



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: 16.11.2020
Studiengang: Wirtschaftsinformatik, Bachelor of Science
Eingereicht bei: Prof. Dr. Martin G. Möhrle (Universität Bremen)
Prof. Dr. Jutta Günther (Universität Bremen)

Inhaltsverzeichnis

Abkürzungsverzeichnis	III
Abbildungsverzeichnis	IV
Tabellenverzeichnis	V
1 Einleitung	1
1.1 Thema	1
1.2 Motivation und Zielsetzung	1
1.3 Methodisches Vorgehen	2
2 Begriffliche Grundlagen	4
2.1 Latent Dirichlet Allocation	4
2.2 Dynamic Latent Dirichlet Allocation	7
2.3 Hierarchical Latent Dirichlet Allocation	7
3 Methodik und Ergebnisse	8
3.1 Patentdatensatz	8
3.2 Preprocessing	8
3.3 Durchführung des Topic Modeling	9
3.3.1 LDA	9
3.3.2 Hierarchisches LDA	13
3.3.3 Dynamisches LDA	13
4 Analyse der Ergebnisse	16
4.1 Analyse der Ergebnisse des Latent Dirichlet Allocation (LDA) . . .	16

4.2	Analyse der Ergebnisse des Hierarchical Latent Dirichlet Allocation (HLDA)	25
4.3	Analyse der Ergebnisse des Dynamic Latent Dirichlet Allocation (DLDA)	26
4.4	Diskussion der Ergebnisse	29
5	Zusammenfassung und Ausblick	31
5.1	Zusammenfassung	31
5.2	Ausblick	31
6	Anhang	i
7	Eidesstattliche Erklärung	iii
	Literaturverzeichnis	iv

Abkürzungsverzeichnis

CVT	Continuously Variable Transmission	21, 23–25
DCT	Dual Clutch Transmission	21, 25
DLDA	Dynamic Latent Dirichlet Allocation II, 7, 13, 15, 16, 26, 27, 29, 31, 32	
EVT	Electrically Variable Transmission	23, 25
GM	General Motors	25, 31
HLDA	Hierarchical Latent Dirichlet Allocation	II, IV, 16, 25, 26, 31
LDA	Latent Dirichlet Allocation	I, 4, 15–17, 21, 25–27, 29, 31
pyLDavis	Python LDA Visualization	16, 17, 21, 31

Abbildungsverzeichnis

2.1	LDA als graphisches Modell, von (vgl. Blei 2012, S. 23)	5
2.2	LDA als graphisches Modell, von (vgl. Blei 2012, S. 25)	6
3.1	Veränderung der Kohärenz- und Distanzwerte der Themen	11
3.2	Kohärenz und Distanz der Themen mit Bigrammen	12
3.3	HLDA Unigram Baumdiagramm	14
4.1	Interthematische Distanz Karte erstellt mit multidimensionaler Ska- lierung und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 1,0$	18
4.2	Interthematische Distanz Karte erstellt mit multidimensionaler Ska- lierung mit einer Themenverteilung welche von dem Term counters- haft abhängt und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 0,6$	20
4.3	Interthematische Distanz Karte erstellt mit multidimensionaler Ska- lierung und qualitativer Methode zur Clusterung der Themen	22
4.4	Der <i>engine</i> Ausschnitt des HLDA Baums	26
4.5	Der <i>fork</i> Ausschnitt des HLDA Baums	26
4.6	Trends der Terme die ihren Cluster am besten beschreiben	28
4.7	Trends der 3 Terme die das Thema 50 am besten beschreiben	28
4.8	Interthematische Distanz Karte der Bigramme, erstellt mit multidi- mensionaler Skalierung bei einem von $\lambda = 1,0$	30
6.1	Distanz zwischen den top 50 Unigrammen der Themen	i
6.2	Distanz zwischen den top 50 Bigrammen der Themen	ii

Tabellenverzeichnis

3.1	Wörterbuch	9
3.2	Korpus	9
3.3	HLDA Parameter	13
4.1	Legende zu Abbildung 4.3	22
4.2	Kohärenzen des Unigramm- und Bigrammmodells, niedriger ist besser	30

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

zitat
fin-
den

beispiel
ge-
ben
zu
LDA

es aufwändig ist diese Themen zu interpretieren und die Distanz bei zu hoher Themenzahl sinkt.

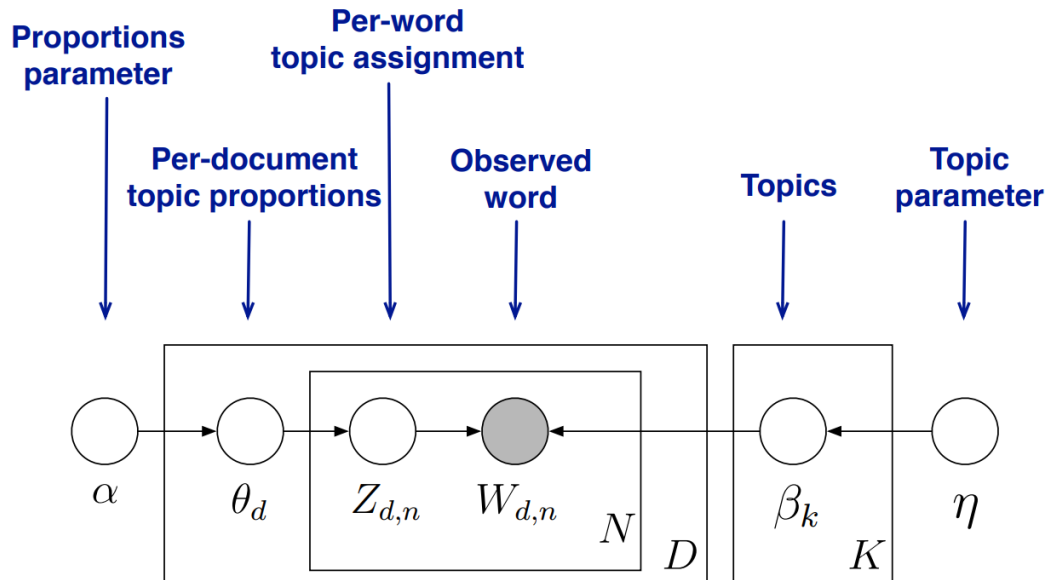


Abbildung 2.1: LDA als graphisches Modell, von (vgl. Blei 2012, S. 23)

W ist das Wort aus N Wörtern eines Dokuments i . Dieses Dokument i ist eines aus allen Dokumenten M . Alle folgenden Parameter sind latent. 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.

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

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.

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 $\frac{1}{2}$ und an einen unbesetzten Tisch mit der Wahrscheinlichkeit $\frac{1}{2}$.

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 Continuously 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. Dieser 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 jedem Dokument und jedem Term eine Zugehörigkeitswahrscheinlichkeit für jedes Thema zugeordnet, auch wenn die Wahrscheinlichkeit 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.

