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

1.1 Contributions of this Thesis 1
1.2 Related Work 1
1.3 Structure of this Thesis 1
1.4 Motivation 1
1.5 Ziel 1

2 Grundlagen 3

2.1 Notation 3
2.2 Modellvergleich 3
2.3 Grundlagen der Latent Dirichlet Allocation (LDA) 4

3 Konzept 8

3.1 Daten 8
3.2 Anwendungsfall ZVO 9

4 Implementierung 13

4.1 Topic Modeling Methode 13
4.2 Toolauswahl 13
4.3 Umsetzung Konzept 13

5 Analyse 19

6 Zusammenfassung und Ausblick 20

1

Einleitung

1.1 Contributions of this Thesis

1.2 Related Work

1.3 Structure of this Thesis

1.4 Motivation

Die digitalisierte Welt generiert täglich riesige Mengen an neuen Informationen. Die Kapazitäten, die ein Mensch aufbringen kann, um solche Massen an Daten zu organisieren und zu verstehen, sind schon lange übertroffen. 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. 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 Themenmodelle statistische Modelle dar, die Verwendung in der Inferenz abstrakter Themen in unsortierten Datenmengen finden. In einer Welt von exponentiell wachsenden Datenmengen finden Methoden der Themenmodellierung stetig eine breitere Anwendung. Bereits heute wird Themenmodellierung in vielen Bereichen der Wirtschaft, Wissenschaft und Informationstechnologie verwendet. 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 einem Cluster gruppiert.

1.5 Ziel

Diese Arbeit wird die Theorie der Themenmodellierung anhand des Beispiels des Zweckverband Ostholstein (ZVO) implementieren und die bezüglichen Parameter im Sinne der

Auswertung bewerten. Der ZVO erhält jährlich ein große Menge an Kundenanfrage. Diese werden momentan händisch an die jeweils zuständige Abteilung weitergeleitet. 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 verschiedener Themenmodellierung verglichen wird. Das Ergebnis einer Themenmodellierung 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.

2

Grundlagen

2.1 Notation

- $\mathcal{K}$ ist die Anzahl der Themen 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 Dirichlet Allocation

Themenmodellierung besteht aus vielen Methoden, die meist verbreitete ist die „Latent Dirichlet Allocation (LDA)“, was als Bag of Word modelliert ist, also keine Kontextinformationen beinhaltet. Dieses Verfahren ist eine Weiterentwicklung des ‚PLSI‘, das durch zwei Dirichlet-Priors ergänzt wurde. LDA liegt ein generierender Prozess zugrunde, den zwei Dirichlet Verteilungen maßgeblich beeinflussen: die Dokument-Themen Verteilung, die die Ausprägungen verschiedener Themen in einem Dokument beschreibt, und die Themen-Wörter 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 Themen ist, während ein Thema 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 Dirchlet- Verteilungen gesucht werden, die ein möglichst ähnliches Dokument generieren.

Latent Semantic Analysis (LSA)

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 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.

Non-Negative Matrix Factorization (NMF)

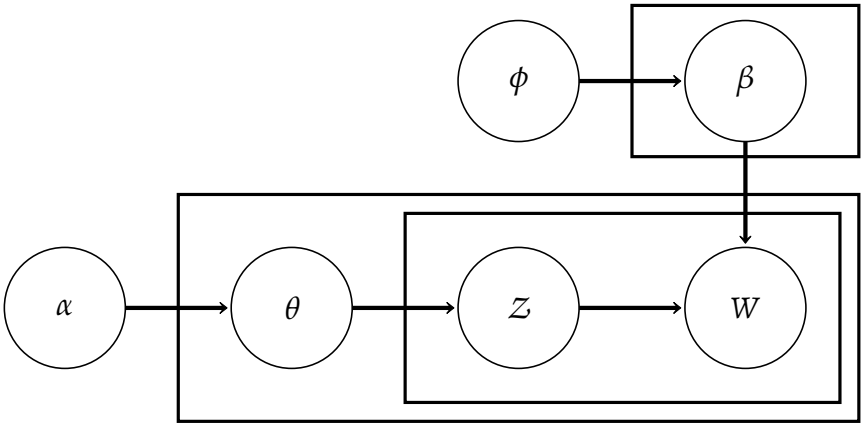
Ein weiteres Verfahren, das auch mit Matrizen funktioniert, wird „Non-Negative Matrix Factorization“ (NMF) genannt. Dabei wird eine Matrix, die Wörter auf Dokumente abbildet, in zwei Teilmatrizen faktorisiert. Die erste Teilmatrix stellt die Topics in Dokumenten, die zweite die Wörter in Topics dar. 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 (LDA)

