätsmaß berechnet. Hierbei werden die Datensätze in einen Trainingsdatensatz (75%) und in einen Testdatensatz (25%) aufgetrennt. Die Perplexität wird im folgenden Unterkapitel beschrieben.

8.1.1 Qualitative Auswertung

Die qualitative Auswertung erfolgt hier dadurch, dass für jedes Thema die Wörter mit den größten Wahrscheinlichkeiten innerhalb des Themas in einer Tabelle aufgelistet werden. Anschließend wird versucht das Thema zu identifizieren, um ihm einen subjektiv passenden Titel zuzuweisen und es ebenfalls in die Tabelle einzutragen. Danach werden erneut die Wörter in der Tabelle betrachtet, und für jedes Wort entschieden, ob es zu dem Thema passt oder fehlerhafter Weise dort aufgeführt ist. Allein die Betrachtung der Anzahl an passenden

Wörtern gibt einen ersten qualitativen Eindruck der Qualität der Auswertung. Zudem wurden die Wörter der Themen in den Tabellen nicht nach absteigender Wahrscheinlichkeit sortiert. Dies würde dazu führen, dass Wörter, die generell häufig auftreten, jedoch nicht aussagekräftig sind (wie die Stoppwörter) in allen Themen weit oben auftreten. Daher wurde durch die folgende Formel die relative Aussagekraft eines Wortes bezogen auf ein Thema berechnet und anschließend absteigend sortiert, sodass in einem Thema die Wörter oben stehen, welche am aussagekräftigsten sind.

$$Rel(\phi_{zw}) = \frac{\phi_{zw}}{\sum_{z'=0}^{T} \phi_{z'w}} \tag{8.1}$$

Diese Umrechnung zur relativen Aussagekraft wird später testweise einmalig weggelassen. Wichtig ist zu erwähnen, dass die Benennung des Themas hier eine gewisse Rolle bei der Frage spielt, ob Wörter als passend bezeichnet werden (siehe Beispiel in 8.1.3.1). Durch grüne Markierungen werden vereinzelt Wörter gekennzeichnet, die subjektiv gesehen zu diesem Thema passen. Hierbei sei angemerkt, dass die Bedeutung arabischer Wörter aufgrund der Menge nur stichprobenartig recherchiert werden konnte. Da durch die Benennung der Topics mit Titeln und die damit verbundene Einfärbung einer gewissen Willkür vorliegt, wurde sie nur exemplarisch bei einzelnen Tabellen durchgeführt, um es bei den anderen auch dem Leser zu ermöglichen, sich unvoreingenommen eine Meinung zu bilden.

8.1.2 Quantitative Auswertung: Perplexitätsberechnung

Da sich die subjektiven Wahrnehmung von Menschen und eine quantitativen Auswertung zur Qualität beim Topic-Modelling erheblich unterscheiden kann, wie von Chang et al., 2009, [54] in einem Experiment veranschaulicht, wird hier auch eine quantitative Auswertung durchgeführt. Die Perplexität gibt an, wie gut die in einem Trainingsdatensatz bestimmten latenten Variablen ein Sample, genannt „Testdatensatz“, vorhersagen können. Im Zusammenhang mit der LDA wird sie mit folgender Formel bestimmt [55]:

$$per(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{M} \log\ p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right) \tag{8.2}$$

Hier werden im Nenner die Wörter jedes Dokumentes aufsummiert und ergeben dementsprechend die Anzahl aller im Testdatensatz vorkommenden Wörter (doppelte Wörter werden doppelt gezählt). Im Zähler werden die logarithmierten Wahrscheinlichkeit für die Entstehung jedes Dokumentes $\mathbf{w}_d$ aus dem Testdatensatz aufsummiert. Bei einer Implementierung ist dabei zu berücksichtigen, dass die Wahrscheinlichkeit für ein Dokument (siehe im Fall vom Link-Content Modell, Kap. 5, Gleichung (5.1) und (5.2)) mit zunehmender Dokumentlänge exponentiell sinkt. Die Auswirkungen erkennt man an folgendem Beispiel: die Wahrscheinlichkeit ein Dokument von User u und Community k mit 1000 Wörtern bei einer Wortwahrscheinlichkeit von $\phi(w_i) = 0.01$ (1%, was schon ziemlich viel ist) für jedes Wort w_i zu erstellen, liegt bei $0.01^{1000} = 10^{-2000}$, siehe Gleichung

(5.2). Der Wertebereich des Datentyps „double“ reicht allerdings nur bis $5.0 \cdot 10^{-324}$. Dementsprechend würde man fälschlicherweise eine Wahrscheinlichkeit von 0 erhalten, bzw sehr große numerische Fehler. Danach müsste nach Gleichung (5.1) noch über k und u summiert werden. Das anschließende Logarithmieren in (8.2) würde die echte Dokumentenwahrscheinlichkeit zwar wieder in verarbeitbare Bereiche bringen ($\log(10^{-2000}) \approx 4605$), allerdings ist die echte Wahrscheinlichkeit durch oben beschriebenes Problem bereits verloren gegangen, sodass sich beim Logarithmieren nur der numerischer Fehler fortpflanzt. Daher müssen die Wahrscheinlichkeiten vorher logarithmiert werden. Für festes u und k, kann man (in vereinfachter Notation) schreiben:

$$\begin{aligned} p_{u,k}(\mathbf{w}_d) &= \prod_{w_i \in \mathbf{w}_d} \phi(w_i) \\ &= \exp\left(\log(\prod_{w_i \in \mathbf{w}_d} \phi(w_i))\right) \\ &= \exp\left(\sum_{w_i \in \mathbf{w}_d} \log(\phi(w_i))\right) \\ &= e^b \end{aligned} \tag{8.3}$$

Sowohl die Wahrscheinlichkeit für ein Wort $\phi(w_i)$ als auch ihr Logarithmus als auch die Summe b über die Logarithmen können numerisch verarbeitet werden. Rechnet man die Exponentialfunktion jedoch aus, steht man erneut vor dem Problem eines nicht verarbeitbaren Wertes. Nach Summation über alle Communities und User erhält man im Zähler jedoch den Ausdruck $\log(\sum_u \sum_k e^b)$, welcher sich durch die Log-Sum-Exp Formel nähern lässt [56].

Im Link-Content Modell wurde die Trennungsgrenze zwischen Trainings- und Testdatensatz durch eine Sortierung der Dokumente bezüglich des Veröffentlichungsdatums gezogen. Die soziale Ebene wird davon nicht beeinflusst, d.h. die Freundschaftslisten in Trainings- und Testdatensatz sind identisch. Bei der LDA hingegen wurden die Wörter aus jedem Dokument auf Test- und Trainingsdatensatz im Verhältnis 1:3 zufällig aufgeteilt.

8.1.3 Auswertung

Im Folgenden wurde der in Kapitel 7 vorgestellte Datensatz mit der ebenfalls in Kapitel 7 erläuterten Datenreduzierung als Vergeichsdatensatz verwendet. Hierauf basierend wurden die Versuche mit den Algorithmen durchgeführt und später einzelne Parameter unter Beibehalten der anderen Parameter variiert. Die Standardparameter, von denen jeweils höchstens einer gleichzeitig variiert wurde, sind im Folgenden aufgelistet. Wird ein Parameter variiert, so sind seine Varianten (in blau) zusätzlich zum Standardwert (in schwarz) angegeben.

- Wortfilter löscht die häufigsten 30 Begriffe und verwendet dann die häufigsten 5%, 10%, 20% der verbleibenden Wörter
- Anzahl der Samples (aufgrund der vielen Versuche und schnell eintretender Konvergenz):
  - Burn-in Samples: 50
  - gespeicherte Samples zur Berechnung der latenten Variablen: 50
  - jeweils zwischen den obigen gespeicherten Samples verworfene Samples: 1
- (keine) relative Aussagekraft
- relative Aussagekraft
- Standardparameter des LDA:
  - Topics: 5, 10
  - α: 0.25
  - β: 0.1
- Standardparameter des Link-Content Modells:
  - Topics: 5, 10
  - Communities: 1, 5, 10
  - $\alpha, \beta, \gamma, \delta$: 0.01 (angelehnt an [6])
  - Anzahl Posts als Freundschaftskriterium: 1, 2, 5 (vgl. Tab.7.2.2)
  - Freundschaften sind (un-)gerichtet

    Bei Natarajan et al., 2013, wurde die Anzahl der Topics auf 10 festgelegt und aus diesen nur die besten mit jeweils den wahrscheinlichsten Wörtern vorgestellt. Hier legen wir die Themenzahl standardmäßig auf 5 fest, treffen allerdings keine Vorauswahl und präsentieren alle Themen und eine größere Menge an zugehörigen Wörtern, um dem Leser die komplette Beurteilung der Qualität der Daten hier zu ermöglichen.

8.1.3.1 Vergleich Link-Content Modell zu LDA

Für die Standardparameter zeigt Tabelle 8.1 die Themeneinordnung. Qualitativ betrachtet macht das Modell seine Sache gut und teilt die Wörter sinnvoll auf. Gleiches kann man auch über die LDA (Tab. 8.2) sagen. Allerdings sieht man schon, dass die Bezeichnung des topics einen gewissen Einfluss bei der qualitativen Betrachtung hat. Beispielsweise könnte man in Tab.8.1 dem Topic 4 auch den Titel „Presse“ oder „Zeitung“ geben. Auch hier

würden sich 6 sehr gut passende Begriffe finden lassen, die sich allerdings von den zu „Auslandspolitik“ markierten Begriffen unterscheiden. Ebenso würden sich beim Titel „Terror“ sogar 10 Begriffe finden lassen. Qualitativ lässt sich nicht sagen, welcher Algorithmus besser geeignet ist. Dagegen zeigt die quantitative Auswertung, dass die Perplexität mit 3926 für das Link-Content Modell und 3133 für die LDA bei der LDA deutlich geringer ist, die LDA also scheinbar besser funktioniert. Dies kann verschiedene Gründe haben: zum einen erlaubt die LDA eine Verteilung über mehrere Themen in einem Dokument, wohingegen beim Link-Content Modell jedes Dokument nur ein Thema haben kann. Dies kann die geringere Perplexität erklären. Zum anderen gehen beim Link-Content Modell auch die gerichteten Freundschaften ein. Im Gegensatz zur ursprünglichen Anwendung auf Twitter sind gerichtete Freundschaften hier im Falle von Online Foren nicht eindeutig definierbar, sodass die Speziallisierung der LDA auf Themen hier im Vorteil sein könnte. Für die erhöhte Perplexität des Link-Content Modells kann es noch weitere Ursachen geben. Bei der Perplexitätsberechnung des Link-Content Modells ist möglicherweise der Datensatz bzw. die Vorverarbeitung nicht geeignet. Falls User nur ein Dokument erstellt haben, erhält dieser User entweder im Test- oder im Trainingsdatensatz kein Dokument. Selbst wenn der User sehr viele Dokumente erstellt hat, kann dieses Problem sehr wohl ebenso auftreten. Andere Probleme ergeben sich bei der Aufteilung in Test- und Trainingsdatensatz im Fall des Link-Content Modells nach dem Thread-Eröffnungsdatums dann, wenn Nutzer sich erst später Anmelden, im Extremfall also in einem der beiden Datensätze also noch gar nicht vorkommen. Dieser Nachteil könnte den ursprünglichen Vorteil der chronologischen Aufteil, nämlich die Unabhängigkeit der Dokumente, überwiegen.