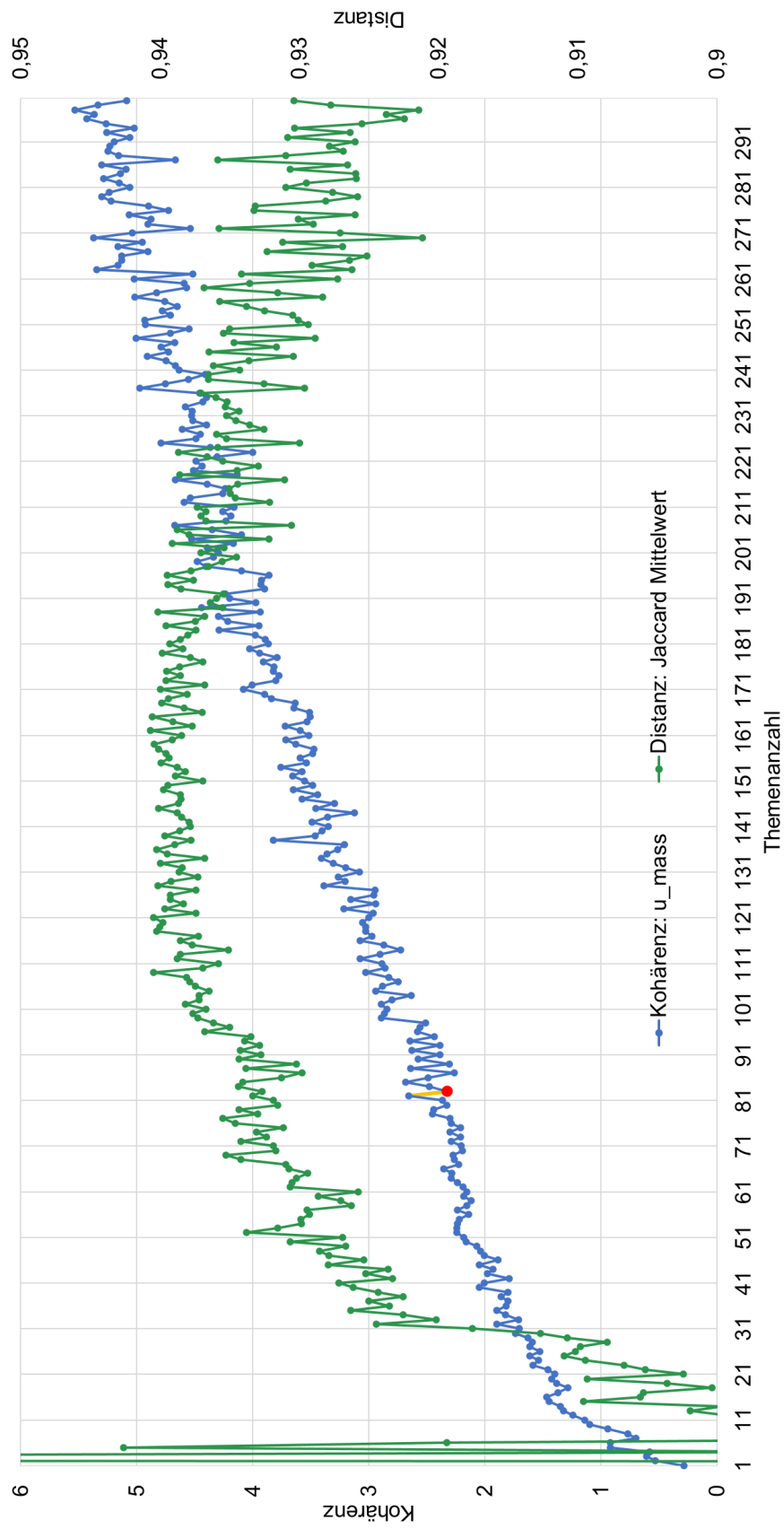


Abbildung 3.1: Veränderung der Kohärenz- und Distanzwerte der Themen

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.

$$1 - \frac{|LDA_n|}{|LDA_{n+1}|}$$

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.

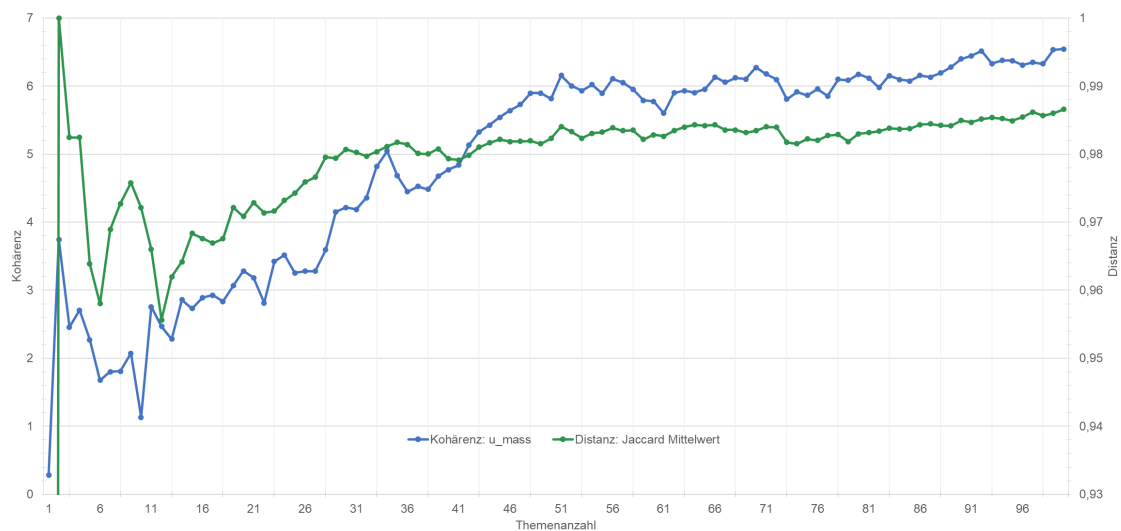


Abbildung 3.2: Kohärenz und Distanz der Themen mit Bigrammen

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.

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 sollen die gefundenen Themencluster bestätigt werden. Der DLDA wird die Entwicklung der vom LDA gebildeten Cluster über die Zeit beschreiben. Abschließend werden die Ergebnisse Diskutiert.

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 Termwahrscheinlichkeit, im ausgewählten Thema Nummer 50 absteigend sortiert und die Häufigkeit im gesamten Korpus.

