HOCHSCHULE
RAVENSBURG-WEINGARTEN
UNIVERSITY
OF APPLIED SCIENCES

Analyse von Geschäftsberichten börsennotierter Unternehmen mit Methoden der Textanalyse

Projektarbeit

Studiengang Wirtschaftsinformatik PLUS

Betreuer
Prof. Dr.-Ing. Wolfram Höpken
Prof. Dr. Christian Lazar

Bearbeitet von: Larissa Kazungu-Igumba
Matrikelnummer: 32723

22. August 2023

Inhaltsverzeichnis

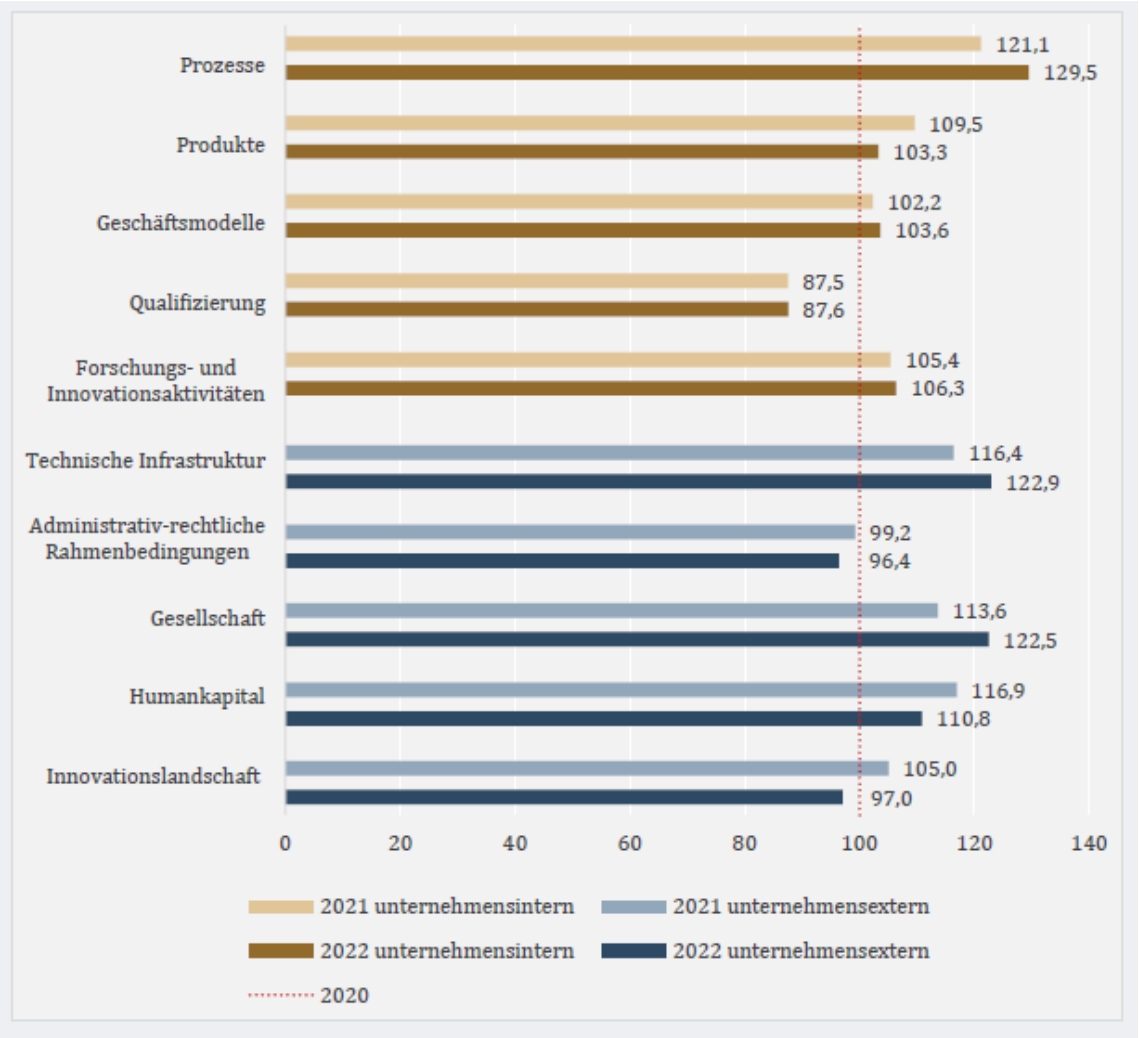
Abbildungsverzeichnis

Tabellenverzeichnis

1 Motivation

Gemäß einer Schätzung der International Data Corporation (IDC) wird erwartet, dass das weltweite Datenvolumen von etwa 33 Zettabyte (ZB) im Jahr 2018 auf 175 ZB im Jahr 2025 ansteigt, was einem jährlichen Wachstum von etwa 27 Prozent entspricht (IWD, 2019). Das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) definiert digitale Produkte, digitale Prozesse, digitale Vernetzung und digitale Geschäftsmodelle als die vier Dimensionen der Digitalisierung (Was ist Digitalisierung?, o. D.). Diese Dimensionen werden durch den nationalen Digitalisierungsindex widerspiegelt, der im Jahr 2020 vom BMWK eingeführt wurde und sich aus 37 externen und internen Indikatoren zusammensetzt. Im Jahr 2020 lag der Indexwert deutschlandweit bei einem Normwert von 100, stieg im Jahr 2021 auf 107,9 und weiter auf den Wert von 108,9 im Jahr 2022.

