UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
THEORETISCHE INFORMATIK

Topic Modellierung zur Zuordnung von Kundenanfragen an Abteilungen

Topic Modeling ...

Bachelorarbeit

verfasst am
Institut für Informationssysteme

im Rahmen des Studiengangs
Informatik
der Universität zu Lübeck

vorgelegt von
Leonard Brenk

ausgegeben und betreut von
Prof. Dr. Ralf Möller

mit Unterstützung von
Dr. Jinghua Groppe, Felix Kuhr, Magnus Bender

Lübeck, den 1. Oktober 2021

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Leonard Brenk

Zusammenfassung

Es ist nicht leicht, eine Abschlussarbeit so zu schreiben, dass sie nicht nur inhaltlich gut ist, sondern es auch eine Freude ist, sie zu lesen. Diese Freude ist aber wichtig: Wenn die Person, die die Arbeit benoten soll, wenig Gefallen am Lesen der Arbeit findet, so wird sie auch wenig Gefallen an einer guten Note finden. Glücklicherweise gibt es einige Kniffe, gut lesbare Arbeiten zu schreiben. Am wichtigsten ist zweifelsohne, dass die Arbeit in gutem Deutsch oder Englisch verfasst wurde mit klarem Satzbau und gutem Sprachrhythmus, dass keine Rechtschreib- oder Grammatikfehlern im Text auftauchen und dass die Argumente der Autorin oder des Autors klar, logisch, verständlich und gut veranschaulicht dargestellt werden. Daneben sind aber auch gut lesbare Schriftbilder und ein angenehmes Layout hilfreich. Die Nutzung dieser LaTeX-Vorlage hilft der Schreiberin oder dem Schreiber dabei zumindest bei Letzterem: Sie umfasst gute, sofort nutzbare Designs und sie kümmert sich um viele typographische Details.

Abstract

It is not easy to write a thesis that does not only advance science, but that is also a pleasure to read. While the scientific contribution of a thesis is undoubtedly of greater importance, the impact of *writing well* should not be underestimated: If the person who grades a thesis finds no pleasure in the reading, that person are also unlikely to find pleasure in giving outstanding grades. A well-written text uses good German or English phrasing with a clear and correct sentence structure and language rhythm, there are no spelling mistakes and the author's arguments are presented in a clear, logical and understandable manner using well-chosen examples and explanations. In addition, a nice-to-read font and a pleasing layout are also helpful. The LaTeX class presented in this document helps with the latter: It contains a number of ready-to-use designs and takes care of many small typographical chores.

Danksagungen

This is the place where you can thank people and institutions, do not try to do this on the title page. The only exception is in case you wrote your thesis while working or staying at a company or abroad. Then you should use the `Weitere␣Unterstützung` key to provide a text (in German) that acknowledges the company or foreign institute. For instance, you could use texts like »Die Arbeit ist im Rahmen einer Tätigkeit bei der Firma Muster GmbH entstanden« or »Die Arbeit ist im Rahmen eines Forschungsaufenthalts beim Institut für Dieses und Jenes an der Universität Entenhausen entstanden«. Do not name and thank individual persons from the company or foreign institute on the title page, do that here.

Inhaltsverzeichnis

1

Einleitung

1.1 Contributions of this Thesis

1.2 Related Work

1.3 Sturktur dieser Arbeit

Der Bereich Data Science und Datenverarbeitung durchläuft aktuell eine starke Welle an Innovation und Veränderung. Deshalb wird in dieser Arbeit zuerst eine Einleitung die Motivation und das Prinzip des Topic Modellings beschreiben. Gefolgt wird die allgemeine Einleitung von der Motivation und Aufgabenstellung in Bezug auf das Anwendungsbeispiel des Zweckverbands Ostholstein (ZVO). Die Anwendung der Algorithmen wird im Laufe der Arbeit mit theoretischen Grundlagen untermauert, die eine klare Notation fordern: diese ist im Teil der Notationen gelistet. Das Prinzip des Topic-Modellings kann mit unterschiedlichen Ansätzen implementiert werden. Beispiele sind: LDA, NNF und LSA. Da diese sich in ihrem Ansatz und Umsetzung unterscheiden, werden im nächsten Abschnitt die bekanntesten dieser Modelle verglichen indem ihre Vorgehens- und Funktionsweise untersucht werden. Die Methode, die in dieser Arbeit genauer beleuchtet wird, ist Latent Dirichlet Allocation. Der dritte Abschnitt wird von der Herkunft, dem Format und der Verarbeitung der Daten der ZVO handeln. Dabei wird der Anwendungsfall und die ZVO genauer erklärt und ein Einblick in die Daten, deren Reinigung und Verarbeitung gegeben. Darauf folgt das Konzept der Lösungsstrategie, um die Aussagekraft und Qualität der Daten für Topic-Modellierung erfolgreich zu untersuchen. Dafür wird der Ablauf des Algorithmus anschaulich dargestellt und die Prozesskette der Analyse im Detail aufgeschlüsselt. Im vierten Abschnitt findet die Implementierung des im dritten Abschnitt erklärten Konzepts statt. Für diese wird eine Methode des Topic-Modellings ausgewählt und in einer vorher definierten und begründeten Programmierumgebung realisiert. Die Umsetzung wird anhand des gelisteten Algorithmus zeilenweise erklärt. Der Output ist in die Arbeit integriert, um dessen zentrale Rolle in der Datenanalyse zu unterstreichen und die Ergebnisse verständlich begründen zu können. Die detaillierte Analyse folgt im nächsten Abschnitt, der sich damit befasst, die Ergebnisse der Implementierung im Anwendungsfall der ZVO zu interpretieren. Dabei wird begründet, wie

gut sich die Daten für eine Topic-Modellierung der ZVO Daten eignen und was die Ergebnisse über die Qualität der Daten aussagt. Die Die Arbeit wird abgeschlossen mit der Zusammenfassung der gesamten Analyse und einem Ausblick für die ZVO in Bezug auf Topic-Modellierung. Die gewonnenen Ergebnisse der Datenqualität sollen der ZVO nach der Ergebnisgewinnung zu einer höheren Effizienz in der Datenverarbeitung verhelfen und Aufschluss über Handlungsbedarf und Optimierungspotential in den Datensätzen geben.

1.4 Motivation

Die digitalisierte Welt generiert täglich riesige Mengen an neuen Informationen. Von E-Books, Blogs über Nachrichten-Websites und Magazinen bis hin zu mobilen Anwendungen auf dem Smartphone, immer mehr Menschen verlassen sich auf und richten ihr Leben nach dem Internet aus. Das Zeitalter des Big-Data ermöglicht es Nutzern unlimitiert viele Daten zu generieren und zu sammeln. Die Kapazitäten, die ein Mensch aufbringen kann, um solche Massen an Daten zu organisieren und zu verstehen, sind schon lange übertroffen. Vorallem durch die steigende digitale Kommunikation und stetig sinkenden Speicherplatzkosten, erhöht sich die Menge an zu speichernden Einsen und Nullen. Laut Statista wurden 2018 33 Zettabyte an Daten generiert mit einer prognostizierten Steigerung bis 2025 um 530% auf 175 Zettabyte. Dieser dramatische Anstieg zeigt die Dringlichkeit für effiziente Algorithmen und Modelle der Datenverarbeitung. Neben der reinen Handhabung solcher Daten steigt aber auch das Bewusstsein, aus diesen Daten Verständnis und Potentiale zu schaffen. Besonders Suchverfahren gewinnen an Bedeutung, wenn in großen, unübersichtlichen Datenmengen bestimmte Informationen gefragt sind. Für die Mehrheit bieten Firmen, wie Google, diese Anwendung an. Zwar kann durch die Keyword-basierten Suche das passenste Dokument gefunden werden, jedoch schlägt die Suchmaschine fehl, wenn nach einer Menge von Dokumente mit einem übergreifenden Thema gefragt ist. Um mehrere Dokumente auf geteilte Themen zu untersuchen, wird das sogenannte Topic-Modelling verwendet. Topic Modeling (dt. Themenmodellierung) beschreibt eine Gruppe von Verfahren, die es ermöglichen, große elektronische Datensammlungen automatisiert zu durchsuchen, organisieren und zu verstehen. Es können Muster innerhalb der Daten entdeckt und Themen extrahiert werden. Dabei stellen Topic Modelle statistische Modelle dar, die Verwendung in der Inferenz abstrakter Topics in unsortierten Datenmengen finden. Topic Modelling gehört zu dem Bereich des Natural Language Processing, also der Verarbeitung natürlciher Sprache. Es verbindet Computerlinguistik, Informatik und Künstliche Intelligenz, um die Potentiale der Sprachverarbeitung mit der heutigen Technik auszuschöpfen. In einer Welt von exponentiell wachsenden Datenmengen finden Methoden des Topic Modeling stetig eine breitere Anwendung. Bereits heute wird Topic Modeling in vielen Bereichen der Wirtschaft, Wissenschaft und Informationstechnologie verwendet. So findet Topic Modeling unter anderem Anwendung bei Zusammenfassungen, Spam Filtern, Internet of Things (IOT), Healthcare, Blockchain, Chatbots, FAQs oder HR. Dies zeigt wie umfangreich das Anwendungsspektrum des Topic Modelings ist. Um semantische Folgerungen aus Datenmengen zu generieren, gibt es verschiedene Ansätze – in dieser Arbeit wird es um das