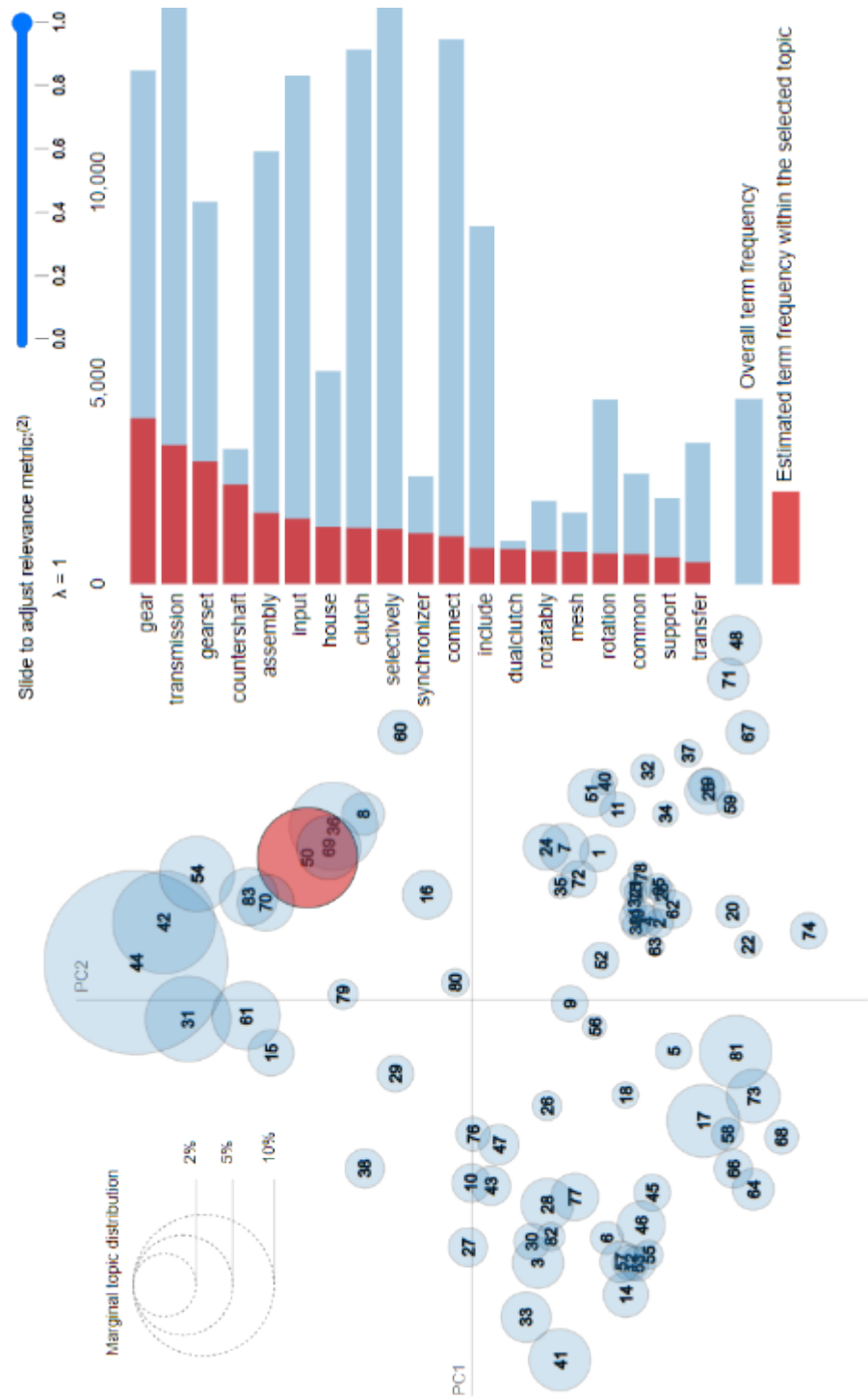


Abbildung 4.1 : Interthematische Distanz Karte erstellt mit multidimensionaler Skalierung und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 1,0$

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. Ein Term ist besonders relevant für ein Thema, wenn er eine möglichst hohe Wahrscheinlichkeit hat zu diesem Thema zu gehören und eine möglichst geringe Wahrscheinlichkeit hat zu allen anderen Themen zu gehören. Nach einer Nutzerstudie sollen Anwender mit einem λ Wert von 0,6 die Themen am besten klassifizieren können (vgl. Sievert et al. 2014, S. 66-68). Rechts wurde der Term *countershaft* (Vorgelegewelle) ausgewählt. Dadurch werden die Themenradien, abhängig von der Verteilung des ausgewählten Terms, skaliert. Die Themen 50, 20 und 83 haben für den Term *countershaft* die höchste Zugehörigkeitswahrscheinlichkeit und sind teil des pink eingekreisten Subcluster des *transmission* Clusters, aus Abbildung 4.3. Vereinfacht gesagt und unter der Annahme, das Modell ist korrekt, kommt der Term *countershaft* sehr häufig in den drei Themen vor und sehr selten in allen anderen Themen. Da *countershaft* der relevanteste und aussagekräftigste Term des 50. Themas ist wird es als das *countershaft* Thema gewertet.

