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

BoW Bag-of-Words 4, 9

CVT Continuously Variable Transmission 9, 22, 24–26

DCT Dual Clutch Transmission 22, 26

DLDA Dynamic Latent Dirichlet Allocation . . II, 3, 7–9, 15–17, 28, 30, 33, 34

EVT Electrically Variable Transmission 24, 26

GM General Motors 1, 8, 26, 33

HLDA Hierarchical Latent Dirichlet Allocation I, IV, 3, 7, 8, 15, 17, 26, 27, 33

IPMI Institut für Projektmanagement und Innovation 9

LDA Latent Dirichlet Allocation . . . . . I, 4–11, 15–18, 22, 26–28, 30, 33

pyLDAvis Python LDA Visualization 3, 15, 17, 18, 22, 33

TDM Term-Dokument Matrix 9

tf-idf Term Frequency-Inverse Document Frequency 15

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Thema

In dieser Bachelorarbeit geht es darum, die relevantesten Themen in einem großen Patentdatensatz von General Motors (GM) zu finden. Dazu werden generativen Wahrscheinlichkeitsmethoden verwendet werden. Diese Themen sollen benannt und graphisch dargestellt werden. Außerdem gilt es herauszufinden welche Themengruppen vorhanden sind und welche Patente zu einem oder mehreren Themen gehören. Außerdem soll der Trend der Themen über die Jahre untersucht werden.

1.2 Motivation und Forschungsfragen

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 umkehrt, funktioniert sie 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 des Preprocessings. Manche Texte haben Schlagworte aber dabei verlässt man sich auf den Autor, diese Richtigen zu wählen und sie werden nicht nach

Relevanz gewichtet. Außerdem könnte man, mit dem Wissen über die Trends der Patentthemen, Vermutungen über die Patentthemen der Zukunft anstellen.

Topic Modeling ist besonders interessant, weil mit relativ geringem Aufwand große Mengen an Dokumenten untersucht werden können. Dadurch könnten, speziell in diesem Fall, die Konkurrenten von General Motors herausfinden, worum es in den Datensatz geht und in welche Richtung sich die Themen der Patente in Zukunft entwickeln könnten. Wodurch sie General Motors bei der Anmeldung neuer Patenten zuvorkommen und Lizenzgebühren verlangen könnten.

Die Forschungsfragen dieser Arbeit sind:

- Welche relevanten Themen befinden sich in dem Datensatz?
- Wie sind diese Themen zu benennen?
- Welchen Gruppen gehören diese Themen an?
- Welche Trends haben diese Themen?

1.3 Methodisches Vorgehen

Um die versteckten Themen zu finden, werden generative Wahrscheinlichkeitsmethoden benutzt. Eine Methode ist die Latent Dirichlet Allocation (vgl. Blei, Ng 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 tausenden Dokumenten viel schneller benennen, als es einem Menschen durch normales

Lesen möglich wäre. Dazu wird das Online Latent Dirichlet Allocation Verfahren angewandt. (vgl. Hoffman et al. 2010) Als Kohärenzmaß wird C_umass verwendet (vgl. Röder et al. 2015).

Der Patentdatensatz von General Motors umfasst 1410 Patente für verschiedene Getriebearten und ist ausreichend groß, um Topic Modeling zu betreiben. Um diese Untersuchungen zu realisieren, wird die Programmiersprache Python verwenden. Mit Hilfe der Programmbibliothek gensim werden die Modelle erstellen und die Kohärenzen ausgewertet (vgl. Řehůřek et al. 2010). Die Ergebnisse werden mit Python LDA Visualization (pyLDAvis) untersucht, um in einem qualitativen Verfahren Gruppen zu bilden und Themen zu benennen (vgl. Sievert et al. 2014). Die Gruppen sollen mit HLDA bestätigt werden (vgl. Griffiths et al. 2004).

Des weiteren soll mit dem DLDA Verfahren herausgefunden werden wie sich die Themen des Datensatzes, entlang der zeitlichen Anmeldedaten der Patente, verändert haben (vgl. Blei und Lafferty 2006). Eine Vorhersage wäre für ein konkurrierendes Unternehmen hilfreich, um Patente vor General Motors anzumelden und Lizenzgebühren verlangen zu können.

Kapitel 2

Grundlagen zur Latent Dirichlet Allocation

Die LDA ist ein generatives Wahrscheinlichkeitsmodell für Textdokumente. (vgl. Blei, Ng 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 (BoW) 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.

$\mathcal{W}$ ist das Wort aus $\mathcal{N}$ Wörtern eines Dokuments i. Dieses Dokument i ist eines aus allen Dokumenten $\mathcal{M}$ bezeichnet als $\mathcal{D}$ (vgl. Blei, Ng et al. 2003, S. 995). Alle

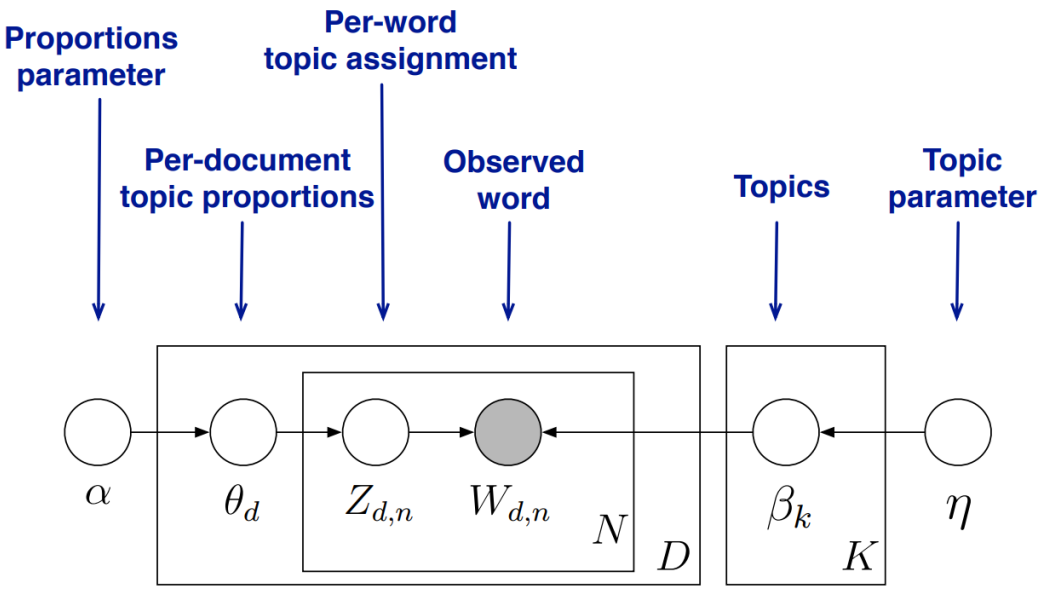

Abbildung 2.1: LDA als graphisches Modell, von (vgl. Blei 2012, S. 23)