generative Modell ‚Latent Dirichlet Allocation' gehen. Dabei werden ähnliche Wörter, die in ähnlichen Kontexten vorkommen in Topics gruppiert. Die Grundlagen des LDA liegen bei der Verallgemeinerung des bereits 1999 veröffentlichten 'Probabilistic Latent semantic Analysis (PLSA)'. Im Gegensatz zu anderen Machine Learning Methoden im Bereich der Datenverarbeitung, hat Topic Modellierung die Besonderheit, dass ein Dokument nicht nur zu einem Topic zugeordnet werden kann, wie z.B. bei Clusteralgorithmen. Bei der Topic Modellierung wird jedes Dokument durch eine Verteilung an Topics beschreiben, das bedeutet in jedem Dokument sind immer alle Topics zu finden - nur zu einem bestimmten Anteil. Genauso sind in einem Topic immer alle Worte zu einer bestimmten Wahrscheinlichkeit vorhanden. Bei einem Artikel, der zu 90% über Sport und 10% über Politik handelt, werden somit neun mal mehr Wörter bezüglich Sport zu finden sein, als über Politik. Topic Modeling wird den unüberwachten Lernmethoden des Data Minings zugeordnet, also der Extraktion von Muster und Trends in Datenmengen durch die Anwendung statistischer Algorithmen. Das bedeutet, dass die Topics ohne die Einwirkung von manuell erzeugten Labels gefunden werden. Im Lernprozess werden dann Verteilungen basierend auf den bislang vorgenommenen Zuordnungen iterativ angepasst und verbessert - jedoch alles ohne menschliches Zutun.

1.5 Ziel

Diese Arbeit wird die Theorie des Topic Modeling anhand des Beispiels des Zweckverband Ostholstein (ZVO) in der Praxis implementieren und die bezüglichen Parameter im Sinne der Auswertung bewerten. Der ZVO ist ein Unternehmen, das in Norddeutschland in der Energie-, Entwässerungs-, Internet- und Entsorgungsbranche tätig ist. erhält jährlich ein große Menge an Kundenanfrage, die in eine oder mehrere der 18 Abteilungen zugeordnet werden müssen. Diese werden momentan händisch an die jeweils zuständige Abteilung weitergeleitet, was sowohl zeitintensiv, als auch fehleranfällig zu Ineffizienz in der Wertschöpfungskette führt. Der Prozess soll zukünftig automatisch durch einen Klassifikationsmechanismus funktionieren. Nach der Implementation eines LDA Algorithmus zur Inferenz verschiedener Abteilungen aus den Kundenanfragen, kann die momentan händische Kategorisierung bewertet werden. Diese Arbeit beschäftigt sich mit der Vorhersage der Qualität des Klassifikators, indem die Qualität der manuell erstellten Kategorien und Kundenanfrage-Gruppen untersucht und mit den Ergebnissen verschiedenen Topic Modellierungen verglichen wird. Das Ergebnis eines Topic Models hängt stark von der Qualität der Daten ab, die sie als Input bekommt. Diese Daten durchlaufen eine Reinigungsphase, bevor sie klassifiziert werden, um sie in eine gut zu verarbeitende Form zu bringen. Ziel dieser Arbeit ist es, Erkenntnisse über die Qualität des Klassifikators zu treffen, in Abhängigkeit zu den verwendeten Daten. Somit wird durch die Nutzung von LDA-Modellen die Qualität der Daten untersucht und Prognosen über einer höheren Qualität der Klassfikiation anhand der Daten gemacht.

2

Grundlagen

2.1 Notation

- $\mathcal{K}$ ist die Anzahl der Topics in einem Topic-Modell $\mathcal{M}$
- Ein Model $\mathcal{M}$ repräsentiert den Korpus $\mathcal{D}$
- Eine Menge von Dokumenten ist ein Korpus $\mathcal{D}$
- Ein Dokument d eines Korpus ist eine Menge von $\mathcal{N}$-vielen Worten

2.2 Modellvergleich

Latent Semantic Analysis

Ein anderes verbreitetes Verfahren ist das „Latent Semantic Analysis“ (LSA), welches auf das Finden von sogenannten Hauptkomponenten in Dokumenten abzielt. Dadurch können sowohl ähnliche Wörter gefunden, als auch Textbereiche, die inhaltliche Überschneidungen mit einem bestimmten Begriff haben, aber das Wort selber nicht enthalten, gefunden werden. Die Methode basiert auf dem Prinzip der Singulärwertzerlegung(SVD). Als Ausgangslage wird aus einer Textsammlung eine Term-Dokument-Matrix erstellt. Diese Matrix wird in der SVD als Produkt von drei Matrizen dargestellt, von denen die mittlere eine Diagonalmatrix darstellt. Die Werte auf der Diagonalen lassen daraus die Topics der Textmenge ablesen. Auf das SVD Verfahren selbst hat der Entwickler wenig Einfluss. Um Rauschen zu verhindern, kann jedoch die Anfangsmatrix mithilfe der term-frequency und inverse-document-frequency (tf-idf) verbessert werden, was sich auf das Gesamtergebnis auswirkt. LSA stellt sich als ein attraktives Verfahren heraus, da es Synonyme besser erkennen kann, als LDA und wird heutzutage unter anderem intensiv in dem Bereich des Digital Marketings genutzt.

Probabilistic Latent Semantic Analysis

Im Gegensatz zu LSA, was Singulärwertzerlegung für die Topic-Modellierung nutzt, verfolgt Probabilistic Latent Semantic Analysis (PLSA) einen anderen Ansatz. Die Kernidee ist, ein Wahrscheinlichkeitsmodell zu finden, das ein Dokument erzeugt, dass zu der

Dokument-Term Matrix passt. PLSA hat als erstes Modell dem Konzept des Topic-Modellings Wahrscheinlichkeiten hinzugefügt, sodass sich der Fall, dass Dokument d über Topic z handelt, mit einer Wahrscheinlichkeit von $P(z\|d)$ beschreiben lässt. Dazu berechnet $P(w\|z)$, wie wahrscheinlich es ist, dass das Wort w aus einem gegebenen Dokument d zu Topic z gehört. Somit führt PLSA die Gesamtwahrscheinlich ein, dass ein Wort in einem gegebenen Dokument zu finden ist: $P(D, W) = P(D) \sum_Z P(Z\|D)P(W\|Z)$. Die Gleichung verbindet die Wahrscheinlichkeit ein bestimmtes Dokument zu erhalten mit der eines Wortes in diesem bedingt durch die Topic-Verteilung in diesem Dokument. Zusammenfassend ist PLSA sehr ähnlich zu LSA, auch wenn es einen anderen Ansatz nutzt, mit einer additiven pobabilistischen Komponente.

Latent Dirichlet Allocation

Topic Modeling besteht aus vielen Methoden, die meist verbreitete ist die „Latent Dirichlet Allocation (LDA)“. Dieses Verfahren geht ursprünglich auf PLSA zurück, ergänzt durch zwei Dirichlet-Verteilungen. LDA liegt ein generierender Prozess zugrunde, den zwei Dirichlet Verteilungen maßgeblich beeinflussen: die Dokument-Topic-Verteilung, die die Ausprägungen verschiedener Topics in einem Dokument beschreibt, und die Topic-Word-Verteilung, die die Wahrscheinlichkeit beschreibt, dass ein bestimmtes Wort in einer gewissen Regularität in einem Themenbereich vorkommt. Dabei geht man davon aus, dass ein Dokument eine Verteilung von Topics ist, während ein Topic als eine Verteilung über Wörter betrachtet wird. Die Wahrscheinlichkeit, dass ein bestimmtes Dokument generiert wird, ist das Produkt der Wahrscheinlichkeiten der beiden Verteilungen mit den Wahrscheinlichkeiten zweier multinomialen Verteilungen, die erst zufällig Topics, wie in der Dirichlet-Verteilung definiert, auswählen und aus diesen dann, mithilfe der zweiten Dirichlet-Verteilung, Wörter aus diesen Topics herleiten, wodurch das Enddokument entsteht. Das Enddokument wird höchstwahrscheinlich stark von dem gegebenen Dokument abweichen, jedoch kann durch anpassen der Dirichlet-Verteilungen ein Optimierungsproblem formuliert werden, nach dem die Dirichlet-Verteilungen gesucht werden, die ein möglichst ähnliches Dokument generieren.

Non-Negative Matrix Factorization

Ein weiteres Verfahren, das auch mit Matrizen funktioniert, wird „Non-Negative Matrix Factorization“ (NMF) genannt. Dabei wird eine Matrix, die Wörter auf Dokumente abbildet, in zwei Teilmatrizen faktorisiert. Dabei müssen alle Werte positiv sein. Faktorisierung bezeichnet die Darstellung einer Matrix durch die Multiplikation zweier anderer Matrizen. Dabei ist es bei NNM nicht immer möglich die originale Matrix mit der Multiplikation zu regenerieren, deshalb wird diese bestmöglich approximiert.

Die erste Teilmatrix stellt die Topics in Dokumenten, die zweite die Wörter in Topics dar. Dadurch kann Speicherplatz gespart, und Themen aufgedeckt werden. Das Verfahren beginnt mit zwei möglichen faktorisierten Matrizen und verbessert sich durch die Errorfunktion iterativ, bis das Ergebnis gut genug ist. Dabei werden die errechneten Werte mit der gegebenen Matrix verglichen und angepasst.