Abbildung 1: Ergebnisse des Digitalisierungsindex für Deutschland in Kategorien

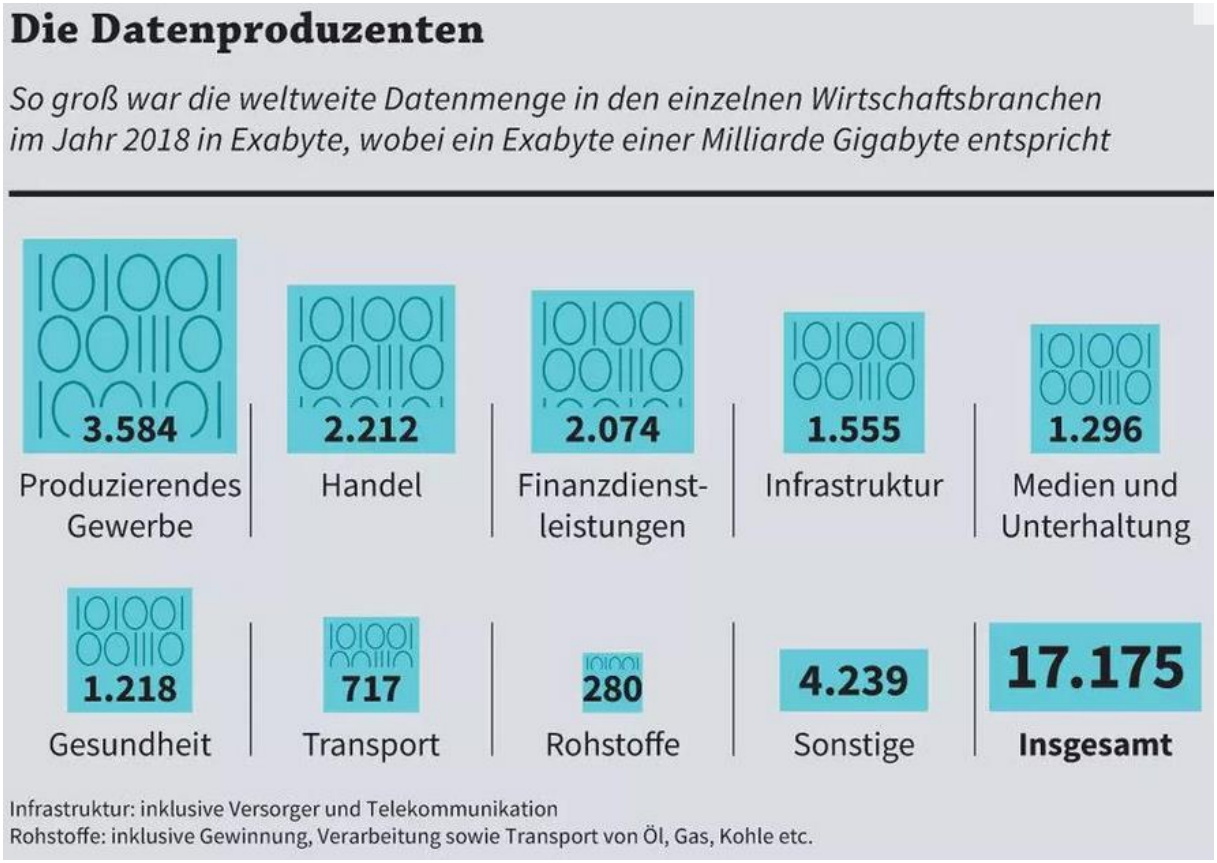


Quelle: Büchel und Engels (2023)

Abbildung 1 stellt die Indexpunkte in Deutschland dar, aufgeschlüsselt nach Kategorien. Es wird deutlich, dass die Kategorie "Gesellschaft" im unternehmensexternen Bereich das stärkste Wachstum verzeichnet. Dies bedeutet, dass die deutsche Gesellschaft eine höhere Affinität zur Digitalisierung entwickelt und vermehrt digitale Produkte und Dienstleistungen nutzt. Immer mehr digitale Geräte wie Smartwatches, Smartphones, Tablets, Smart TVs, virtuelle Sprachassistenten und Smart Homes finden Einzug in den Alltag. Zusätzlich erfreuen sich digitale Dienstleistungen, wie Cloudspeicher oder cloudbasiertes Gaming, und digitale Inhalte, wie YouTube oder Musik-Streaming, einer wachsenden Beliebtheit.
Die Kategorie "Technische Infrastruktur" im unternehmensexternen Bereich verzeichnet ebenfalls eine deutliche Steigerung der Punktzahl. Sie erhöht sich um 6,5 Punkte und erreicht insgesamt 122,9 Punkte. Diese Kategorie ist seit 2020 die am stärksten wachsende unternehmensexterne Kategorie, was hauptsächlich auf die signifikanten Verbesserungen in der Breitbandverfügbarkeit für Unternehmen zurückzuführen ist.