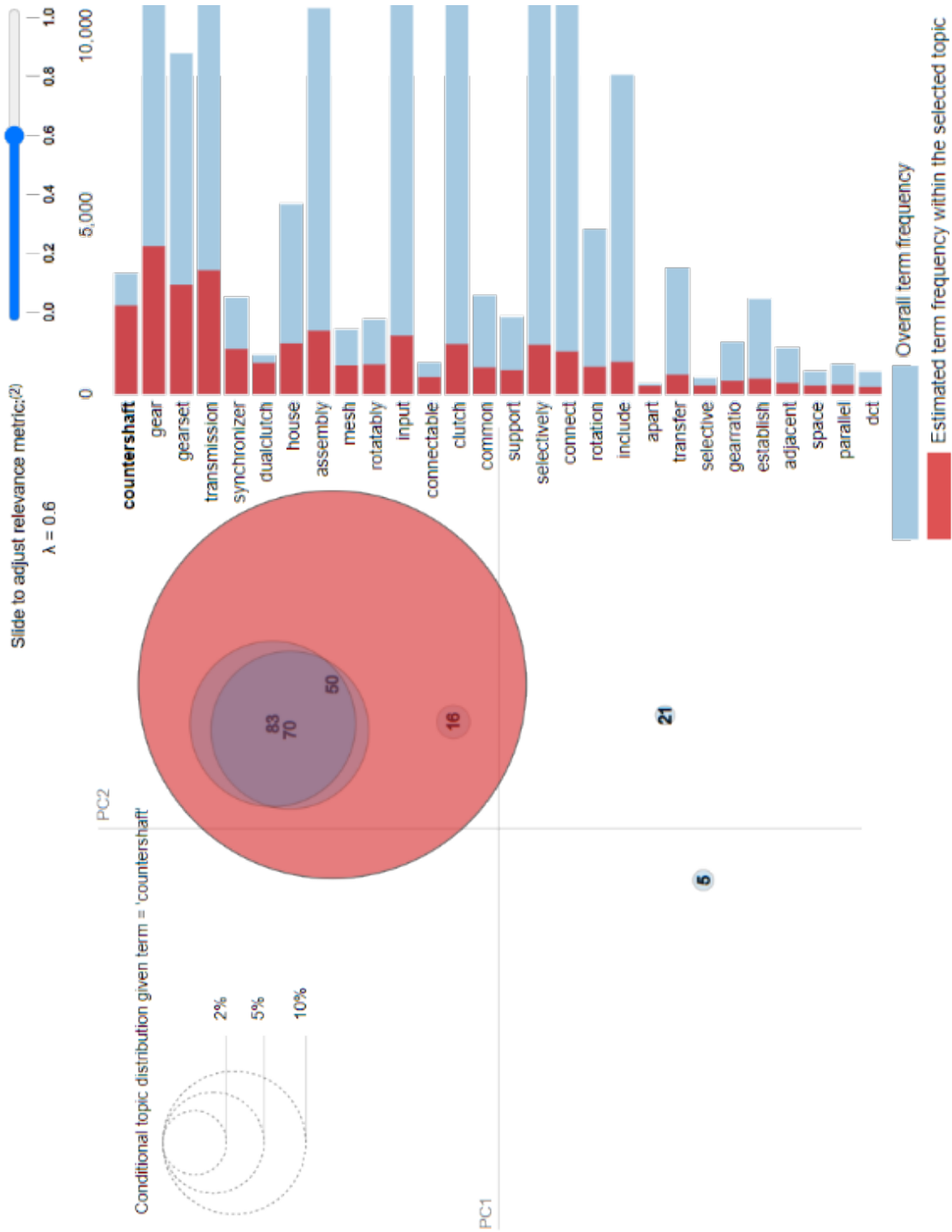


Abbildung 4.2: Interthematische Distanz Karte erstellt mit multidimensionaler Skalierung mit einer Themenverteilung welche von dem Term countershaft abhängt und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 0,6$

Außerdem kommen die äquivalenten Terme *dualclutch*, *Dual Clutch Transmission (DCT)* und *automatictransmission* 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 Wellen (*shafts*) hin, beispielsweise mit dem ausgewählten Gang. 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 nur in dem *method* Cluster? Eigentlich ist es ein rein mechanisches Bauteil und der *method* Cluster enthält Themen zu elektronischen Steuerung und Regelung. Für genauere Einblicke in Themen wurde mit LDA eine Patent-Themen-Matrix erstellt. Das US-Patent 9,989,146 passt am besten zu Thema 77. Es beschreibt eine Methode, welche den optimalen Druck (*pressure*) einer Kupplung (*clutch*) in einem Stufenlosem Getriebe (*Continuously Variable Transmission (CVT)*) erlernt, damit sie ein Verrutschen des Riemens (*pulley slip*) verhindern kann. Daher kommt in diesem Thema der Term *pressure* ohne *fluid* oder *hydraulik* vor.

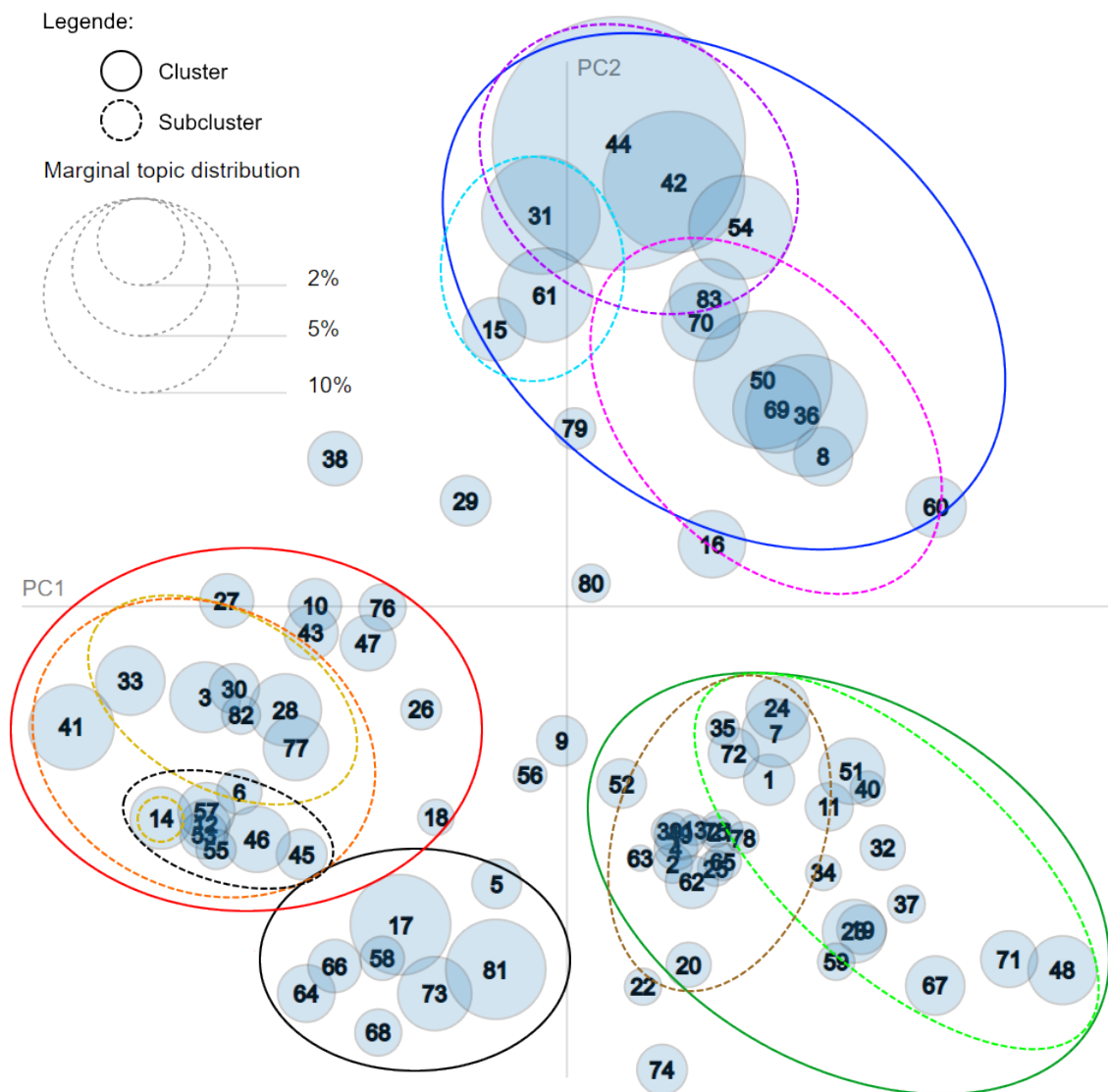


Abbildung 4.3: Interthematische Distanz Karte erstellt mit multidimensionaler Skalierung und qualitativer Methode zur Clusterung der Themen

Tabelle 4.1: Legende zu Abbildung 4.3

transmission, selectively	method, predetermine	fluid, valve
planetarygearset	monitor	surface, side, extend
gear, gearset	command	axis, house
motorgenerator, EVT	temperature, sensor	mount

Durch dieses Verfahren wurden vier Cluster und acht Subcluster gebildet. Diese Cluster werden, mit ihren Subclustern, im Uhrzeigersinn analysiert.

