Universität Bremen

FACHBEREICH 3
FAKULTÄT FÜR MATHEMATIK UND INFORMATIK

Topic Modeling basierte Analyse eines Patentdatensatzes von General Motors

Abschlussarbeit im Studiengang
Bachelor of Science Wirtschaftsinformatik
der Universität Bremen

Name,Vorname:	Tietjen, Hauke
Matrikelnummer:	4224296
Datum:	XX.XX.XXX
Studiengang:	Wirtschaftsinformatik, Bachelor of Science
Eingereicht bei:	Prof. Dr. Martin Möhrle (Universität Bremen)
	Prof. Dr. XXX (Universität Bremen)

Inhaltsverzeichnis

Abkürzungsverzeichnis

BoW	Bag-of-Words	5
CRP	Chinese Restaurant Process	8
CVT	Continiuously Variable Transmission	11
DLDA	Dynamic Latent Dirichlet Allocation	8
GM	General Motors	10
HLDA	Hierarchic Latent Dirichlet Allocation	8
IPMI	Institute of Project Management and Innovation	10
LDA	Latent Dirichlet Allocation	5
TDM	Term-Dokument Matrix	11
tf-idf	Term Frequency-Inverse Document Frequency	15

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Thema

In dieser Bachelorarbeit geht es darum die versteckten Themen, in einem Patentdatensatz von General Motors, zu finden. Diese Themen sollen benannt und graphisch dargestellt werden, um herauszufinden welche Themengruppen es gibt und welche Patente zu einem oder mehreren Themen gehören. Außerdem soll die Entwicklung der Themen über die Jahre untersucht werden.

1.2 Motivation und Zielsetzung

Uns standen noch nie so viele Informationen zur Verfügung wie heute und jeden Tag kommen neue hinzu. Wir durchsuchen schriftliche Informationen nach Stichwörtern, mit der Hilfe von Suchmaschinen. So lassen sich zu einem Thema schnell mehrere Texte finden.

Man beschreibt ein Thema aus Stichwörtern und sucht Texte, welche diese enthalten. Wenn man diese Suche umdreht funktioniert dies nicht mehr. Man hat einen Datensatz aus Texten und möchte alle darin enthaltenen Themen herausfinden. Intuitiv denkt man hier an den Titel aber der reicht nicht aus, um alle Themen eines Textes zu beschreiben. Allein der Titel dieser Arbeit verschweigt das Thema

der Programmiersprache Python. Manche Texte haben Schlagworte aber hier verlässt man sich auf den Autor, die Richtigen zu wählen und sie werden nicht nach Relevanz gewichtet. Außerdem könnte man, mit dem Wissen über die Entwicklung der Patentthemen, Vermutungen über die Patentthemen der Zukunft anstellen.

Topic Modeling finde ich besonders interessant, weil man mit relativ geringem Aufwand große Mengen an Dokumenten untersuchen kann. Dadurch könnte man, speziell in diesem Fall, für die Konkurrenten von General Motors herausfinden worum es in den Patenten geht und in welche Richtung sich die Themen der Patente in Zukunft entwickeln könnten. Wodurch man General Motors bei der Anmeldung von neuen Patenten zuvorkommen und Lizenzgebühren verlangen könnte.

Also wie findet man in einem Textdatensatz die enthaltenen Themen und ihren zeitlichen Verlauf?

1.3 Methodisches Vorgehen

Um die versteckten Themen zu finden werden generative Wahrscheinlichkeitsmethoden benutzt. Eine Methode ist die Latent Dirichlet Allocation. (vgl. Blei, Ng, Jordan und Lafferty 2003) Zuerst wird eine bestimmte Zahl an Themen festgelegt. Wörter die häufig gemeinsam vorkommen werden einem gemeinsamen Thema zugeordnet. Nachdem alle Wörter mindestens einem Thema zugeordnet wurden, wird der Vorgang für eine höhere Zahl an Themen wiederholt bis man genug Modelle hat, um sie zu vergleichen. Aus den Modellen wird das mit der höchsten

Kohärenz ausgewählt. (vgl. Röder, Both und Hinneburg 2015)

Die wahrscheinlichsten Wörter eines Themas könnten lauten Ventil, Hydraulik und Flüssigkeit. Dieses Thema kann dann wiederum Texten zugeordnet werden. Mit dieser Methode lassen sich die Themen eines Datensatzes von hunderten Dokumenten viel schneller herausfinden, als es einem Menschen allein möglich wäre.

Am Beispiel des Patentdatensatzes von General Motors werde ich die Modelle des Online Latent Dirichlet Allocation Verfahrens (vgl. Hoffman, Blei und Bach 2010) und des MALLET Verfahrens (vgl. McCallum 2002) auf Kohärenz vergleichen. Dabei werde ich auch die Kohärenzmaße C_v und C_umass vergleichen. (vgl. Röder, Both und Hinneburg 2015)

Der Patentdatensatz von General Motors umfasst über 1400 Patente für verschiedene Getriebearten und ist ausreichend groß um Topic Modeling zu betreiben.