2.3 Grundlagen der Latent Dirichlet Allocation

Latent Dirichlet Allocation ist ein grundlegendes und bekanntes Verfahren aus der Natürlichen Sprachverarbeitung. Begründet wird dies unter anderem auf der Komplexität der damals bestehenden Techniken der Textverarbeitung. So waren Clusteralgorithmen zu starr in ihrem Anwendungsumfeld, während Dimensionsreduktionen, wie die Hauptkomponentenanalyse Ergebnisse lieferte, die sehr schwer zu interpretieren waren. Das Prinzip des Topic Modeling basiert auf einer Menge an Dokumenten, die den Korpus darstellen. Dabei werden bei LDA alle Dokumente als Menge von Wörtern angenommen, die als Bag of Words modelliert sind. Dabei hat weder die Reihenfolge, noch die Groß- und Kleinschreibung Einfluss auf das Ergebnis. Die Themen werden allein an der Vorkommenswahrscheinlichkeit der Wörter ohne Reihenfolgen- oder Kontextinformationen erkannt. Durch die Reduktion der Dimension wird die Effizienz gesteigert. Somit wird also jedes Dokument durch eine Verteilung der enthaltenen Wörter repräsentiert.

Bezüglich der Namensgeben, steht Latentfür alles, was wir im Vorhinein nicht kennen. Im Fall LDA handelt es sich um die Themen, die in einem Dokument zu einem bestimmten Teil vertreten sind. "Dirichlet"beschreibt eine Verteilung von Verteilungen. Dies ist vergleichbar mit einem Würfel, bei dem regulierbar ist, wie gleichmäßig die Zahlen gewürfelt werden. Dabei ist der Würfel eine Verteilung und die Aufteilung der Gleichmäßigkeit auch. Beim Topic Modeling bedeutet Dirichlet eine Verteilung von Topics in Dokumenten und eine Verteilung von Wörtern in Topics. Die Ällocation"weist mithilfe der errechneten Dirichlet-Verteilungen Topics Wörter und Dokumenten Topics zu. Eine Besonderheit bei der Themenerkennung mit LDA ist, dass die Anzahl der gesuchten Themen K vorgegeben werden muss. Oft ist diese vorher jedoch nicht bekannt und muss über Hilfsverfahren, wie der Perplexitätsberechnung ermittelt werden. Die Funktionsweise von LDA ist über folgende graphische Abbildung beschrieben:

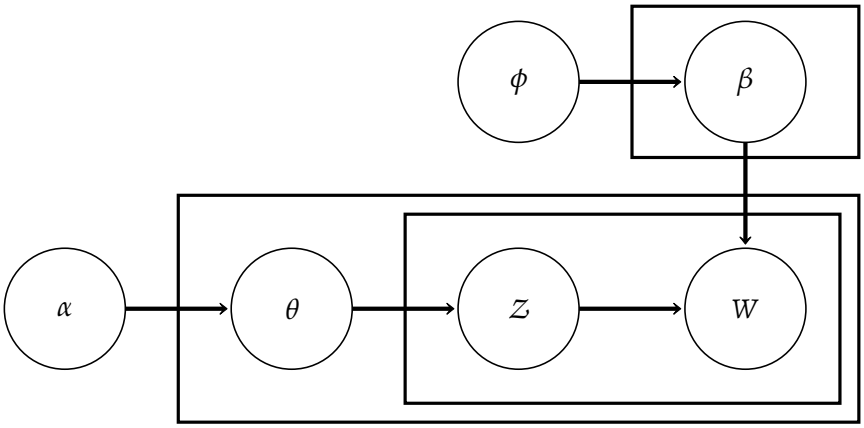


Abbildung 2.1: Graphische Darstellung von LDA

Dabei beschreibt W als einzige nicht verborgene Variable eines von N Wörtern des Dokuments. Das Wort ist semantisch einem Thema Z zugeordnet. Das Thema wieder-

um hängt von der Themen-Verteilung theta des Dokuments ab, das als ein Elment der M vorliegenden Dokumente betrachtet wird. Neben dem Thema, wird jedes Wort auch von der jeweilgen Thema-Wort-Verteilung der K Themen beeinflusst. Das Modell und dessen Verteilungen kann durch die Parameter α und β angepasst werden. α kann bestimmt die Intensität der Dokument-Themen-Verteilung, während β die der Topic-Wort-Verteilung beeinflusst. Bei einem großen α ist die Verteilung der Topics in einem Dokument ähnlicher. Zusätzlich werden bei LDA zwei Bedingungen verfolgt, die von den beiden Parametern beeinflusst werden können. Erstens strebt man für alle Wort eines bestimmten Dokuments so wenig zugeordnete Themen an, wie möglich. Zweitens soll ein Thema über so wenig relevante Wörter wie möglich verfügen. Die beiden Ziele stehen in einer Wechselbeziehung zueinander, da eine minimale Anzahl an vertretenen Topics in einem Dokument zu maximal vielen Wörtern in diesen Topics führt. Die minimale Anzahl an Topics wäre erreicht, wenn man alle Wörter eines Dokuments einem Thema zuweist. Dadurch verfügt das Topic jedoch über alle Wörter des Dokuments. α befindet sich in dem Bereicht $[0, 1]$ mit sinnvollen Werten zwischen $[0.01, 0.1]$, während $\beta = 0.01$ durchschnittlich die besten Ergebnisse liefert. Große Werte führen zu einer Gleichverteilung, die wiederum eine Verschlechterung der Perplexität bedeutet. Somit bietet die Perplexität ein Mittel, um α und β optimal für die individuelle Anwendung zu finden.

Bei LDA werden zwei Verteilungen aus den Dokumenten $d \in \mathcal{D}$ und $k \in \mathcal{K}$gelernt: die Dokument-Topic-Verteilung θ und die Topic-Wort-Verteilung ϕ. Dabei gibt die Dokument-Topic-Verteilung an, mit welcher Wahrscheinlichkeit das Dokument zu jedem Themen gehört. Die Topic-Wort-Verteilung berechnet die Wahrscheinlichkeit, dass ein Wort einem Thema angehört. $\mathcal{M} = LDA(\mathcal{D})$ beschreibt ein LDA Modell, das auf der Dokumentenmenge/Korpus $\mathcal{D}$ trainiert wurde.

Der generative Prozess

Bei der Klassifikatorenerstellung gibt es zwei unterschiedliche Herangehensweisen: den deskriptiven und den generativen Ansatz. Bei der deskriptiven, oder auch beschreibenden, Statistik geht es um die sinnvolle und übersichtliche Darstellung empirischer Daten durch zum Beispiel Tabellen oder Kennzahlen. Betrachtet man zwei Variablen, ein bekanntes $\mathcal{X}$ und eine gesuchte Variable $\mathcal{Y}$, dann wird im deskriptiven Modell die bedingte Wahrscheinlichkeit von $\mathcal{Y}$ bedingt durch $\mathcal{X}$ betrachtet. Im Gegensatz dazu ist bei generativen Modellen die Wahrscheinlichkeit von $\mathcal{X}$ und $\mathcal{Y}$ gemeinsam relevant. Bei generativen Modellen besteht die Möglichkeit, Ausgabeinstanzen zu erstellen. Beispiele dafür sind das "Hidden Markov Modell", Bayes Netze oder das "Gaussian mixure model". Bei LDA handelt es sich um einen generativen Algorithmus. Somit können theoretisch im Fall einer neuen Dokumentzuordnung zu Abteilungen für jede Abteilung zufällig Dokumente generiert werden und diese mit dem neuen Dokument verglichen werden. Der Algorithmus hinter LDA generiert neue Dokumente mithilfe von Dirichletverteilungen $Dir(\gamma)$ und Multinomialverteilungen $Multinom(\delta)$. Dabei sorgen Multinomialverteilungen dafür, dass ein Dokument Teil mehrerer Topics sein kann. Die Verteilungen θ und ϕ werden errechnet, indem iterativ neue Dokumente über andere Verteilung generiert werden, bis das generierte Dokument die Anforderungen befriedigt, dann können die

Verteilungen abgelesen werden, mit denen das Dokument erstellt wurde. Der Prozess verläuft folgendermaßen:

1. Wähle ein θ als $Dir(\alpha)$
2. Wähle ein ϕ als $Dir(\beta)$
3. Für jedes Wort w and Stelle $i = 1, ..., N$ im Dokument d:
   3.1 Wähle ein Thema $z_{d,i}$ als $Multinom(\theta_d)$
   3.2 Wähle ein Wort $w_{d,i}$ als $Multinom(\phi_{zd,i})$

Somit kann der Algorithmus nun neue Dokumente erstellen und das Ergebnis durch die Parameter, wie Alpha und Beta, anpassen, bis das Ergebnis ähnlich genug zu dem Anfangsdokument ist. Dann ist die Verteilung der Themen in diesem Dokument bekannt. Bei der Anwendung von LDA für praktische Problemstellungen, geht LDA das Prinzip rückwärts durch, d.h. für bestehende Gruppen an Dokumenten werden Verteilungen gesucht, durch die das Dokument generiert hätte werden können.

$$P(w, z, \theta, \phi, \alpha, \beta) = \prod P(\theta, \alpha) \cdot \prod P(\alpha, \beta) \cdot \prod P(z,) \cdot \prod P(w \mid \phi) \tag{2.1}$$

Die Formel beschreibt die totale Wahrscheinlichkeit des LDA Modells. Sie setzt sich zusammen aus den Produkten der Dirichlet Verteilung der Topics und der Wörter zusammen mit den multinomialen Verteilungen der Topics und Wörter. Die Schwierigkeit des Algorithmus besteht in der Berechnung der θ-Verteilung der gegebenen Dokumente für latente Variablen. Dies lässt sich durch folgende Wahrscheinlichkeitsverteilung ausdrücken:

$$P(\theta, z \mid w, \alpha, \beta) = \frac{P(\theta, z, w \mid \alpha, \beta)}{P(w \mid \alpha, \beta)} \tag{2.2}$$

Die Formel berechnet die Wahrscheinlichkeit der Verteilung unter einem bestimmten Topic gegeben der α und β Parameter und dem bekannten Wort. Die Wahrscheinlichkeit kann nicht exakt bestimmt werden, weshalb Verfahren wie Gibbs Sampling diese approximieren.