folgenden Parameter sind latent. $\mathcal{Z}$ ist das Thema für das Wort j aus besagtem Dokument i (vgl. Blei, Ng et al. 2003, S. 996-997) (vgl. Blei 2012, S. 27). Jedem Wort wird ein Thema zugeordnet. Wodurch jedes Dokument eine Mischung aus allen Themen ist. Die Verteilung der Themen für Dokument i ist θ. α und β sind Hyperparameter des LDA (vgl. Blei 2012, S. 32). 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.

Die Abbildung 2.2 zeigt die Wahrscheinlichkeit ein Dokument zu generieren, mit den Einstellungen des LDA Modells. Die Wahrscheinlichkeit ist gering aber je

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

Abbildung 2.2: LDA als graphisches Modell, Knoten sind zufällige Variablen, Kanten zeigen Abhängigkeit, der graue Knoten wird beobachtet, die Rechtecke sind Wiederholungen (vgl. Blei 2012, S. 25)

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 (vgl. Blei 2012, S. 53-54). 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 (vgl. Blei, Ng et al. 2003, S. 996-999).

Der dritte Faktor ist eine Multinomialverteilung des ersten Faktors (vgl. Blei, Ng et al. 2003, S. 999-1002). 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.

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 und Lafferty 2006)

HLDA erweitert LDA durch eine Hierarchie aus Unterthemen, mit beliebiger Tiefe (vgl. Griffiths et al. 2004). Diese Hierarchien 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.

Kapitel 3

Methodik und Ergebnisse

Um den Inhalt des Patentdatensatzes zu erforschen müssen die Patente zuerst im Preprocessing in eine maschinenlesbare Form gebracht werden. Diese Daten werden dann zu Modellen verarbeitet und die besten Modelle anhand von Kennzahlen ausgewählt. Der Algorithmus LDA erzeugt die Themen welche später in der Analyse einem qualitativen Verfahren gruppiert und benannt werden. Mit dem HLDA Algorithmus können die Gruppen bestätigt werden. Der DLDA Algorithmus zeigt die Trends der LDA Terme auf.

3.1 Patentdatensatz

Der Patentdatensatz wurde mit der Software Query4Files in der Datenbank Lucene erstellt. Es wurden 1410 Dokumente gesammelt. Der Datensatz enthält ausschließlich Patente von GM, die durch das Tochterunternehmen GM Global Technology Operations angemeldet wurden. Ein Patent ist ein Schutzrecht für eine Erfindung (vgl. Organization 2004, S. 17). Das Patent beschreibt die Erfindung detailliert und kann vom Erfinder beim Staat beantragt werden. Wenn es genehmigt wird, kann der Erfinder anderen Bedingungen für die Nutzung seines Patents stellen.

3.2 Preprocessing

Das Preprocessing dient dazu die Dokumente so zu filtern, dass sie ein Programm lesen kann. Das Preprocessing wurde mit dem PatVisor®, das Patentanalyse-

werkzeug vom Institut für Projektmanagement und Innovation (IPMI), durchgeführt (vgl. Walter et al. 2016, S. 159-164). Dazu wurde vom IPMI ein themenbezogener Synonymfilter bereitgestellt. Aus den Patenten wurde nur der Titel, der Abstract und die Claims als Text 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 BoW Modell irrelevant ist. Außerdem wurden manuell Abkürzungen erfasst wie Continuously Variable Transmission (CVT). So werden die Worte zu Unigrammen und werden Terme genannt. Bigramme bestehen aus zwei Termen. Die 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 mit der Programmiersprache Python implementiert: Wörterbuch- und Korpuserzeugung, LDA, Evaluation, Visualisierung. Dieser Vorgang muss für Unigramme und Bigramme einzeln Durchgeführt werden. Der Vorgang ist bei beiden sehr ähnlich. Gensim ist eine Python library für Textanalyse. Die Skripte zur Durchführung von LDA und DLDA wurden vom IPMI bereitgestellt. Diese Skripte wurden von mir modifiziert, um LDA und DLDA vergleichen zu können. Im ersten Schritt wird aus der TDM des Preprocessings ein Wörterbuch und ein Korpus erstellt, ein Ausschnitt ist in Tabelle 3.1 und 3.2 dargestellt. 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. 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 (vgl. Mimno et al. 2011, S. 265-266). 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 universal nutzbar ist und aufgrund des fehlenden Schiebefensters auch bei Bigrammen funktioniert. Das ist hier kein Nachteil, weil die bereinigten Patentexte recht kurz sind im Vergleich zu Büchern. Dort würde ein Schiebefenster sinnvoll sein.

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 (vgl. Kosub 2019, S. 1).

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

Im zweiten Schritt 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 Zugehörigkeitswahrscheinlickeit 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.