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

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 Verwandtschaft mit dem pinken Subcluster, weil er diese Terme ebenfalls häufig 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.

Der rote Cluster enthält besonders viele Terme wie *method*, *command* und *request*. Der Cluster besteht daher aus Steuerungs- und Regelungsthemen von mechanischen Bauteilen. 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 *monitor* gleichmäßiger im gesamten orangen Subcluster, einschließlich des gelben Subclusters, verteilt ist. Das Thema Nummer 14 ist eine Ausnahme, weil es fast gleich viele Terme von *command* und *monitor* für die *hydraulic pressure* enthält. Deshalb ist es extra gelb umkreist. Die Rutschkupplung (*frictionclutch*) ist hauptsächlich dem Thema 77 zuzuordnen. Der Term *slip* kommt in Verbindung mit der *clutch* in den umliegenden Themen häufig vor. Auch die Klauenkupplung (*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* (US-Patent 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* (Thema 14), die *hydraulic pump* (Thema 45) und die *temperature* des *coolant* im Kühlkreislauf der *electricmachine* (US-Patent 8,167,773), der *engine* und des *radiator* (Thema 57). Durch die Steuerung des Kühlkreislauf können Komponenten wie die *transmission* auf Betriebstemperatur gebracht werden (US-Patent 10,161,501). Zu dem Thema 12 passt am besten das US-Patent 9,404,403, es beschreibt eine Methode um das Öllevel zu beobachten (*monitor*). Auf Grund der vielen Flüssigkeits- und Regelungsthemen befindet sich der schwarze Subcluster in der Nähe des schwarzen Hauptclusters aber innerhalb des roten Clusters.

Der schwarze Hauptcluster enthält fast alle Themen die in Verbindung mit Flüssigkeit stehen. Er teilt die Häufigkeit des Terms *control* mit dem roten Cluster aber unterscheidet sich durch die Verwendung von den Termen *communication* und *communicate*, die sonst nur selten auftreten. Besonders häufig ist die Kombination Elektromagnet (*solenoid*), *hydraulic*, *valve* und *fluid*. Das Thema 17 beschreibt in mehreren US-Patenten (8,820,185, 8,382,639) die Steuerung einer *dualclutch*, mithilfe von *hydraulic* und *solenoids*. Mit einem *solenoid* wird ein Verschluss aus der *valve* gezogen. Dieser Verschluss wird nach dem Ausschalten des *solenoids* von 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 Aktuatorgabel (*actuator fork*) kontrolliert (*control*) werden kann (US-Patent 9,605,755).

Die Ergebnisse des LDA zeigen das in dem Patentdatensatz von General Motors (GM) hauptsächlich um Getriebe geht. Darunter ist größtenteils das Doppelkupplungsgetriebe (*DCT*) und Stufenlose Getriebe wie *CVT* und die elektrische Variante *EVT*. Manuelle Getriebe kommen seltener vor. Außerdem sind viele Patente vorhanden zu den verschiedenen Kupplungen dieser Getriebe. Darunter sind die Rutschkupplung (*friction clutch*) und Klauenkupplung (*dog clutch*). Auch die verschiedenen Wellen kommen häufig vor, die durch Kupplungen und Zahnräder verbunden werden. Die Wellen sind unter anderem die Antriebswelle (*inputshaft*), Abtriebswelle (*outputshaft*), Vorgelegewelle (*countershaft*) und die Kurbelwelle (*crankshaft*). Häufig geht es in Verbindung mit dem Wellen auch um das Differentialgetriebe (*differential*). Es kommen auch besondere Zahnradkonstruktionen wie das Planetengetriebe (*planetarygear*) vor. Die Getriebe werden in Gehäusen (*house*) verbaut. Die Steuerung der Getriebe erfolgt elektrisch mit Aktuatoren (*actuator*) oder hydraulisch mit Elektromagneten (*solenoid*).

4.2 Analyse der Ergebnisse des HLDA

Die Ergebnisse des HLDA werden im Vergleich mit den LDA Ergebnissen analysiert, um diese zu bestätigen. Zwei Subcluster des HLDA Baums werden mit den Ergebnissen des LDA verglichen und interpretiert. Der ganze HLDA Baum ist zu groß um leserlich abgebildet zu werden. Er kann im digitalen Anhang mit einem Programm geöffnet werden, welches das graphml-Format 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.

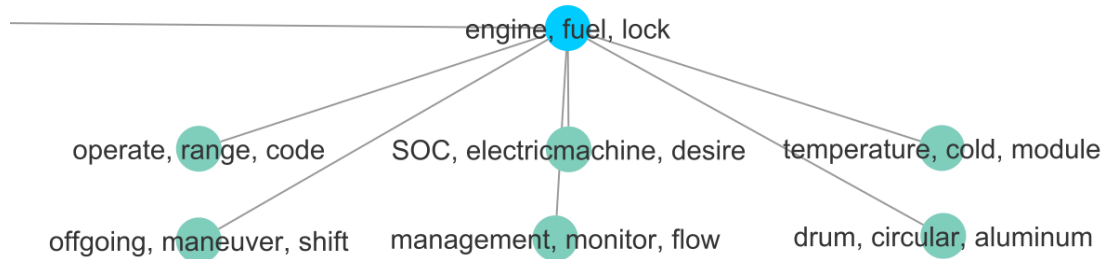


Abbildung 4.4: Der *engine* Ausschnitt des HLDA Baums

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 US-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*.

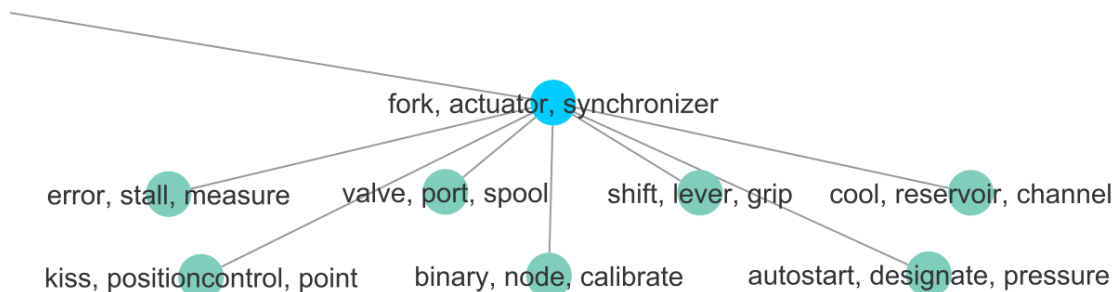


Abbildung 4.5: Der *fork* Ausschnitt des HLDA Baums

Die Cluster und Termwahl des HLDA sind ähnlich zum LDA. Selbst die Subcluster sind meistens passend zum LDA gebildet worden, obwohl der HLDA recht unterschiedlich zum LDA. Beispielsweise wird keine genaue Themenanzahl vorgegeben. Das eine unterschiedliche Methode zu ähnlich nachvollziehbaren Ergebnissen kommt ist ein Indiz für die Güte beider Methoden.