8.1.3.2 Vergleich LDA und Link-Content mit mehr Themen

Auch bei der Verwendung von jeweils 10 Themen, sehen die qualitativen Ergebnisse sowohl für das Link-Content Modell (Tab.8.3), als auch für die LDA (Tab.8.3) gut aus. Es lassen sich in den meisten Themen entsprechende Zusammenhänge wiederfinden, teilweise sind die Themen sogar sehr ähnlich zum Fall mit nur 5 Themen. Beispielsweise lassen sich die topics 3 und 9 in Tabelle 8.3 gut mit topics aus der Analyse mit nur 5 Themen beim Link-Content Modell identifizieren. Quantitativ ist auch bei 10 Themen die Perplexität mit 3893 beim Link-Content Modell erneut größer als bei der LDA mit 2706. Im Vergleich zu den Perplexitäten bei nur 5 Themen ist die Perplexität für beide Algorithmen bei 10 Themen (bei der LDA mehr, beim Link-Content Modell weniger) gesunken.

8.1.3.3 Variation der Communities

In diesem Teil wurde getestet, wie das Link-Content Modell auf eine Änderung der vorgegebenen Anzahl an Communities reagiert. Tabelle 8.5 zeigt die Ergebnisse für nur eine Community und Tabelle 8.6 zeigt die Ergebnisse für 10 Communities. In beiden Fällen scheint das

Unbekannt	Handel	Gebete	Auslandspolitik	Körperpflege
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	verkaufe	rakah	weiterlesen	faerben
unglauben	biete	raka	islamist	zupfen
herrschaft	installieren	iqama	isaf	augenbrauen
erschuf	esselamu	niederwerfungen	tagesschau	cola
ablehnung	rahmatullahu	mitternacht	auslan	henna
goetter	hotmail	witr	kandahar	essig
nusra	abaya	nachmittagsgebet	staatsanwaltschaft	nass
schwaeche	amazon	adhaan	kundus	wunde
jabhat	software	leise	focus	haare
rawafid	downloads	taschahhud	islamistischen	awrah
unglaubens	guenstig	sahw	islamisten	zaehne
fakt	melden	tashahhud	attentaeter	kaese
messias	flyer	layl	pakistans	lab
goettlichen	shop	daemmerung	kaida	verzehr
ungerechtigkeit	versandkosten	taslim	erschossen	waschen
kalif	format	gemeinschaftsgebet	karikaturen	urinieren
polytheismus	schaa	adhkar	politik	substanzen
autoritaet	adressen	zuhr	sueddeutsche	berauscht
faehigkeit	wassalamu	fatiha	vorsitzende	glied

Tabelle 8.1: Link-Content Modell mit Standardparametern. Die Perplexität beträgt 3926.

Terror	Zutaten	Unbekannt	Fasten	Englisch
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
euronews	gelatine	takfir	ramada	push
cet	schwarzkuemmeloel	imamat	djuma	blessings
isis	warahmutuallah	rafidi	dawu	scholars
nato	barakhatu	alsalafway	bukha	prayer
washington	bier	istiwa	hija	yayinlari
zivilisten	essig	multim	fastete	messenger
liveleak	alhamdoulillah	rafidah	khuschu	sayfa
nusra	abitur	rawafid	gefaess	rapidshare
taliban	subhnallah	mutawatir	wusch	narrated
syrische	produkten	ahlulbayt	speichel	mercy
obama	ueberlege	schia	ihra	kitaabun
jabhat	psychologie	muschrikin	saha	pbuh
panzer	schwesternbereich	noebaum	fastens	asked
russland	praktizierenden	hujja	fastenden	things
zawahiri	praktizierende	evangelien	blutung	pleased
demonstranten	hmmm	thora	fastende	nda
israelische	filme	jesu	nachtgebet	meaning
nordkorea	flyer	paulus	suehne	produc
aleppo	soo	saas	gesang	kag

Tabelle 8.2: LDA mit Standardparametern. Die Perplexität beträgt 3133 und ist somit geringer als beim Link-Content Modell mit vergleichbaren Parametern.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
cet	geb	rakah	bubenheim	augenbrauen	groups	warahmutuallah	rapidshare	weiterlesen	schnurrbart
wohlzufrieden	geistig	takbir	biete	ihra	american	barakhatu	downloads	isaf	tore
erschuf	osman	nachmittagsgebet	installieren	tayammum	dog	gelatine	rar	tagesschau	rahmani
eigenschaft	ethik	mitternacht	product	speichel	head	hochzeit	ram	auslan	nursi
goetter	augenlicht	adhaan	versandkosten	nass	countries	yassir	adnani	staatsanwaltschaft	llahi
isis	blinder	iqama	waschen	children	restaurant	files	islamisten	islamisten	muslimbruderschaft
nusra	dateianhang	rakat	kitapyurdu	unreinheit	states	verlobung	ami	islamistischen	bismillahi
himmeln	verfuegbar	leise	amana	glied	brothers	opa	hidden	politik	hud
schwaeche	verschont	sabw	abaya	zaehne	year	cola	megaupload	kaida	beantwortung
existenz	schmerzhafte	ruku	software	ghusl	don	sihr	zip	kundus	benutzer
jabhat	weigert	niederwerfungen	woerterbuch	kuessen	country	beruf	platzieren	attentaeter	entzogen
rawafid	jah	layl	kitaabun	fastende	fight	beruf	bewegungen	soldat	rahmatu
unglaubens	ahaadeeth	daemmerung	baende	bricht	cilt	brautgabe	praesentiert	erschossen	zakaah
offenbarung	neid	taslim	rezitator	zupfen	namaz	freigeschaltet	ata	radikalen	vierzig
moses	angenehm	sonnenaufgang	ideale	nase	brother	helal	behaupteten	pakistans	fuenfzig
messias	maertyrer	pflichtgebete	basari	wunde	including	geschenke	idah	karikaturen	highlight
kalif	hegt	assr	grammatik	verband	tevhid	kaese	language	bundeswehr	beschneidung
herrschaft	kampagne	zuhr	amazon	unterhalb	rand	berufe	sound	anschlag	kreis
anbeten	befallen	nachtgebet	guraba	faste	turn	zins	translation	reuters	evolution
polytheismus	fuehlst	heben	exemplar	urinieren	gehindert	geheiratet	nuzul	vorsitzende	forenregeln
imamat	rahmatullahu	fatiha	qudsi	gewaschen	bni	drogen	verschwoerung	spiegel	highlight
autoritaet	accept	mittagsgebet	sira	wusch	ruling	spenden	esselamu	sueddeutsche	persoenliche
liveleak	aswad	horizont	inkl	mash	years	schweinefleisch	wahb	burka	tawassul
gest	maajah	qunut	rassoul	mund	truth	geschlechter	makki	merkel	veroeffentlichen
evangelium	ergehen	aufsagen	shop	streichen	lead	bank	solidaritaet	festgenommen	reduzieren
muschrik	verstorbene	verbeugung	bestellen	najis	ile	schaeme	ansari	zivilisten	taqlid
abraham	echte	taschahhud	baender	magen	family	miete	feel	dpa	rechtschaffener
alsalafway	hamid	tarawih	yayinlari	fluessigkeit	fire	suessigkeiten	hid	innenminister	silber
glaubten	alajkum	segenswuensche	isbn	stirn	pay	getraenke	german	verfassungsschutz	gemachten
muschrikin	gehorsamkeit	zeitspanne	hochgeladen	waescht	making	wohnung	hussain	szene	furcht
wohlergehen	grundsaetze	abendgebet	schick	hija	government	schlachtet	rules	soldaten	summe
mutawatir	begierden	qiyaam	bestellt	zahn	claim	lebensmittel	vorm	gaddafi	passende
multim	helfe	praying	vokabeln	gebetswaschung	related	produkten	archive	terroristischen	dam
khawarij	fortsetzung	adhkar	produc	religious	djinn	aris	italien	warnungen	
zuteil	gedenkt	maghrib	hotmail	reinigen	support	zahlt	haupt	berliner	idris
assadisten	verstarb	adhan	tauschen	blutung	fear	eed	getoeteten	empfehlung	
luegner	blinden	tashahhud	books	socken	told	geburtstag	tafseer	salafisten	schari
irah	naehert	lail	bestellung	kleidungsstueck	important	eingeladen	audio	luftangriff	regenten
gehorcht	muhamed	ischa	gedruckt	jibri	oldug	scheidung	download	islamistische	relation
auslegung	begehren	untergeht	zaidan	fastentage	party	hergestellt	dialekt	bundesregierung	azm
nordkorea	wohlstand	dhuhr	buecher	wud	permissible	gramm	kopie	kuendigte	thabit
polytheisten	zusatz	rezitiere	play	unbedeckt	body	riba	nachgedacht	tal	ight
ghu	erscheinung	isha	verkaufe	streicht	back	verstorben	befuerworten	sicherheitskraefte	lasset
ablehnt	ahbasch	niederwerfung	lade	urin	called	lab	nbsp	stuetzpunkt	tahrir
radi	inneren	eingetroffen	shopping	ramada	middle	visum	arabic	provinz	bartes
daula	bereut	salaah	paltalk	brechen	military	abitur	ismail	offenbar	asch
saas	muhammads	rezitieren	runterladen	ausgetreten	change	gluecksspiel	file	nachrichtenagentur	wiederlegen
herabgesandt	ueberlieferungskette	sonne	guenstiger	kohl	mother	vertrag	free	faz	abhaengigkeit
goetzendienst	mist	betest	breite	beruehren	tefsir	geschlachtet	twitter	militaer	beitragen
entsandt	weisheit	raka	tejma	austritt	iraq	fett	juz	afghanistans	zeichnen

Tabelle 8.3: Link-Content Modell mit 10 Topics. Die Perplexität beträgt 3893 und ist somit ein wenig kleiner als bei 5 Topics.