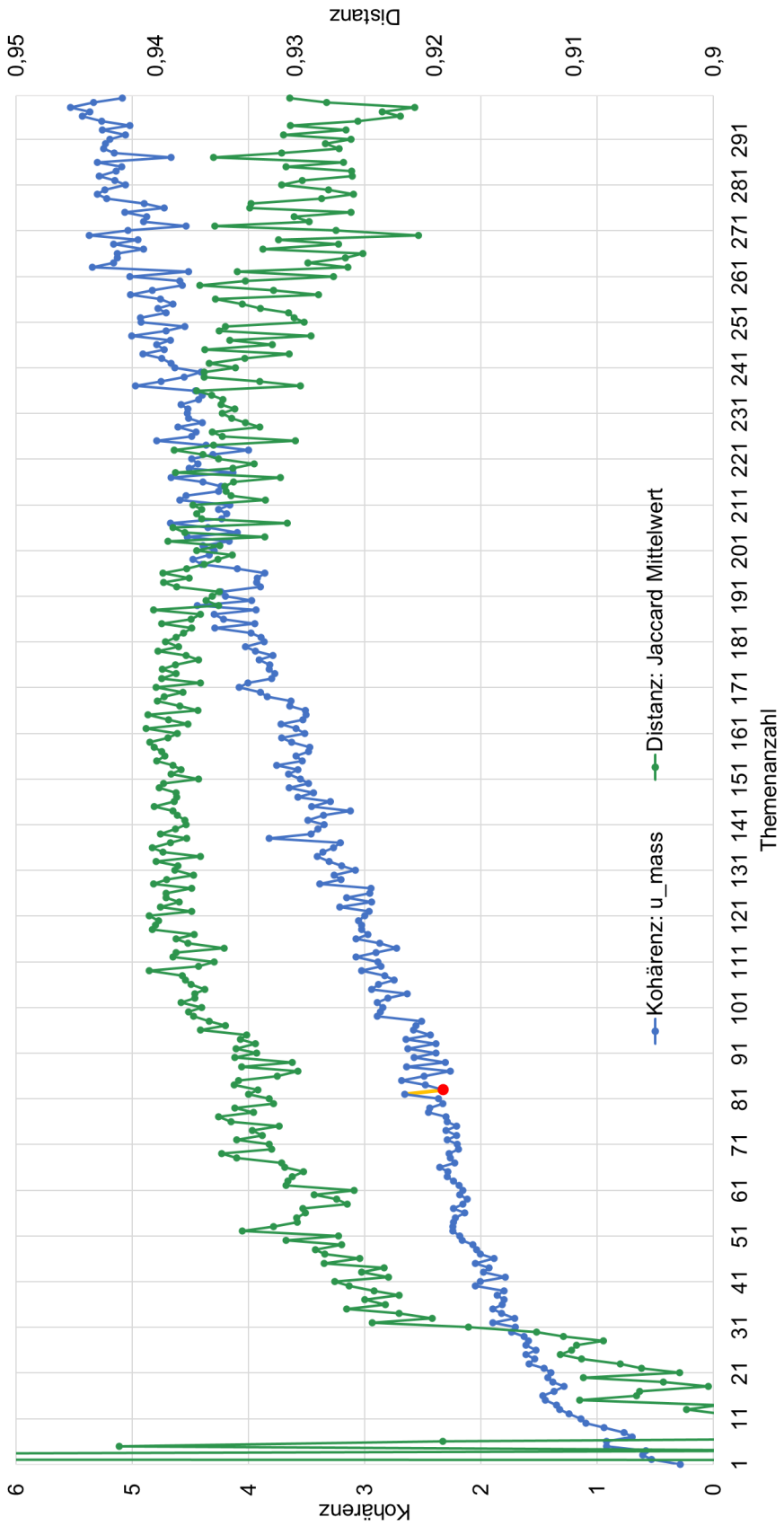

Abbildung 3.1: Veränderung der Kohärenz- und Distanzwerte der Themen

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.

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

Bei den Unigrammen sind es in Abbildung 3.1 bei dem rot markiertem Punkt 83 Themen. Die absolute Kohärenzsteigerung ist gelb markiert. 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 lokale Kohärenz und Distanz bei der Themenanzahl von 51 erreicht. Das Thema 51 wurde rot markiert. Mit der Anzahl wurde ein deutlich granulareres Modell erstellt.