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, im Uhrzeigersinn ihrer Cluster aus Abbildung 4.3 ,analysiert.

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. Der Trend der Cluster wird durch die relative Häufigkeit ihrer beiden relevantesten Terme pro Jahr bestimmt.

Der größten Cluster heißt *transmission*. In Abbildung 4.6 verzeichnet er einen leichten Abwärtstrend. Der alternative Term *selectively* bestätigt diesen Trend. Der Subcluster von *transmission* ist *gear*, *gearset* und steigt stark an. In diesem Subcluster befindet sich auch das Thema 50 und weist in Abbildung 4.7 eigene Trends auf. Der Term *countershaft* ist von 2004 bis 2007 mit 12% bis 14,7% in diesem Thema sehr dominant und verliert bis 2010 stark an Häufigkeit. Er verbleibt dann bei 6% bis 7%. Der Term *transmission* hingegen nimmt über den gesamten Zeitraum von 7% bis 10,5% zu. Der Term *gear* schwankt stark und nimmt über den gesamten Zeitraum von 9,5% bis 8,2% leicht ab. Es ist also durchaus möglich das innerhalb der Cluster unterschiedliche Trends vorkommen. In diesem Fall sind die Trends des steigenden *transmission* Themas und fallendem *gear* Themas sogar gegensätzlich zum Trend ihres Clusters. Auch das *ringgear* und *sungear* folgen diesem Trend. Allerdings spiegelt sich der Abwärtstrend des Terms *countershaft* von Thema 50 im gesamten Datensatz wieder. Dort sinkt die Häufigkeit von 0,21% auf 0,14%. Der *inputshaft* und *outputshaft* folgen diesem Trend. Die verwandten Terme folgen gemeinsamen Trends, das deutet auf eine korrektes Modell hin.

Der Cluster *surface*, *side* zeigt kaum Veränderung. Die dort am häufigsten auftretenden Terme wie *house* und *body* stagnieren. Allerdings zeigt der Term *axis* ein Wachstum von 2004 (0,36%) bis zum Höhepunkt 2011 (0,51%). Bis 2017 sinkt die Häufigkeit wieder auf 0,35%. Dieser Term kommt genau wie *house* und *shaft*

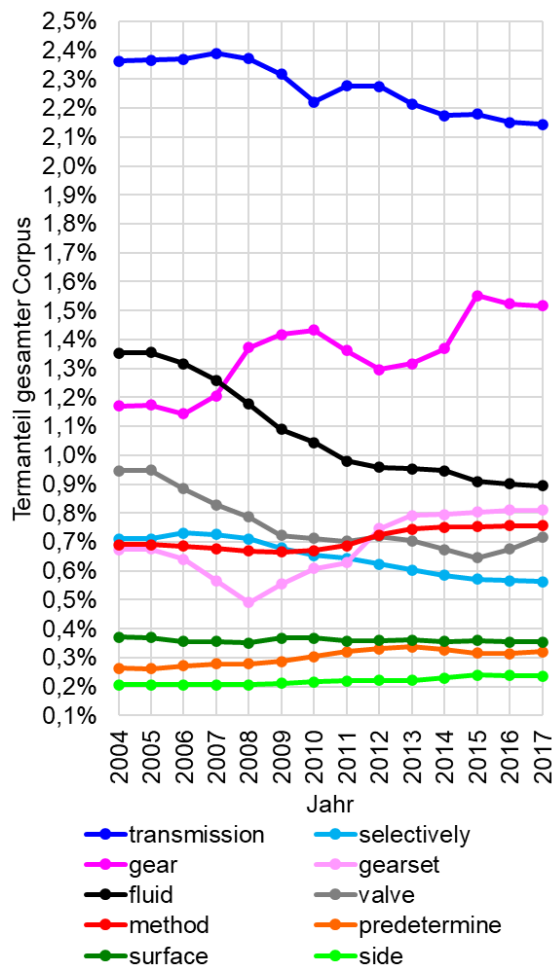


Abbildung 4.6: Trends der Terme die ihren Cluster am besten beschreiben

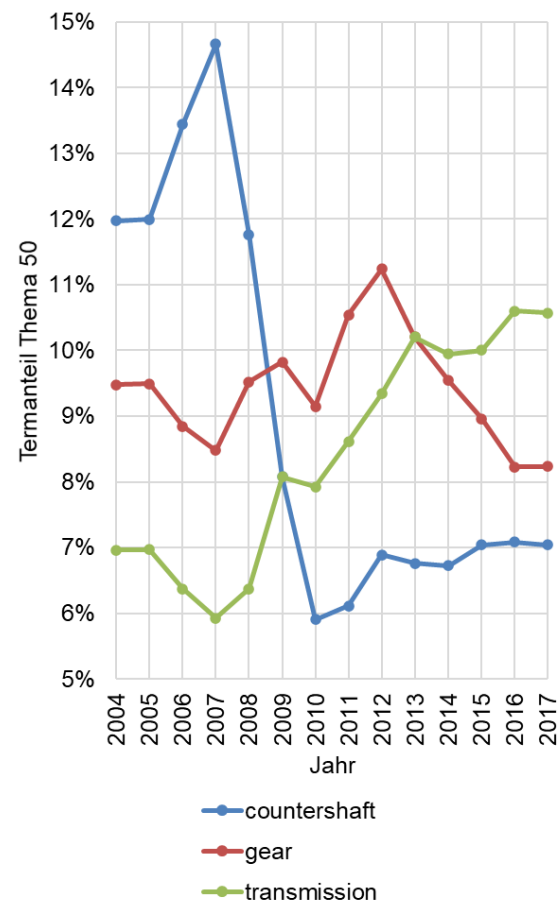


Abbildung 4.7: Trends der 3 Terme die das Thema 50 am besten beschreiben

häufig im pinken Subcluster vor. Allerdings gibt es hier keinen gemeinsamen Trend. Die gehäufte Anmeldung von Patenten im Achsenthema um das Jahr 2011 ist ein eigener Trend.

Die relative Häufigkeit des Terms *fluid*. Der Term *valve* bestätigt die sinkende Zahl von Patentanmeldungen mit Bezug zu Flüssigkeiten. Die Häufigkeit verringerte sich von 2004 bis 2017 von 0,95% auf 0,72%.