Generell weist die unternehmensinterne Kategorie "Prozesse" den höchsten Wert auf. Sie erhöht sich um 8,3 Punkte und steigt auf 129,5 Punkte. Diese Kategorie umfasst nicht nur den digitalen Reifegrad der internen Unternehmensprozesse, sondern auch die digitale Vernetzung mit anderen Unternehmen. Dies verdeutlicht, dass nicht nur im privaten Umfeld eine zunehmende Nutzung und Generierung von Daten zu beobachten ist, sondern auch in der Wirtschaft.

Abbildung 2, basierend auf Daten der International Data Corporation aus dem Jahr 2018 (IWD, 2019), zeigt, dass das produzierende Gewerbe weltweit das größte Datenvolumen von fast 3,6 ZB in Bezug auf die verschiedenen Wirtschaftsbranchen aufweist. Angesichts der Tatsache, dass etwa 80 % der Daten in der Praxis in unstrukturierter Form vorliegen, beispielsweise als Bilder, Berichte oder Rezensionen, spielt Text Mining eine äußerst wichtige Rolle in Unternehmen (IBM Technology, 2022). Dies ist vor allem auf die enorm wachsende Datenmenge, auch als Big Data bezeichnet, zurückzuführen, die es einer einzelnen Person in angemessener Zeit unmöglich macht, die schnell wechselnde Masse an Daten effizient zu verarbeiten (Litzel & Luber, 2019a).

Abbildung 2: Datenproduzenten 2018



Quelle: IWD (2019)

Die Textanalyse mittels Text Mining ermöglicht die Umwandlung unstrukturierter Dokumente in ein strukturiertes Format, das eine Analyse und eine hochwertige Erkenntnisgewinnung ermöglicht. Durch den Einsatz maschinellen Lernens, Statistik und Data Mining werden in unstrukturierten Daten Textmuster und Trends identifiziert.