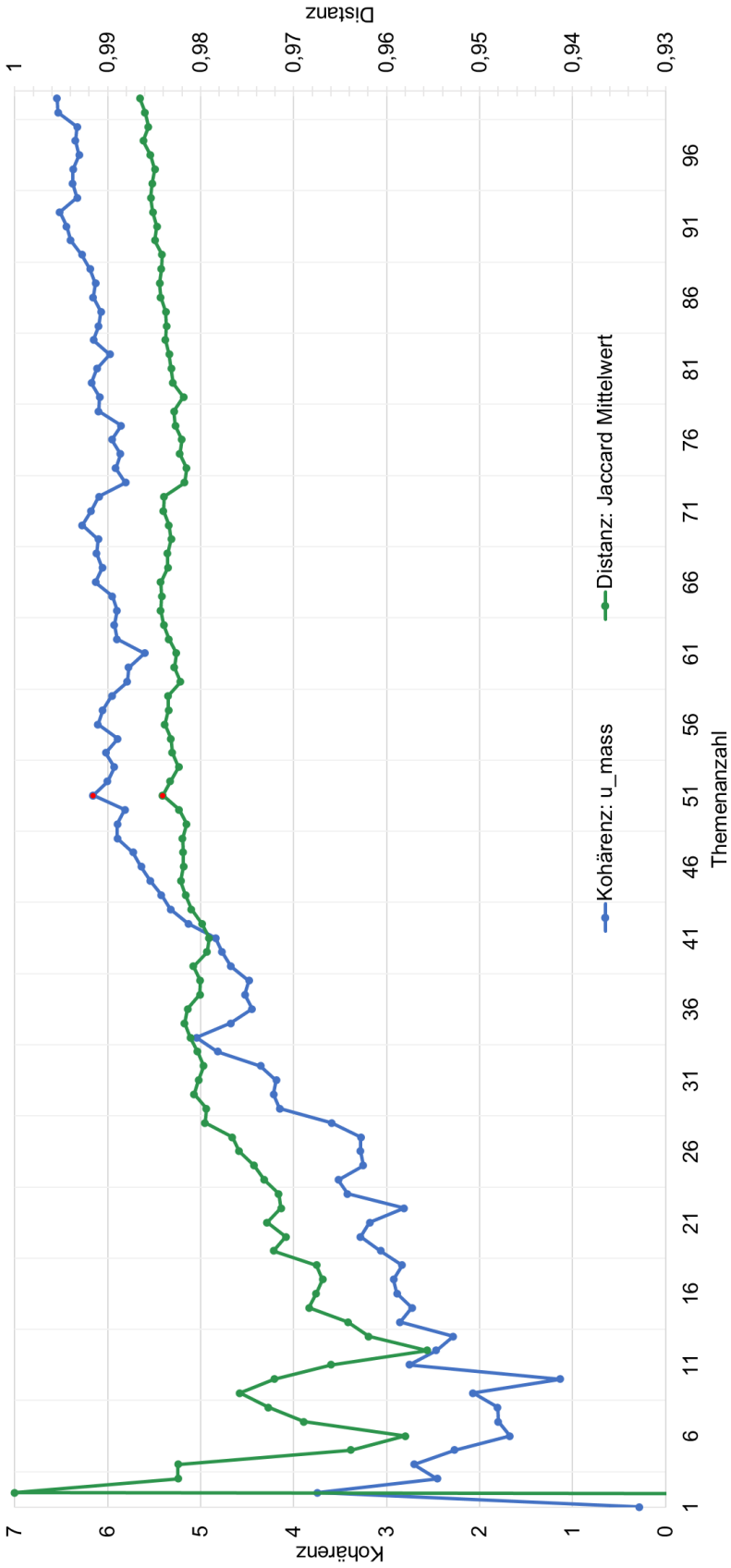

Abbildung 3.2: Kohärenz und Distanz der Themen mit Bigrammen

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

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

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 (vgl. Luhn 1957) (vgl. 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 Anmeldedatum 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. Besonders vorteilhaft ist, dass der Code so modifiziert werden konnte, dass die Ergebnisse des LDA mit denen des DLDA vergleichbar sind. Das LDA Modell wurde als Eingabewert für den DLDA verwendet. So sind die Themen von LDA und DLDA die gleichen.

Kapitel 4

Analyse der Ergebnisse

Die Analyse der Unigrammergebnisse 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 und Lösungen gesucht, um die Bigrammergebnisse zu verbessern.

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 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.

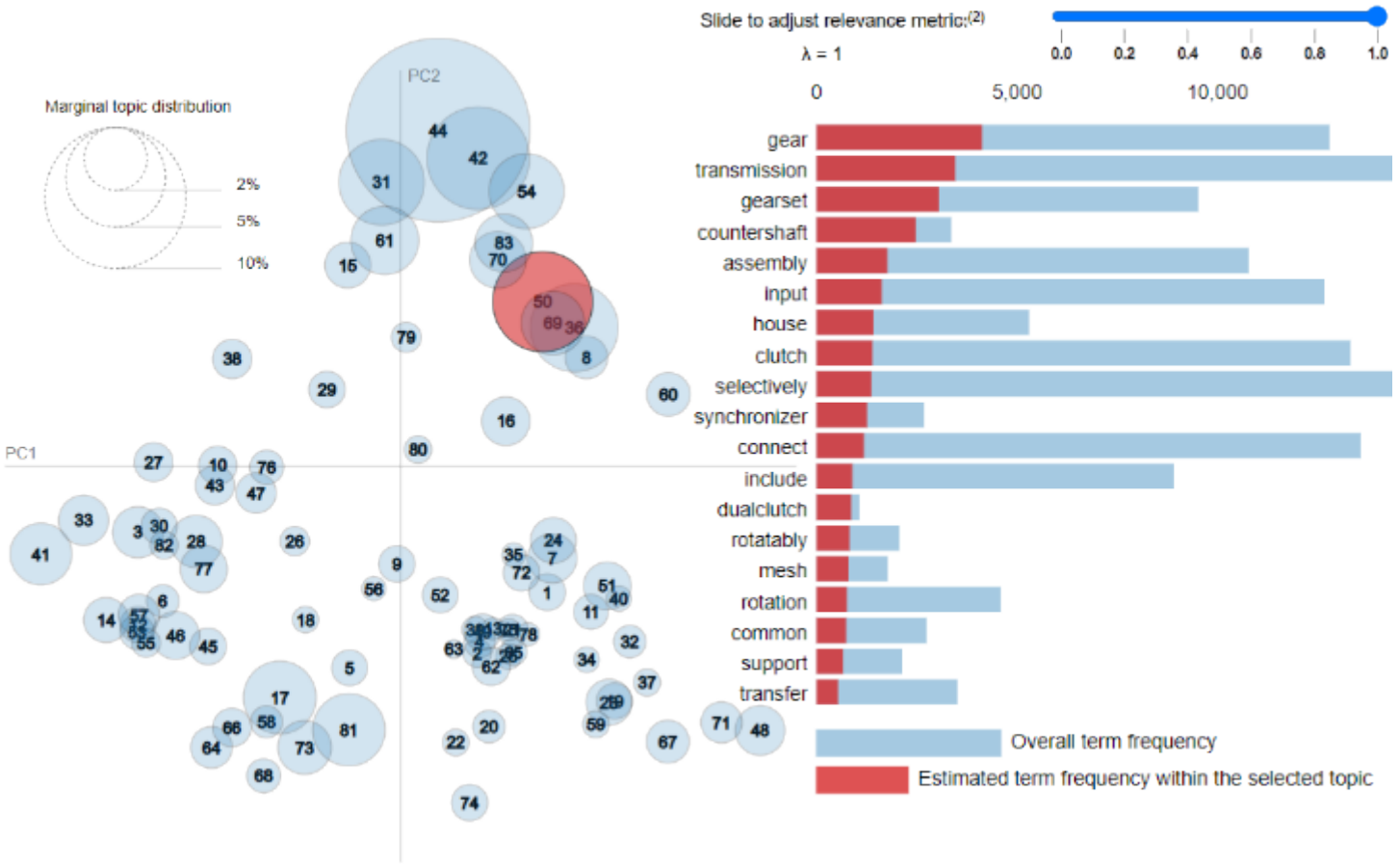

Abbildung 4.1: Interthematische Distanzkarte erstellt mit multidimensionaler Skalierung und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 1,0$

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.