Heirat	**Aussenpolitik**	**Unbekannt**	**Unbekannt**	**Englisch**	**Christentum**	**Osmanisch**	**Wissenschaft**	**Handel**	**Gebete**
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
mehrehe	euronews	schayta	bayt	scholars	alsalafway	mujahedin	tahakum	versand	push
brautgabe	cet	wohlzufrieden	alis	blessings	multim	atatuerk	aslu	yayinlari	witr
eheschliessung	isis	istawa	ahlulbayt	asked	noebaum	dawlah	takfirieren	sayfa	taslim
ehemannes	liveleak	taqiyyudin	majlisi	messenger	evangelien	osmanen	ibaada	versandkosten	ghusl
ehevertrag	nato	lauteren	hadisen	laws	jesu	osmanische	hujja	barakhatu	dhuhr
produkten	washington	segensreiche	kafi	mercy	paulus	isil	ikrah	kag	taschahhud
gatten	jabhat	scholastiker	ghadir	clear	matthaeus	laizisten	evolutionstheorie	fiyat	nachzuholen
sklavinnen	russland	allhoerende	adit	things	bibel	kemal	mehrdeutige	tuerkc	ramada
scheidung	israelische	zauberei	genussehe	narrated	testaments	jihads	habashi	warahmutuallah	rakat
polygamie	nordkorea	majestaetische	fadak	permissible	luther	verbuenden	mehrdeutigen	bestellen	sahw
mahr	israelischen	maryams	rafidi	meaning	trinitaet	osmanischen	atheismus	kontodaten	tahajjud
heiratet	irib	herabkommen	baqir	good	mose	laizismus	takfir	dili	rakah
lab	ahrar	kharawij	majusi	matter	kreuzigung	nationalismus	universums	hardcover	tikaaf
ehemann	assadisten	ibli	nasibi	follow	psychologie	armeen	taghut	ytl	gebetswaschung
zutaten	syrischen	lehm	khum	pleased	cms	shaam	theorien	kapak	sichtung
beschneidung	moskau	weltenbewohner	qummi	worship	christus	weltgeschehen	weltbild	cilt	socken
heirat	demonstranten	majestaet	radiyallahu	give	jesus	krieger	urteilt	format	mondsichtung
verlobung	obama	erschuf	ahlel	praise	christentums	mujahidin	takfiri	kitap	fastens
ehegatten	syrische	demgemaess	bihar	evidence	hallo	alyy	takfirs	seminar	zuhr
beschnitten	anschlag	tahawiya	radi	seek	johannes	schahid	evolution	kitapyurdu	ramad
mahram	auslan	gepriesene	fatimah	evil	testament	verbuendet	yussuf	teymiyye	fastentage
ehemaenner	aussenminister	auferweckt	bakrs	punishment	christentum	reiches	materie	shaydzmi	nachgeholt
ringe	cia	segensreichen	othman	glorious	adila	ayman	lossagt	shop	abendgebet
heiraten	assad	dschami	taqiyah	regard	silvester	schlachtfeld	erlaesst	hotmail	fastenden
homosexuelle	gaza	zumar	schia	correct	juenger	munafiqin	wissenschaftlich	ebay	raka

Tabelle 8.4: LDA mit 10 Topics. Die Perplexität ist mit 2706 erneut deutlich geringer als beim Link-Content Modell, aber auch kleiner als bei der LDA mit 5 Topics.

Link-Content Modell weiterhin eine vernünftige Einteilung zu liefern. Allerdings fällt auf, dass die Topic 2 bis 5 sehr ähnlich in beiden Fällen sind, aber die Begriffe isis und himmeln, welche bei nur einer Community noch am 3.- und 5.-häufigsten vorkommen, im Fall von 10 Communities deutlich weiter hinten landen. Einmalig werden hier auch 2 Parameter verwendet, da in Tab.8.7 sowohl Communities, also auch topics auf 10 gestellt wurden. Auch hier sind manche im direkten Vergleich zwischen 10 topics und 10 Communities und 10 topics bei nur 5 Communities manche topics (z.B. 3 und 9) nahezu identisch, während sich andere sehr unterscheiden. Es lässt sich keine qualitive Aussage treffen, welche Anzahl an Communities vernünftiger ist. Quantitativ bringt sowohl eine Verringerung der Communities auf 1, als auch eine Erhöhung auf 5 mit Perplexitäten von 3915 (für k=1) und 3892 (für k=5) eine minimal niedrigere Perplexität, die allerdings immer noch deutlich über dem Wert bei der LDA liegt. Werden gleichzeitig zur Erhöhung der Communities auf 10 auch die Themen auf 10 erhöht, so ergibt sich erneut eine kaum verringerte Perplexität für das Link-Content Modell mit 3880.

8.1.3.4 Keine themenbezogene relativen Häufigkeiten

In diesem Abschnitt wird betrachtet, wie sich die Zuteilungen ändern, wenn man die oben beschriebene themenbezogene Normierung nicht durchführt. Tabelle 8.8 zeigt die entsprechenden Ergebnisse für das Link-Content Modell und Tabelle 8.9 für die LDA. Ohne die nachträgliche Normierung kommen mehrere Wörter in mehreren topics sehr häufig vor, wie man in beiden Tabellen 8.8 und 8.9 erkennt. Der Effekt ist beim Link-Content Modell allerdings für dieses Sample sichtbarer. Quantitativ ergeben sich keine Änderungen durch die Nachnormierung, da diese nach der Berechnung der für die Perplexität wichtigen Wahrscheinlichkeiten durchgeführt wird, sodass die Änderungen der Perplexität nur statistisches Rauschen sind.

8.1.3.5 Variation des Wortfilters

Hier wird untersucht, welchen Einfluss die Einstellung des vorgeschalteten Wortfilters hat. Tabellen 8.10 und 8.11 zeigen die Ergebnisse, wenn anstatt 10 % der häufigsten Wörter (nach Stopword-Filter und Verwerfen der 30 nächst-häufigsten Wörter) nur 5% bzw 20% der häufigsten Wörter verwendet werden für das Link-Content Modell.
Zunächst lässt sich feststellen, dass die jeweils häufigsten Wörter in allen 3 betrachten Fällen in allen Topics jeweils unterschiedlich sind. Das ist insofern nachvollziehbar, da das zur Verfügung stehende Vokabular unterschiedlich groß ist und manche Wörter aus Tabelle 8.11 für 20% in den anderen Fällen gar nicht erst vorkommen. Außerdem scheint es schwieriger zu sein, einen gemeinsamen Titel für die Themen in 8.11 für 20% zu finden (außer für Topic 1 mit dem Titel Politik) als für 5% oder 10%, welche sich in ihren Titeln recht ähnlich sind. Der Wortfilter scheint also einen recht großen Einfluss auf das Ergebnis

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	verkaufe	rakah	weiterlesen	faerben
unglauben	biete	witr	isaf	zupfen
isis	installieren	raka	tagesschau	augenbrauen
herrschaft	esselamu	iqama	islamist	cola
himmeln	amazon	niederwerfungen	auslan	henna
ablehnung	hotmail	mitternacht	kandahar	essig
goetter	software	nachmittagsgebet	staatsanwaltschaft	wunde
gehorcht	downloads	gesichtet	kundus	zaehne
nusra	rahmatullahu	sonnenaufgang	focus	getraenke
schwaeche	melden	adhaan	islamistischen	substanzen
jabhat	flyer	leise	islamisten	nass
rawafid	shop	isha	attentaeter	kaese
unglaubens	versandkosten	dhuhr	pakistans	berauscht
fakt	format	layl	kaida	waschen
messias	schaa	sahw	erschossen	lab
goettlichen	adressen	aufzustehen	karikaturen	gelatine
ungerechtigkeit	versand	morgendaemmerung	politik	haare
kalif	anschreiben	gefastet	sueddeutsche	glied
verborgene	guenstiger	ruku	qaida	alcohol
polytheismus	hochgeladen	daemmerung	radikalen	schnurrbart
autoritaet	bubenheim	asr	reuters	schaedlich
faehigkeit	biografie	maghrib	vorsitzende	tayammum
liessen	kitapyurdu	fajr	terroristischen	rasieren
verborgenen	flash	zuhr	islamistische	najis
evangelium	tejma	taslim	luftangriff	wein
muschrik	hardcover	gemeinschaftsgebet	merkel	zahn
alsalafway	amana	adhkar	bundeswehr	unreinheit
glaubten	melde	rakat	anklage	produkten
muschrikin	moderatorin	fatiha	taliban	geschmack
wohlergehen	basari	taschahhud	anschlag	seide
attribute	runterladen	verkuerzen	gaddafi	verband
mutawatir	kontaktieren	adhan	soldat	rasiert
verleugnet	grafiken	dhur	festgenommen	gewaschen
multim	adresse	niederwerfung	getoeteten	brot
schwoeren	seminar	fardh	kuendigte	verzehr
zuteil	guenstig	aufsagen	panorama	unrein
assadisten	miniaturansichten	freiwilliges	afghanischen	unterhalb
irah	angehaengter	jumu	demokratismus	zufuegt
deutung	megaupload	verbeugung	zivilisten	reinigt
nordkorea	anmelden	tarawih	bundesregierung	haaren
polytheisten	nikah	lail	festnahme	tieres
ghu	verschenken	assr	nato	riechen
homosexualitaet	abaya	untergeht	geb	mengen
daula	rezitatoren	pflichtgebet	tal	urin
saas	product	mittagsgebet	innenminister	ihra
goetzendienst	anbieten	nachtgebet	luftwaffe	speichel
entsandt	seyyid	heben	afp	schwein
ahlulbayt	didi	qunut	government	schmeckt
froemmigkeit	zip	arafat	terror	substanz
ida	baende	horizont	dpa	schneidet