Des weiteren werde ich mit dem dynamischen Latent Dirichlet Allocation Verfahren herausfinden wie sich die Themen des Datensatzes, entlang der zeitlichen Anmeldedaten der Patente, verändert haben. (vgl. Blei und Lafferty 2006) Besonders interessant wäre hier eine Veränderung des Themenschwerpunktes. Auch eine Vorhersage zu welchen Themen in Zukunft Patente angemeldet werden könnte möglich sein. Eine Vorhersage wäre für ein konkurrierendes Unternehmen hilfreich, um Patente vor General Motors anzumelden und Lizenzgebühren verlangen zu können.

Um diese Untersuchungen zu realisieren werde ich die Programmiersprache Python verwenden. Mit Hilfe der Programmbibliothek gensim (vgl. Řehůřek und Sojka 2010) werde ich die Modelle erstellen und die Kohärenzen auswerten. Die Ergebnisse werde ich entsprechend ihrer Art visualisieren. Für die am häufigsten vorkommenden Themen werde ich LDAvis verwenden. (vgl. Sievert und Shirley 2014)

2 Begriffliche Grundlagen

2.1 Latent Dirichlet Allocation

Herr Möhrle hat gesagt bei längeren Zitaten, Verweis nach dem ersten Satz

Bei Erstverwendung einer Abkürzung (ABK) in Klammern?

Die Latent Dirichlet Allocation (LDA) ist ein generatives Wahrscheinlichkeitsmodell für Textdokumente. (vgl. Blei, Ng, Jordan und Lafferty 2003, S. 996) Dokumente werden als zufällige Mischverteilungen über latente Themen dargestellt, wobei jedes Thema eine Wahrscheinlichkeitsverteilung über Worte ist.

Vereinfacht gesagt werden alle Dokumente mit einer Wahrscheinlichkeit zu vorher unbekannten Themen zugeordnet. Die Themen werden also durch den Algorithmus gefunden. Ein Thema besteht aus der Menge aller in den Dokumenten vorkommenden Wörtern und ihrer Wahrscheinlichkeit das sie zu diesem Thema gehören. Die Reihenfolge der Dokumente ist nicht relevant. Auch die Reihenfolge der Wörter in den Dokumenten wird nicht beachtet, sondern nur die Häufigkeit, es gilt das Bag-of-Words Modell. (vgl. Harris 1954, S. 155-156) Die Anzahl der latenten Themen muss vorher gegeben sein. Um die Anzahl an versteckten Themen zu approximieren werden alle LDA Modelle mit den Themenanzahlen von 1 bis 100 erstellt. Diese Modelle werden anhand ihrer Kohärenz innerhalb der Themen und anhand ihrer Distanz zwischen den Themen verglichen. Mithilfe dieser Daten sucht man ein Modell aus, das eine möglichst geringe Themenanzahl, hohe Kohärenz

zitat finden

beispiel geben zu LDA

und hohe Distanz aufweist. Die Themenanzahl sollte möglichst gering sein, weil es aufwändig ist diese Themen zu interpretieren und die Distanz bei zu hoher Themenzahl sinkt.

