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 G. Möhrle (Universität Bremen)
	Prof. Dr. Jutta Günther (Universität Bremen)

Inhaltsverzeichnis

Abkürzungsverzeichnis

CVT Continuously Variable Transmission 20, 22

DLDA Dynamic Latent Dirichlet Allocation . . . . . II, 6, 13, 15, 24, 28

EVT Electrically Variable Transmission 20

HLDA Hierarchical Latent Dirichlet Allocation . . . . . II, IV, 15, 23, 24

LDA Latent Dirichlet Allocation . . . . . I, 4, 13, 15, 16, 20, 23, 24

pyLDAvis Python LDA Visualization 15, 16, 20

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 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 et al. 2014)

Kapitel 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 et al. 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 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.

zitat finden

beispiel geben zu LDA

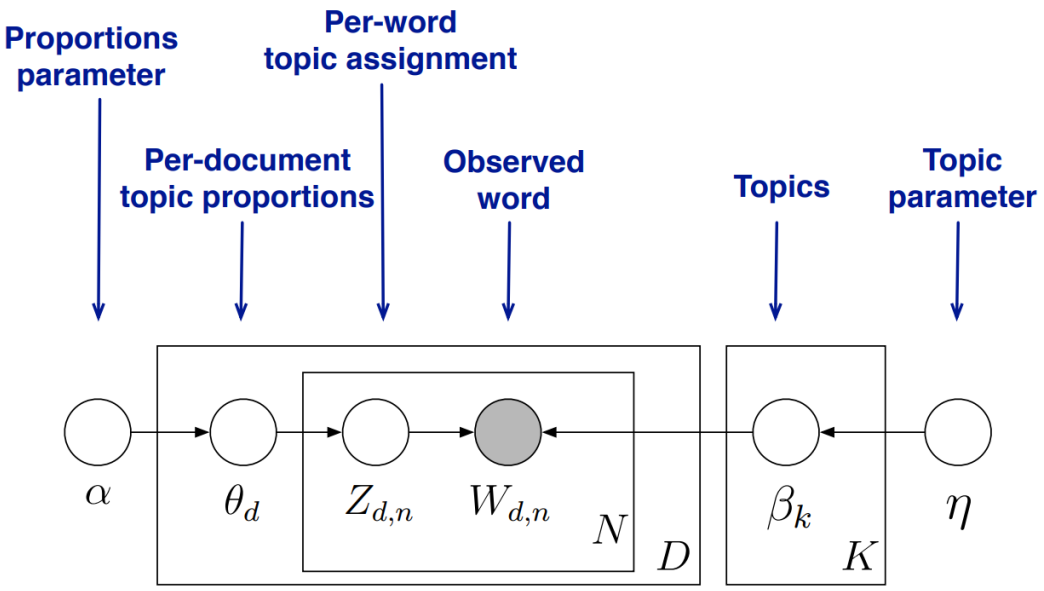

Abbildung 2.1: LDA als graphisches Modell, von (vgl. Blei 2012, S. 23)

$\mathcal{W}$ 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 hoher β Wert lässt die Themen ähnlicher aussehen. Mit diesem Algorithmus lässt sich ein Model erstellen, das jedes Wort mit Wahrscheinlichkeit zu jedem Thema zuordnet.

Dies ist die Wahrscheinlichkeit ein Dokument zu generieren, mit den Einstellungen des LDA Modells. Die Wahrscheinlichkeit ist gering aber je höher sie ist desto

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

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

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 DLDA, welche die chronologische Reihenfolge der Dokumente berücksichtigt. Dadurch ist

es möglich die Veränderung der Themenschwerpunkte über den Zeitraum zu betrachten. (vgl. Blei et al. 2006)

2.3 Hierarchical Latent Dirichlet Allocation

Hierarchisches LDA (HLDA) erweitert LDA, um eine beliebig tiefe Hierarchie aus Unterthemen. (vgl. Griffiths et al. 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 ().

Kapitel 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 Durchführung des Topic Modeling

3.3.1 LDA

Das Topic Modeling 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 et al. 2011, S. 265-266) Mit der 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 wie das c_v Maß, die eine bestimmte Anzahl an Wörtern in einem Schiebefenster betrachten. Dadurch wird ein feinerer Kontext betrachtet anstatt das gesamte Dokument. Allerdings wird hier u_mass verwendet, weil es aufgrund des fehlenden Schiebefensters auch bei Bigrammen funktioniert.

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.