Tabelle 8.5: Link-Content Modell mit nur einer Community. Die Perplexität beträgt 3915.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	installieren	rakah	weiterlesen	faerben
herrschaft	biete	niederwerfungen	tagesschau	zupfen
unglauben	downloads	raka	isaf	augenbrauen
erschuf	software	mitternacht	islamist	cola
nusra	format	nachmittagsgebet	auslan	lab
jabhat	versandkosten	iqama	staatsanwaltschaft	kaese
rawafid	amazon	witr	kandahar	alcohol
unglaubens	verkaufe	adhaan	islamistischen	essig
messias	hotmail	leise	kundus	gelatine
goettlichen	zip	takbir	focus	najis
kalif	flyer	rakat	islamisten	getraenke
polytheismus	abaya	sahw	attentaeter	substanzen
evangelium	guenstig	daemmerung	pakistans	verzehr
muschrik	flash	taslim	erschossen	henna
alsalafway	shop	isha	kaida	berauscht
muschrikin	bubenheim	layl	vorsitzende	haare
wohlergehen	runterladen	sonnenaufgang	politik	rasieren
attribute	versand	gemeinschaftsgebet	karikaturen	produkten
mutawatir	kitapyurdu	taschahhud	sueddeutsche	zaehne
multim	amana	zuhr	radikalen	rasiert
zuteil	hardcover	mittagsgebet	qaida	schweinefleisch
assadisten	biografie	ruku	soldat	schnurrbart
irah	hochgeladen	assr	reuters	schaedlich
faehigkeit	rahmatullahu	fatiha	gaddafi	unrein
nordkorea	rezitatoren	tashahhud	terroristischen	speisen
ghu	downloaden	adhkar	luftangriff	alkohol
daula	megaupload	aufsagen	islamistische	kuessen
saas	adresse	maghrib	innenminister	schwein
goetzendienst	schaa	fardh	merkel	seide
entsandt	basari	dhuhr	anklage	geschlachtet
ahlulbayt	product	horizont	verfassungsschutz	unbedeckt
froemmigkeit	melden	segenswuensche	taliban	wein
ida	tejma	qunut	anschlag	ihra
nachkommenschaft	didi	heben	italien	speichel
goetter	seyyid	tahajjud	kuendigte	kopfes
gesetzgeber	baende	nachtgebet	nato	schmeckt
khomeini	moderatorin	praying	demokratismus	brot
neuerungen	esselamu	rezitiere	festgenommen	tier
evangelien	kalamullah	rezitieren	bundeswehr	nass
islamiyya	woerterbuch	pflichtgebet	bundesregierung	hergestellt
polytheisten	grafiken	qiyaam	afghanischen	wunde
beigesellt	adressen	dhur	zivilisten	saa
tazila	miniaturansichten	untergeht	festnahme	geschmack
qadar	angehaengter	zeitspanne	panorama	schlachtung
brachten	kostenlos	asr	dpa	getrunken
schwaeche	kitap	adhan	tal	tieres
paulus	tevhid	hamd	radikale	salz
asha	baender	verbeugung	luftwaffe	kopfschmerzen
menschengemachten	windows	falaq	afp	fleisch
boeses	bestellt	tarawih	government	konsum

Tabelle 8.6: Link-Content Modell mit 10 Communities und 5 Themen. Die Perplexität beträgt 3893.

Unbekannt	**Englisch**	**Gebete**	**Handel**	**Hygiene**	**Unbekannt**	**Essen**	**Internet**	**Terror**	**Unbekannt**
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
cet	groups	rakah	installieren	augenbrauen	geb	gelatine	megaupload	weiterlesen	schnurrbart
unglauben	killed	raka	biete	zupfen	rahmatullahu	restaurant	platzieren	isaf	nursi
eigenschaft	dog	iqama	product	gesichtet	alajkum	opa	rar	tagesschau	tore
erschuf	year	adhaan	kitapyurdu	speichel	tevhid	verlobung	translation	auslan	trauern
goetter	countries	nachmittagsgebet	amana	ihra	cilt	getraenke	downloads	staatsanwaltschaft	muslimbruderschaft
nusra	accept	mitternacht	versandkosten	tayammum	ndan	beruf	zip	islamistischen	qudsi
schwaeche	father	leise	basari	waschen	namaz	geschlachtet	ram	islamisten	zeichnen
isis	head	ruku	baender	glied	bni	medikamente	adnani	kaida	ighlight
jabhat	states	sahw	runterladen	zaehne	osman	geschieden	ami	politik	tartusi
unglaubens	children	niederwerfungen	abaya	kuessen	ile	freigeschaltet	archive	kundus	beantwortung
messias	called	mittagsgebet	software	ghusl	gehindert	brautgabe	herrschern	attentaeter	highlight
offenbarung	fight	daemmerung	woerterbuch	fastende	oldug	geschenke	praesentiert	soldat	benutzer
goettlichen	country	niederwerfung	baende	arafat	risale	kaese	bewegungen	radikalen	entzogen
existenz	including	layl	ideale	faste	eature	warahmutuallah	behaupteten	pakistans	evolution
kalif	written	taslim	guraba	wunde	related	helal	rapidshare	vorsitzende	rahmani

Tabelle 8.7: Link-Content Modell mit 10 Topics und 10 Communities. Die Perplexität beträgt 3880.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
segen	aleykum	beten	html	fasten
frauen	salam	salam	politik	erlaubt
muhammad	alaikum	aleykum	afghanistan	aleykum
waere	geschwister	salamu	soldaten	wudu
buch	salamu	allahu	taliban	salamu
erlaubt	allahu	alaikum	spiegel	wasser
religion	buch	gebete	verfuegbar	alaikum
glauben	selam	moschee	dateianhang	salam
leben	liebe	threads	aleykum	ghusl
qur	rahmatullahi	ahlu	getoetet	essen
person	ahlu	betet	selam	alkohol
gesandten	fragen	fajr	welt	fatwa
salam	barakatuh	geschwister	weiterlesen	antwort
kufr	aleikum	raka	ausland	haram
weg	youtube	aleikum	deutschland	ref
geben	inshallah	ref	news	allahu
quran	pdf	fatwa	polizei	verboten
aussage	threads	pflicht	usa	geschwister
seite	buecher	verrichten	quelle	beten
einfach	schwester	lam	islamisten	lam
allahu	seite	erlaubt	leben	threads
meinung	alaykum	laut	angaben	ahlu
ali	assalamu	gebetet	qaida	haare
dafuer	waere	hab	verletzt	waehrend
fragen	moechte	waehrend	deutsche	liebe
antwort	barakallahu	rezitieren	auslan	rahmatullahi
waehrend	hab	selam	nato	waschen
aleykum	arabisch	fatiha	deutschen	urteil
lassen	inshaallah	witr	regierung	selam
sehen	brueder	barakatuh	haetten	bzw
gehoert	leider	liebe	moschee	ramadan
gesandte	video	islamqa	laut	person
vers	quran	rak	land	absicht
frieden	gerne	assalamu	video	tragen
bedeutung	feekum	rahmatullahi	bild	islamqa
genau	helfen	alaykum	kinder	barakatuh
ahlu	kennt	isha	online	alaykum
taten	deutsch	antwort	zivilisten	halal
liebe	forum	barakallahu	pakistan	gehoert
sahih	koennt	nacht	krieg	aleikum
geschwister	koennte	bzw	isaf	fleisch
darueber	geben	person	terror	fragen
thema	akhi	fragen	frauen	hand
moechte	barak	waere	wsalam	einfach
beweis	einfach	einfach	festgenommen	barakallahu
swt	watch	dua	armee	hab
push	bzw	bete	irak	quelle
wahrheit	wsalama	haende	focus	thema
berichtet	rahmatullah	maghrib	afghanischen	waere
sogar	thread	gehoert	angriff	assalamu

Tabelle 8.8: Standard Link-Content Modell ohne nachträgliche Normierung der Wörter auf Themenrelevanz. Die Perplexität ist mit 3926 identisch mit dem Wert ohne nachträgliche Normierung. Das ergibt Sinn, da die Perplexität *vor* der nachträglichen Normierung bestimmt wird.
Man erkennt gut, dass mehrere Wörter wie „salam “ , „allahu “ oder „buch “für mehrere Themen im oberen Bereich vorkommen. Daran sieht man den Nutzen einer nachträglichen Normierung, um die Themenrelevanz zu erhöhen.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
besuche	ermoeglicht	erlaeuterte	gerichtes	admin
beides	hueten	geliebter	adila	fielen
gezogen	handel	countries	aegypter	gebuehrt
betracht	grundsatz	englischer	buergerkrieg	individuum
beruehmte	chapter	angefangen	achtung	bruederbereich
aufenthalt	einkaufen	islamistische	gering	gerechtfertigt
islamist	geklaert	fundament	ausgefuehrt	geehrt
aleykoum	absichtlich	angefuehrt	ghusl	gezogen
ibadat	griff	fanatiker	gewohnheiten	ausgefuehrt
exemplar	ernte	hand	abwendung	hinrichtung
ernte	japan	isst	abtruennige	ego
entzogen	ayaat	bay	heiratete	brueder
beleidigungen	emotional	harith	deg	highlight
fundament	duesseldorf	botschaft	bestraft	abul
articles	bemerken	erlaeutern	eigendlich	anschliessend
brueder	erkenne	heim	ausfuehrungen	betonte
hatim	einschaetzung	erben	adresse	countries
fitna	einfluss	gleichwertig	bay	anh
fliessen	ausgewandert	great	erschwernis	geeignet
bedeutung	einfuegen	ernte	gefordert	angefangen
ausnahmen	gaenzlich	berge	ausgelassen	hijab
halten	geduld	jahrhundert	adha	issue
fuqaha	allsehende	auslaendische	fehlerhaft	achi
herrscher	heutige	aufgeschrieben	fuerchtete	geruechte
halte	harun	bestraft	eig	ght
jahl	gruppierung	geeignet	allerhabene	biete
erwarten	ira	hubschrauber	erwarten	abdallah
bab	england	aufgerufen	ght	articles
berichterstattung	falaq	beides	islamistische	begangen
bewohner	entzogen	ausnahmen	actions	direkte
allverzeihend	dead	bewirkt	ankunft	folgendem
hundes	entfernung	inschaallah	gleiche	abdurrahman
indien	einhaltung	bestaetigt	afghanen	direktor
aufenthaltsort	geschaefte	ghazali	amal	geruch
erneuern	erniedrigen	besatzung	bewirkt	bekommt
bereue	islamist	groessere	anklage	download
khairan	erledigen	festzuhalten	erstens	bat
anschliessend	einigung	haqq	bestimmter	hijjah
entziehen	aktionen	kafirun	heim	geprueft
deutlicher	islamqa	erneuern	ausgestattet	hans
essens	fik	gift	geeignet	anschliessen
festgehalten	deal	aussprechen	brot	absoluter
difference	gattin	ausgestattet	karim	kafirun
bewirkt	ablage	hierueber	ermoeglichen	abtruennig
beruhen	gang	hukm	gewohnt	ausgerechnet
dawu	altes	bund	erfolgen	erhaelt
diene	erledigt	behaupten	ira	fuehrt
fanatiker	herabsandte	council	diin	copy
burka	imaam	bestimmter	erteilen	aas
bezieht	demuetig	empfiehlt	hukm	aufgerufen