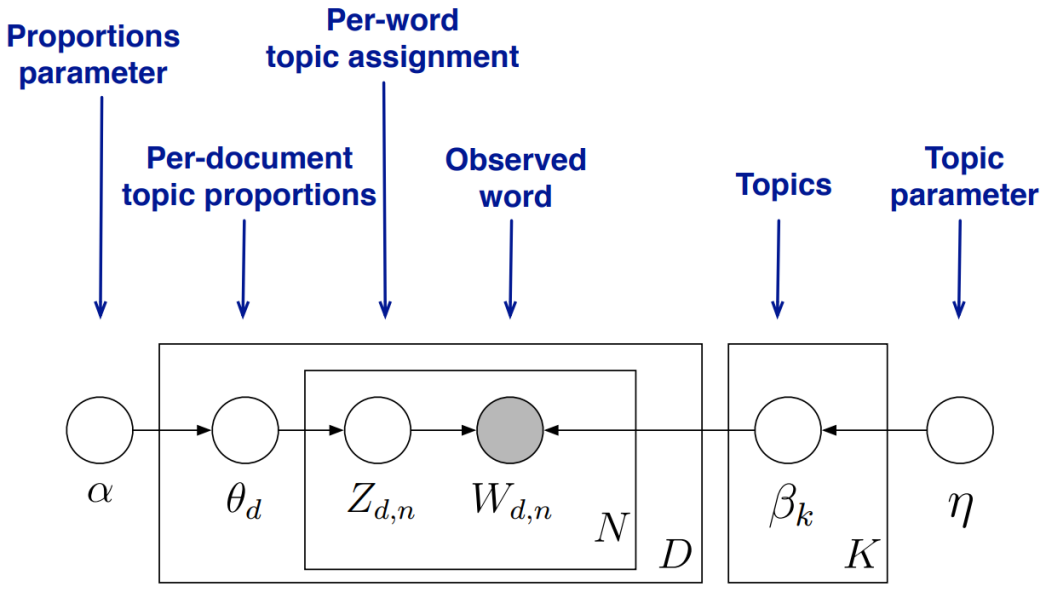

Abbildung 2.1: *LDA als graphisches Modell, von (vgl. Blei 2012, S. 23)*

$\mathcal{W}$ ist das Wort aus $\mathcal{N}$ Wörtern eines Dokuments i. Dieses Dokument i ist eines aus allen Dokumenten $\mathcal{M}$. Alle folgenden Parameter sind latent. $\mathcal{Z}$ ist das Thema für das Wort j aus besagtem Dokument i. Jedem Wort wird ein Thema zugeordnet. Wodurch jedes Dokument eine Mischung aus allen Themen ist. Die Verteilung der Themen für Dokument i ist θ. Die Hyperparameter α und β der Latent Dirichlet Allocation. α bestimmt die Dokument-Themen Verteilung und die Wort-Themen Verteilung. Ein hoher α Wert erhöht die Wahrscheinlichkeit dafür das einem Dokument mehr Themen zugeordnet werden. Ein niedriger α Wert verringert die Wahrscheinlichkeit das einem Dokument mehrere Themen zugeordnet werden. Ein hoher β Wert erhöht die Wahrscheinlichkeit das einem Thema mehr Wörter zugeordnet werden. Ein niedriger β Wert erhöht die Wahrscheinlichkeit das einem Thema weniger Wörter zugeordnet werden. Vereinfacht gesagt lässt ein großer α Wert die Dokumente ähnlicher aussehen und ein ho-

her β Wert lässt die Themen ähnlicher aussehen. Mit diesem Algorithmus lässt sich ein Model erstellen, das jedes Wort mit Wahrscheinlichkeit zu jedem Thema zuordnet.

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Abbildung 2.2: *LDA als graphisches Modell, von (vgl. Blei 2012, S. 25)*

Dies ist die Wahrscheinlichkeit ein Dokument zu generieren, mit den Einstellungen des LDA Modells. Die Wahrscheinlichkeit ist gering aber je höher sie ist desto besser ist das Modell. Die vier Komponenten der Formel sind die Einstellungen des LDA Modells als Faktoren. Diese ergeben wiederum eigene Wahrscheinlichkeiten. Der Erste Faktor ist eine Dirichletverteilung von Dokumenten zu Themen. Eine Dirichletverteilung kann man sich als n-Simplex vorstellen, mit n gleich der Anzahl von Themen. Jedes Dokument hat eine Wahrscheinlichkeit für die Zugehörigkeit zu jedem Thema. Die Dirichletverteilung ist also eine Verteilung von Verteilungen. Der zweite Faktor ist eine Dirichletverteilung von Themen zu Wörtern und verhält sich analog zur ersten.

Der dritte Faktor ist eine Multinomialverteilung des ersten Faktors. Eine Multinomialverteilung ist wie eine Urne mit mehreren verschiedenen Themen, die mit Wahrscheinlichkeiten gezogen werden können. Diese zweite Multinomialverteilung ist also eine von Themen. Der Vierte Faktor ist eine Multinomialverteilung des zweiten Faktors mit Worten. Diese verhält sich analog zur ersten.

Kombiniert man diese Multinomialverteilungen miteinander, indem man immer ein Thema aus der ersten zieht und zu dem Thema passend ein Wort aus der zweiten, generiert man ein neues Dokument. Dies wird wiederholt bis gleich viele Dokumente generiert wurden wie verarbeitet wurden. Die Wahrscheinlichkeit

das man mit dieser Methode die gleichen Dokumente erzeugt ist wie gesagt gering.

Die Dirichletverteilungen werden mit den α und β Werten beeinflusst. Es werden viele verschiedene Werte getestet und das Modell mit der höchsten Wahrscheinlichkeit die gleichen original Dokumente zu erzeugen gewinnt.

warum LDA? in der conclusion: stevens2012e

2.2 Dynamic Latent Dirichlet Allocation

Die Dynamic Latent Dirichlet Allocation (DLDA) ist eine Version des LDA, welche die chronologische Reihenfolge der Dokumente berücksichtigt. Dadurch ist es möglich die Veränderung der Themenschwerpunkte über den Zeitraum zu betrachten. (vgl. Blei und Lafferty 2006)

2.3 Hierarchical Latent Dirichlet Allocation

Hierarchisches LDA (HLDA) erweitert LDA, um eine beliebig tiefe Hierarchie aus Unterthemen. (vgl. Griffiths, Jordan, Tenenbaum und Blei 2004) Diese lassen sich als Baumdiagramm darstellen. Dadurch erhält man noch mehr Informationen zu einem Thema, um es genauer zu benennen. Auch Cluster lassen sich dadurch erkennen. HLDA benutzt den Chinese Restaurant Process (CRP). Angenommen es gibt ein chinesisches Restaurant mit unendlich vielen Tischen, an denen unendlich viele Gäste sitzen können. Der erste Gast setzt sich an den ersten Tisch. Der zweite Gast setzt sich an den ersten Tisch mit der Wahrscheinlichkeit () und an einen unbesetzten Tisch mit der Wahrscheinlichkeit ().

2.4 Hierarchical Dirichlet Process

3 Methodik und Ergebnisse

3.1 Patentdatensatz

Der Patentdatensatz enthält ausschließlich Patente von General Motors (GM), die durch das Tochterunternehmen GM Global Technology Operations angemeldet wurden. Mit der Suchanfrage ((AN/„GM Global“ AND ((ICL/F16H$ AND APD/20040101->20121231) OR (CPC/F16H$ AND APD/20130101->20181231))) AND ISD/20040101->20181231)

diese Gänsefüßchen muss man in der Suchanfrage bis jetzt manuell ersetzen

lassen sich die 1411 Dokumente auf der Internetseite des United States Patent and Trademark Office einsehen.