Alpha und Beta

Die Dirichlet Verteilungen werden durch die beiden Parameter α und β bestimmt. Diese formulieren die mathematische Bedeutung der beiden Ziele von LDA:

1. Ein Dokument wird so wenigen Themen wie möglich zugewiesen (α)
2. Jedes Thema hat so wenig relevante Wörter wie möglich (β)

Dabei kann 1) erreicht werden, wenn alle Worte eine Topic wären, was jedoch nicht mit 2) übereinstimmen würde. Für ein erfülltes 2) gibt es nicht die minimale Anzahl an Topics. Die Funktionsweise von verschiedenen *alpha*-Werten zeigen folgende Abbildungen:

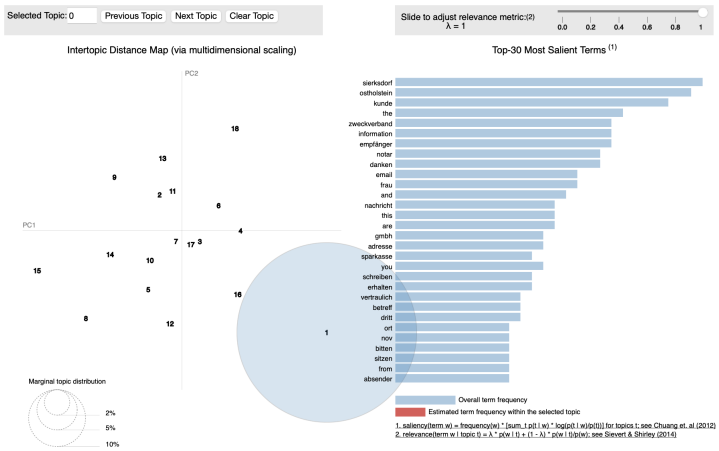


Abbildung 2.2: Ein kleines Alpha

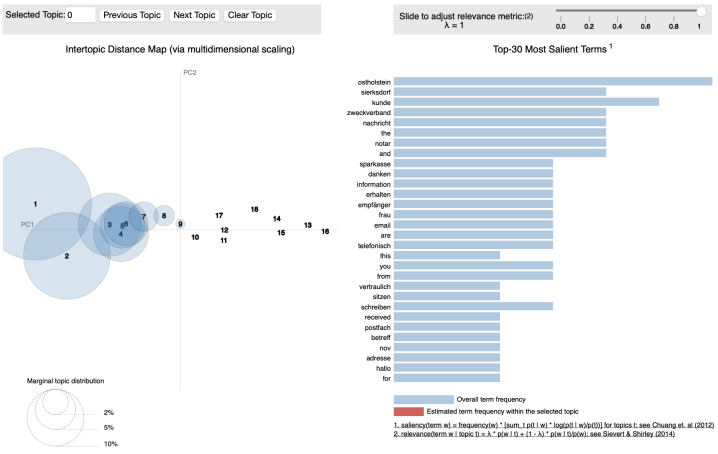


Abbildung 2.3: Ein. großes Alpha

In der ersten Abbildung führt ein kleiner $\alpha = 0.01$ zu einer sehr eindeutigen Themenverteilung. Bei der unteren Abbildung hingegen haben wir ein $\alpha = 1$, was eine gleichmäßigere Verteilung zur Folge hat. Der Trade-off zwischen den beiden Zielen ist der Grund für das funktionieren von LDA. Dadurch wird eine

3

Konzept

3.1 Daten

Die Daten liegen in folgendem Format vor:

	filname	subject-message	Abt0	Abt1	Abt2	Abt3	...	Abt15	Abt16	Abt17
0	FILE0	content0	0	0	0	1	...	0	0	0
1	FILE1	content1	0	0	0	0	...	1	0	0
2	FILE2	content2	1	0	0	0	...	0	0	0
3	FILE3	content3	0	0	0	1	...	0	1	0
4	FILE4	content4	0	1	0	0	...	0	0	1
...	...	...	...	...	...	...	...	...	...	...
133044	FILE133044	content133044	0	0	1	0	...	0	0	0

Abbildung 3.1: Daten TODO...

Relevant für die Auswertung sind die subject-message und die jeweilige Abteilung. Die Tabelle verfügt über eine Matrix mit 18 Abteilungen, von denen pro subject-message eine oder mehrere mit einer 1 versehen ist bzw. sind. Dies beschreibt die Abteilung bzw. Abteilungen, der bzw. denen diese Anfrage manuell zugeordnet wurde. Die Daten in subject-message sind bereits bereinigt, also liegen wie in diesem künstlichen Beispiel vor:

OUTPUT:

```
'wasser verbraucht amt deutschland ablesung zaehlen strom voll ort
    luebeck art straße messung verband nummer platz markieren wechsel
    lieferant stelle verbrauch kunde kunden anrede mann sommer
    beschwerde schrift allgemein kommunikation datenmanagement fern'
```

Um die Einträge in eine computer-lesbare Form zu verwandeln, muss ein Dictionary erstellt werden, dass alle Wörter auf eine Anzahl ihrer Vorkommen abbildet. Dafür müs-

sen die Wörter als alleinige Listeneinträge einlesbar sein:

```
OUTPUT split:
['wasser', 'verbraucht', 'amt', 'deutschland', 'ablesung', 'zaehlen',
    'strom', 'voll', 'ort', 'luebeck', 'art', 'straße', 'messung',
    'verband', 'nummer', 'platz', 'markieren', 'wechsel', 'lieferant',
    'stelle', 'verbrauch', 'kunde', 'kunden', 'anrede', 'mann',
    'sommer', 'beschwerde', 'schrift', 'allgemein', 'kommunikation',
    'datenmanagement', 'fern']
```

Datenreinigung

Bevor eine Themenmodellierung auf Daten durchgeführt werden kann, müssen die Daten einem Prozess unterzogen werden. Dieser beginnt mit der Datenaquise, also der Akquirierung bestimmter relevanter Daten. Im Falle der ZVO bedeutet dies, dass es genügend Kundenanfragen gibt, die verarbeitet werden können. Wenn diese Daten bestehen, werden sie auf die relevanten Wörter reduziert, aus denen eine bedeutsame Inferenz von Informationen möglich ist, sodass unter anderem die sogeneannten „Stop-Words“, also eine Menge von Verbindungswörtern entfernt werden. Ein anderer Schritt der Datenreinigung ist das Transponieren aller Wörter in kleine Buchstaben, um eine Einheitlichkeit zu erlangen, da das Bag of Words Modell keine Reihenfolge mehr beachtet und somit große Satzanfänge irrelevant werden. Wenn die Daten in der gewünschten Form vorliegen, beginnt der Schritt des Featureengineerings. Für einen Computer sind Wörter nicht so leicht zu verarbeiten, wie Zahlen, weshalb in diesem Schritt eine Quantisierung der Wörter und Überführung dieser in eine zahlenbasierte Form vorgenommen wird. Dies kann zum Beispiel in Form eines Bag-of-Words Modells, Dictionary oder TF-IDF, also einer relativen Vorkommensauflistung verschiedner Wörter über Dokumente umgesetzt werden. Nachdem die Daten in eine für den Computer kompatiblen Form gebracht wurden, kann das Themenmodell entwickelt werden.