Tabelle 8.9: LDA Ergebnisse ohne themenrelevante Nachnormierung der Ergebnisse. Die Perplexität ist hier ebenfalls mit 3130 nur minimal unterschiedlich zu der ohne Nachnormierung, da hier lediglich eine neue Sample verwendet wurde. Hier ist es schwieriger als beim Link-Content Modell, ähnliche Begriffe in verschiedenen topics wie „ausgefuehrt “ (topic 4 und 5) oder „gezogen “ (topic 1 und 5) zu finden. Der gleiche Effekt wie beim Link-Content Modell ist aber erkennbar.

zu haben. Das sieht man auch daran, dass das Wort „verkaufe“ bei 5% nur an Position 6 in Topic 2 vorkommt. Bei 10%, und somit einem größeren Vokabular, fällt es jedoch durch die neuen Wörter nicht weiter zurück, sondern rückt auf Position 1 vor. Die Themenzuordnung könnte sich durch das größere Vokabular also geändert haben, oder aber die Erkennung des Wortes „verkaufe “ im Zusammenhang „Handel“ wird besser erkannt. Analog dazu zeigen die Tabellen 8.12 und 8.13 das 5%-ige bzw 20%-ige Vokabular für die LDA. Das 5%-ige hat mit den Topics 1,2 und 3 sehr ähnliche Topics, wie sie auch in der Standard 10% Variante vorkommen. Bei Topic 4 passen zumindest die deutschen Wörter inhaltlich zueinander, Topic 5 besteht fast ausschließlich aus fremdsprachigen Wörtern. Die objektive Qualität scheint aber auch hier für die 20% Variante nachzulassen. Quantitativ ergibt sich beim mit 20% größten verwendeten Vokabular sowohl für die LDA mit 4450, also auch für das Link-Content Modell mit 5764 eine deutlich erhöhte Perplexität, somit ein schlechteres Ergebnis. Das deckt sich schon mit den qualitativen Beobachtungen und den Schwierigkeiten, den scheinbar zusammengehörigen Wörtern ein gemeinsames Thema zu geben. Eine Reduzierung des Vokabulars (5%) verringert dabei die Perplexitäten sowohl beim Link-Content Modell (2465), als auch bei der LDA (2032) erneut deutlich.

8.1.3.6 Variation der Freundschaftsdefinition

Wie in Kapitel 7.2.2 erläutert, werden Freundschaften zwischen Nutzern über die Anzahl der Posts in Threads des Threaderstellers bestimmt. Tabelle 8.14 zeigt die Ergebnisse des Link-Content Modells, wenn anstatt 2 Antworten nur eine Antwort als Threshold für eine Freundschaft verwendet wird. Dadurch entstehen mehr, allerdings eventuell weniger aussagekräftige Freundschaften. Umgekehrt zeigt Tabelle 8.14 die Ergebnisse, wenn für eine Freundeschaften 5 Antworten benötigt werden. Dagegen zeigt Tabelle 8.16 die Ergebnisse, wenn eine durch 2 Antworten erzeugte Freundschaft nicht mehr gerichtet ist, sondern gleichzeitig auch eine Freundschaft in die entgegengesetzte Richtung erzeugt wird. Alle Freundschaften sind in diesem Fall beidseitig gerichtet. Die Ergebnisse sind für das für das Link-Content Modell qualitativ sehr vergleichbar mit denen für nur gerichtete Freundschaften. Vergleicht man die Tabellen für den Threshold 1 und 5 miteinander, erkennt man, dass 3 Topics sehr ähnlich zueinander sind, wobei sich die 2 anderen unterscheiden. Die Freundschaftsdefinition hat also auch einen gewissen Einfluss auf die Ergebnisse. Quantitativ ergeben sich allerdings bei den zwischen den verschiedenen hier verwendeten Varianten kaum unterscheidbare Perplexitäten. Mit 3906 ist die Perplexität für das Link-Content Modell bei beidseitig gerichteten Freundschaften ein wenig kleiner als bei einseitig gerichteten.

Da für die LDA die Freundschaften sowieso nicht eingehen, also im Prinzip nur eine andere Sample Stichprobe entseht, wurden die Tabellen für verschiedene Thresholds bzw beidseitig gerichtete Freundschaften bei der LDA nicht abgebildet. Dafür wurden seperate

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	hidden	witr	auslan	gelatine
goetter	versand	ghusl	weiterlesen	geschenke
messias	bubenheim	raka	kaida	essig
kalam	dateianhang	gemeinschaftsgebet	islamisten	zinsen
goettlichen	format	taslim	politik	halal
polytheismus	verkaufe	takbir	qaida	hochzeit
imamat	runterladen	rakat	zivilisten	helal
kalif	versandkosten	pflichtgebet	nato	schulden
evangelium	didi	socken	focus	geheiratet
alsalafway	verfuegbar	ruku	bundeswehr	schweinefleisch
muschrikin	rassoul	dhuhr	erschossen	riba
wohlergehen	rapidshare	fajr	york	spiele
multim	archive	maghrib	anschlag	benutzer
mutawatir	download	gefastet	taliban	heiraten
zuteil	eur	sujud	spiegel	spenden
irah	inkl	leise	soldat	scheidung
nordkorea	sira	wudu	dpa	verstorben
daula	bestellen	ungueltig	afghanischen	beschneidung
ghu	grafiken	isha	nachrichtenagentur	musik
saas	miniaturansichten	niederwerfung	festgenommen	spendet
ahlulbayt	angehaengter	augenbrauen	anschlaege	spende
gesetzgeber	audio	gebetswaschung	bush	zimmer
neuerungen	arabic	nachholen	israelische	wohnung
evangelien	files	knoechel	truppen	kennenlernen
islamiyya	rar	heben	cia	onkel
beigesellt	uebersetzungen	asr	terror	wein
paulus	uebertragung	mustahabb	israelischen	bank
menschengemachten	nda	niederwerfen	sicherheitskraefte	haram
asha	laenge	rak	jaehrige	scheiden
noebaum	hochladen	gebetet	osama	vertrag
zeitehe	grammatik	tarawih	geheimdienst	fuehle
universums	datei	jumu	soldaten	abgeben
beigesellen	buecher	bewegungen	streitkraefte	rauchen
yunus	shop	bricht	salafisten	nasiha
jabha	word	waschen	haft	zina
snowden	kostenlos	ischa	offenbar	alkohol
theorien	yayinlari	streichen	afghanische	verheiratet
khalif	isbn	freiwilligen	terroristen	freundin
offensichtliche	auflage	liki	galt	eltern
noah	auszuege	verrichtet	hauptstadt	erfahrungen
veranlagung	books	fastens	palaestinenser	geschlachtet
jabir	verlag	nachtgebet	zufolge	zauberei
yazid	anhang	fatiha	teilte	hijab
tambali	book	qiblah	afghanistan	gewinn
wahhabiten	bestellt	aufstehen	provinz	mahram
ijmaa	programm	rezitieren	kaempfer	warahmutuallah
istawa	gedicht	sonne	bomben	schlimme
erschuf	file	fasten	sprecher	fleisch
isis	erlaeuterung	nase	internationale	konto
fuersprache	preis	walla	polizei	drogen

Tabelle 8.10: Link-Content Modell mit 5% der häufigsten Wörter. Die Perplexität beträgt 2465 und ist die niedrigste für das Link-Content Modell gefundene.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
euronews	height	passwort	abendroete	tahzan
cet	erstellers	telefonnummer	dusche	tadschwid
unglauben	uebermittelten	freischalten	rakaa	baender
obama	param	bernhard	dhuhur	reader
fuersprache	fatrah	duisburg	rakah	neuwertig
nato	gluecksbringer	photoshop	nachfasten	kitapyurdu
rebellen	weapons	muntasir	nachbeten	ebook
amerikaner	rtl	dawaffm	nafila	yayinevi
erschuf	kuenftige	feeq	wasche	kostenloser
himmeln	police	raqi	ghusl	zip
truppen	forenleitung	geschwiester	tasli	hardcover
erdogan	sari	visum	mitternacht	mahmood
goetter	islamist	standesamtlich	rakaat	biete
washington	hidden	bafoeg	witr	stk
hamas	reality	lidl	adhaan	tefsiri
tawassul	hellblauen	alykom	mondsichtung	zzgl
provinz	bisherige	kredit	iqaamah	eingottglaube
nusra	centre	eintragen	sujuud	amana
abba	president	ruqia	guiding	darulfirdaus
schwaeche	amulette	hochzeit	nachmittagsgebet	versand
politischen	beworfen	stream	aah	harakat
jabhat	central	spam	ganzkoerperwaschung	woerterbuch
rawafid	meere	skype	sperma	nlar
unglaubens	rib	rente	basmala	roducts
saba	radikale	wiesenhof	niederwerfungen	tajwid
fakt	visit	anmelden	wuduu	guraba
baum	south	schokolade	taslim	basari
gehorcht	kuehl	zeichnen	leisen	frank
messias	related	bewerben	blutet	biography
goettlichen	tretet	fuehrerschein	rekat	muqaddima
festhalten	entzogen	sparkasse	raka	kesir
anschlag	src	helal	gesichtet	tak
kalif	amulett	rindergelatine	gebetszeit	verkaufe
ungerechtigkeit	warheit	sidr	tashahud	rezitatoren
panzer	beschneidung	yassir	rakat	wehr
verborgene	vergaenglich	termine	sutrah	spohr
israelischen	wall	bank	sonnenaufgang	kat
polytheismus	behinderte	kontonummer	suhur	ryadussalihindownload
autoritaet	registrierung	clips	tih	bucher
israelische	tore	ausprobieren	eiter	darulkitab
imamat	eature	flaschen	rukn	kitaabun
anfuehrer	samira	harram	zehen	miniaturansichten
raketen	wmv	praktikum	mittagsgebet	angehaengter
faehigkeit	dateianhang	lebensmitteln	hamidah	ausverkauft
vertreter	schwoeren	warahmutuallah	abendgebet	thalatha
verborgenen	geistig	versicherungen	dhur	darussalam
evangelium	berechtigten	gmbh	kandil	azr
zog	verhaften	traeumt	ungeraden	cilt
fuerwahr	hayya	fernstudium	pusten	braun
gekaempft	abwehren	nagi	spuelen	baende