Latent Dirichlet Allocation ist ein grundlegendes und bekanntes Verfahren aus der Natürlichen Sprachverarbeitung. Das Prinzip der Themenmodellierung basiert auf einer Menge an Dokumenten, die den Korpus darstellen. Dabei werden 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. Bei der Topic-Modellierung 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:



Dabei beschreibt W als einzige nicht verborgene Variable eines von N Wörtern des Dokuments. Das Wort ist semantisch einem Thema Z zugeordnet. Das Thema wiederum 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 Themen-Wort-Verteilung beeinflusst. Bei einem großen α ist die Verteilung der Themen 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 Themen in einem Dokument zu maximal vielen Wörtern in diesen Themen führt. Die minimale Anzahl an Themen wäre erreicht, wenn man alle Wörter eines Dokuments einem Thema zuweist. Dadurch verfügt das Thema 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 je-

dem 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

Der Algorithmus hinter LDA generiert neue Dokumente mithilfe von Dirichletverteilungen $Dir(\gamma)$ und Multinomialverteilungen $Multinom(\delta)$. 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 Themen 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:

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

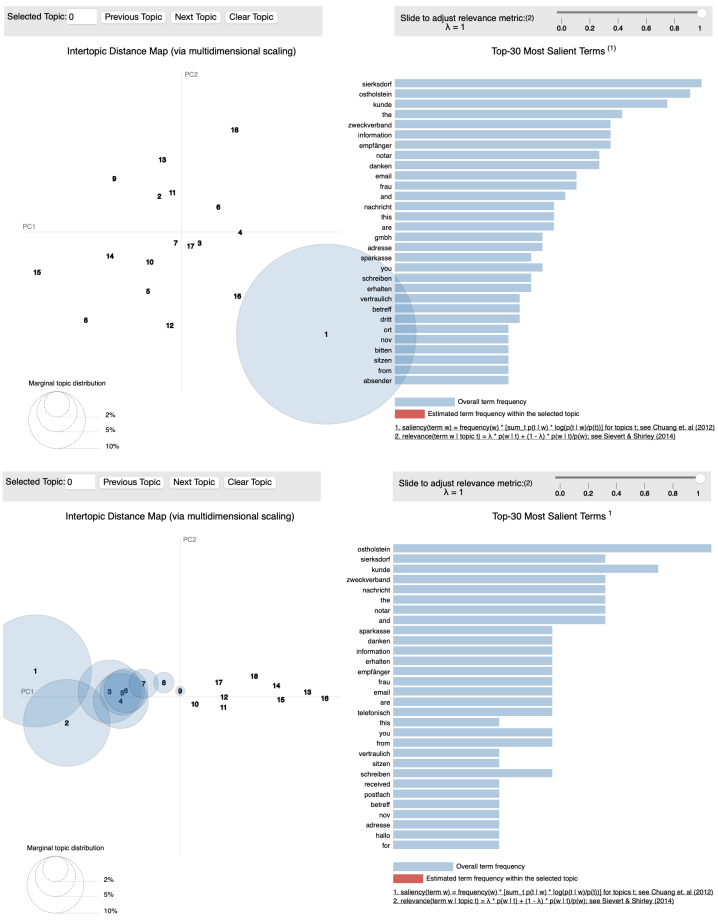
Die Formel berechnet die Wahrscheinlichkeit der Verteilung unter einem bestimmten Thema gegeben der α und β Parameter und dem bekannten Wort. Die Wahrscheinlichkeit kann nicht exakt bestimmt werden, weshalb Verfahren wie Gibbs Sampling diese approximieren.