3.2 Preprocessing

Das Preprocessing wurde mit dem PatVisor®, das Patentanalysewerkzeug vom Institute of Project Management and Innovation (IPMI), durchgeführt. Dazu wurde vom IPMI ein themenbezogener Synonymfilter bereitgestellt. Aus den Patenten wurde nur der Titel, der Abstract und die Claims als Text verwendet. Die Anmeldedaten wurden als Metadaten für DLDA verwendet. Die Texte wurden mit dem Patvisor lemmatisiert. Das Lemma ist die Grundform eines Wortes und

wird hier verwendet damit die Häufigkeit des Wortes bestimmt werden kann, einschließlich aller Varianten. Herausgefiltert wurden Artikel, Pronomen und Ähnliches das nur im Kontext eine Bedeutung hat und daher im Bag-of-Words Modell irrelevant ist. Außerdem wurden manuell Abkürzungen erfasst wie Continiuously Variable Transmission (CVT). Bigramme wurden in einem Fenster von fünf Worten erstellt, das über den Text rolliert. Die Worte eines Fensters wurden ohne Wiederholung permutiert. Die Wörter in einer Term-Dokument Matrix (TDM) gespeichert.

3.3 Implementierung in Python

3.3.1 Gensim

Das Topicmodeling wurde nach dem Preprocessing in vier Schritten implementiert: Wörterbuch- und Korpuserzeugung, LDA, Evaluation, Visualisierung. Gensim ist eine Python library für Textanalyse. Ein Teil des Codes wurde vom IPMI bereitgestellt. Zuerst wird aus der TDM des Preprocessings ein Wörterbuch und ein Korpus erstellt. Das Wörterbuch indiziert jedes Wort und speichert die Häufigkeit des Wortes aus dem gesamten Korpus. Der Korpus verbindet die Indizes der Wörter mit den Indizes der Dokumente und speichert die Häufigkeit der Wörter pro Dokument.

Tabelle 3.1: *Wörterbuch*

Dokument ID	Wort ID	Häufigkeit
1	5	65
1	10	20
2	11	11

Tabelle 3.2: *Korpus*

Wort ID	Wort	Häufigkeit
1923	ability	3
2049	aboard	3
1404	abort	5

Ein Thema wird für Menschen durch die wahrscheinlichsten Wörter ersichtlich. (vgl. Mimno, Wallach, Talley, Leenders und McCallum 2011, S. 265-266) Mit der

besser Mimno et al.?

Kohärenz eines Themas ist der semantische Zusammenhang zwischen diesen Wörtern gemeint. Diese Kohärenz kann man durch das gemeinsame Auftreten von Wörtern in einer Gruppe berechnen. Das u_mass Maß funktioniert nach diesem Prinzip, benannt nach der Universität von Massachusetts. Es gibt auch andere Kohärenzmaße, die eine bestimmte Anzahl an Wörtern in einem Schiebefenster betrachten. Dadurch wird ein feinerer Kontext betrachtet anstatt das gesamte Dokument. Das c_v Maß benutzt ein Schiebefenster.

c_v Maß Zitat?

Die Distanz zwischen zwei Themen ist die Unterschiedlichkeit der Wörter zweier Themen. Eine Methode der Berechnung ist der Jaccard-Koeffizient. Diese ist die Mächtigkeit der Schnittmenge dividiert durch die Mächtigkeit der Vereinigungsmenge zweier Themen.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Danach wird LDA angewandt. Die Hyperparameter Alpha und Beta werden auf der Einstellung auto belassen, um die Werte selbst zu erlernen. Die Iterationen werden auf 20.000 gesetzt und die minimale Wahrscheinlichkeit beträgt null. Dadurch wird jedes Dokument zu jedem Thema mit einer Wahrscheinlichkeit zugeordnet, auch wenn diese gering ist. LDA benötigt eine vorgegebene Anzahl an Themen. Um eine möglichst kohärente und interpretierbare Anzahl an Themen zu finden, werden für die Unigramme alle Modelle bis zu 300 Themen erstellt. Da die Kohärenz der Bigramme bereits ab 51 Themen ein Plateau erreicht wurde die Themenerstellung ab 100 abgebrochen. Mit zunehmender Themenanzahl steigt zwar auch die Kohärenz aber so viele Themen sind nicht sinnvoll interpretierbar. Der Vorteil des LDA ist schließlich die Zeit, welche benötigt wird einen Datensatz zu verstehen, zu verringern. Mit zunehmender Zahl an Themen sinkt außerdem die Distanz zwischen den Themen, was zu ähnlichen Themen führt. Eigentlich ist eine hohe Kohärenz beim u_mass negativ. Damit Kohärenz und Distanz in einem

Diagramm dargestellt werden können wurde von der Kohärenz der absolute Wert genommen.