Tabelle 8.11: Link-Content Modell mit 20% der häufigsten Wörter. Die Perplexität beträgt 5764 und ist die höchste bei allen Untersuchten Parametern.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
ramada	euronews	takfir	versand	push
djuma	cet	alsalafway	yayinlari	wohlzufrieden
fastenden	taliban	multim	sayfa	gest
gesang	isis	atheisten	shop	tawassul
fastens	usa	menschengemachten	rahmatu	imamat
urin	nato	bibel	permissible	istiwa
augenbrauen	washington	evangelien	rapidshare	dschami
ehevertrag	obama	hujja	versandkosten	tarikh
schmuck	russland	noebaum	nda	thron
aurah	liveleak	paulus	blessings	ahlulbayt
hara	gaza	tambali	barakhatu	awiyah
hija	nusra	murjia	kheir	dhahabi
ehemann	syrische	behauptest	warahmutuallah	hafidh
ehemannes	zivilisten	schariah	scholars	tazila
dhuhr	panzer	takfiris	narrated	istawa
gefastet	israelische	kufrs	pbuh	mustadrak
taslim	afghanischen	scheinst	yay	alis
dawu	anschlag	kufr	things	ueberlieferungskette
musikinstrumente	jabhat	lossagung	worship	kabir
gebetswaschung	ahrar	takfiri	waswaas	hadschar
ghusl	demonstranten	wahhabiten	bestellen	bayt
essig	israelischen	takfi	salamun	mirza
pflichtgebet	irib	taqlid	anschreiben	kafi
nachholen	cia	kuffr	evil	isnad
knoechel	nordkorea	atheist	fiikum	yazid
gemeinschaftsgebet	aussenminister	urteilt	good	tahawi
waschung	hauptstadt	jesu	eur	jahmiyyah
bedecken	moskau	gesetzgebung	evidence	asqalani
fasten	assadisten	mushrik	fiekum	siyar
silber	zawahiri	kafir	nasheed	anh
datteln	assad	irja	prayer	mutawatir
waschen	afghanistan	sachlich	llahi	ahad
mustahabb	russischen	apostasie	kheiran	bari
geschlechtsverkehr	daula	materie	feekum	razi
ischa	aleppo	gesetzen	format	hussayn
raka	raketen	muschrik	barakatuh	hatim
sichtung	rebellen	christentum	insallah	zeitehe
witr	york	muschrikin	mercy	gabriel
bedeckt	kaida	shirk	asked	radhiallahu
honig	sicherheitskraefte	evolution	great	saba
nachtgebet	provinz	aleviten	arabic	tabari
sahi	journalisten	redest	runterladen	ueberlieferer
fastet	syrischen	taghut	inshaa	qadi
speisen	dpa	massstab	end	noblen
wohlergehen	fsa	jahil	meaning	fuersprache
unreinheit	auslan	universums	rar	prophetentum
scheidung	liwa	batil	knowledge	izz
scheiden	snowden	tawaghit	dujana	altvorderen
keuschheit	mujahedin	mushrikin	verkaufe	sufyan
socken	jabha	laien	rahmatullah	hanbal

Tabelle 8.12: LDA mit 5% der häufigsten Wörter. Die Perplexität ist mit 2032 die niedrigste in dieser Arbeit gefundene Perplexität.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
blutungen	arbeitest	deutschem	effekt	blaehungen
angewiesen	andersglaeubigen	ablehnt	empfang	einzunehmen
arzt	deutschland	bestimme	butter	ausgestattet
ausfuehren	bedeutender	aehnliche	daff	dhimma
asa	dschilbaab	beschneidung	blossstellen	basteln
dschami	betrifft	ameer	bihi	beduerftigen
afa	einkaufszentrum	arbeiteten	bni	anhaengerschaft
anmassen	bewirken	baraktuhu	effektiv	devil
chip	bspw	bewegung	betruegen	dient
affairs	benzin	bier	amr	abgehalten
aya	aufgelistet	deutschlands	aidh	ahki
bahnhof	bewertung	bestimmter	begrenzung	angenehme
anliegen	desto	alba	einheitlich	authubillah
anluegen	englischen	category	christlicher	dhul
angenommen	bibeln	deutscher	betreffend	beduerftige
aufpassen	beduerfen	bekam	biete	besuchten
dha	abessinien	chair	empfand	abholen
angeblicher	abwesend	bekaempfung	dienstag	divine
bestreichen	beeilt	berechnet	elende	befuerchtete
aufgetreten	abgelegt	deutschen	einladen	auszuziehen
ausmacht	angegeben	eingeschlichen	allies	akhee
aras	beduerfnis	deutsche	bukhari	beschuss
bildete	bedeckung	deutschsprachigen	abwesenheit	befugt
beten	aneignet	bekaempften	arbeitslosigkeit	bedroht
behaupte	beweise	aehnlich	aspekten	empfangen
aufzwingen	agieren	dschilani	aussenpolitik	deutlicher
einkaufen	ayni	befestigt	arba	beende
akhy	bestaetigten	ausgepraegt	arbeitslos	advise
artillerie	aegypter	bestrafung	aspects	danken
damascus	empfunden	befunden	betreiben	abteilung
angeben	aegyptens	bedenke	birmingham	betaeubung
einigung	abfaellt	bucharyy	beobachtung	abgabe
betaeuben	bsp	achtzig	begriff	aim
ausgedacht	angebeteten	besatzenden	asche	bombardiert
bedeuten	abgrund	ausgenutzt	atheismus	barnabas
blutige	blutung	bestrafen	aib	aym
dajja	bonn	andre	bayyah	endeffekt
einzigartigen	empfehlen	bestraft	beschraenkung	aufgehaengt
akida	began	daf	adolf	blockade
barakallahufik	dahlawi	arbeitsplaetze	ausspucken	bucha
betitelt	bittgebet	abhacken	add	detailliert
buergerkrieg	befolgten	boy	brutal	distrikt
asad	destroy	deutlichen	beabsichtigte	blasen
datum	angreifer	besagten	anschuldigung	beweggruende
einzuladen	aufeinmal	beeintraechtigen	beschaedigt	aufgebaut
dreimal	afrika	dreckige	bnu	bittest
certainty	beobachten	ajr	behaupten	beduinen
centre	daraqutni	dhilal	anhoert	ablenkt
beglichen	bestenfalls	bestreben	early	derartig
daraufhin	beeilen	bekomme	blumen	awla

Tabelle 8.13: LDA mit 20% der häufigsten Wörter. Die Perplexität beträgt 4450 und ist die höchste bei allen LDA Auswertungen.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	gelatine	zip	tagesschau	rakah
herrschaft	hochzeit	kitapyurdu	isaf	witr
erschuf	helal	hardcover	weiterlesen	faerben
goetter	opa	amana	islamist	niederwerfungen
schwaeche	cola	baender	auslan	mitternacht
jabhat	freigeschaltet	rezitatoren	staatsanwaltschaft	gemeinschaftsgebet
rawafid	seminar	bubenheim	kandahar	raka
unglaubens	sihr	kitap	kundus	adhaan
messias	hidden	tevhid	focus	nachmittagsgebet
goettlichen	verlobung	basari	islamistischen	sonnenaufgang
kalif	kunden	ideale	attentaeter	zupfen
polytheismus	brautgabe	guraba	pakistans	ghusl
imamat	daff	megaupload	islamisten	gesichtet
autoritaet	nachnamen	seyyid	kaida	wuduu
evangelium	lab	biete	sueddeutsche	iqama
muschrik	virus	kitaabun	karikaturen	tayammum
abraham	geschieden	woerterbuch	politik	rakat
alsalafway	geschenke	kitabu	vorsitzende	socken
muschrikin	kaese	baende	radikalen	henna
wohlergehen	zins	miniaturansichten	soldat	takbir
attribute	miete	angehaengter	reuters	dhur
mutawatir	ablage	versand	terroristischen	wud
multim	beruf	cilt	erschossen	augenbrauen
zuteil	madkhalis	versandkosten	merkel	mittagsgebet
assadisten	flaggen	verkaufe	luftangriff	isha
irah	suessigkeiten	grafiken	innenminister	taslim
nordkorea	zinsen	anfaenger	anschlag	gebetszeit
polytheisten	alhamdoulillah	didi	qaida	sahw
ghu	gramm	teymiyye	kuendigte	dhuhr
daula	silvester	product	demokratismus	nachgeholt
saas	geschlechter	sira	getoeteten	morgendaemmerung
yazid	forenregeln	rahmatullahu	faz	layl
goetzendienst	medikamente	downloaden	bundesregierung	ruku
entsandt	ringe	shared	bundeswehr	leise
ahlulbayt	berufe	language	afghanischen	daemmerung
froemmigkeit	djinn	zaidan	nato	nachzuholen
verurteilen	spenden	kutub	verfassungsschutz	wudu
ida	bedanken	grammatik	zivilisten	pflichtgebete
nachkommenschaft	produkt	kayyim	dpa	nachholen
gesetzgeber	lebensmittel	oducts	islamistische	waescht
khomeini	visum	rar	taliban	maghrib
neuerungen	weint	abaya	festnahme	verbeugung
asked	bruederbereich	dateianhang	festgenommen	aufsteht
evangelien	glueckspiel	handbuch	luftwaffe	pflichtgebet
islamiyya	produkten	alajkum	afp	segenswuensche
beigesellt	halal	produc	government	rasiert
osmanen	braut	format	tal	verpassten
qadar	beschnitten	rassoul	radikale	verband
brachten	geheiratet	risala	stuetzpunkt	streicht
ijtihad	drogen	yayinlari	george	freiwilliges

Tabelle 8.14: Link-Content Modell mit nur einer benötigten Antwort für eine Freundschaft. Die Perplexität beträgt 3928.