In dieser Projektarbeit wird der jährlich erstellte Geschäftsbericht eines börsennotierten Unternehmens untersucht, in dem die Geschäftsentwicklung des vergangenen Geschäftsjahres aufgeführt wird. Diese Geschäftsentwicklung kann positiv, neutral oder negativ ausfallen. Hierfür werden die Geschäftsberichte der 40 DAX-Konzerne aus dem Jahr 2021 mittels Textanalyse genauer untersucht. Dabei sollen die Zusammenhänge zwischen den Geschäftsberichten börsennotierter Unternehmen und ihrer finanziellen Performance erforscht werden. Durch die Anwendung von Textanalyse-Methoden, insbesondere des Topic Modeling, wird untersucht, ob Unternehmen mit positiven oder negativen Jahresabschlüssen ähnliche oder unterschiedliche Themen in ihren Berichten behandeln.

2 Zielsetzung

Die vorliegende Ausarbeitung befasst sich mit der Analyse von Geschäftsberichten börsennotierter Unternehmen mithilfe von Textanalyse-Methoden. Im Fokus steht dabei die Anwendung des Topic Modeling, um Informationen über die Gemeinsamkeiten und Unterschiede in den Geschäftsberichten der 40 DAX-Unternehmen aus dem Jahr 2021 zu erlangen. Ziel dieser Untersuchung ist es, Zusammenhänge zwischen den finanziellen Ergebnissen der Unternehmen und den behandelten Themen in ihren Geschäftsberichten zu identifizieren.

Das Hauptziel dieser Arbeit besteht darin, herauszufinden, ob Unternehmen, die im Jahr 2021 positive Jahresabschlüsse erwirtschaftet haben, ähnliche Themen und Schwerpunkte in ihren Geschäftsberichten aufweisen. Ebenso soll untersucht werden, ob Unternehmen, die negative Abschlüsse verzeichneten, andere Themen behandeln. Hierbei wird die Methodik des Topic Modeling mithilfe von RapidMiner angewandt, um die Geschäftsberichte zu analysieren und relevante Themenclusters zu extrahieren. Die folgenden Forschungsfragen werden in dieser Arbeit behandelt:

- Welche gemeinsamen Themen und Schwerpunkte lassen sich in den Geschäftsberichten der Unternehmen mit positiven Jahresabschlüssen identifizieren?
- Gibt es Unterschiede in den behandelten Themen und Schwerpunkten zwischen den Unternehmen mit positiven Abschlüssen und denen mit negativen Abschlüssen?
- Inwiefern können die identifizierten Themen und Schwerpunkte als Indikatoren für die finanzielle Performance der Unternehmen dienen?

Diese Darlegung trägt zur Erweiterung des Verständnisses über die Beziehung zwischen Geschäftsberichten und finanzieller Performance von börsennotierten Unternehmen bei. Die Ergebnisse dieser Untersuchung können Unternehmen und Investoren dabei unterstützen, relevante Informationen aus den Geschäftsberichten zu gewinnen und potenzielle Zusammenhänge zwischen den behandelten Themen und der finanziellen Performance zu erkennen. Darüber hinaus bietet die Arbeit eine praktische Anwendung der Textanalyse-Methoden, insbesondere des Topic Modeling, im Bereich der Unternehmensanalyse.

Vorliegende Arbeit gliedert sich in mehrere Kapitel, um eine klare Struktur und einen logischen Aufbau zu gewährleisten. Nach der Motivation und Zielsetzung werden im dritten Kapitel die theoretischen Grundlagen zu Geschäftsberichten und im vierten Kapitel zur Textanalyse, insbesondere des Topic Modeling, erläutert. Im fünften Kapitel erfolgt eine Darstellung der verwendeten Methodik, einschließlich der Datenbeschaffung und der Durchführung und Auswertung der Textanalyse mit RapidMiner. Anschließend werden im sechsten Kapitel Verbindungen zur Performance der Unternehmen und den Analyseergebnissen gezogen und diskutiert. Es werden die Zusammenhänge zwischen den identifizierten Topics und dem finanziellen Erfolg der Unternehmen untersucht. Dabei werden mögliche Korrelationen und Beziehungen analysiert und interpretiert. Abschließend erfolgt eine Zusammenfassung der wichtigsten Erkenntnisse sowie ein Ausblick auf mögliche zukünftige Forschungsrichtungen in diesem Bereich.