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:



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.

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

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 müssen 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.

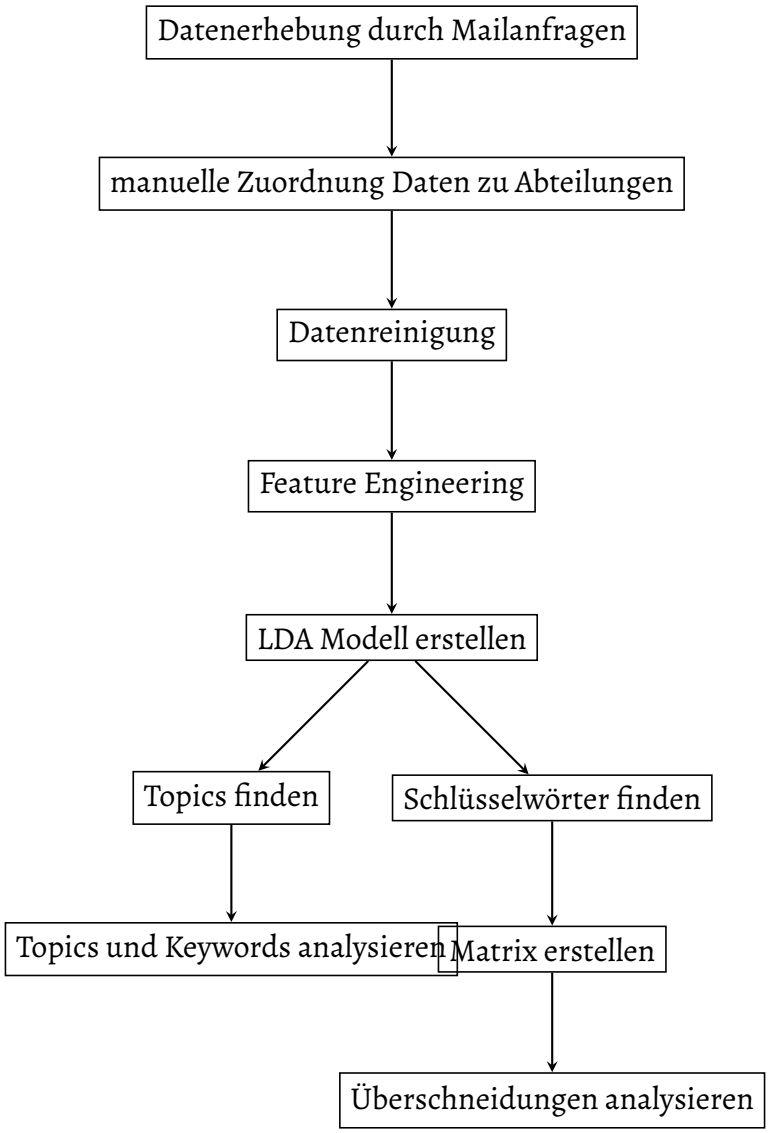
Feature Engineering

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).

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 Da-

ten 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:

Ist das Modell generiert und die Topics auslesbar, können die Daten evaluiert werden. Dazu werden zwei Ansätze verfolgt. Zum Einen werden Topics und zugehörigen

```
data ←
for d in Anzahl Dokumente do
    data ← 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
```

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 To-

pics 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. 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. Bei LDA Modellen ist die Anzahl der Topics ein individueller Input, durch den sich das Ergebnis schwerwiegend verändern kann. 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

Als Bibliothek wird Gensim verwendet, die für die Verarbeitung von unstrukturierten Daten und Anwendung von unüberwachten Algorithmen entwickelt wurde. Algorithmen, wie word2vec, LSI oder LDA entdecken automatisch Strukturen durch das Prüfen von gemeinsam auftretenden Mustern im Korpus der Trainingsdaten. Gensim erlangte in der Vergangenheit Bekanntheit durch seine hochoptimierten Implementationen bekannter Algorithmen und der Schnelligkeit und Verlässlichkeit, mit der diese ausgeführt wurden.