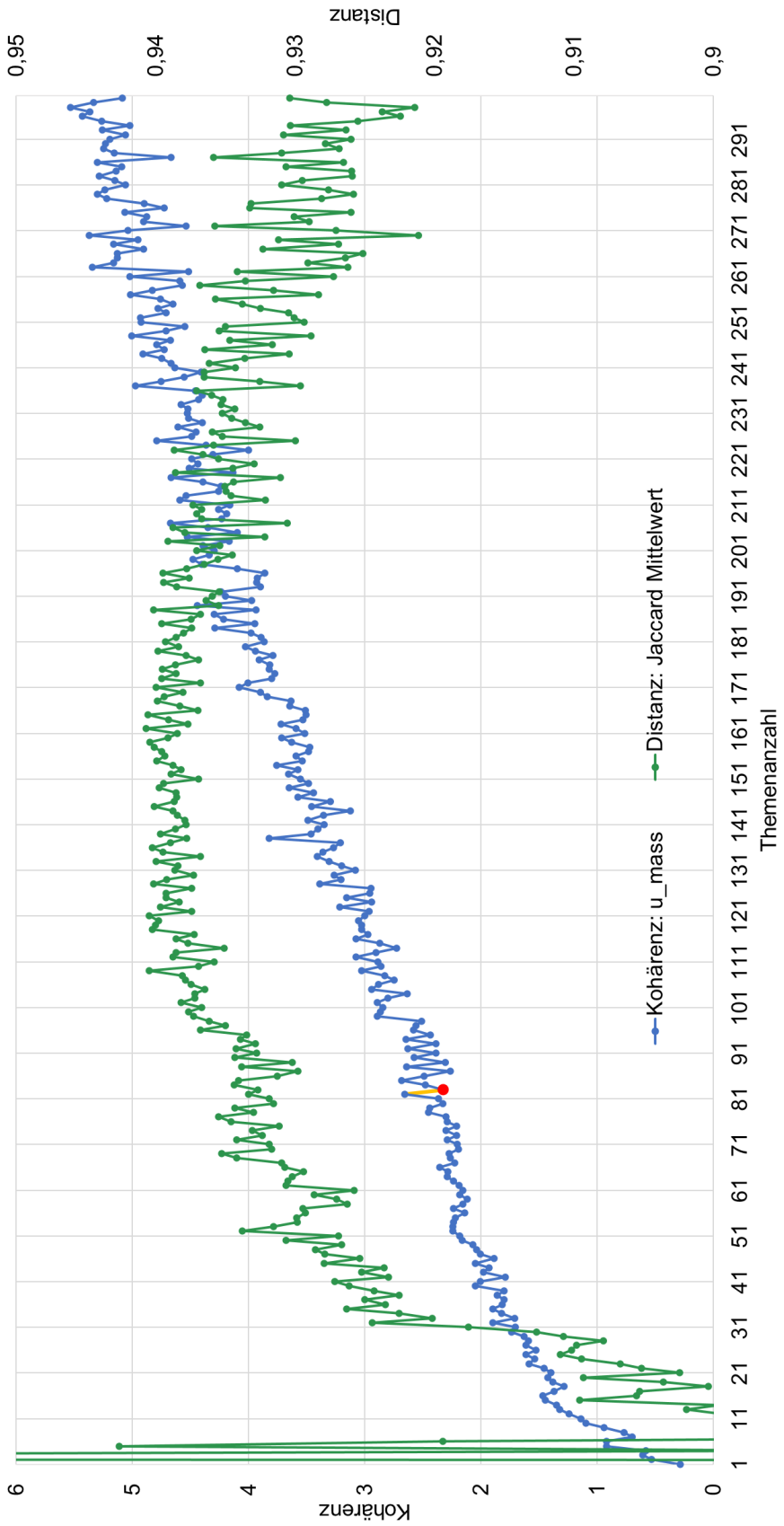

Abbildung 3.1: Veränderung der Kohärenz- und Distanzwerte der Themen

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.

Tabelle 3.3: HLDA Parameter

Name	iter.	seed	TW	α	η	γ	depth	rm_top	burn_in
Unigramme	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

3.3.2 Hierarchisches LDA

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 et al. 2004, S. 6)

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 stark ein Wort zu einem Dokument gehört in einer Menge von Dokumenten. (Luhn 1957) (Jones 1972) Der wert steigt mit mit der Frequenz des Wortes in einem Dokument und sinkt mit der Anzahl an Dokumenten in denen das Wort vorkommt.

3.3.3 Dynamisches LDA

Für das DLDA wurden die 1410 Patente des Datensatzes in die 14 Zeitabschnitte von 2004 bis 2017 aufgeteilt. Jeder Zeitabschnitt ist ein Jahr lang und enthält die Patente deren APD in jenes Jahr fällt. Jeder Zeitabschnitt enthält 1 bis 173 Patente. Die Jahre 2004 und 2017 enthalten mit 1 und 13 Patenten die wenigsten im Datensatz. Damit die Ergebnisse des DLDA denen des LDA vergleichbar sind wurden für die Uni- und Bigramme wieder Modelle mit 83 und 51 Themen erstellt.