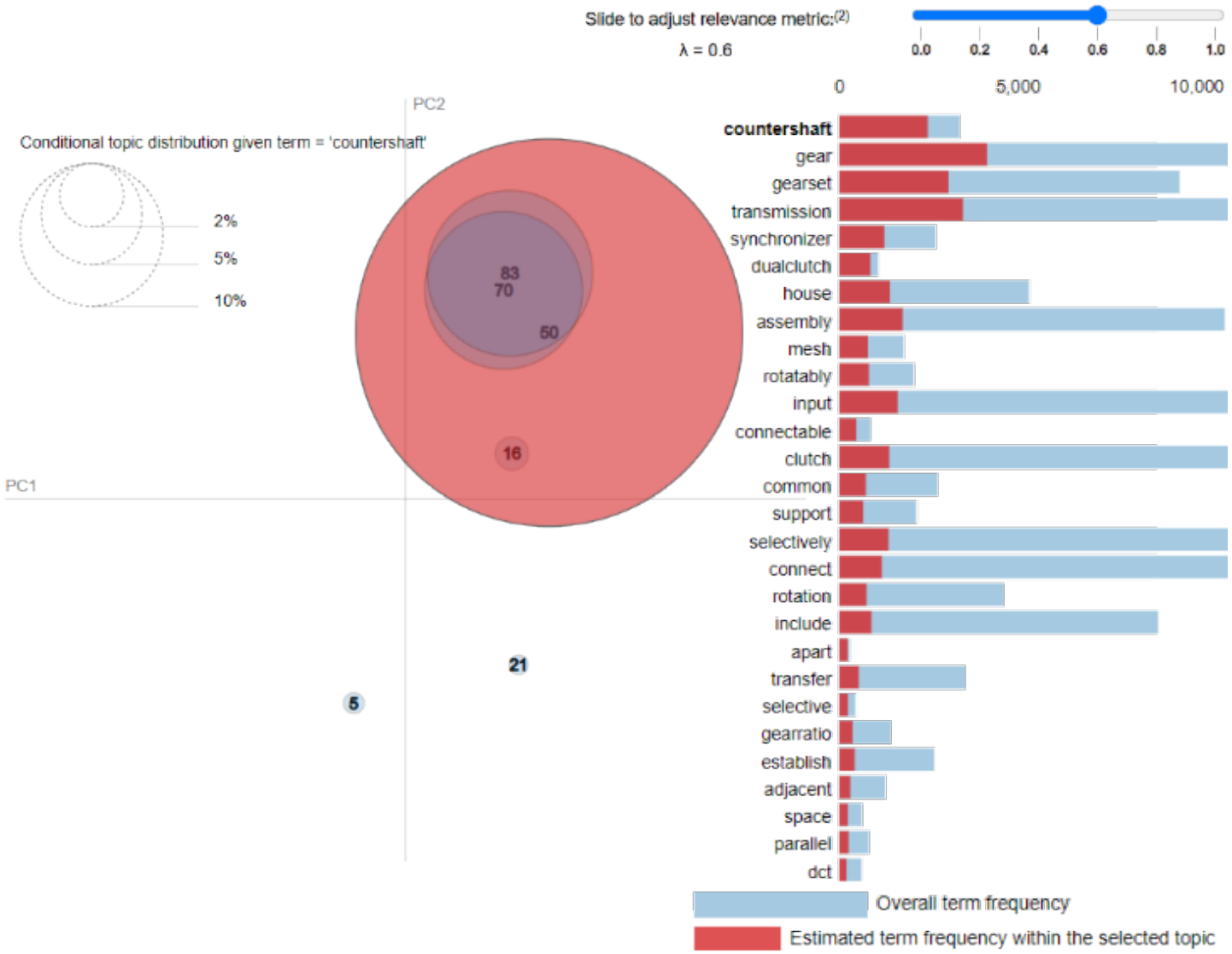

Abbildung 4.2: Interthematische Distanzkarte erstellt mit multidimensionaler Skalierung mit einer Themenverteilung welche von dem Term countershaft abhängt und die relevantesten Terme für das Thema Nummer 50 bei einem $\lambda = 0,6$

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 (*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.