3 Grundlagen zum Geschäftsbericht

Im folgenden Abschnitt werden grundlegende Begriffe um das Wesen des Geschäftsberichts kurz erklärt. Ergänzende Erläuterungen zum Jahrabschluss und dem IFRS folgen, sowie zum Nutzen der Geschäftsberichtsanalyse allgemein und mittels maschineller Textanalyse (Text Mining).

Der Geschäftsbericht eines Unternehmens ist ein freiwilliges Instrument, das seit 1985 existiert und nicht gesetzlich normiert ist. Er informiert über die Geschäftsentwicklungen des vorangegangenen Geschäftsjahres. Gemäß dem Handelsgesetzbuch (HGB) in Deutschland sind bestimmte Berichtselemente je nach Rechtsform und Größe des Unternehmens, wie Bilanz, Gewinn- und Verlustrechnung (GuV), Anhang, Kapitalflussrechnung, Eigenkapitalspiegel und Lagebericht im elektronischen Bundesanzeiger, offenzulegen. Viele Unternehmen erstellen zusätzlich freiwillig umfangreichere Geschäftsberichte, die weitere Informationen enthalten. Der Geschäftsbericht erfüllt damit sowohl gesetzliche Berichtspflichten als auch die Funktion der externen Kommunikation mit relevanten Stakeholdern wie Aktionären, Banken und Kunden (*Geschäftsbericht Definition*, o. D.).
Gemäß §242 HGB ist der Jahresabschluss ein Einzelabschluss, der die finanzielle Situation und die Ergebnisse einer einzelnen juristischen Einheit, also eines Unternehmens innerhalb einer Konzernstruktur, darstellt. Der Einzelabschluss ist verpflichtend für alle Kaufleute und setzt sich aus Bilanz und GuV zusammen, siehe dazu auch Tabelle 1. Neben seiner Funktion zum Schutz der Gläubiger hat der Einzelabschluss auch eine Ausschüttungsbemessungs- und Besteuerungsfunktion sowie eine Feststellungsfunktion. Es ist wichtig zu beachten, dass lediglich der Einzelabschluss geprüft oder betrachtet wird und nicht der Konzernabschluss (*Geschäftsbericht Definition* | *finanzen.net Wirtschaftslexikon*, o. D.).

Tabelle 1: Aufbau Jahresabschluss

Jahresabschluss	**=**	**Bilanz & GuV (gem. § 242 HGB für alle Kaufleute)**
	+	Anhang (wenn es sich um Kapitalgesellschaft handelt und Publizitätsgesetz greift)

Quelle: eigene Darstellung

Der Konzernabschluss hat - unter Beachtung der Grundsätze ordnungsgemäßer Buchführung (GoB) - eine Informationsfunktion hinsichtlich Vermögens-, Finanz- und Ertragslage und Zahlungsströme eines Konzerns. Tabelle 2 zeigt den Aufbau eines Konzernabschlusses gemäß HGB.

Tabelle 2: Aufbau Konzernabschluss

Konzernabschluss (gem. § 297 I HGB)	**=**	**Konzernbilanz & Konzern-GuV**
	+	Konzernanhang
	+	Kapitalflussrechnung
	+	Eigenkapitalspiegel

Quelle: eigene Darstellung

3.1 International Financial Reporting Standards

Gemäß § 315a des Handelsgesetzbuchs (HGB) ist ein kapitalmarktorientiertes Mutterunternehmen verpflichtet, den Konzernabschluss nach den International Financial Reporting Standards (IFRS) aufzustellen. Für nicht kapitalmarktorientierte Mutterunternehmen besteht hingegen ein Wahlrecht zwischen der Anwendung des HGB und der IFRS ((Alexander Meneikis [Alexander Meneikis], 2014dayx; Rega et al., 2014, S. 4). Gemäß § 290 HGB sind Mutterunternehmen dazu verpflichtet, einen Konzernabschluss zu erstellen, der gemäß § 294 HGB das Mutterunternehmen und alle Tochterunternehmen einschließt, sofern diese nicht gemäß § 296 HGB von der Verpflichtung befreit sind.