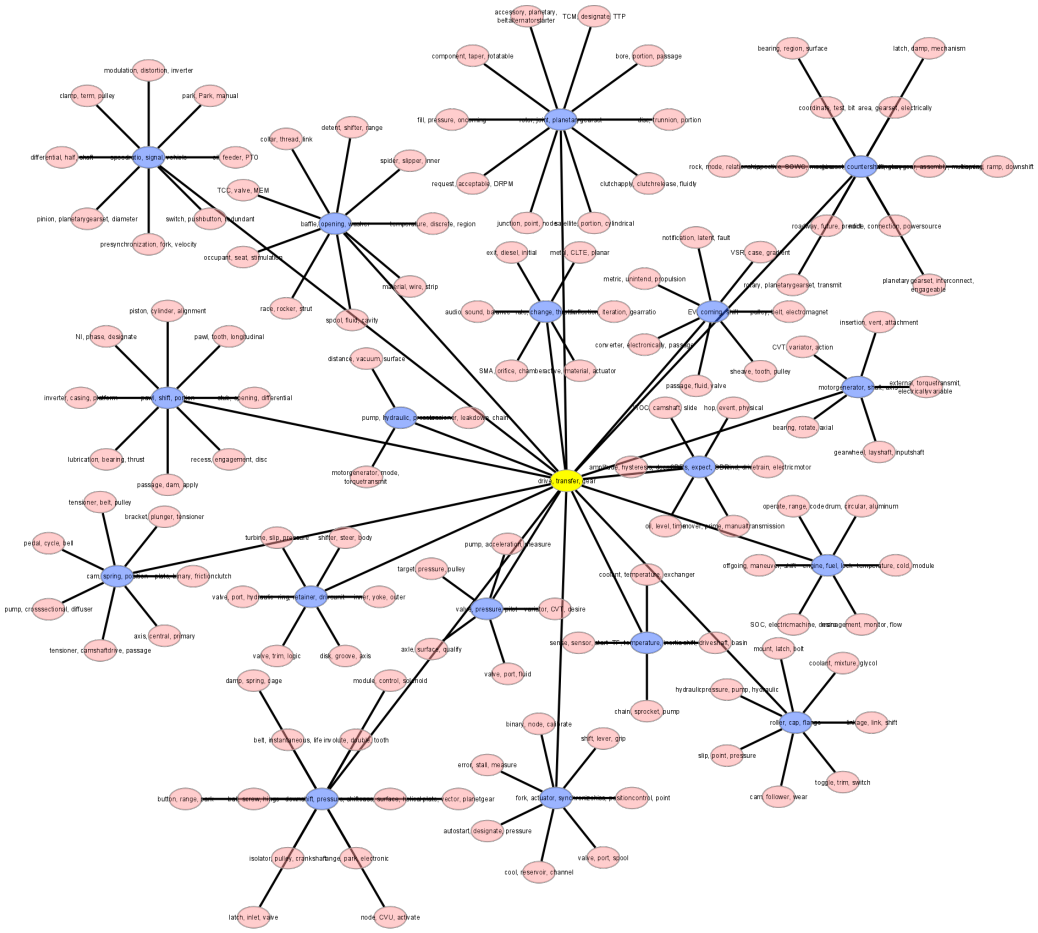

Abbildung 3.3: HLDA Unigram Baumdiagramm

Kapitel 4

Analyse der Ergebnisse

Die Analyse der Ergebnisse erfolgt in vier Schritten. Die Ergebnisse der drei Algorithmen werden einzeln ausgewertet und anhand markanter Beispiele erläutert. Mit dem LDA werden die latenten Themen und Themencluster des Patentdatensatzes benannt und eingegrenzt. Mit dem HLDA werden die gefundenen Themencluster bestätigt. Der DLDA wird die Entwicklung der relevantesten Themen über die Zeit beschreiben. Abschließend werden die Ergebnisse anhand von Kennzahlen verglichen.

4.1 Analyse der Ergebnisse des LDA

Zuerst werden die vier Schritte des qualitativen Verfahrens zur Benennung und Gruppierung der Themen aufgelistet. Danach werden die Schritte an Beispielen erklärt und wie das Verfahren durch quantitative Daten vom LDA unterstützt wird. Dies wird mit Abbildungen aus Python LDA Visualization (pyLDAvis) verdeutlicht. Die interaktive Version von pyLDAvis befindet sich im digitalen Anhang. Danach werden die Themen benannt und gruppiert.

Das Verfahren zur Benennung und Gruppierung der Themen besteht aus vier Schritten:

1. In pyLDAvis Themencluster auswählen
2. Die relevantesten Terme der Themen nacheinander auswählen
3. Themenradien beobachten und Terme die in fast allen Themen des ausgewählten Clusters häufig vorkommen aber außerhalb nur selten vorkommen benennen den Cluster

4. Bei diesem Vorgehen werden häufig Subcluster entdeckt, die ebenfalls nach dieser Methode benannt werden

Die Themen welche durch LDA gefunden wurden, werden mit Hilfe von pyLDAvis benannt und visualisiert (vgl. Sievert et al. 2014, S. 63). pyLDAvis ist ein Programm, das die Themen multidimensional skaliert und interaktiv darstellt. In Abbildung 4.1 werden links die Themenradien nach Termanzahl skaliert. Die Tabelle rechts zeigt die Termhäufigkeiten im ausgewählten Thema Nummer 50 und im gesamten Korpus.