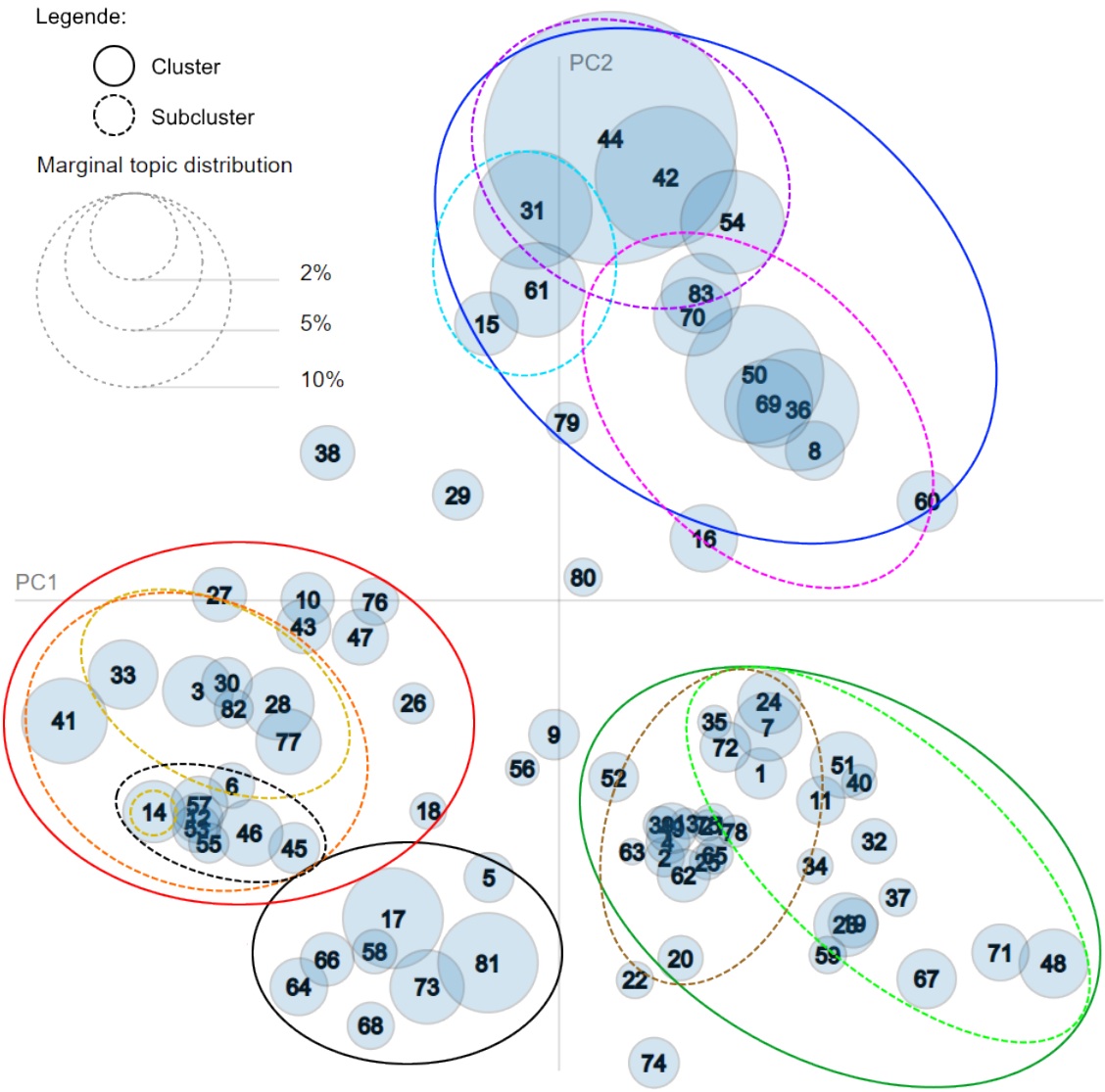

Abbildung 4.3: Interthematische Distanz Karte erstellt mit multidimensionaler Skalierung und qualitativer Methode zur Clusterung der Themen

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 Verwandschaft 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 *montior* 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 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 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.

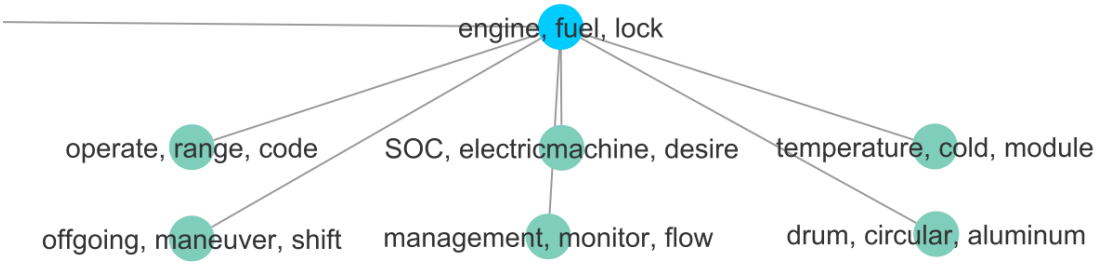

Abbildung 4.4: Der *engine* Ausschnitt des HLDA Baums

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*.

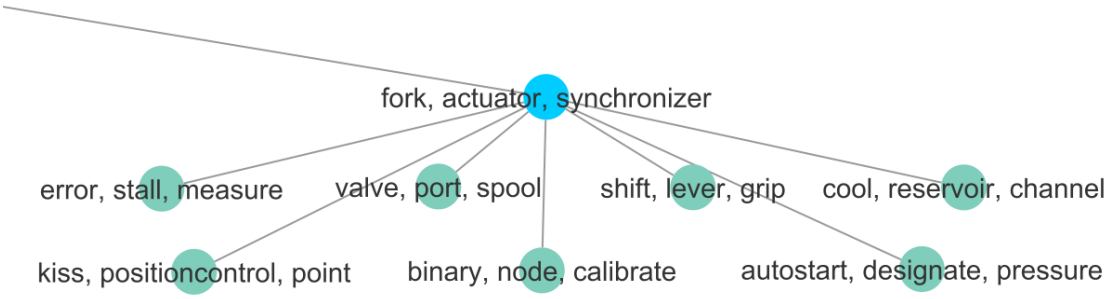

Abbildung 4.5: Der *fork* Ausschnitt des HLDA Baums

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 singt 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.