Der benachbarte Cluster *method*, *predetermine* gewinnt moderat an Häufigkeit. Er enthält hauptsächlich Patente mit Methoden zur elektronischen Steuerung und Regelung von mechanischen Bauteilen. Der Term *command* bestätigt den Aufwärtstrend mit einer Erhöhung von 0,2% auf 0,23%. Terme wie *control* und

system wurden explizit nicht ausgewählt, weil sie große Überschneidungen mit dem Flüssigkeitsthema haben.

Der DLDA zeigt deutliche Trends, wie stark sinkende Patentanmeldungen zu Flüssigkeitsthemen und steigende Anmeldungen zu Methoden Themen. Die Schlussfolgerung liegt nahe das mehr Patente mit elektrischen Aktuatoren (*actuator*) zur Steuerung und Regelung der Getriebe angemeldet werden und weniger hydraulische Patente mit Elektromagneten (*solenoid*). Das ist ein Trugschluss. Die Trends der großen Cluster dürfen nicht verallgemeinert werden, denn das Gegenteil ist der Fall, die Aktuatorpatente werden seltener und die Hydraulikpatente werden häufiger angemeldet. Auch die Trends innerhalb der Cluster zeichnen sich deutlich ab. Während der Cluster *transmission* kleiner wird, wächst der Subcluster *gear*. Selbst in einzelnen Themen lassen sich Trends abbilden. Thema 50 ist anfangs ein *counershaft* Thema und wird zu einem *transmission* Thema. Daher ist es notwendig die Trends der einzelnen Terme zu beobachten, die Trends der großen Cluster können andere sein.

4.4 Diskussion der Ergebnisse

LDA ist ein robustes Verfahren, um die versteckten Themen großer Textmengen herauszufinden, ohne alle Texte selbst lesen zu müssen. Allerdings müssen die Ergebnisse immer überprüft werden. Denn die Ergebnisse sind stark abhängig von den Daten und Parametern. Selbst Patente bringen sprachliche Ungenauigkeiten mit sich, wie Synonyme die man im Preprocessing nur sehr schwer voll umfänglich erfassen kann. Dadurch behandelt das Modell *dualclutch* und *binaryclutch* unterschiedlich, obwohl sie die selbe Bedeutung haben.

Das Modell der Bigramme wurde nicht weiter analysiert, weil es sich bei weitem nicht so gut interpretieren ließ wie das der Unigramme. Obwohl die Kennzahlen des Bigrammodells deutlich besser sind als die des Unigrammodells. Das Bigrammodell sollte viel kohärenter und die Themen viel distanzierter sein als die des Unigrammodells wie in der Tabelle 4.2 abzulesen ist. Die Vergleiche der Distanz für jedes Thema befinden sich im Anhang als Abbildung Unigramme 6.1

Tabelle 4.2: Kohärenzen des Unigramm- und Bigrammmodells, niedriger ist besser

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

und Bigramme 6.2. Die Abbildung 4.8 zeigt das sich die Themen nicht gut in Cluster unterteilen lassen, weil die meisten sich direkt nebeneinander befinden. Das könnte an unvorteilhaft zusammengeschriebenen Termen wie *clutch dualclutch* liegen, weil sie wahrscheinlich zwei Themen hervorrufen, anstatt nur das Thema *dual clutch*.

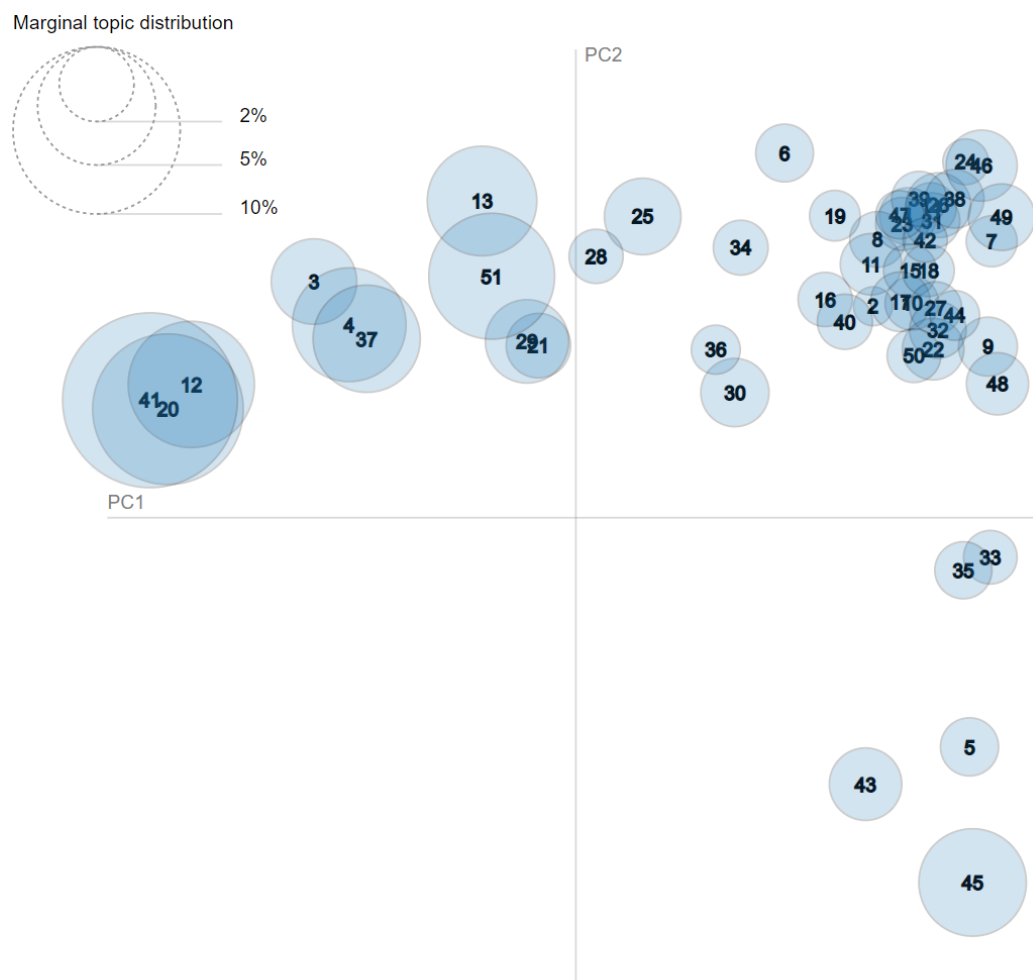


Abbildung 4.8: Interthematische Distanz Karte der Bigramme, erstellt mit multidimensionaler Skalierung bei einem von $\lambda = 1,0$

Kapitel 5

Zusammenfassung und Ausblick

5.1 Zusammenfassung

Das Ziel dieser Arbeit war herauszufinden welche relevanten Themen sich in dem Patentdatensatz von GM befinden, sie zu benennen und darzustellen wie sie sich über einen Zeitraum von 14 Jahren entwickelt haben.