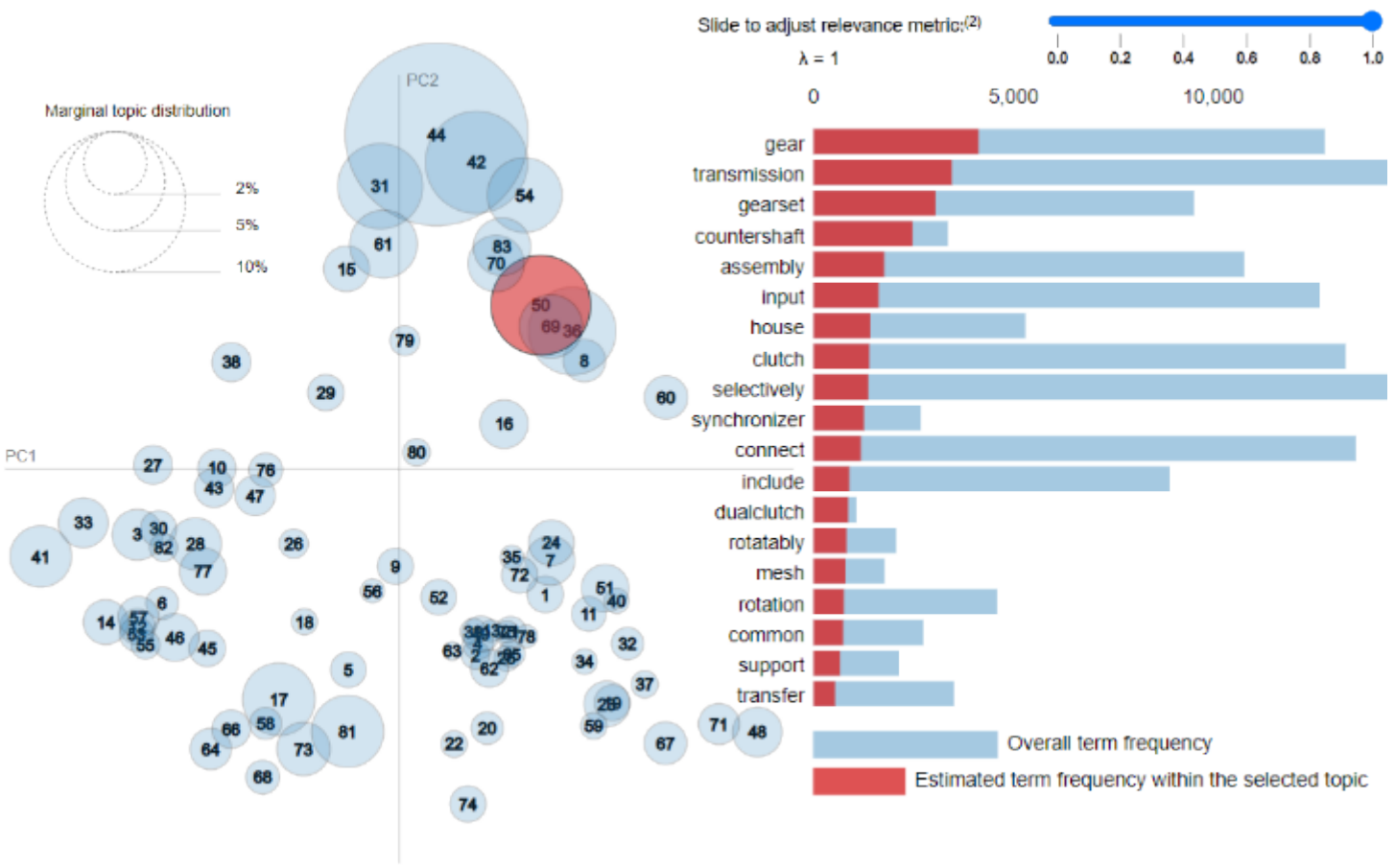

Abbildung 4.1: Interthematische Distanz Karte erstellt mit multidimensionaler Skalierung und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 1,0$

In Abbildung 4.2 wurde das λ von $1,0$ auf $0,6$ herabgesetzt. Dadurch werden die Terme im ausgewählten Thema absteigend nach Relevanz sortiert. Der Parameter λ soll mit dem Wert $0,6$ den Anwender die korrektesten Themen finden lassen (vgl. Sievert et al. 2014, S. 66-68). Rechts wurde der Term *countershaft* ausgewählt. Dadurch werden die Themenradien, abhängig von der Verteilung des ausgewählten Terms, skaliert. Die Themen 50, 20 und 83 enthalten die meisten der *countershaft* Terme und sind teil des pink eingekreisten Subcluster des *transmission* Clusters, aus Abbildung 4.3. Deshalb sind sie als *countershaft* Themen zu werten.

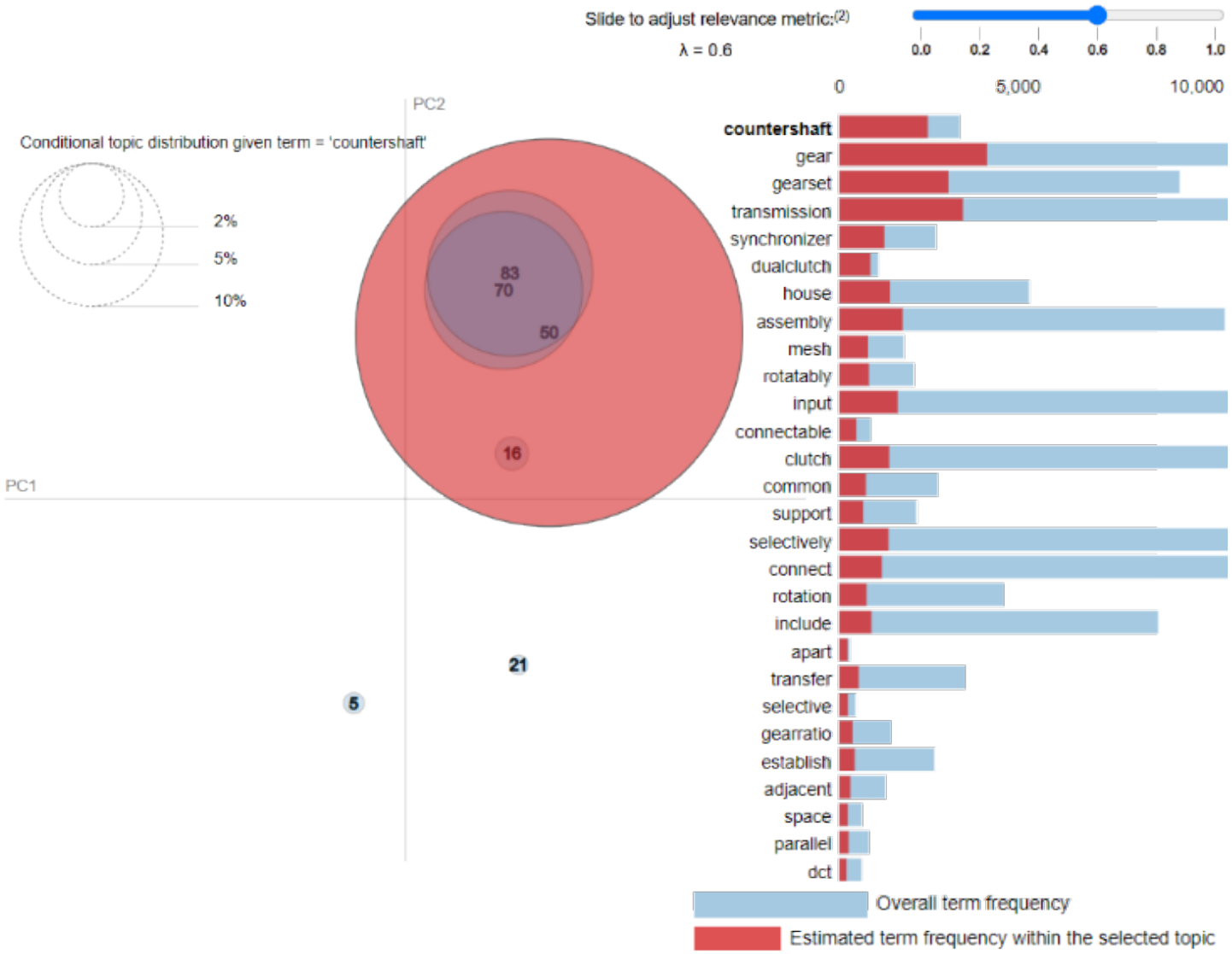

Abbildung 4.2: Interthematische Distanz Karte erstellt mit multidimensionaler Skalierung mit einer Themenverteilung welche von dem Term countershaft abhängt und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 0,6$