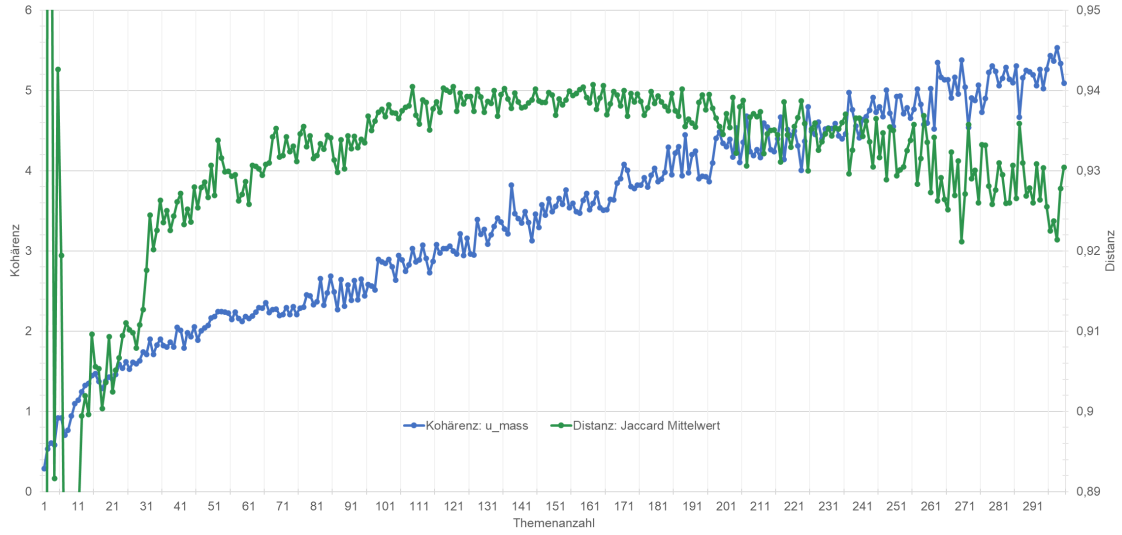

Abbildung 3.1: *Kohärenz und Distanz der Themen mit Unigrammen*

Die Kohärenz der LDA Modelle wird mit dem u_mass Maß bestimmt. Von eins wird der absolute u_mass Wert vom LDA Modell mit n Themen subtrahiert und durch den absoluten u_mass Wert des LDA Modells mit n + 1 Themen dividiert. Diese Berechnung wird für jedes Modell durchgeführt. Dadurch lässt sich die größte absolute Kohärenzsteigerung zum Vorgänger finden.

$$1 - \frac{|LDAn|}{|LDAn+1|}$$

Bei den Unigrammen sind es in Abbildung 3.1 83 Themen. Bei den Bigrammen funktioniert diese Methode nicht so gut, um ein Plateau zu finden. Sie schlägt zehn Themen vor, was zu einem sehr groben Modell führt. Wie die Abbildung 3.2 zeigt wird eine hohe Kohärenz und Distanz bei der Themenanzahl von 51 erreicht. Mit der Anzahl wurde ein deutlich granulareres Modell erstellt.

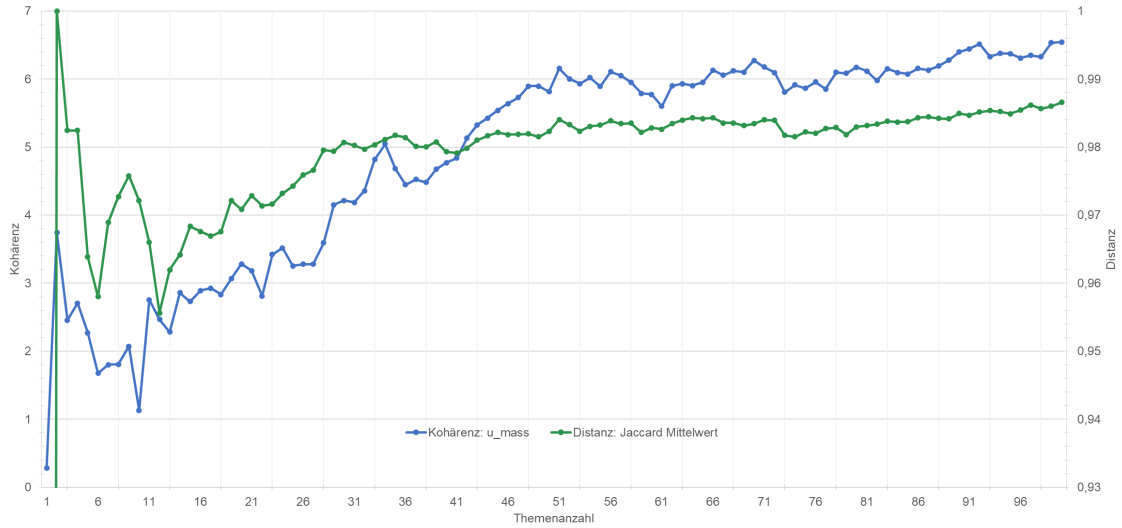

Abbildung 3.2: *Kohärenz und Distanz der Themen mit Bigrammen*

Im letzten Schritt werden die Daten als Themenliste, Dokument-Themen und Themen-Wort Matrizen gespeichert. Mit pyLDAvis wird eine interaktive multidimensional skalierte Visualisierung erstellt. Diese Visualisierung berücksichtigt die Distanz und die Größe der Themen.