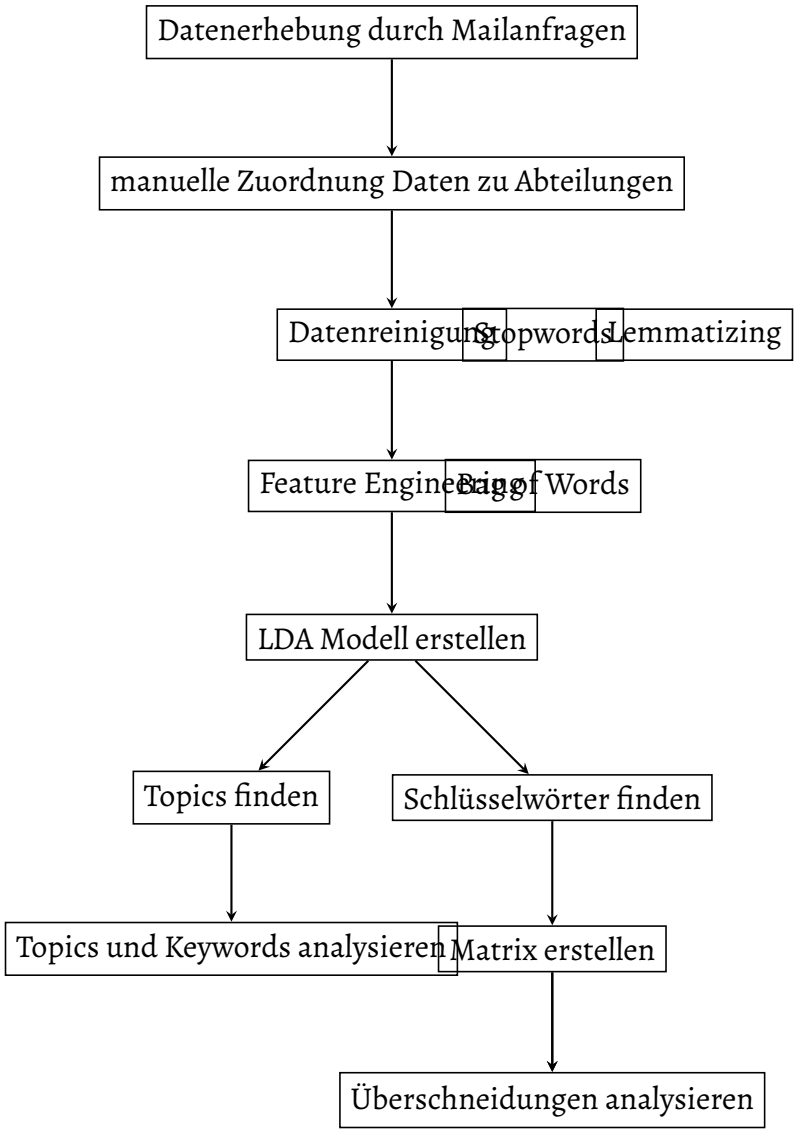
Feature Engineering

Das alleinige Reinigen der Daten reicht nicht aus, um das LDA Topic Modell auf diesen zu generieren. Für eine maschinelle Verarbeitung sind die Wörter in der Form nicht adressierbar. Bei LDA ist ein elementarer Bestandteil, wie oft ein Wort vorkommt. Das bedeutet im Feature Engineering müssen die Daten in ein Format übersetzt werden, das für jedes Wort ohne Duplikate die Anzahl mit einer Wort-ID referenziert. Die Auflistung der Wörter zusammen mit ihrer Vorkommensanzahl und laufenden Indexnummer kann als Input für ein LDA Modell verwendet werden. Dies nennt sich das Wörterbuch (o.a. Dictionary).

TF-IDF und Word Embedding Die Art, wie Wörter repräsentiert werden kann sich unterschieden. Das Dictionary wird bei beiden kreiert. TODO ERKLÄRE TF-IDF UND WORD EMBEDDIN

3.2 Anwendungsfall ZVO

Bei der ZVO sollen jährlich händisch aufgenommene Anfragen in Zukunft maschinell klassifiziert werden. Für die korrekte Klassifikation ist die Qualität der vorliegenden Daten immens wichtig. Die Qualität der ZVO Daten werden in dieser Arbeit untersucht. Dafür wird ein LDA Modell generiert, das als Datengrundlage alle verfügbaren ZVO Daten verwendet. Die Qualitätsuntersuchung der Daten wird durch zwei Methoden durchgeführt. Dies zeigt das folgende Prozessablaufdiagramm:



Nach der Erhebung der Daten können die einzelnen Anfragen manuell in die vorgegebenen Abteilungsgruppen eingeteilt werden. Dort wird die Datenreinigung, wie in 3.2 beschrieben vorgenommen. Sind die Daten bereinigt, kann das Feature Engineering beginnen, wonach die Daten für den Computer verständlich formatiert sind. Die Erstellung des LDA Modells wird vorgenommen, sobald der Korpus aus den feature engineerten Daten erfolgt ist. Die Modellerstellung ist durch folgenden Pseudocode beschrieben:

data ←
for d in Anzahl Dokumente **do**
  $data \leftarrow data + str(d)$
end for
Teile data in einzelne items einer Liste auf
Erstelle ein Dictionary aus der Liste
Wandle ID aus Dictionary in Wörter um
Erstelle den Korpus
Erstelle das Modell
Gibt die Topic-Wort-Verteilungen für alle Topics aus

Ist das Modell generiert und die Topics auslesbar, können die Daten evaluiert werden. Dazu werden zwei Ansätze verfolgt. Zum Einen werden Topics und zugehörigen Schlüsselwörter untersucht, um u.a. Aufschluss über mögliche Verbesserungspotentiale bei der Datenreinigung oder dem Feature Engineering feszustellen. Als zweites werden die vom Modell erfassten Gruppen betrachtet und den gegebenen manuell klassifizierten Abteilungen zugeordnet. Die folgenden Punkte beschreiben den Prozess im Detail:

1. **Gruppen und Wörter finden**
   Ein LDA Modell besteht aus zwei Verteilungen, die die zugrundeliegenden Daten semantisch darstellbar machen: die Dokument-Topic-Verteilung und die Topic-Wort-Verteilung. Als Ausgabe des Modells ist also zu erkennen, welche Topics die Dokumentmenge durchschnittlich hauptsächlich beschreiben und welche Wörter in den Topics jeweils dominant vorkommen. Das Modell kann Topics nicht inhaltlich benennen, sondern nur die Verteilungen darstellen. Somit ist nicht eindeutig, welches Topic welche Abteilung der ZVO darstellt. Dafür betrachten wir die Topic-Wort-Verteilungen und schließen von dieser auf die Qualität der Daten. Ist über die dominanten Wörter in einem Topic zu erkennen, welche Abteilung dieser repräsentiert, scheint das Modell und die Daten gut genug zu sein, um die Daten zu klassifizieren. Sollte die Abteilung nicht an den Wörtern ablesbar sein, sind die Daten nicht optimal für eine Klassifikation geeignet.
2. **Zuordnung Abteilung zu Topic** Das LDA Modell clustert die Daten in 18 Topics. Diese Topics sollten im optimalen Fall sehr ähnlich zu den händisch klassifizierten Abteilungen sein. Ist dies nicht der Fall, kann man auf eine schlechte Klassifikation schließen. Dies kann durch eine schlechte Qualität der Daten als auch durch eine ineffiziente manuelle Einteilung der Topics bedingt sein. Die Fähigkeit, Topics auf Abteilungen zu mappen, gibt Aufschluss über die Qualität der Daten. Für die Zuordnung werden zwei Matrizen verwendet: gruppen_LABEL und gruppen_LDA. Die erste Liste sortiert alle Dokumentindizes als Listenelemente in die jeweilige Zeile der Matrix, sodass der Index eines händisch in Abteilung 3 eingeordneten Dokumentes in gruppen_LABEL[4] zu finden ist. Die Matrix gruppen_LDA beinhaltet alle

Indizes der Dokumente, die vom Modell klassifiziert wurden, in gleicher Struktur. Dafür wird für jedes Dokument, das den Korpus ausmacht, eine dokumentseigene Dokument-Topic-Verteilung errechnet. Das Topic, für das das Dokument am wahrscheinlichsten ist, bestimmt, welcher Teilliste der Dokumentindex angehängt wird. Beide Matrizen verfügen nun über die Indizes der Dokumente in den jeweiligen Topics bzw. Abteilungen und können anhand der einzigartigen Indizes auf Überschneidungen geprüft werden. Die Anzahl der Überschneidungen werden in einer Matrix gespeichert, die jedes Element von gruppen_LDA auf jedes Element von gruppen_LABEL abbildet und deren Überschneidung zählt. Eine optimale Zuordnung von Topic auf Abteilung ist möglich, wenn jede Zeile ein Maximum in einer Spalte hat, die nicht auch das Maximum einer anderen Zeile enthält. Ist dies jedoch nicht der Fall, sind die Daten in der aktuellen Form nicht optimal für die Klassifizierung.

TODO Welcher Algorithmus, wie implementiert?
TODO Abbildung mit flow chart
TODO pseudocode von topic model

4

Implementierung

4.1 Topic Modeling Methode

Zur Untersuchung der Qualität der ZVO-Daten wird in dieser Arbeit die LDA Methode verwendet. Dabei wird nur der Text als bekannt angenommen. Weder die Meta-Daten, noch die Anmerkungen oder Labels sind zu Beginn bekannt. Als grundlegendes Topic Modellierungsverfahren findet es Verwendung in einem breiten Anwendungsspektrum. Durch die Bekanntheit von LDA sind bereits viele Pakete und Bibliotheken in Programmierumgebungen vorzufinden und einfach zu implementieren. Seit LDAs Veröffentlichung in 2000 wurde eine umfassende und detailreiche Dokumentation entwickelt, die neben vielen Forenbeiträgen, die Arbeit mit LDA stark erleichtern. Zusätzlich hat LDA in diesem Anwendungsfall den Vorteil, dass es nicht wie zum Beispiel LSA direkt Dokumentähnlichkeiten ausgibt, sondern das Ergebnis in Form einer Matrix dartellt, die Wörter auf Dokumente abbildet. Damit ergeben sich als Werte der Matrix die Topics, denen die Wörter jeweils angehören. Bei LDA Modellen ist die Anzahl der Topics ein individueller Input, durch den sich das Ergebnis schwerwiegend verändern kann. Die Werte in der Matrix können somit von 0 bis zu der individuellen Anzahl der Topics reichen. Die optimale Anzahl an Topics ist grundsätzlich ein nicht einfaches Problem bei Anwendungen. Im Fall der ZVO werden als Anzahl der Topics 18 gewählt, da dies die Anzahl der bereits erstellten Abteilungen ist.