Außerdem kommen die äquivalenten Terme *dualclutch*, *dct* und *automatic-transmission* besonders häufig in den Themen 50, 83, 69 und 10 vor. Das ist ein weiterer Teil des Subclusters *transmission*. Des weiteren sind die Terme *synchronizer* und *mesh* beide in den Themen 50, 70 und benachbarten Themen häufig zu finden. Das deutet auf das synchronisieren der *shafts* hin. Durch dieses qualitative Verfahren werden die Themen benannt und Cluster gebildet. Doch pyLDAvis reicht allein nicht immer aus. Thema 77 deutet mit den Termen *clutch* und *slip* auf eine Slipper clutch hin aber warum befindet es sich dann in dem *method* Cluster? Für genauere Einblicke in Themen wurde mit LDA eine Patent-Themen-Matrix erstellt. Das Patent 9,989,146 passt am besten zu Thema 77. Es beschreibt eine Methode, welche den optimalen Druck einer *clutch* in einem Continuously Variable Transmission (CVT) erlernt, damit die sie einen *pulley slip* verhindern kann. Daher kommt in diesem Thema der Term *pressure* ohne *fluid* oder *hydraulik* vor.

Der blaue Cluster besteht hauptsächlich aus Themen zu *transmission* Komponenten. Fast alle Themen des blauen Clusters enthalten die Terme *transmission*, *sungear* und *ringgear*. Der lila Subcluster enthält das größte Thema des Datensatzes Nummer 44. Die Themen beinhalten das *planetarygearset*, die *multispeedtransmission* und werden zum *torquetransmit* genutzt. Der türkise Cluster zeigt das Thema Electrically Variable Transmission (EVT) mit einem *motorgenerator* für Hybridautos. Der pinke Cluster beinhaltet die *dualclutch*, die *automatic-transmission*, den *countershaft*, den *inputshaft* und die *synchronizer* welche die *shafts* verbinden (*mesh*). Dies wird im Patent 8,240,224 beschrieben.

Der Rote Cluster enthält besonders viele Terme wie *method*, *command* und *request*. Der Cluster besteht daher aus Steuerungs- und Regelungsthemen von Fahrzeugen.

Im gelben Subcluster geht es hauptsächlich um die elektronische Steuerung von Kupplungen, dem Motor und dem CVT. Er ist eine Teilmenge des orangen Subclusters, weil in dem gelben Subcluster viel häufiger der Term *command* vorkommt und *montior* gleichmäßiger im gesamten orangen Subcluster einschließlich des gelben Subclusters verteilt ist. Das Thema Nummer 14 ist eine Ausnahme, weil