3.3.2 Tomotopy

Tomotopy ist ebenfalls eine Python library für Textanalyse. Sie ist ähnlich zu Gensim aber ist besonders performant und unterstützt zusätzlich HLDA, allerdings keine Kohärenzmaße. Deshalb werden hier beide librarys verwendet, um die jeweiligen Funktionen zu nutzen.

Die Wörter der Dokumente werden in eine Liste aus Listen geladen. Für HLDA wird keine Themenanzahl benötigt aber einige Parameter aus der Tabelle 3.3 die LDA in Gensim selbst erlernt. Der HLDA verwirft die ersten 10.000 Iterationen und erstellt danach zehn Modelle mit einem Abstand von jeweils 100 Iterationen. (vgl. Griffiths, Jordan, Tenenbaum und Blei 2004, S. 6)

Name	iterations	seed	TermWeight	α	η	γ	depth	rm_top	burn_in
Unigrame	1000	100	tf-idf	0,3	0,6	0,15	3	1	10.000
Bigramme	1000	100	tf-idf	0,3	0,6	0,15	3	1	10.000

Tabelle 3.3: *HLDA Parameter*

25 random restarts to avoid local maxima, take highest posterior likelihood

In Abbildung 3.3 hat sich ein drei Ebenen tiefes Baumdiagramm als übersichtlich erwiesen, um Überthemen zu finden und Unterthemen zu clustern.

Die Term Frequency-Inverse Document Frequency (tf-idf) wird benutzt, um herauszufinden wie relevant ein Wort für ein Dokument ist in einer Menge von Dokumenten. Der wert steigt proportional zu der Frequenz des Wortes in einem Dokument und sinkt mit der Anzahl an Dokumenten in denen das Wort vorkommt.

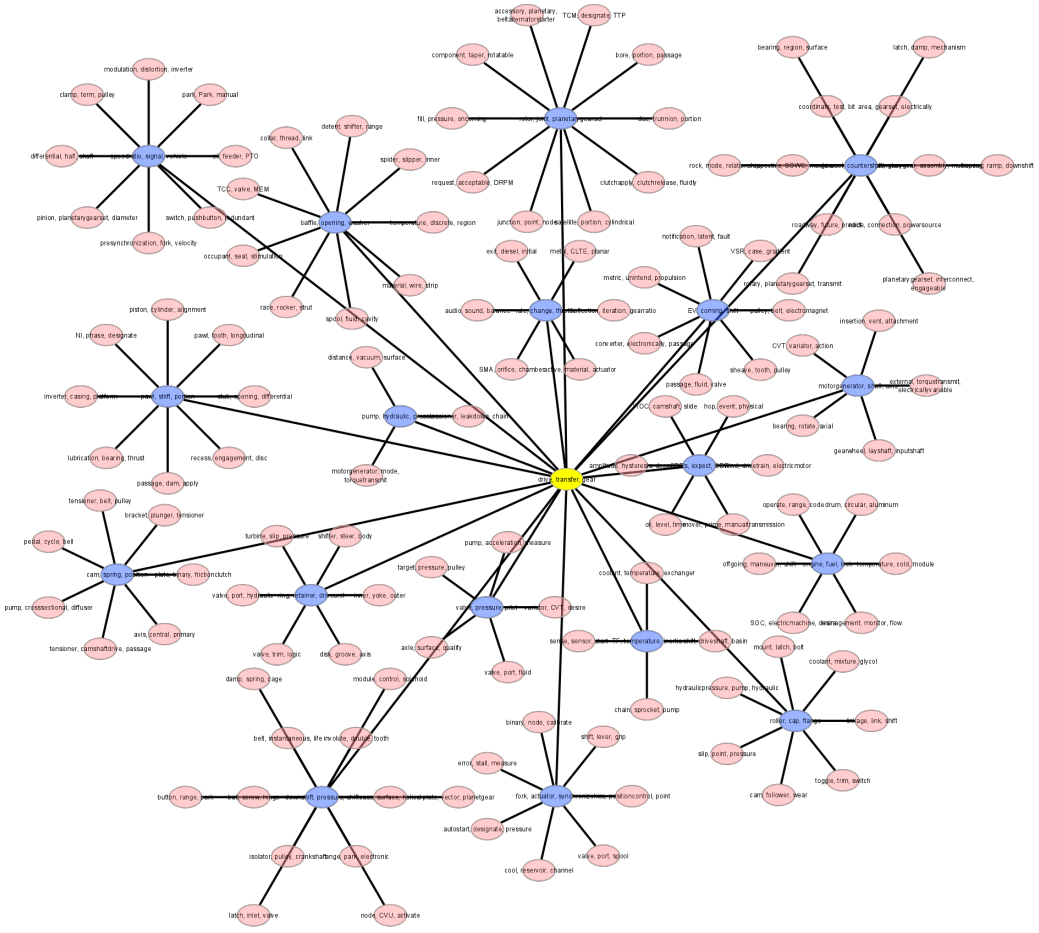

Abbildung 3.3: *HLDA Unigram Baumdiagramm*

4 Vergleich der Ergebnisse

4.1 Kennzahlen

Die Unigrammmodelle weisen eine niedrigere Distanz zueinander auf als die Bigrammmodelle. In Abbildung 4.1 sind schon Muster und Hotspots aus Unigrammthemen zu erkennen, die besonders Ähnlich oder unähnlich sind. Diese werden später geclustered. In Abbildung 4.2 gibt es ebenfalls Muster und Hotspots. Allerdings sind manche Bigrammthemen disjunkt.