4.2 Toolauswahl

Bekannte Frameworks zum Topic-Modelling implementieren grundsätzlich ähnliche Algorithmen. Betrachtet wurden in dieser Arbeit Mallet, Gensim und Sci-kit Learn. Wichtige Schritte sind die Vorbereitung der Daten, die Implementierung, die Auswertung und die Visualisierung. Als Bibliothek wird in dieser Arbeit Gensim verwendet, die für die Verarbeitung von unstrukturierten Daten und Anwendung von unüberwachten Algorithmen bekannt ist. Algorithmen, wie word2vec, LSI oder LDA entdecken automatisch Strukturen durch das Prüfen von gemeinsam auftretenden Mustern im Korpus der Trainingsdaten. Viele bekannte Algorithmen sind bereits in diesem Framework implementiert, was die Umsetzung der Topic-Modellierung vereinfacht. Gensim erlangte in der Vergangen-

heit Bekanntheit durch seine hochoptimierten Implementationen bekannter Algorithmen und der Schnelligkeit und Verlässlichkeit, mit der diese ausgeführt wurden. Außerdem wird Gensim in Python verfasst, was sich sehr gut für Probleme im Bereich der Data Science eignet.

4.3 Umsetzung Konzept

1. Alle Dokumente ergeben einen Korpus. Der Korpus generiert eine Topic-Verteilung für die Gesamtheit aller Dokumente. Dabei werden zuerst alle Anfragedaten in einen String zusammengefügt, der als Grundlage für das Wörterbuch und den Korpus dient. Um diesen in ein Dictionary, also eine numerierte Auflistung aller Wörter und dessen Anzahl, zu verwandelt, muss der String in eine Liste mit voneinander getrennten Items gesplittet werden. Hier wird ein Bag of Words Prinzip verfolgt, die Reihenfolge ist irrelevant für das Ergebnis des Modells. Aus der Liste wird dann das Dictionary erstellt. Durch den Aufruf des LDA Modells wird aus dem Bag of Words mithilfe des Dictionary eine vorgegebene Anzahl an Themen aus der Wortenge modelliert, basierend auf häufig zusammen auftretenden Wörtern. Dadurch ergibt sich neben einer Verteilung der Topics in einem Modell die Verteilung der Wörter, die ein Topic besonders beeinflussen. Ist das Modell generiert, können die Topics ausgelesen werden mit diesem Aufruf:pprint(lda. print_topics ())

INPUT:

```
data = ''

for x in range(0,106000):
    data += df.at[x,'subject_message']

list = data.split()

dictionary = corpora.Dictionary([list])
temp = dictionary[0]
id2word = dictionary.id2token

corpus = [dictionary.doc2bow(text) for text in [list]]

lda = LdaModel(corpus, num_topics=18, id2word = id2word)

pprint(lda.print_topics())
```

OUTPUT:

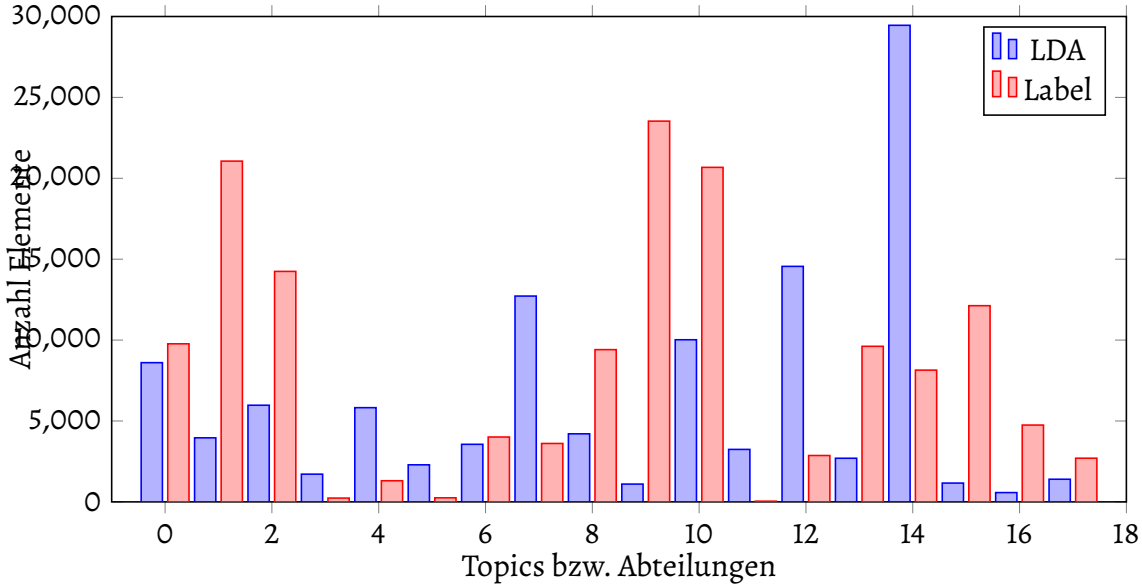
```
[(0,
  '0.012*"ostholstein" + 0.011*"nachricht" + 0.010*"sierksdorf" + '
  '0.009*"zweckverband" + 0.008*"betreff" + 0.008*"danken" +
     0.007*"email" + '
  '0.007*"hra" + 0.007*"datum" + 0.007*"hyperlink"'),
```

```
(1,
 '0.014*"ostholstein" + 0.011*"nachricht" + 0.010*"zweckverband" + '
 '0.009*"sierksdorf" + 0.008*"sitzen" + 0.008*"danken" +
    0.007*"wagrienring" '
 '+ 0.007*"lübeck" + 0.007*"betreff" + 0.006*"the"'),
(2,
 '0.014*"sierksdorf" + 0.013*"zweckverband" + 0.011*"nachricht" + '
 '0.010*"ostholstein" + 0.008*"danken" + 0.008*"betreff" +
    0.007*"sitzen" + '
 '0.006*"the" + 0.006*"homepage" + 0.006*"frau"'),
(3,
 '0.014*"sierksdorf" + 0.011*"ostholstein" + 0.011*"nachricht" + '
 '0.010*"zweckverband" + 0.010*"betreff" + 0.008*"the" + 0.007*"frau" + '
 '0.006*"danken" + 0.006*"denken" + 0.006*"lübeck"'),
(4,
 '0.012*"zweckverband" + 0.011*"frau" + 0.011*"sierksdorf" + '
 '0.011*"ostholstein" + 0.009*"nachricht" + 0.008*"betreff" +
    0.007*"the" + '
 '0.007*"danken" + 0.006*"lübeck" + 0.006*"öffentlich"'),
(5,
 '0.015*"zweckverband" + 0.011*"sierksdorf" + 0.011*"ostholstein" + '
 '0.009*"the" + 0.008*"betreff" + 0.007*"lübeck" + 0.007*"danken" + '
 '0.007*"nachricht" + 0.007*"hyperlink" + 0.006*"sitzen"'),
(6,
 '0.013*"nachricht" + 0.012*"ostholstein" + 0.012*"zweckverband" + '
 '0.010*"betreff" + 0.009*"sierksdorf" + 0.007*"danken" + 0.006*"the" + '
 '0.006*"lübeck" + 0.006*"frau" + 0.005*"hyperlink"'),
(7,
 '0.012*"ostholstein" + 0.012*"sierksdorf" + 0.011*"zweckverband" + '
 '0.010*"nachricht" + 0.008*"danken" + 0.008*"the" + 0.007*"hra" + '
 '0.007*"email" + 0.007*"betreff" + 0.006*"lübeck"'),
(8,
 '0.012*"ostholstein" + 0.009*"sierksdorf" + 0.008*"nachricht" + '
 '0.008*"danken" + 0.008*"betreff" + 0.008*"frau" + 0.008*"zweckverband"
    + '
 '0.008*"hyperlink" + 0.007*"email" + 0.007*"homepage"'),
(9,
 '0.013*"sierksdorf" + 0.012*"zweckverband" + 0.011*"ostholstein" + '
 '0.009*"nachricht" + 0.009*"danken" + 0.008*"frau" + 0.007*"the" + '
 '0.007*"lübeck" + 0.006*"hyperlink" + 0.006*"sitzen"'),
(10,
 '0.011*"zweckverband" + 0.011*"nachricht" + 0.010*"sierksdorf" + '
 '0.010*"betreff" + 0.009*"ostholstein" + 0.009*"lübeck" + 0.007*"the" +
    '
 '0.007*"sitzen" + 0.007*"danken" + 0.007*"hyperlink"'),
(11,
 '0.013*"sierksdorf" + 0.009*"ostholstein" + 0.009*"zweckverband" + '
 '0.008*"betreff" + 0.008*"nachricht" + 0.007*"danken" +
    0.007*"hyperlink" + '
 '0.007*"lübeck" + 0.006*"email" + 0.006*"datum"'),
```

```
(12,
 '0.013*"ostholstein" + 0.012*"sierksdorf" + 0.009*"nachricht" + '
 '0.009*"zweckverband" + 0.008*"betreff" + 0.008*"danken" +
    0.007*"email" + '
 '0.006*"frau" + 0.006*"lübeck" + 0.006*"wagrienring"'),
(13,
 '0.013*"sierksdorf" + 0.010*"the" + 0.010*"zweckverband" + '
 '0.009*"ostholstein" + 0.009*"nachricht" + 0.008*"danken" +
    0.007*"lübeck" + '
 '0.007*"betreff" + 0.007*"frau" + 0.007*"öffentlich"'),
(14,
 '0.012*"zweckverband" + 0.011*"danken" + 0.011*"sierksdorf" + '
 '0.010*"ostholstein" + 0.010*"betreff" + 0.008*"nachricht" + '
 '0.007*"hyperlink" + 0.007*"gmbh" + 0.006*"lübeck" +
    0.006*"kundennummer"'),
(15,
 '0.014*"ostholstein" + 0.011*"sierksdorf" + 0.008*"zweckverband" + '
 '0.008*"danken" + 0.008*"frau" + 0.007*"betreff" + 0.007*"the" + '
 '0.007*"nachricht" + 0.006*"amtsgericht" + 0.006*"sitzen"'),
(16,
 '0.014*"zweckverband" + 0.013*"ostholstein" + 0.011*"nachricht" + '
 '0.009*"danken" + 0.008*"sierksdorf" + 0.007*"hyperlink" +
    0.007*"lübeck" + '
 '0.006*"sitzen" + 0.006*"the" + 0.006*"betreff"'),
(17,
 '0.012*"ostholstein" + 0.011*"danken" + 0.011*"sierksdorf" +
    0.009*"betreff" '
 '+ 0.009*"zweckverband" + 0.008*"nachricht" + 0.007*"the" +
    0.007*"frau" + '
 '0.006*"hyperlink" + 0.006*"sitzen")]
```