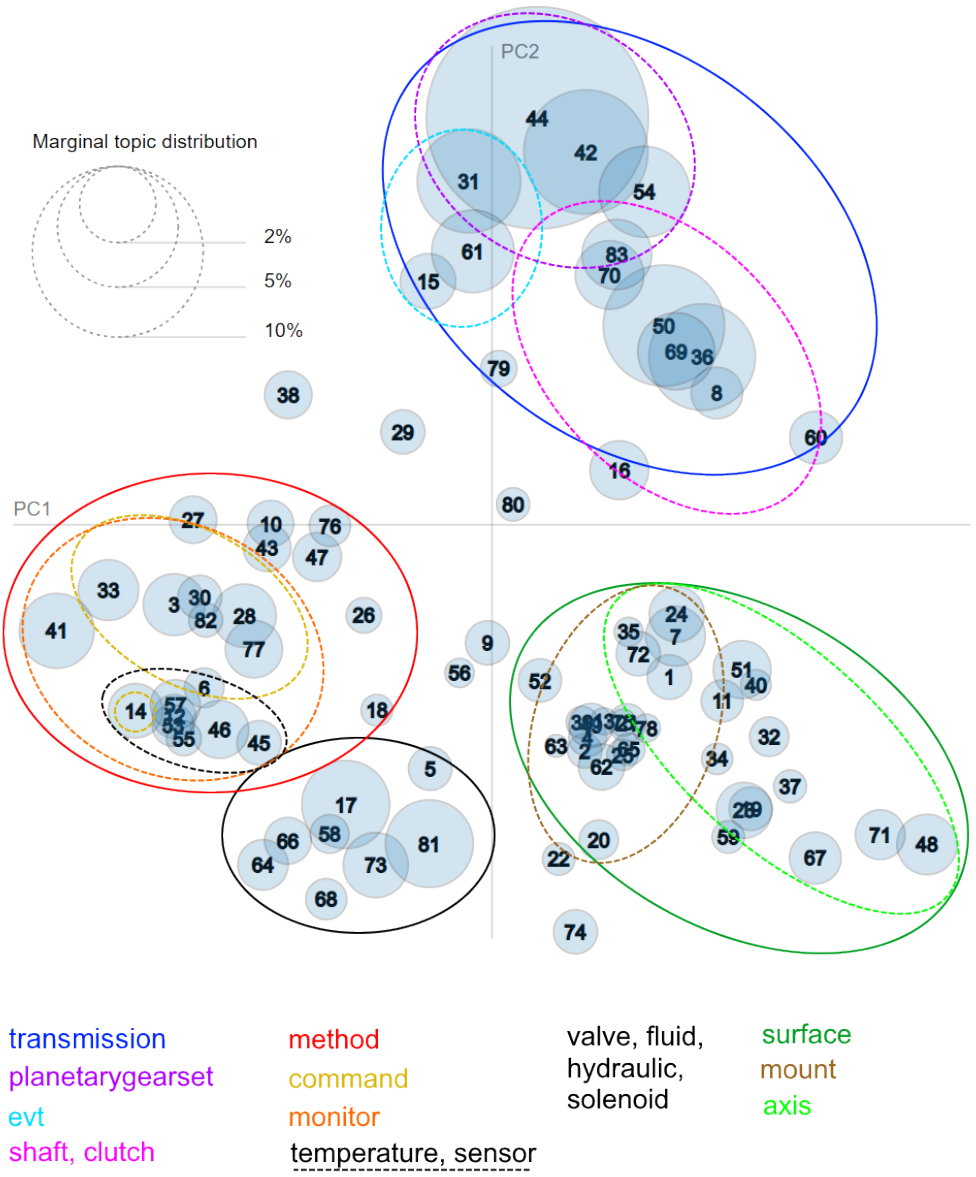

Abbildung 4.3: Themengruppen der LDA Unigramme multidimensionale Skalierung

es fast gleich viele Terme von *command* und *monitor* für die *hydraulic pressure* enthält. Deshalb ist es extra gelb umkreist. Die *frictionclutch* ist hauptsächlich dem Thema 77 zuzuordnen. Der Term *slip* kommt in Verbindung mit der *clutch* in den umliegenden Themen häufig vor. Auch die *dogclutch* ist in dem gelben Subcluster zu finden obwohl sie hauptsächlich im Thema 5 vorkommt. Im Thema 28 geht es um die Steuerung der *binary clutch* (9,061,675), die im Thema 50 schon *dualclutch* genannt wurde. Das Thema 3 umfasst das besagte CVT. Fast alle Themen des roten Clusters sind eng verbunden mit dem *engine*, besonders die Themen 41 und 33.

Der schwarze Subcluster enthält viele Themen zur Beobachtung (*monitor*, *sensor*) der *temperature* und der *pressure*. In diesem schwarzen Subcluster geht es hauptsächlich um Themen die mit Flüssigkeit in Verbindung stehen. Es geht um die *hydraulic pressure* (14), die *hydraulic pump* (45) und die *temperature* des *coolant* im Kühlkreislauf der *electricmachine* (8,167,773), der *engine* und des *radiator* (57). Durch die Steuerung des Kühlkreislauf können Komponenten wie die *transmission* auf Betriebstemperatur gebracht werden (10,161,501). Zu dem Thema 12 passt am besten das Patent 9,404,403, es beschreibt eine Methode um das Öllevel zu beobachten (*monitor*). Auf Grund der vielen Flüssigkeitsthemen befindet sich der schwarze Subcluster in der Nähe des schwarzen Hauptclusters.

Der schwarze Hauptcluster enthält fast alle Themen die in Verbindung mit Flüssigkeit stehen. Er teilt die Häufigkeit des von *control* mit dem roten Cluster aber unterscheidet sich durch die Verwendung von die Terme *communication* und *communicate* die sonst nur selten auftreten. Besonders häufig ist die Kombination *solenoid*, *hydraulic*, *valve* und *fluid*. Das Thema 17 beschreibt in mehreren Patenten (8,820,185 , 8,382,639) die Steuerung einer *dualclutch* mithilfe von *hydraulic* und *solenoids*. Mit einem Elekrtomagnet (*solenoids*) wird ein Verschluss aus der *valve* gezogen. Dieser Verschluss wird nach dem nach dem Ausschalten des *solenoids* mit einer Feder zurück in die *valve* geschoben. Mit einem *solenoid* kann auch Druck erzeugt werden. Daher kommt der Term *pressure* ebenfalls häufig vor.

Diese Methode findet Verwendung in Thema 68, dort wird beschrieben wie eine *actuator fork* kontrolliert (*control*) werden kann (9,605,755).

Der dunkel grüne Cluster besteht aus sehr vielen kleinen Themen die nicht gänzlich durch thematische Nähe gruppiert wurden. Ein gemeinsamer Term ist *bias* was auf Zahnräder hindeutet. Das ist leider unspezifisch. Der hellgrüne Subcluster hingegen hat zwei beschreibende Terme. *shaft*, *axis* und *house* zeigen eine Verwandschaft mit dem pinken Subcluster, der ebenfalls *shaft house* enthält. Das *house* deutet auf das gear housing hin was auch zu den Termen *surface*, *side* und *body* passt die den gesamten dunkelgrünen Cluster bilden.

4.2 Analyse der Ergebnisse des HLDA

Die Ergebnisse des HLDA werden im Vergleich mit den LDA Ergebnissen analysiert. Zwei Subcluster des HLDA Baums werden mit den Ergebnissen des LDA verglichen und analysiert. Der ganze HLDA Baum ist zu groß um leserlich abgebildet zu werden. Er kann im digitalen Anhang mit einem Programm geöffnet werden, das graphml unterstützt, zum Beispiel Cytoscape.

Der *engine* Subcluster aus Abbildung 4.4 passt gut zu dem orangefarbenen Subcluster aus der LDA Abbildung 4.3. Die Terme *engine* und *fuel* passen zu dem Motorthema 41. Die Subthema *monitor* passt genau zu dem orangefarbenen Subcluster und *temperature* gehört mit *electricmachine* zum Thema 57. Dadurch werden die mit LDA benannten Themen bestätigt. Die Subthemen *operate*, *shift* und *drum* sind im LDA Modell in dieser Form nicht in der Nähe aufzufinden. Allerdings passen sie thematisch zu den anderen.