Der Datensatz wurde mit einem Preprocessing zu Dokumenten mit einzelnen Termen vereinfacht. Die Terme wurden mit LDA einer möglichst optimalen Zahl an Themen zugeordnet. Die größten und relevantesten Themen wurden mit einem qualitativen Verfahren benannt und gruppiert. Dieses Verfahren wurde durch das Programm pyLDavis unterstützt. Die Gruppierungen wurden mit der alternativen Methode HLDA größtenteils bestätigt. Die Trends der relevantesten Themen und größten Gruppen wurden mithilfe von DLDA gefunden und dargestellt.

Die Ergebnisse dieser Arbeit ermöglichen es jedem die Hauptthemen und die meisten Unterthemen dieses Patentdatensatzes zu erfassen ohne ein einziges Patent lesen zu müssen. Außerdem ermöglichen die Ergebnisse das Filtern des Datensatzes nach bestimmten Themengruppen, Themen, Patenten und Termen. So könnten beispielsweise die relevantesten Patente für ein beliebiges Thema angezeigt werden. Des weiteren kann der Trend aller Terme visualisiert werden.

5.2 Ausblick

Wie in der Diskussion beschrieben würde es sich lohnen am Preprocessing der Bigramme zu arbeiten. Es wäre möglich das sich die Themen breiter verteilen, wenn mit den Erkenntnissen aus den Unigrammen Terme aufgespalten oder zusammengelegt werden würden. Terme wie *dualclutch* könnten in *dual clutch*

aufgespalten werden und *binary clutch* sollte zu *dual clutch* werden. So könnten Bigramme wie *clutch dualclutch* vermieden werden. Dadurch sollte es beispielsweise nur noch ein Thema *dual clutch* geben und nicht mehr das redundante Thema *clutch dualclutch*. Mit dem vermeiden von Synonymen würde das Thema *binary clutch* ebenfalls wegfallen.

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 der Themen direkt zu messen wäre eine noch genauere Methode.

Kapitel 6

Anhang

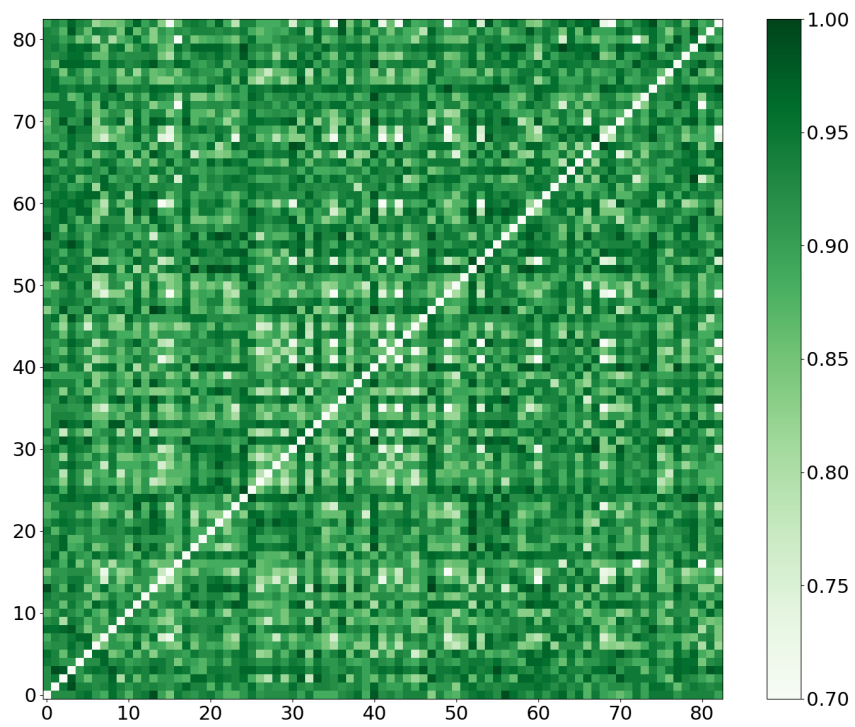


Abbildung 6.1: Distanz zwischen den top 50 Unigrammen der Themen

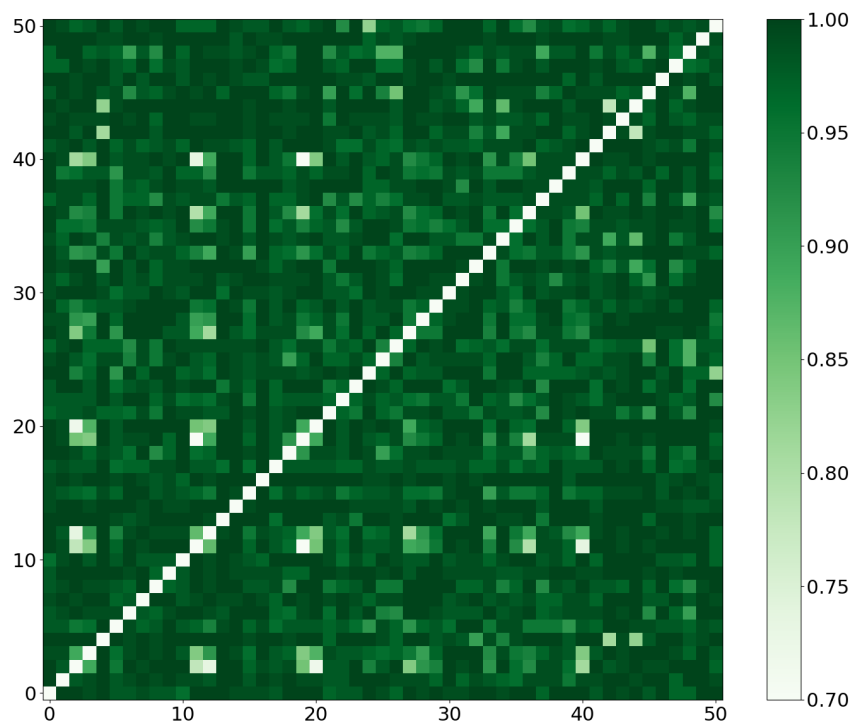


Abbildung 6.2: Distanz zwischen den top 50 Bigrammen der Themen

Kapitel 7

Eidesstattliche Erklärung

Erklärung zur Abschlussarbeit

Ich versichere, den Bachelor-Report oder den von mir zu verantwortenden Teil einer Gruppenarbeit*) ohne fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, sind als solche kenntlich gemacht.

*) Bei einer Gruppenarbeit muss die individuelle Leistung deutlich abgrenzbar und bewertbar sein und den Anforderungen entsprechen.

Bremen, den 16.11.2020

Hauke Tietjen

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.