4.2 Themengruppen

LDA

DLDA

HLDA

4.3 Interpretation der Themen

LDA

DLDA

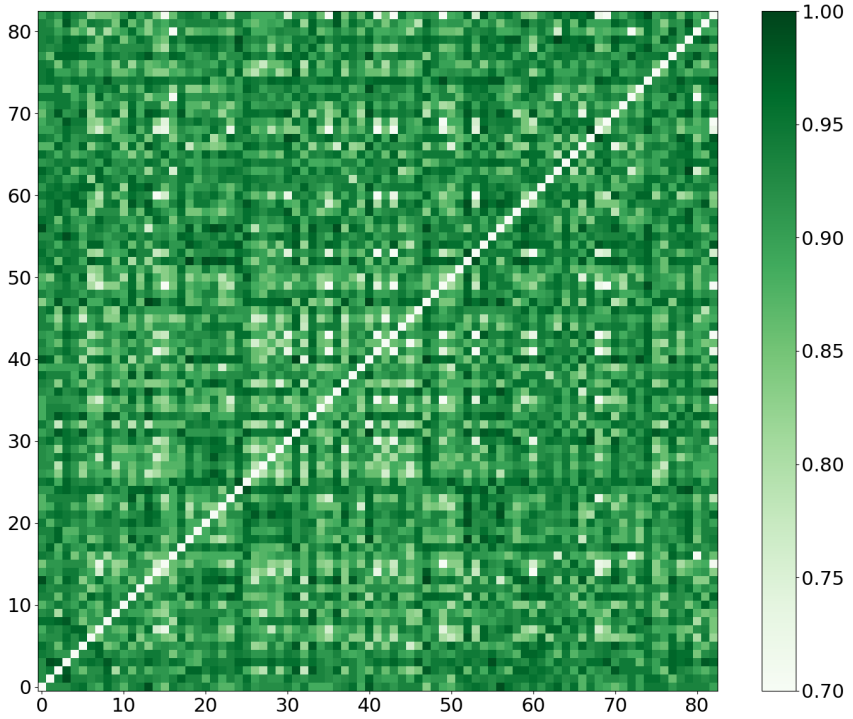

Abbildung 4.1: *Distanz zwischen den top 50 Unigrammen der Themen*

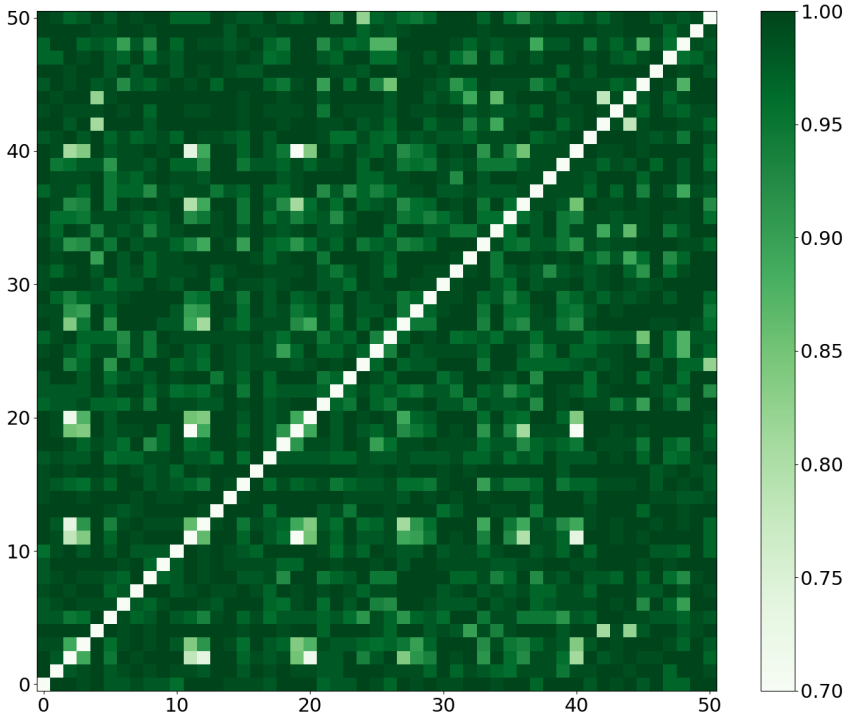

Abbildung 4.2: *Distanz zwischen den top 50 Bigrammen der Themen*

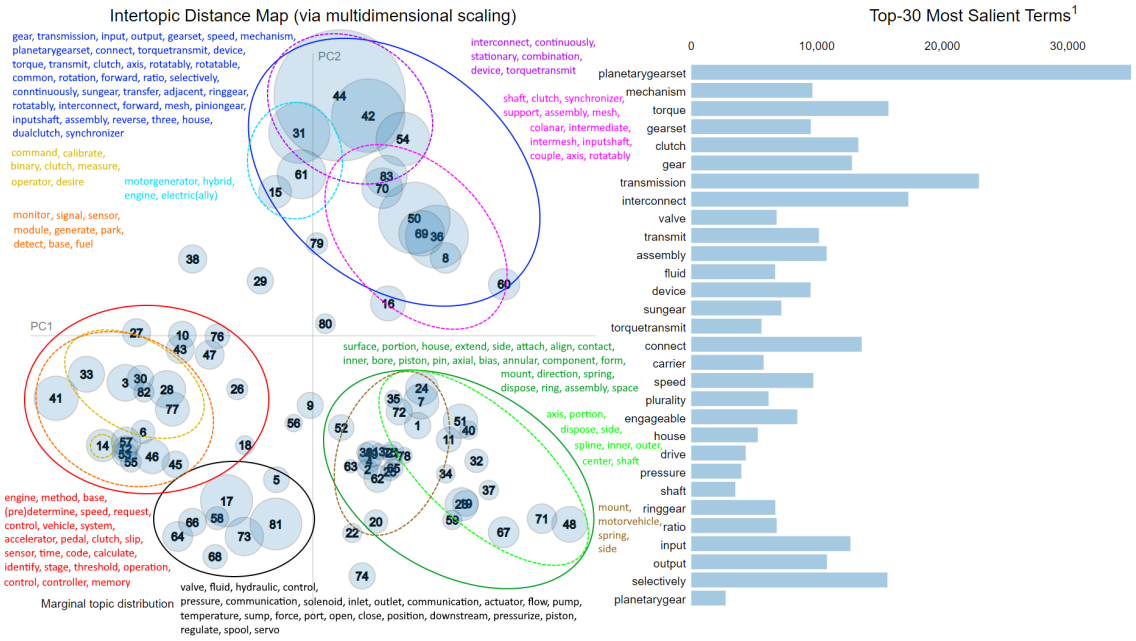

Abbildung 4.3: *Themengruppen der LDA Unigramme*

HLDA

5 Diskussion

5.1 Grenzen von TDM

GM-3-1-1 Topic 1 10,066,722 63,53% Limited slip differentials werden mit LSD abgekürzt und daher kommen die Wörter nicht so oft vor wie sie eigentlich verwendet werden.

6 Zusammenfassung und Ausblick

Anhang

1 Anhang 1

2 Anhang 2

Eidesstattliche Erklärung

Erklärung zur Abschlussarbeit

Name:	XXX XXX
Matrikel-Nr:	XXX
Fach:	XXX
Modul:	Masterarbeit

Ich erkläre, dass ich die vorliegende Abschlussarbeit mit dem Thema

Thema der Abschlussarbeit

selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe. Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich, inhaltlich oder sinngemäß entnommenen Stellen als solche den wissenschaftlichen Anforderungen entsprechend kenntlich gemacht. Die Versicherung selbstständiger Arbeit gilt auch für Zeichnungen, Skizzen oder graphische Darstellungen. Die Arbeit wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht. Mit der Abgabe der elektronischen Fassung der endgültigen Version

der Arbeit nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate überprüft und ausschließlich für Prüfungszwecke gespeichert wird.

Ort, den XX.XX.XXX

Vorname Nachname

Literaturverzeichnis

[BL06] David M. Blei und John D. Lafferty. „Dynamic Topic Models“. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, 113–120. ISBN: 1595933832. DOI: 10.1145/1143844.1143859. URL: https://doi.org/10.1145/1143844.1143859.