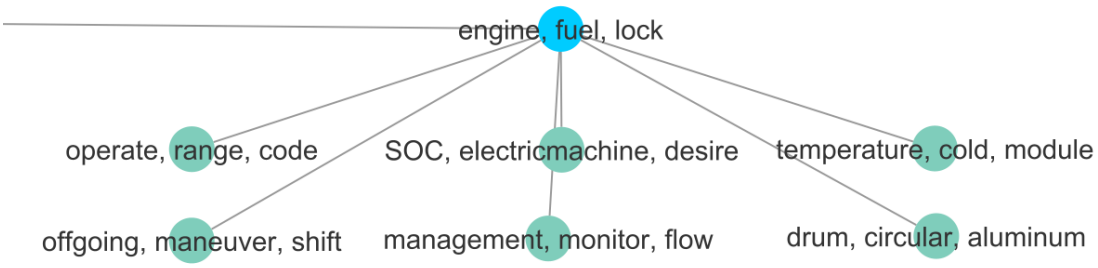

Abbildung 4.4: Der *engine* Ausschnitt des HLDA Baums

Der *fork* Subcluster aus Abbildung 4.5 passt genau zu dem schwarzen Cluster aus Abbildung 4.3. Dort beschreibt das Thema 68 ebenfalls eine *synchronizer actuator fork* aus dem Patent 9,605,755. Der HLDA Subcluster enthält außerdem die Subthemen *valve*, *spool*, *cool*, *pressure*, *calibrate*, *position* und *control*. Diese kommen auch alle im LDA Cluster vor und bestätigen erneut die Ergebnisse. Die *spool* ist eine Spule und daher ein Bestandteil des Elektromagneten *solenoid*.

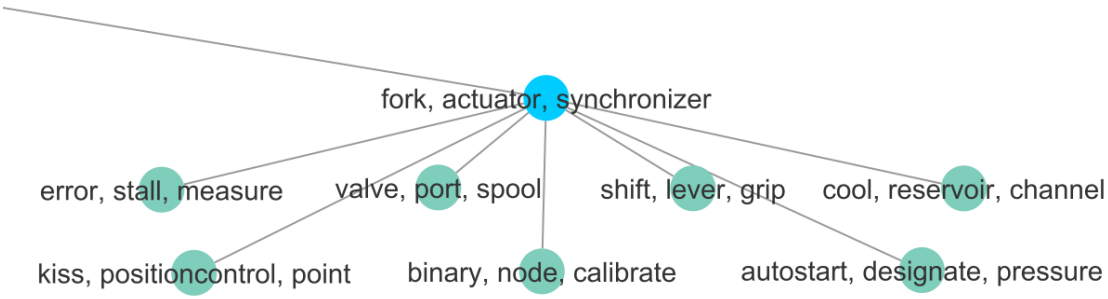

Abbildung 4.5: Der *fork* Ausschnitt des HLDA Baums

4.3 Analyse der Ergebnisse des DLDA

Die Ergebnisse des DLDA werden zuerst eingeordnet und die Wahl der Terme, welche die ausgewählten Cluster und Themen repräsentieren, erläutert. Dann werden die Trends der Terme festgestellt.

Der DLDA wurde mit den Ergebnissen des LDA und den Anmeldedaten der Patente gespeist. Daher sind die Ergebnisse des DLDA direkt vergleichbar mit denen des LDA. Die Terme wurden so gewählt, dass sie die Cluster möglichst genau beschreiben. Außerdem sollen sie relativ selten in anderen Clustern vorkommen, damit ihr Trend nicht verfälscht wird. Diese Bedingungen der Clusterbenennung wurden in 4.1 berücksichtigt.

4.4 Vergleich der Ergebnisse anhand von Kennzahlen

Die Unigrammmodelle weisen eine niedrigere Distanz zueinander auf als die Bigrammmodelle. Das liegt daran, das es deutlich mehr Bigramme gibt und diese auch als unterschiedlich gewertet werden wenn sie sich nur teilweise unterscheiden. Ein Beispiel wäre gear gearset und gear gear. In 3.1 wird ersichtlich das sich die Kohärenz mit steigender Themenzahl verbessert bis sie gleich der Anzahl an

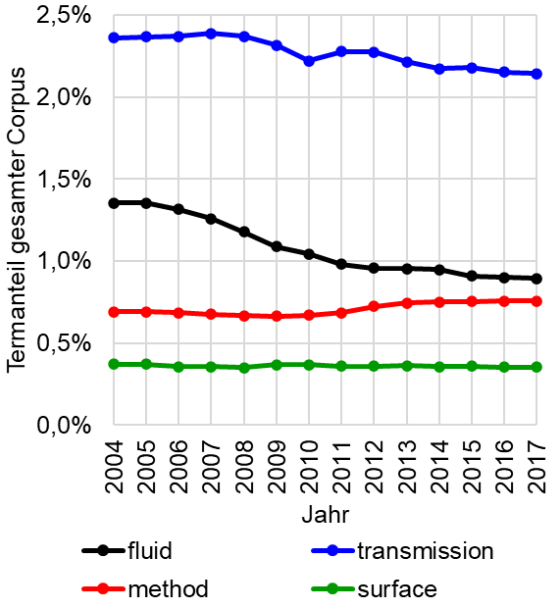

Abbildung 4.6: Trends der Terme die ihren Cluster am besten beschreiben

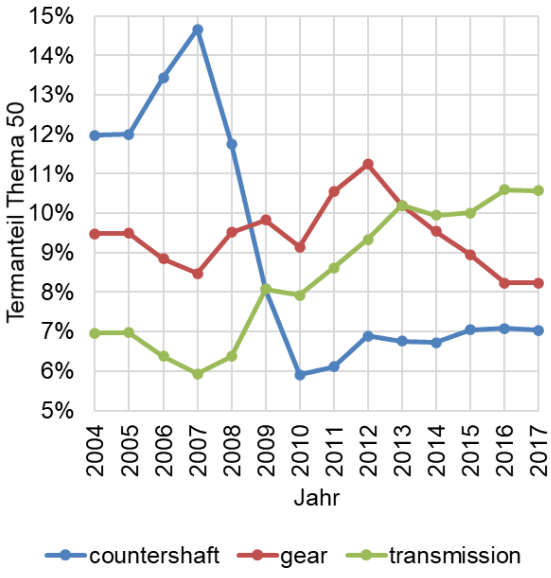

Abbildung 4.7: Trends der 3 Terme die das Thema 50 am besten beschreiben

Tabelle 4.1: Kohärenzen

Modell	Unigramm	Bigramm
LDA	-2,03	-5,51
HLDA	-4,30	-5,90

Wörtern im Datensatz ist. Die Distanz hingegen sinkt nachdem sie einen Hochpunkt erreicht. In Abbildung 4.1 sind 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, wodurch sie eine Distanz von 1 haben.