Das LDA hat erfolgreich auf allen verfügbaren Daten ein Modell gebaut, das die Daten in Topics $[0, 17]$ eingeteilt hat. Jedes Topic hat dominante Wörter, die zusammen mit ihren jeweiligen Auftretenswahrscheinlichkeiten aufgeslistet sind. Die Wahrscheinlichkeiten für das Auftreten von Wörtern liegt im Intervall $[1.5\%, 0\%]$. Es ist auffällig, dass sich die Themen viele der wahrscheinlichsten Wörter teilen.

2. Die Qualität der Daten und des Klassifikators kann alternativ über die Zuordnung der Topics zu den Abeteilungen untersucht werden. Der wohl naivste Ansatz ist, Topic x auf Abteilung x für alle x in $[0, 17]$ abzubilden. Die folgende Graphik zeigt diesen Ansatz durch die Gegenüberstellung der jeweils zugeordneten Dokumente. Dabei wird nur die Anzahl der zugeordneten Dokumente überprüft, nicht, die Dokumente, die in beiden vorliegen. Überschneidungen werden noch nicht betrachtet.



Das Säulendiagramm zeigt, dass die Topics nicht optimal auf die Abteilungen abgebildet sind. Die Anzahl der enthaltenen Dokumente sollte ähnlicher sein. Betrachtet man zum Beispiel Topic 14, kann diese nicht Abteilung 14 darstellen, da es sich um einen Unterschied von 21312 Dokumenten, also 72%, handelt. Die genaue Anzahl der Zuteilungen zu Topics bzw Abteilungen ist in folgenden Tabellen gelistet:

Topic	Counts LDA
0	8605
1	3959
2	5974
3	1714
4	5823
5	2291
6	23558
7	12721
8	4207
9	1100
10	10021
11	3243
12	14556
13	2697
14	29456
15	1162
16	575
17	1400

Abteilung	Counts Label
0	9770
1	21061
2	14245
3	235
4	1307
5	251
6	4009
7	3610
8	9410
9	23533
10	20676
11	43
12	2866
13	9616
14	8144
15	12126
16	4748
17	2700

Die gleichen Daten sind hier in absteigender Reihenfolge dargestellt. Dabei wurde der Anteil an den Gesamtdaten hinzugefügt, um einen Vergleichswert zu haben:

Topics	Counts LDA	Anteil[%]
14	29456	22.13
6	23558	17.7
12	14556	10.94
7	12721	9.56
10	10021	7.53
0	8605	6.47
2	5974	4.49
4	5823	4,38
8	4207	3.16
1	3959	2.98
11	3243	2.44
13	2697	2.03
5	2291	1.72
3	1714	1.29
17	1400	1.05
15	1162	0.87
9	1100	0.83
16	575	0.43

Abteilung	Counts Label	Anteil[%]
9	23533	15.86
1	21061	14.20
10	20676	13.94
2	14245	9.6
15	12126	8.17
0	9770	6.59
13	9616	6.48
8	9410	6.34
14	8144	5.49
16	4748	3.20
6	4009	2.7
7	3610	2.43
12	2866	1.93
17	2700	1.82
4	1307	0.88
5	251	0.17
3	235	0.16
11	43	0.03

Nun wird durch die quantitativere Darstellung deutlich, dass sich nicht alle Topics eindeutig einer Abteilung zuordnen lassen. Das Ergebnis zeigt, dass die Topic-Verteilung bei LDA das 14. Thema stärker erkennt, als die ZVO die Abteilung mit den meisten zugeordneten Dokumenten. Diese unterschieden sich durch 5923 Dokumente. Die Betrachtung der unterschiedlichen Anteile dient somit nicht gut als Zuordnungsmethode, kann jedoch Aufschluss darüber geben, dass die Daten von der ZVO gleichmäßiger zugeteilt wurden, als sie von dem Topic-Modelling erkannt werden. Dies bedeutet, dass die Wörter in den Dokumenten in ihrer Bedeutung und syntaktischen Umgebung zu ähnlich sind, als dass sie sich optimal in die von der ZVO vorgegebenen Abteilungen einteilen lassen. Dies könnte auch an bestimmten Wörtern liegen, die in vielen Abteilungen vorkommen, wie zu, Beispiel 'Ostholstein' oder 'Zweckverband'. Die Wörter können dafür sorgen, dass Dokumente in die gleiche Topic sortiert werden, die sich eigentlich durch andere Wörter semantisch stärker voneinander unterscheiden und manuell anderen Abteilungen untergeordnet wurden.

Den Topics ist also keine Verbindung zu den gegebenen Abteilungen direkt zuzuordnen. Diese muss über andere Weise herausgefunden werden, damit die Zuordnung der Dokumente evaluiert werden kann. Die Zuordnung der Topics zu Abteilungen kann approximiert werden, indem gezählt wird, in wie vielen Dokumenten sich die Topics und Abteilungen jeweils überschneiden. Genau wie bei der Gruppeneinteilung wird für diese Überschneidung jedes Dokument erneut auf dem Korpus geprüft, um dessen Topic-Wort-Verteilung zu erstellen. Die Dokumente werden anhand ihrer einzigartigen Zeilen-ID adressiert. Die IDs werden der Topic zugeordnet, die die höchste Wahrscheinlichkeit in der Verteilung des Dokuments erreicht. Die Zuordnung findet in einer großen Liste statt, die über 18 Unterlisten verfügt, in die

die IDs jeweils hinzugefügt werden. Die Überschneidungen werden in einer Matrix dargestellt, die für jedes Topic des Modells die Überschneidungen zu jeder Abteilung auflistet. Somit kann für jede Zeile die passende Zuordnung gefunden werden. Dabei stellt eine Zeile jeweils die Dokumente dar, die aufsummiert die Zahl der Tabelle X (vorherige) ergeben:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	555	977	**2150**	21	66	19	211	251	910	960	902	2	210	596	302	758	318	250
1	512	580	**650**	3	36	7	139	101	370	453	570	0	66	378	91	196	123	53
2	508	755	263	9	67	21	254	145	282	**1498**	1207	1	119	300	479	527	177	130
3	108	179	98	2	15	3	33	57	236	**432**	218	0	34	55	153	198	74	27
4	493	867	442	5	63	8	134	180	451	**1111**	1033	8	73	348	446	489	191	89
5	91	467	98	5	25	2	60	75	162	**480**	309	1	55	127	169	351	113	47
6	1263	3993	1738	28	127	48	811	587	1384	4139	**4730**	6	573	2096	1620	2080	680	495
7	**3376**	1298	870	17	524	16	221	338	662	1904	1740	9	185	381	664	891	537	133
8	310	585	216	10	27	5	67	114	203	771	**1013**	0	81	323	264	440	205	73
9	58	126	47	0	3	1	17	16	131	174	**400**	0	15	55	63	69	59	17
10	286	**1610**	1733	12	51	27	284	299	858	1567	1331	4	239	893	674	673	304	248
11	194	409	377	8	27	3	88	122	180	**678**	470	0	80	173	167	395	142	118
12	486	2355	2495	16	80	24	564	327	1005	**2672**	1513	3	311	1254	886	1377	579	327
13	151	393	198	5	23	4	64	83	160	414	**639**	1	68	192	123	275	151	68
14	1008	6084	2681	86	153	61	990	820	2223	**5730**	3832	8	701	2327	1860	3093	993	552
15	270	95	36	3	10	1	29	32	46	228	**284**	0	19	31	61	81	23	21
16	81	82	25	1	3	0	18	38	27	**105**	90	0	14	29	37	76	24	15
17	20	206	128	4	7	1	25	25	120	**217**	395	0	23	58	85	157	55	37