[Ble+03] David M. Blei, Andrew Y. Ng, Michael I. Jordan und John Lafferty. „Latent Dirichlet allocation“. In: *Journal of Machine Learning Research* (2003), S. 993–1022.

[Ble12] David M. Blei. *Probabilistic Topic Models*. http://www.cs.columbia.edu/~blei/talks/Blei_ICML_2012.pdf. Accessed: 2020-08-17. 2012.

[Gri+04] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum und David M Blei. „Hierarchical topic models and the nested chinese restaurant process“. In: *Advances in neural information processing systems*. 2004, S. 17–24.

[HBB10] Matthew D. Hoffman, David M. Blei und Francis Bach. „Online learning for latent dirichlet allocation“. In: *In NIPS*. 2010.

[Har54] Zellig S Harris. „Distributional structure“. In: *Word* 10.2-3 (1954), S. 146–162.

[McC02] Andrew Kachites McCallum. „MALLET: A Machine Learning for Language Toolkit“. http://mallet.cs.umass.edu. 2002.

[Mim+11] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders und Andrew McCallum. „Optimizing semantic coherence in topic models“. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, S. 262–272.

[RBH15] Michael Röder, Andreas Both und Alexander Hinneburg. „Exploring the Space of Topic Coherence Measures“. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: Association for Computing Machinery, 2015, 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[ŘS10] Radim Řehůřek und Petr Sojka. „Software Framework for Topic Modelling with Large Corpora“. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, Mai 2010, S. 45–50.

[SS14] Carson Sievert und Kenneth Shirley. „LDAvis: A method for visualizing and interpreting topics“. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, Juni 2014, S. 63–70. DOI: 10.3115/v1/W14-3110. URL: https://www.aclweb.org/anthology/W14-3110.