Für börsennotierte Unternehmen in der Europäischen Union wurde durch die Verordnung Nr. 1606/2002, die vom Europäischen Parlament und dem Rat der Europäischen Union im Juli 2002 erlassen wurde, festgelegt, dass ab spätestens dem 1. Januar 2005 die IFRS für Konzernabschlüsse gelten: **I**nternational **F**inancial **R**eporting **S**tandards (*IDL Wissenswert: Wer muss nach IFRS bilanzieren?*, 2021). In Deutschland haben alle Unternehmen, die nicht am Kapitalmarkt teilnehmen, gemäß § 315e Abs. 3 HGB das Wahlrecht, ihren Konzernabschluss nach IFRS aufzustellen (*Bilanz nach IAS / IFRS BETROFFENE UNTERNEHMEN*, o. D.). Sowohl kapitalmarktorientierte als auch nicht kapitalmarktorientierte Unternehmen haben beim Einzelabschluss die Möglichkeit, ein Wahlrecht auszuüben (Redaktion RWP, 2023; Rega et al., 2014, S. 4), siehe dazu Tabelle 3.

Tabelle 3: Übersicht Wahlrecht IFRS- und/ oder HGB-Vorgaben

	nicht kapitalmarktorientierte Unternehmen	**kapitalmarktorientierte Unternehmen**
Jahresabschluss	Wahlrecht*	Wahlrecht*
Konzernabschluss	IFRS- oder HGB-Vorgaben	IFRS-Vorgaben

* Pflicht Abschluss nach HGB-Vorgaben, zusätzlich freiwillig ein Abschluss nach IFRS-Vorgaben

Quelle: eigene Darstellung

IFRS sind globale Standards für die Finanzberichterstattung, um Vergleichbarkeit zu gewährleisten. Auf Grund der Globalisierung entstehen immer mehr internationale Geschäftsbeziehungen und grenzübergreifende Transaktionen. Um verlässliche Informationen für die damit verbundenen Geschäftsentscheidungen bereitzustellen, waren einheitliche Rechnungslegungsstandards erforderlich, da der bisherige Flickenteppich aus unterschiedlichen Rechnungslegungsanforderungen dieses Ziel oft nicht erreichen konnte (*Why global accounting standards?*, o. D.).

Das International Accounting Standards Board (IASB) ist für die Festlegung der IFRS verantwortlich (Rega et al., 2014, S. 4). Derzeit sind 36 Standards in Kraft. Gemäß dem Rahmenwerk des IASB, dem sogenannten Conceptual Framework, besteht das Hauptziel der IFRS darin, einen Jahresabschluss zu erstellen, der als Grundlage für wirtschaftliche Entscheidungen dient. Sowohl Unternehmenseigentümer als auch internationale Investoren oder Kleinanleger an der Börse sollten in der Lage sein, die wirtschaftliche Lage eines international tätigen Unternehmens anhand von Unternehmensabschlüssen im Vergleich zu Wettbewerbern zu beurteilen. Ebenfalls haben Arbeitnehmer, Kreditgeber, Lieferanten oder Kunden Interesse an Unternehmensinformationen (Alexander Meneikis, 2014; Thiele, o. D., Folie 13). Die IFRS schaffen eine einheitliche Informationsbasis, um fundierte Entscheidungen zu ermöglichen. Darüber hinaus, legt die Unternehmensleitung durch den Jahresabschluss Rechenschaft über die Ergebnisse ihres Handelns ab (*IFRS - Who uses IFRS Accounting Standards?*, o. D.). Das Ziel der IFRS ist es, weltweit Jahresabschlüsse nach denselben Regeln zu erstellen, um Transparenz, Rechenschaftspflicht und wirtschaftliche Effizienz sicherzustellen (*Why global accounting standards?*, o. D.). Laut der Website des IFRS werden diese Standards in 166 Ländern angewandt (*IFRS - Who uses IFRS Accounting Standards?*, o. D.).