Diese Matrix bildet die Überschneidungen aller 133044 Dokumenten dar. Dabei sind die 18 LDA Topics auf der vertikalen Achse und die 18 LABEL Abteilungen auf der horizontalen Achse aufgetragen. Die Ausführung auf diesen Daten hat bereits mehrere Tage gedauert. Bezüglich der Zuordnung ist der naivste Ansatz, jeder Zeile (also LDA Topic) die Spalte (also Label Abteilung) mit der maximalen Überschneidung zuzuordnen. Dabei ist der Ziel die Anzahl der Gesamtüberschneidungen zu maximieren. Wendet man diesen Algorithmus an, sieht der Output wie folgt aus:

```
maxmatrix = [2,2,9,9,9,9,10,0,10,10,1,9,9,10,9,10,9,9]
```

Wie zu erkennen ist, sind nur die Zahlen 0, 1, 2, 9, 10 in der Liste vertreten. Eine optimale Zuordnung wäre jedoch erst erreicht, wenn alle Zahlen von Eins bis 18 ohne Duplikate in der Liste in einer beliebigen Reihenfolge vorkommen. Dies liegt daran, dass eine Topic nur genau eine Abteilung darstellen darf. In der Liste wird jedoch deutlich, dass viele Themen des LDA Modells mit dem neunten Thema der ZVO kompatibel wären. Dabei werden 9/18 Topics Abteilung 9, 1/18 Topic Abteilung 1, 5/18 Topics Abteilung 10 und 2/18 Topics Abteilung 2 zugeordnet. Hier fällt auf, dass diesen

Abteilungen die meisten Dokumente von der ZVO händisch zugeordnet wurden. Somit können sie sich auch mit vielen Dokumenten aus den LDA Topics überschneiden. **Dies lässt zusätzlich vermuten, dass zum Beispiel die Abteilung 9, die laut Tabelle X die meisten Dokumente enthält, ein zu breites inhaltliches Spektrum abdeckt und in noch weitere Unterthemen unterteilt werden könnte. Dies wird dadurch begründet, dass das Topic-Modell viele der Dokumente aus Abteilung 9 in verschiedene Topics eingeteilt hat.**

Die Zuteilung anhand der maximalen absoluten Überschneidungen ist nicht injektiv. Für die Verarbeitung der Daten ist die injektive Zuordnung jedoch fundamental wichtig, aber schwierig zu erreichen. Für die Beurteilung der Matrix führen wir einen weiteren Parameter ein, die durchschnittliche Überschneidung. Dadurch kann die Kompatibilität der Abteilungen im Bezug auf das Matching mit Topics besser analysiert werden. Vorallem bei Abteilungen, die eine sehr unausgeglichene Überschneidungsmengen haben, da sie deutlich mehr Dokumente als andere enthalten, bietet der Durchschnitt eine alternative Sichtweise. Für die durchschnittliche Überschneidung wird zuerst die Summe aller Dokumente einer Abteilung errechnet, indem alle Werte einer Spalte addiert werden. Daraufhin entsteht eine neue Matrix, in der jeder Wert jeweils durch die Summe seiner Spalte dividiert wird. Dabei wird die Anzahl an Dokumenten in jeder Abteilung irrelevant für das Endergebnis. Somit beschreibt jeder Wert, wie groß der Anteil der aus dieser Topic überschneidenden Dokumente bezogen auf die gesamte Abteilung ist in %:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
0	5.68	4.64	15.09	8.94	5.05	7.57	5.26	6.95	9.67	4.08	4.36	4.65	7.33	6.2	3.71	6.25	6.7	9
1	5.24	2.75	4.56	1.28	2.75	2.79	3.47	2.8	3.93	1.92	2.76	0	2.3	3.93	1.12	1.62	2.59	1
2	5.2	3.58	1.85	3.83	5.13	8.37	6.34	4.02	3	6.37	5.84	2.33	4.15	3.12	5.88	4.35	3.73	4
3	1.11	0.85	0.69	0.85	1.15	1.2	0.82	1.58	2.51	1.84	1.05	0	1.19	0.57	1.88	1.63	1.56	
4	5.05	4.12	3.1	2.13	4.82	3.19	3.34	4.99	4.79	4.72	5	18.6	2.55	3.62	5.48	4.03	4.02	
5	0.93	2.22	0.69	2.13	1.91	0.8	1.5	2.08	1.72	2.04	1.49	2.33	1.92	1.32	2.08	2.89	2.38	1
6	12.93	18.96	12.2	11.91	9.72	19.12	20.23	16.26	14.71	17.59	22.88	13.95	19.99	21.8	19.89	17.15	14.32	18
7	34.55	6.16	6.11	7.23	40.09	6.37	5.51	9.36	7.04	8.09	8.42	20.93	6.45	3.96	8.15	7.35	11.31	4
8	3.17	2.78	1.52	4.26	2.07	1.99	1.67	3.16	2.16	3.28	4.9	0	2.83	3.36	3.24	3.63	4.32	
9	0.59	0.6	0.33	0	0.23	0.4	0.42	0.44	1.39	0.74	1.93	0	0.52	0.57	0.77	0.57	1.24	0
10	2.93	7.64	12.17	5.11	3.9	10.76	7.08	8.28	9.12	6.66	6.44	9.3	8.34	9.29	8.28	5.55	6.4	9
11	1.99	1.94	2.65	3.4	2.07	1.2	2.2	3.38	1.91	2.88	2.27	0	2.79	1.8	2.05	3.26	2.99	4
12	4.97	11.18	17.51	6.81	6.12	9.56	14.07	9.06	10.68	11.35	7.32	6.98	10.85	13.04	10.88	11.36	12.19	1
13	1.55	1.87	1.39	2.13	1.76	1.59	1.6	2.3	1.7	1.76	3.09	2.33	2.37	2	1.51	2.27	3.18	2
14	10.32	28.89	18.82	36.6	11.71	24.3	24.69	22.71	23.62	24.35	18.53	18.6	24.46	24.2	22.84	25.51	20.91	20
15	2.76	0.45	0.25	1.28	0.77	0.4	0.72	0.89	0.49	0.97	1.37	0	0.66	0.32	0.75	0.67	0.48	0
16	0.83	0.39	0.18	0.43	0.23	0	0.45	1.05	0.29	0.45	0.44	0	0.49	0.3	0.45	0.63	0.51	0
17	0.2	0.98	0.9	1.7	0.54	0.4	0.62	0.69	1.28	0.92	1.91	0	0.8	0.6	1.04	1.29	1.16	1
∑	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1

Wählt man aus jeder Zeile das maximale Element aus und notiert die assoziierten Abteilungen in einer Liste, dann ergibt sich das folgende Ergebnis. Der Unterschied bei der Darstellung ist nun, dass die Zahlen pro Zeile nicht mehr als absolutes Maß betrachtet werden, sonder als relative Häufigkeit in der jeweiligen Abteilung. Somit macht es keinen Unterschied, dass die Abteilungen verschieden viele Dokumente von der ZVO zugeordnet bekommen haben.

```
average_row = [2 ,0 ,5 ,8 ,11 ,15 ,10 ,4 ,10 ,10 ,2 ,17 ,2 ,16 ,3 ,0 ,7 ,10]
```

Man erkennt bereits eine Verbesserung zu Figure X (absolute Max Zuteilung), da in der Zuteilung mehr Abteilungen zugeordnet wurden. Insgesamt sind 12/18 in der Liste vertreten. Wie bei der ersten Liste, bei der 0, 1, 2, 9, 10 mehrfach zugeweisen wurden, sind hier 2 und 10 drei mal zugewiesen und 0 zwei mal. Somit unterstreicht das Ergebnis die Problematik der Zuteilung der ZVO in Bezug auf diese Abteilungen.

5

Analyse

Die Implementierung des LDA Modells wurde im vorausgegangenen Teil umegsetzt. Die Ausgaben werden nun analysiert....

1. **Gruppen und Wörter finden**
   In der Implementierung wurden 18 Topics gefunden. Diese Topics sind durch die Wörter der Topic-Wort-Verteilung definiert. Als Anwendungsziel in Bezug auf die ZVO sollen die Topics des Modells in Zukunft die Abteilungen darstellen. Betrachtet man dafür die Wörter der Topics, um die Topics semantisch den Abteilungen zuzuordnen, wird dies nicht gelingen. Das liegt daran, dass sich die dominaten Wörter in den Topics zu stark überschneiden. Zum Beispiel gehört östholstein
   in allen Topics zu den Top 5 Wörtern. Das bedeutet, dass sich die Themen semantisch nicht stark genug von einander abgrenzen lassen, da die Wörter zu ähnlich sind. Dies lässt auf die Folgerung schließen, dass die Daten bei der Datenreinigung noch stärker um die häufigen Wörter reduziert werden sollten, um eine bessere Qualität zu erreichen. Dabei ist es jedoch wichtig, nicht die entscheidenen Topic-relevanten Wörter auszuschließen.

2. **Zuordnung Abteilung zu Topic**
   Die durchschnittliche Überschneidung zwischen Abteilung und Topic spiegelt wider, wie stark die Abteilung mit einem oder mehreren Topics übereinstimmt. Ein hoher Durschnitt kann entweder eine sehr hohe Überschneidung der Abteilung mit einer Topic oder relativ hohe Überschneidungen mit mehreren Topics bedeuten. Konkret für die Abteilungen bedeutet das:

6

Zusammenfassung und Ausblick

...