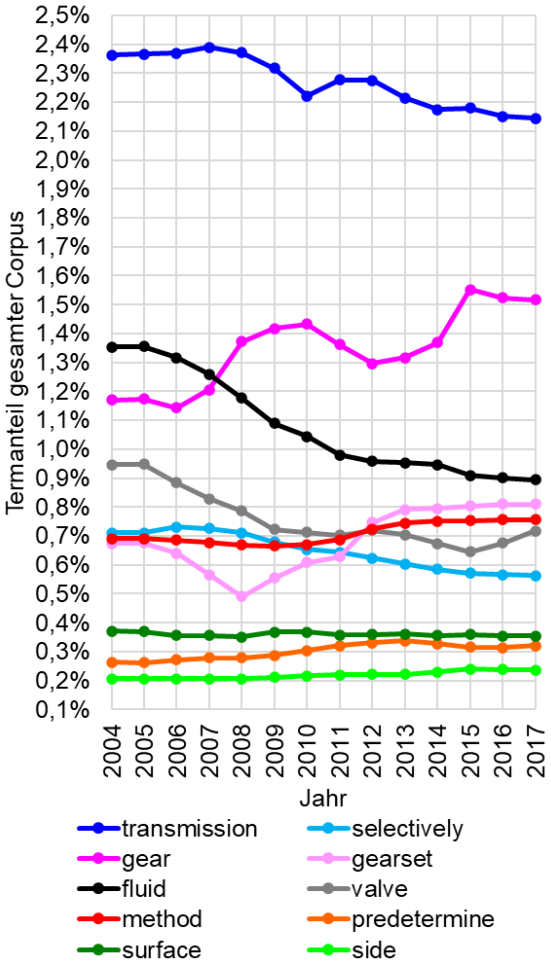

Abbildung 4.6: Trends der Terme die ihren Cluster am besten beschreiben

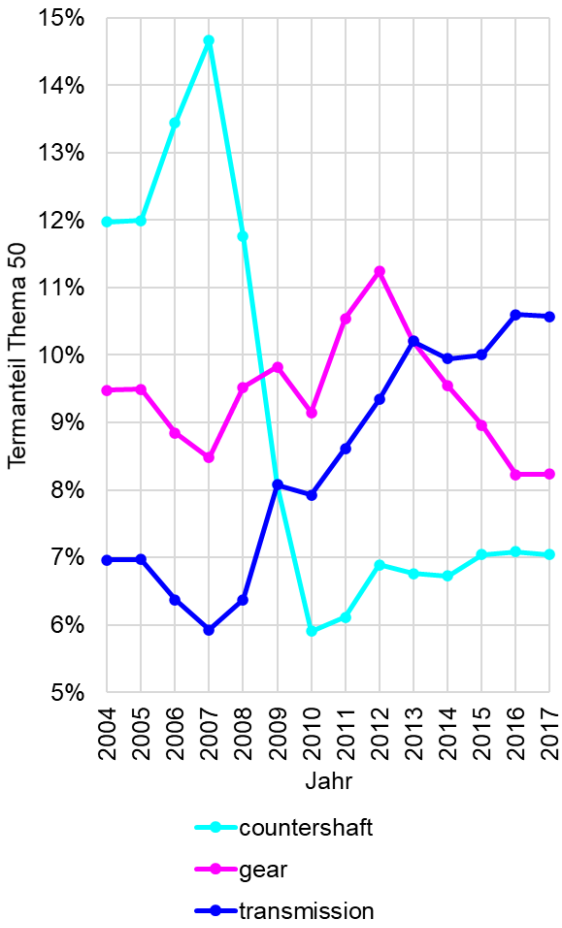

Abbildung 4.7: Trends der 3 Terme die das Thema 50 am besten beschreiben

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* 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 Reglung 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. Währen 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 verstecken 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.

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

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

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 Bigrammmodells deutlich besser sind als die des Unigrammmodells. Das Bigrammmodell sollte viel kohärenter und die Themen viel distanzierter sein als die des Unigrammmodells wie in der Tabelle 4.2 abzulesen ist. Die Vergleiche der Distanz für jedes Thema befinden sich im Anhang als Abbildung Unigramme 6.1 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*.

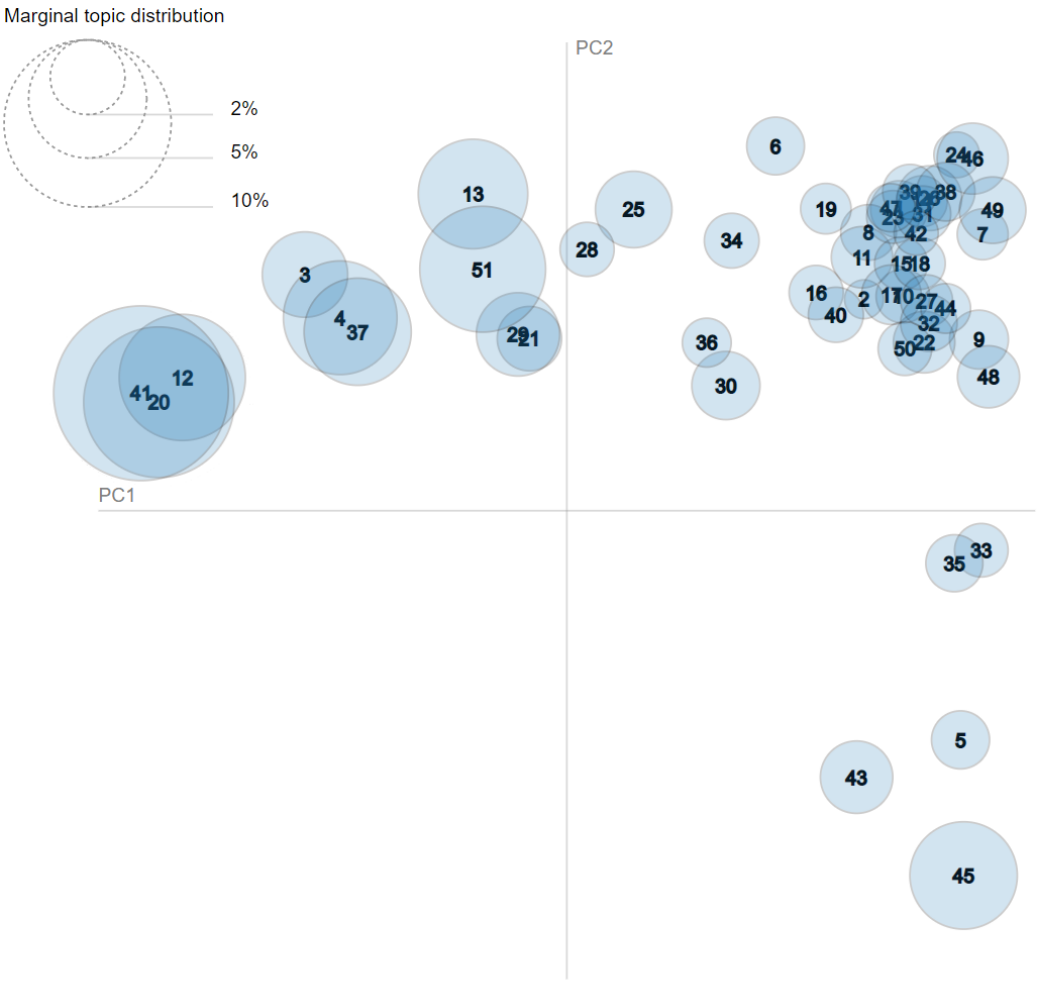

Abbildung 4.8: Interthematische Distanzkarte der Bigramme, erstellt mit multidimensionaler Skalierung bei einem von $\lambda = 1,0$

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. Mit diesen Möglichkeiten könnte sich ein Konkurrent von GM einen Überblick über die Patente und deren Trends schaffen. Der Konkurrent könnte bestehende Patente verbessern oder mit einer Trendanalyse GM bei der Anmeldung von neuen Patenten zuvorkommen, um Lizenzgebühren zu verlangen.