3.2 Geschäftsberichtsanalyse unter Hilfe maschineller Textanalyse

Eine Geschäftsberichtsanalyse, auch als Bilanzanalyse oder Jahresabschlussanalyse bezeichnet, stellt ein Verfahren der Informationsgewinnung und -auswertung dar, das mittels Kennzahlen Erkenntnisse über die aktuelle und zukünftige finanzielle und geschäftliche Lage eines Unternehmens zu gewinnen sucht. Das übergeordnete Ziel besteht darin, ein objektives Gesamturteil über die wirtschaftliche Situation des analysierten Unternehmens zu treffen (Thiele, o. D., Folie 32f). Die Analyse umfasst den Jahresabschluss (bestehend aus Bilanz, Gewinn- und Verlustrechnung, Kapitalflussrechnung, Eigenkapitalspiegel, Segmentberichterstattung und Anhang), den Lagebericht, den Vorstandsbericht sowie sonstige Unternehmensveröffentlichungen und -berichte (*BfJ - Bestandteile des Jahresabschlusses*, o. D.).

Um die finanzielle und geschäftliche Leistung des Unternehmens zu bewerten, wird der Geschäftsbericht in seine einzelnen Bestandteile zerlegt. Die Daten des Jahresabschlusses werden aufbereitet und Vergleiche durchgeführt, um genauere Aussagen über den Erfolg oder Misserfolg des Unternehmens zu ermöglichen. Verschiedene Analysemethoden, wie die Bilanzlesung, um sich einen Überblick und ersten Eindruck der vorliegenden Daten zu verschaffen, der Zahlenvergleiche, um Trends zu erkennen, die Datenumstellung und -gliederung sowie die Bildung von Kennzahlen, werden zur Informationsgewinnung angewandt (Thiele, o. D., Folie 21).

Eine Geschäftsberichtsanalyse kann aufgrund der Komplexität der Unternehmensaktivitäten umfangreich sein und bis zu 400 Seiten umfassen. Die Analyse erfordert einen hohen Grad an Detailgenauigkeit und kann schnell komplex werden („KMPG - Geschäftsberichte lesen und verstehen“, 2014).

Durch den Einsatz von Algorithmen und maschinellen Lernverfahren kann die Auswertung großer Mengen an Textdaten in skalierbarer, konsistenter und objektiver Weise deutlich schneller und effizienter erfolgen (Chen, 2020). Die maschinelle Textanalyse bietet eine hohe Genauigkeit, da sie klar definierte Schlüsselbegriffe und Algorithmen verwendet (Winter, 2023). Sie kann bei der Entdeckung verborgener Themen und Trends über mehrere Seiten oder Abschnitte hinweg unterstützen, Ton und Stimmung bewerten, Gemeinsamkeiten und Unterschiede in den Geschäftsberichten verschiedener Jahre oder Unternehmen identifizieren und automatisierte Zusammenfassungen generieren, die auf wichtige Themen und

Erkenntnisse hinweisen. Unterschiedliche Interessengruppen, wie beispielsweise Investoren oder Kreditgeber, haben unterschiedliche Schwerpunkte und Interessen bei der Analyse des Geschäftsberichts börsennotierter Unternehmen (Thiele, o. D., Folie 13). Durch die flexible Anpassung der Methoden und Parameter kann der Fokus der maschinellen Textanalyse entsprechend angepasst werden.

Im Rahmen dieser Projektarbeit erfolgt eine computergestützte Untersuchung von Geschäftsberichten mithilfe des Topic Modeling Verfahren. Die Analyse zielt darauf ab, Zusammenhänge zwischen den behandelten Themen in den Geschäftsberichten und der finanziellen Performance der Unternehmen aufzuzeigen. Hierbei wird geprüft, ob die identifizierten Topics als Indikatoren für die wirtschaftliche Leistungsfähigkeit der Unternehmen dienen können. Eine vergleichende Betrachtung der Themen zwischen den erfolgreich und weniger erfolgreich performanten Unternehmen ermöglicht es, potenzielle Zusammenhänge und Muster zu erkennen und zu bewerten. Durch die Anwendung des Topic Modeling Verfahren und die Verknüpfung mit der wirtschaftlichen Leistung der Unternehmen wird eine umfassende Analyse durchgeführt, um potenzielle Korrelationen zwischen den behandelten Themen in den Geschäftsberichten und der ökonomischen Unternehmensperformance aufzudecken.