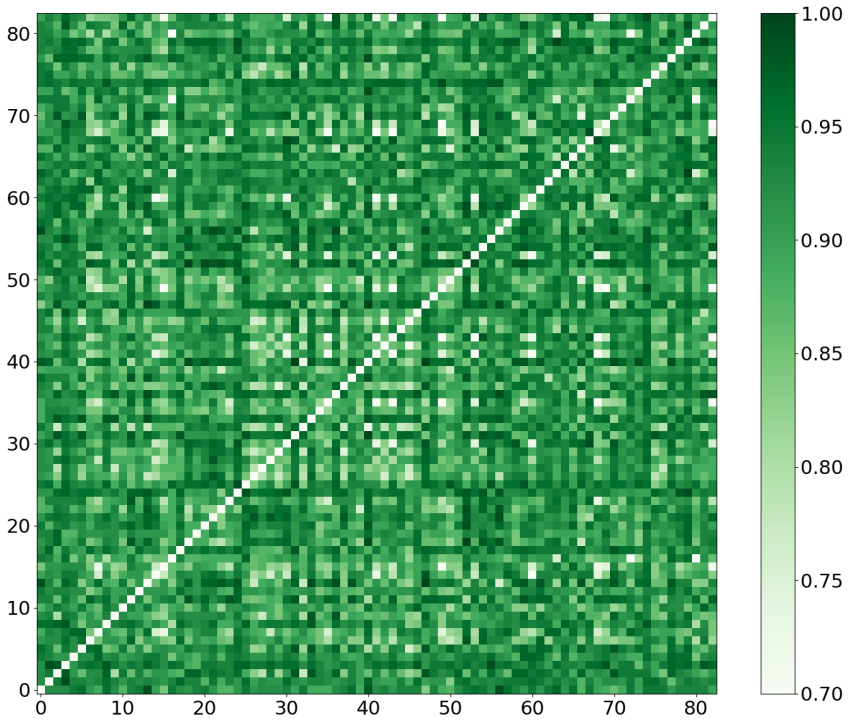

Abbildung 4.8: Distanz zwischen den top 50 Unigrammen der Themen

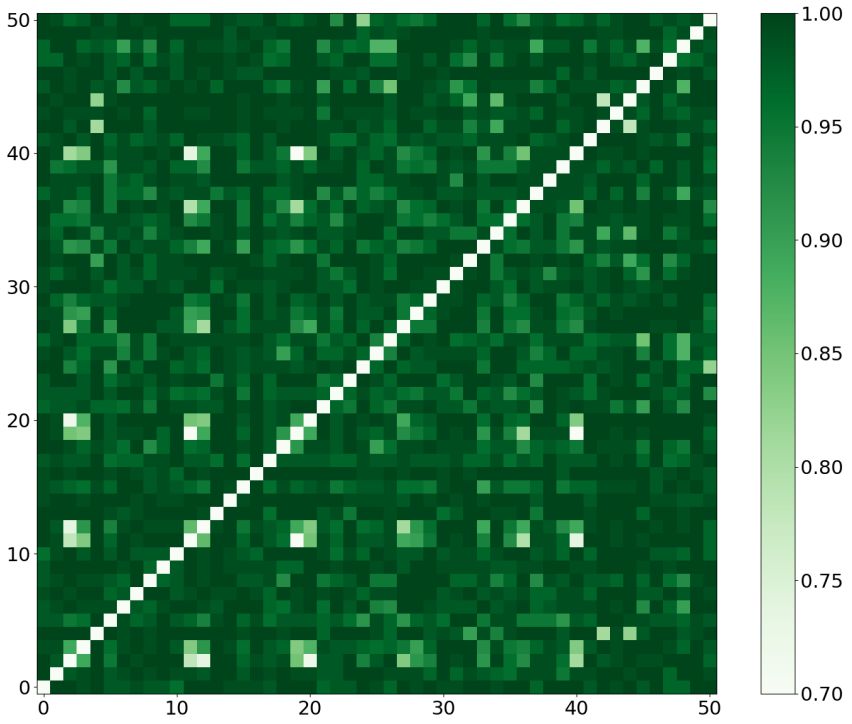

Abbildung 4.9: Distanz zwischen den top 50 Bigrammen der Themen

Kapitel 5

Diskussion

5.1

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.

Es wäre gut die Trends der Cluster und Themen direkt mit DLDA zu messen und nicht nur die Trends der Terme. Die gewählten Terme beschreiben die Cluster und Themen zwar gut und grenzen sie voneinander ab aber die Trends direkt zu messen wäre eine genauere Methode.

Kapitel 6

Zusammenfassung und Ausblick

Anhang A

Anhang

A.1 Anhang 1

A.2 Anhang 2

Kapitel 7

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.

[Jon72] Karen Sparck Jones. „A statistical interpretation of term specificity and its application in retrieval“. In: *Journal of documentation* (1972).

[Luh57] Hans Peter Luhn. „A statistical approach to mechanized encoding and searching of literary information“. In: *IBM Journal of research and development* 1.4 (1957), S. 309–317.

[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*. 2014, S. 63–70.