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 Unigrammmen Terme aufgespalten oder zusammengelegt werden würden. Terme wie *dualclutch* könnten in *dual clutch* aufegspalten werden und *binary clutch* sollte *dual clutch* heißen. 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

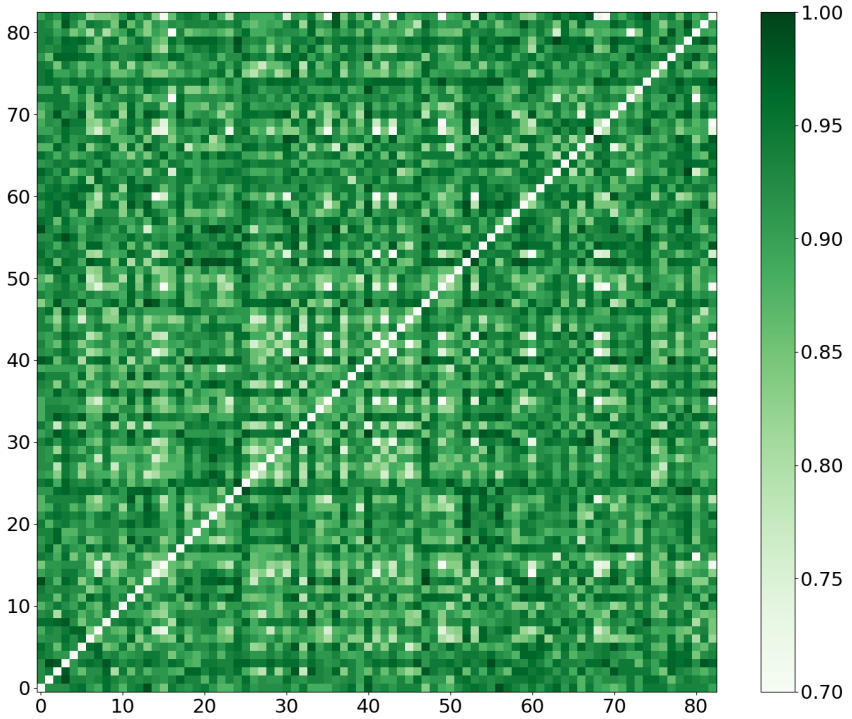

Abbildung 6.1: Distanz zwischen den top 50 Unigrammen der Themen

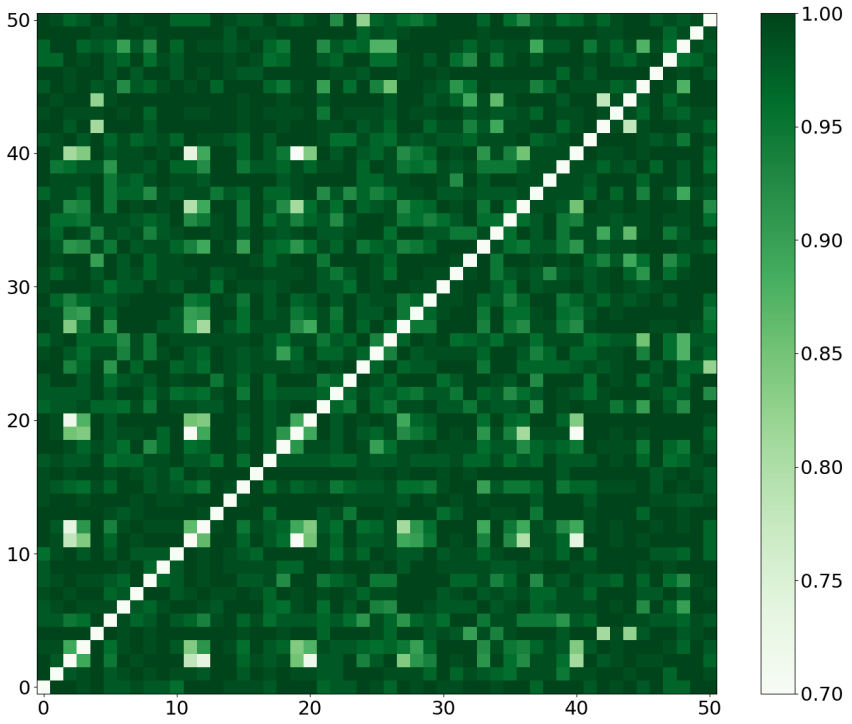

Abbildung 6.2: Distanz zwischen den top 50 Bigrammen der Themen

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, S. 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.

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

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

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

[Kos19] Sven Kosub. „A note on the triangle inequality for the Jaccard distance“. In: *Pattern Recognition Letters* 120 (2019), S. 36–38.

[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.

[OCh18] M. O'Church. *Why agile and especially SCRUM are terrible*. 22. Sep. 2018. URL: https://michaelochurch.wordpress.com/2015/06/06/why-agile-and-especially-scrum-are-terrible/.

[Org04] World Intellectual Property Organization. *WIPO intellectual property handbook: Policy, law and use*. Bd. 489. WIPO, 2004.

[OT17] M. Oppitz und P. Tomsu. *Inventing the Cloud Century: How Cloudiness Keeps Changing Our Life, Economy*. Berlin, Deutschland: Springer, 2017.

[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, S. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[Rei08] T. Reichhardt. *Bedürfnisorientierte Marktstrukturanalyse für technische Innovationen: Eine empirische Untersuchung am Beispiel Mobile Commerce*. Wiesbaden, Deutschland: Gabler Verlag, 2008.

[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.

[Ste+12] Keith Stevens, Philip Kegelmeyer, David Andrzejewski und David Buttler. „Exploring topic coherence over many models and many topics“. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics. 2012, S. 952–961.

[Tan02] A. S. Tanenbaum. *Computer Networks*. New Jersey, USA: Pearson Education Deutsch, 2002.

[Tho10] I. Thomas. „Reliable Digital Identities for SOA and the Web“. In: *Proceedings of the 4th Ph.D. Retreat of the HPI Research School on Service-oriented Systems Engineering*. Hrsg. von C. Meinel, H. Plattner, J. Döllner, M. Weske, A. Polze, R. Hirschfeld, F. Naumann und Giese H. Potsdam, Deutschland: Hasso Plattner Institut, Universität Potsdam, Apr. 2010, S. 61–72.

[WS16] Lothar Walter und Frank C Schnittker. *Patentmanagement: recherche, analyse, strategie*. Walter de Gruyter GmbH & Co KG, 2016.