Durchläufe für die LDA durchgeführt, um über die daraus resultierenden Perplexitäten 3131, 3129 und 3119 ein Gefühl für das statistische Rauschen zu erlangen.

8.2 Community Detection

8.2.1 Vergleich zwischen Link-Content und Louvain Algorithmus

Für eine Bewertung der Community Detection auf Basis des Link-Content Modelles wurden die Daten mit Ergebnissen einer Analyse durch einen Louvain Algorithmus [57, 58, 59], also einer reinen Community Detection, verglichen. Hierzu wurden die Knoten und Kan-

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	verkaufe	weiterlesen	hidden	ghusl
herrschaft	biete	tagesschau	dog	rakah
himmeln	abaya	isaf	lar	rakat
erschuf	hotmail	auslan	flash	raka
herzens	amazon	staatsanwaltschaft	megaupload	iqama
goetter	software	kandahar	rapidshare	witr
nusra	shop	islamistischen	rules	sonnenaufgang
jabhat	flyer	focus	dateianhang	niederwerfungen
rawafid	versandkosten	kundus	sound	gemeinschaftsgebet
unglaubens	adressen	islamisten	geb	mitternacht
fakt	exemplar	attentaeter	rar	faerben
messias	basari	pakistans	tevhid	zupfen
goettlichen	schaa	erschossen	archive	dhur
kalif	salamalaikum	kaida	zip	nachmittagsgebet
polytheismus	guenstiger	politik	true	augenbrauen
autoritaet	bestellen	karikaturen	verfuegbar	gebetszeit
faehigkeit	adresse	vorsitzende	files	adhaan
liessen	ebay	soldat	language	gesichtet
evangelium	melde	sueddeutsche	cilt	leise
fuerwahr	kitapyurdu	radikalen	data	taslim
muschrik	amana	gaddafi	ndan	wudu
alsalafway	baender	reuters	deckt	wuduu
glaubten	moderatorin	bundeswehr	download	tayammum
muschrikin	kontaktieren	terroristischen	translation	socken
wohlergehen	anmelden	luftangriff	alajkum	wud
attribute	anbieten	innenminister	rahmatullahu	dhuhr
mutawatir	hochzeit	italien	flv	mittagsgebet
verleugnet	baende	anschlag	osman	layl
multim	warahmutuallah	anklage	shared	nachholen
zuteil	teuer	panorama	rashid	henna
assadisten	kontodaten	spiegel	abdulwahab	waescht
irah	verschenken	festgenommen	part	vergesslichkeit
unglauben	tastatur	kuendigte	downloads	waschen
auslegung	internetseite	nato	party	sahw
nordkorea	bestellt	zivilisten	namaz	essig
polytheisten	paltalk	demokratismus	file	gefastet
ghu	bestellung	islamistische	ram	maghrib
daula	barakhatu	bundesregierung	bni	asr
saas	schicke	merkel	oldug	tarawih
goetzendienst	paket	dpa	kayyim	pflichtgebet
allweise	windows	terror	scholar	ruku
entsandt	weitergeben	tal	american	morgendaemmerung
ahlulbayt	verlag	afp	title	unreinheiten
froemmigkeit	schick	verfassungsschutz	related	nachgeholt
vergehen	cent	luftwaffe	fire	qiyaam
ida	mail	stuetzpunkt	feature	lail
apostasie	erstellen	anschlaegen	tekfir	wudhu
nachkommenschaft	ideale	szene	gharib	aufsteht
gesetzgeber	guraba	presse	children	daemmerung
khomeini	freigeschaltet	festnahme	langes	wunde

Tabelle 8.15: Link-Content Modell mit 5 benötigten Antworten für eine Freundschaft. Die Perplexität beträgt 3928.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cet	faerben	installieren	warahmutuallah	weiterlesen
isis	lab	basari	barakhatu	isaf
herrschaft	henna	software	adhaan	tagesschau
unglaube	ihra	downloads	rakah	auslan
sham	essig	runterladen	yassir	islamist
unglaeubiger	augenbrauen	audios	sonnenaufgang	kandahar
existenz	kaese	downloaden	raka	islamistischen
erschuf	verzehr	tejma	witr	kundus
himmeln	gelatine	amazon	opa	islamisten
herzens	produkten	biete	gebetszeit	focus
goetter	zupfen	format	mitternacht	qaida
rawafid	annulliert	hotmail	leise	attentaeter
schwaeche	kohl	abaya	nachmittagsgebet	pakistans
zwang	verpoent	flash	isha	sueddeutsche
jesu	schaedlich	rezitator	verlobung	kaida
universum	fasting	kalamullah	qiyaam	staatsanwaltschaft
unglaubens	kuessen	miniaturansichten	takbir	erschossen
jabhat	shafii	angehaengter	layl	radikalen
messias	azi	bubenheim	aufsagen	terroristischen
argumentation	wusch	shop	betest	politik
fakt	zaehne	versand	niederwerfungen	anklage
herab	auswirkung	rezitatoren	brautgabe	reuters
nusra	hergestellt	megaupload	mittagsgebet	dog
goettlichen	schnurrbart	kitapyurdu	istikhara	luftangriff
kalif	nass	salafimedia	zakaah	taliban
alsalafway	grundprinzip	amana	abbrechen	islamistische
evangelium	fastende	guenstig	pflichtgebet	festgenommen
ahrar	zahn	verkaufe	hochzeit	karikaturen
multim	lajnah	flv	begehe	bundeswehr
interessen	waescht	grafiken	aufgestanden	anschlag
faehigkeit	unreinheiten	kanal	beruf	radikale
nordkorea	unbedeckt	versandkosten	nikah	soldat
thora	ueblicherweise	rashid	ruku	nato
liessen	imah	lade	assr	szene
voelker	menstruierende	kitap	sujud	tal
mutawatir	ahnaf	erhaeltlich	iqama	zivilisten
erhob	beeintraechtigt	eur	aufzustehen	including
ghu	alcohol	product	rakat	afghanen
assadisten	jibri	anfaenger	aufsteht	bundesregierung
zuteil	cola	hidden	zinsen	vorsitzende
polytheismus	hija	bestellt	gesichtet	afghanischen
glaubten	haare	didi	erneuern	innenminister
abraham	ramada	hochladen	gelehrtenaussagen	stuetzpunkt
mushrikin	einsetzt	tajweed	khutbah	army
jahrhundert	reinigt	rapidshare	pflichtgebete	afp
attribute	mengen	programm	erhoert	spiegel
schwoeren	lik	windows	gefastet	terror
goetzendienst	geschmack	paltalk	tarawih	geschossen
ehren	verbrennt	hochgeladen	gebetsruf	getoeteten
saas	wud	tevhid	berufe	faz

Tabelle 8.16: Link-Content Modell mit ungerichteten Kanten, d.h. eine Freundschaft ist automatisch auch beidseitig. Die Perplexität ist hier mit 3906 etwas kleiner als bei den gerichteten Freundschaften.

ten in das Programm Gephi geladen, um die Ergebnisse zu verbildlichen. Gephi ist ein Programmpaket zur Darstellung von Knotendiagrammen, welches auch direkt die Anwendung des Louvain Algorithus erlaubt. Die Farbe der Knoten repräsentiert die zugeordnete Community. Abbildungen 8.1 (Louvain) und 8.2 (Link-Content) zeigen exemplarisch die gefundenen Unterschiede, welche für alle Versuche repräsentativ sind.
Für die graphische Darstellung wurden Knoten ohne Kante nicht berücksichtigt. Auch wurde bei der Anwendung des Louvain Community Modelling die Kanten als gerichtet betrachtet. Betrachtet man zunächst die Ergebnisse der Louvainanalyse in Bild 8.1, so zeigen sich einige Besonderheiten. Die Analyse endete mit der Berechnung von vier Communities. Die größte Community mit 67,3% (pink) gruppiert sich um die Moderatoren (Knoten 0, 1, 2 und 1556). Die mit den Moderatoren verbundenen Knoten weisen teilweise keine Kanten zu andern Nutzern des Forums auf. In Abhängigkeit von der Anzahl und der Ausrichtung der Kanten zu den verschiedenen Moderatoren treten Häufungen auf. Die mit 16,1% zweitstärkste Gruppe (grün) hat aber bereits mehr Knoten unabhängig von den Moderatoren und zeigt im Bild entsprechend eine geringere Nähe. Auch hier zeigen sich Häufungen, welche aber deutlich geringer ins Auge fallen. Die beiden letzten Communities mit 9,6% (orange) und 7,0% zeigen nur eine leichte Gruppierung. Ursache hierfür ist, das diese sehr stark auch mit andern Communities verbunden sind.

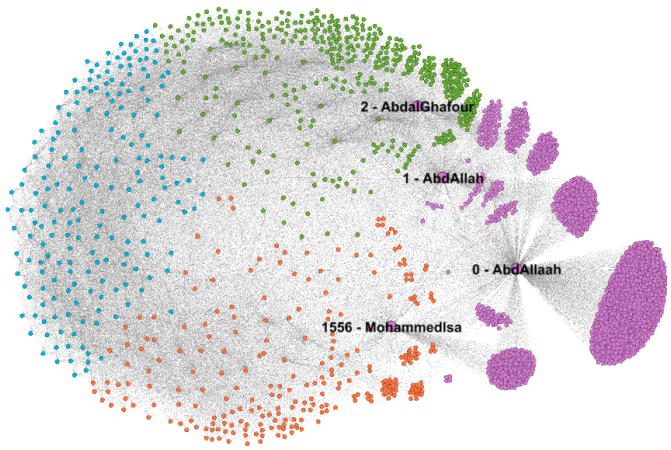

Abbildung 8.1: Knotendarstellung Louvain Algorithmus

Die Ergebnisse des Link-Content Modells in Abbildung 8.2 zeigen nun, dass die Communities gleichmäßig über die Fläche verteilt sind. Die User mit nur einer Kante zu einem Moderator werden gleichverteilt zu den Communities zugeordnet.
Dies ist noch nachvollziebar, da die Moderatoren nicht unbeding einer Community zuzuordnen sind. Die gleiche Erscheinung tritt aber auch bei allen andern Usern auf. Dies ist be-

merkenswert, da man eigentlich, unter der Annahme, dass Personen, die sich für gleichen Themen interessieren, meist zur auf Threats aus der eigenen Community antworten. Aber hier erscheint es so, als ob alle User unabhängig von den Themen auf Posts antworten. Hier stellt sich wieder die Frage, ob das gewählte Kriterium zur Definition einer Kante sinnvoll ist. Man könnte es aber auch so verstehen, dass bei der Anwendung des Link-Content Modelles auf diese Forum beziehungweise auf Foren allgemein die Bindung zu Themen stärker ist als zu Personen. Schließlich werden im Internet auch eher andere Plattformen als Foren (z.B. Facebook) verwendet, um mit anderen Menschen in Kontakt zu bleiben.