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. An den Topics ist keine Verbindung zu den gegebenen Abteilungen ersichtlich. Diese muss über andere Weise herausgefunden werden, damit die Zuordnung der Dokumente evaluiert werden kann. Die Zuordnung der Topics zu Abteilungen kann nur approximiert werden, indem gezählt wird, in wie vielen Dokumenten sich die Topics und Abteilungen jeweils überschneiden. Für diese Überschneidung wird jedes Dokument erneut auf dem Korpus geprüft, um dessen Topic-Wort-Verteilung zu erstellen. Die Dokumente werden anhand ihrer einzigartigen Zeilen-ID adressiert. 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. Das folgende Beispiel zeigt eine Zurodnung mithilfe der ersten 10000 Dokumente:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	536	91	67	1	12	1	12	6	23	145	69	0	22	12	38	87	24	6
1	15	22	11	0	2	0	0	1	7	32	77	0	6	3	5	16	3	1
2	6	50	81	1	3	1	31	6	28	61	30	0	18	15	22	33	4	16
3	14	69	58	3	3	0	19	11	37	104	63	0	18	20	21	51	33	5
4	7	81	104	1	3	0	7	9	56	112	97	1	11	28	36	42	15	12
5	0	4	3	1	0	0	1	0	2	5	3	0	1	1	0	1	0	0
6	2	7	24	0	0	0	2	0	7	25	23	0	2	5	2	13	8	1
7	19	331	90	3	16	3	49	22	111	206	200	2	29	49	61	127	36	44
8	3	20	33	0	1	0	2	1	11	41	78	0	3	5	7	15	8	1
9	4	21	17	0	2	0	1	2	26	18	33	0	3	9	5	4	2	0
10	13	137	115	0	6	0	25	16	53	183	136	3	6	34	75	57	39	9
11	27	429	221	8	34	4	129	34	291	312	271	2	48	137	75	149	50	106
12	0	12	8	0	1	0	3	2	5	24	13	0	2	1	3	10	5	1
13	12	89	73	0	9	1	9	18	25	106	124	0	4	14	29	69	52	5
14	10	174	110	4	11	1	15	5	42	226	137	1	19	38	39	181	21	27
15	8	67	70	1	4	0	11	4	21	85	115	0	8	16	20	47	21	14
16	19	33	44	0	5	0	4	4	18	56	63	0	6	9	16	18	7	3
17	39	47	52	0	5	1	15	10	203	74	71	1	12	13	16	33	17	6

Diese Matrix bildet die Überschneidungen der ersten 1000 Dokumenten dar. Dabei sind die 18 LDA Topics auf der vertikalen Achse und die 18 LABEL Abteilungen auf der horizontalen Achse aufgetragen. Die Schwierigkeit besteht darin, ein Mapping zu finden, das in Summe die meisten Überschneidungen ergibt und gleichzeitig jede Abteilung nur genau einem Topic zugeordnet wird. Würde man die zweite Prämisse vernachlässigen, würde sich diese Zuordnung der Topics zu Abteilungen ergeben:

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

Eine optimale Zuordnung ist erreicht, wenn alle Zahlen von Eins bis 18 ohne Duplikate in der Liste in einer beliebigen Reihenfolge vorkommen. Das würde bedeuten, dass einem Thema von LDA genau eine Abteilung der ZVO am wahrscheinlichsten zugeordnet ist. In der Liste wird jedoch deutlich, dass viele Themen des LDA Modells mit dem 9. und 10. Thema der ZVO kompatibel wären. Für die Verarbeitung der Daten ist die injektive Zuordnung funamental wichtig, jedoch 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, bietet der Durchschnitt eine alternative Sichtweise. Die durchschnittliche Überschneidung ergibt sich aus der Summe aller Werte in einer Spalte geteilt durch die Anzahl der Werte. Somit ist die durchschnittliche Anzahl, die eine bestimmte Abteilung mit allen Topics hat gegeben:

```
sum_[734, x,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x]
average_cut = [40.78,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x,x]
```

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

...