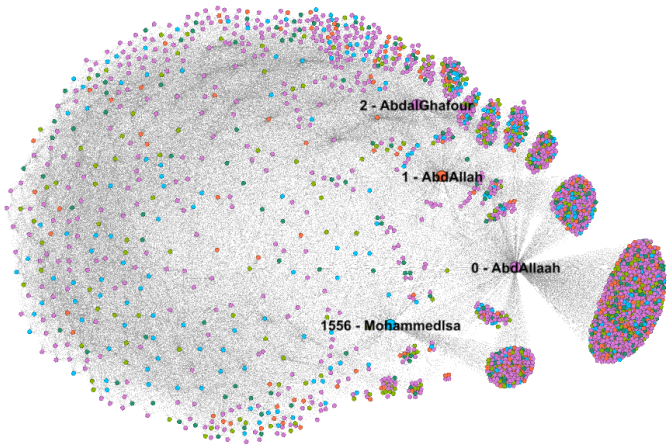

Abbildung 8.2: Knotendarstellung Link-Content Modell

Kapitel 9

Zusammenfassung und Ausblick

In dieser Arbeit wurden die computergestützte Themen- und Communityerkennung anhand eines Dokumentendatensatzes aus Online Foren untersucht. Dazu wurden sowohl die Wörter als auch die Metadaten aus einem zur Verfügung gestellten Datensatz extrahiert, umgewandelt und auf die relevantesten Inhalte gefiltert. Anschließend wurde dieser so aufbereitete Datensatz mit den statistischen Modellen des Link-Content Modells und der Latent Dirichlet Allocation untersucht, automatisch Themen zu erkennen und die dazugehörigen Wörter zuzuordnen. Parallel dazu wurde eine entsprechende Untersuchung unter Einbeziehung des Louvain Algorithmus für die Communitys, also für den sozialen Zusammenhang der Nutzer durchgeführt. Die Ergebnisse wurden sowohl zwischen den verschiedenen Methoden, als auch innerhalb der Methoden durch Variation verschiedener Parameter ebenso qualitativ wie quantitativ verglichen.
Thematisch war qualitativ zwischen der LDA und dem Link-Content Modell kein Unterschied erkennbar, quantitativ schien die LDA aber effektiver zu sein. Mögliche Gründe dafür wurden in Kapitel 8.1.3.1 aufgeführt. Bei der Variation der Variablen zeigte eine Veränderung der Datenvorverarbeitung zur Erzeugung des Vokabulars den größten Einfluss auf die Perplexität.
Bezogen auf die Communitys zeigten sich ebenfalls Unterschiede. Die scheinbar bessere Performance des Louvain Algorithmus kann dadurch erklärt werden, dass das dem Link-Content Modell zugrunde liegende Modell der gerichteten Freundschaften nicht einfach von Followern auf Twitter auf Beziehungen zwischen Usern in Foren übertragen werden kann.
Im Anschluss an diese Arbeit wäre eine weitere Untersuchung in Hinblick auf den vorgeschalteten Wortfilter interessant, da gezeigt wurde, dass die vorgeschaltete Filterung einen großen Einfluss auf die Ergebnisse hat. Diverse andere neben den hier verwendeten Methoden wurden in Kapitel 6 vorgestellt. Insbesondere eine Untersuchung unter Verwendung der TF-IDF Gewichtung wäre interessant. Außerdem hat die Definition des Freundschaftsbegriffes im Link-Content Modell einen gewissen Einfluss, wie in Kapitel 8.1.3.6 gezeigt. Da das Link-Content Modell ursprünglich für Fälle wie Twitter mit gerichteten Freund-

schaften entwickelt wurde, ist eine Übertragung des Begriffes für den Fall von Online Foren nicht eindeutig und trivial. Hier wären Tests von weiteren gänzlich anderen Definitionen neben der hier verwendeten aufschlussreich.

Abbildungsverzeichnis

2.1 Datensatz: Knoten mit relativen Entfernungen 8
2.2 Baumdiagramm des Bottom-up Clusterings 9

4.1 Plate Modell für LDA 17

5.1 Zusammenfassende Legende der Notation, übernommen aus [6]. 20
5.2 Platte Notation des Link-Content Modells, übernommen aus [6]. 21
5.3 Gibbs Sampling für das Link-Content Modell, ergäntzte Abb. aus [6]. Für das Verständnis und dem Vergleich zum Paper wurde in grün ein fehlender Subskript ergänzt und in blau die Counts markiert, deren Superskripts ohne Bedeutung sind. 25

6.1 Zipf1Legende 29
6.2 Schematische Darstellung der Häufigkeit eines Wortes und seines Ranges unter gleichzeitiger Betrachtung ihrer Signifikanz, wie von Luhn [32] beschrieben. 30

8.1 Knotendarstellung Louvain Algorithmus 64
8.2 Knotendarstellung Link-Content Modell 65

Literaturverzeichnis

[1] H. Pross, *Publizistik: Thesen zu einem Grundcolloquium*. Luchterhand, 1970.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*, vol. 21. AAAI press Menlo Park, 1996.

[3] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[4] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets:*. 10 2011.

[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," *Expert systems with applications*, vol. 39, no. 12, pp. 11303–11311, 2012.

[6] N. Natarajan, P. Sen, and V. Chaoji, "Community detection in content-sharing social networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 82–89, ACM, 2013.

[7] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[8] G. Heyer and P. Jähnichen, "Topicmodelle.".

[9] S. Tu, "The dirichlet-multinomial and dirichlet-categorical models for bayesian inference," *Computer Science Division, UC Berkeley*, 2014.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[11] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 524–531, IEEE, 2005.

[12] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 352–359, Morgan Kaufmann Publishers Inc., 2002.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[14] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *NIPS*, vol. 6, pp. 1378–1385, 2006.

[15] T. Griffiths, "Gibbs sampling in the generative model of latent dirichlet allocation," *technical report, Stanfort University*, 2002.

[16] D. Blei, "Topic models." http://videolectures.net/mlss09uk_blei_tm, September 2009. abgerufen 10.05.17.

[17] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[18] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[19] R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[20] C. D. Manning, P. Raghavan, H. Schütze, *et al.*, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

[21] A. Henrich, "Information retrieval 1-grundlagen, modelle und anwendungen, vorlesungsskript," 2008.

[22] S. Gerard and J. M. Michael, *Introduction to modern information retrieval.* McGraw-Hill, Inc., 1983.

[23] K. C. Ó Kane, "Implementing information retrieval," 2017.

[24] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: a comparison of retrieval performances," *Lecture Notes on Software Engineering*, vol. 2, no. 3, p. 262, 2014.

[25] H. Saif, M. Fernández, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," 2014.

[26] C. Van Rijsbergan, "Information retrieval buttersmiths," *London,*, 1979.

[27] C. J. Fox, "Lexical analysis and stoplists.," 1992.

[28] G. K. Zipf, *The Psycho-biology of Language: An Introduction to Dynamic Philology*, vol. 463. Houghton Mifflin Company, Boston, 1935.

[29] D. M. Powers, "Applications and explanations of zipf's law," in *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pp. 151–160, Association for Computational Linguistics, 1998.

[30] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic bulletin & review*, vol. 21, no. 5, p. 1112, 2014.

[31] S. Bundesamt, "Großstädte (mit mindestens 100 000 einwohnerinnen und einwohnern) in deutschland am 31.12.2011 auf grundlage des zensus 2011 und früherer zählungen," 2017.

[32] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.

[33] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, *Web information retrieval*. Springer Science & Business Media, 2013.

[34] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[35] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[36] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[37] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[38] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1089–1098, ACM, 2013.

[39] R. B. Yates and B. R. Neto, "Modern information retrieval: the concepts and technology behind search," *Addison-Wesley Professional*, 2011.

[40] P. Spyns, "Natural language processing," *Methods of information in medicine*, vol. 35, no. 4, pp. 285–301, 1996.

[41] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[42] P. Grebe, *Der Duden in 10 Bänden. 4. Duden - Grammatik der deutschen Gegenwartssprache*. Gotha: Bibliographisches Institut, 3. aufl. ed., 1973.

[43] H. Bussmann and H. Lauffer, "Lexikon der sprachwissenschaft," 2008.

[44] J. B. Lovins, "Development of a stemming algorithm," *Mech. Translat. & Comp. Linguistics*, vol. 11, no. 1-2, pp. 22–31, 1968.

[45] A. G. Jivani *et al.*, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[46] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[47] J. Dawson, "Suffix removal and word conflation," *ALLC bulletin*, vol. 2, no. 3, pp. 33–46, 1974.

[48] C. Paice and G. Husk, "Another stemmer," *ACM SIGIR Forum*, vol. 24, no. 3, pp. 56–61, 1990.

[49] J. Mayfield and P. McNamee, "Single n-gram stemming," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 415–416, ACM, 2003.

[50] M. Melucci and N. Orio, "A novel method for stemmer generation based on hidden markov models," in *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 131–138, ACM, 2003.

[51] P. Majumder, M. Mitra, S. K. Parui, G. Kole, P. Mitra, and K. Datta, "Yass: Yet another suffix stripper," *ACM transactions on information systems (TOIS)*, vol. 25, no. 4, p. 18, 2007.

[52] R. Krovetz, "Viewing morphology as an inference process," *Artificial intelligence*, vol. 118, no. 1-2, pp. 277–294, 2000.

[53] E. Russell-Walling, "Das 80/20-prinzip," in *50 Schlüsselideen Management*, pp. 68–71, Springer, 2011.

[54] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, pp. 288–296, 2009.

[55] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 17–24, ACM, 2007.

[56] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, 2006.

[57] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[58] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Generalized louvain method for community detection in large networks," in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pp. 88–93, IEEE, 2011.

[59] L. Waltman and N. J. van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *arXiv preprint arXiv:1308.6604*, 2013.