4 Grundlagen zum Text Mining

In diesem Abschnitt wird zuerst Text Mining definiert und von anderen Begriffen abgegrenzt. Das Ziel von Text Mining und die Anwendungsbereiche werden vorgestellt. Es folgt ein Unterpunkt zu den möglichen Datenstrukturen, sowie eine Erläuterung des Analyseverfahrens Topic Modeling zur Datenaufbereitung.

Auf der zweiten International Conference on Knowledge Discovery & Data Mining im Jahr 1996 präsentierten Usama Fayyad, Gregory Piatetsky-Shapiro und Padhraic Smyth ihr Paper mit dem Titel "Knowledge Discovery and Data Mining: Towards a Unifying Framework". In diesem Paper wird der Begriff Knowledge Discovery of Data (KDD) als ein nicht-trivialer Prozess zur Identifizierung gültiger, neuartiger, potenziell nützlicher und letztlich verständlicher Muster in Daten definiert (Simoudis et al., 1996). Die "Wissensentdeckung" bezieht sich demnach auf einen datenanalytischen Prozess, der darauf abzielt, Muster, Trends und Beziehungen aufzudecken, um neues Wissen zu gewinnen. Das zu entdeckende Wissen muss bestimmte Kriterien erfüllen, um als gültig und relevant angesehen zu werden. Es sollte

- *nicht-trivial sein*, das heißt, es sollte über das hinausgehen, was bereits bekannt oder offensichtlich ist.
- *gültig* sein, das heißt, die zugrundeliegenden Daten werden genau widergespiegelt.
- *neuartig* sein, indem es bisher unbekannte Zusammenhänge oder Erkenntnisse enthüllt.
- *potenziell nützlich* sein, um einen praktischen Nutzen oder Mehrwert zu bieten.
- *verständlich* sein, sodass es von Menschen interpretiert und genutzt werden kann, um informierte Entscheidungen zu treffen oder neue Erkenntnisse zu gewinnen.

Data Mining ist ein Bestandteil des KDD-Prozesses, wobei Text Mining als eine Unterkategorie oder Ergänzung des Data Mining betrachtet wird. Text Mining, auch als Textanalyse, Text-Data-Mining oder Knowledge Discovery in Text (KDT) bezeichnet, bezieht sich auf die Anwendung von Data-Mining-Techniken als automatisierten Prozess zur Wissensentdeckung in Textdaten. Bereits Feldman und Dagan stellt 1995

fest, dass in der Realität ein großer Teil der verfügbaren Informationen nicht in strukturierten Datenbanken, sondern in Sammlungen von Text aus verschiedenen Quellen vorliegt. Text Mining umfasst eine Vielzahl von Methoden und Anwendungen, die über die einfache Informationssuche und -extraktion hinausgehen. Dazu gehören unter anderem Textklassifikation, Sentimentanalyse, Textclustering und Topic Modeling. Es geht also darum, implizites Wissen aus großen Mengen an Textdaten zu extrahieren und Themen, Trends, Muster, Stimmungen oder Beziehungen zu identifizieren (Tiedemann, 2021). Die automatisierte Textanalyse ermöglicht die Gewinnung quantitativ und qualitativ hochwertiger Erkenntnisse aus großen Mengen unverarbeiteter Textdokumente (Winter, 2023).
Der Fokus dieser Arbeit liegt auf dem Prozess der Wissensentdeckung, bei dem mithilfe von Text Mining Techniken und Analysemethoden Muster und Erkenntnisse in den vorliegenden Daten identifiziert werden sollen. Dabei wird angestrebt, die genannten Kriterien der Gültigkeit, Neuheit, potenziellen Nützlichkeit und Verständlichkeit des entdeckten Wissens zu erfüllen

4.1 Anwendungsbereiche

Text Mining kann überall dort Anwendung finden, wo Texte als zentrale Arbeitsgrundlage dienen (Tiedemann, 2021). Es dient dazu, die Erfahrungen von Produktnutzern zu verbessern sowie Geschäftsentscheidungen schneller und fundierter zu treffen. Gemäß der IBM-Quelle "Was ist Text-Mining?" (o. D.) sind exemplarische Anwendungsfelder des Text Mining:

- Kundenservice: Durch die Nutzung von Kundenumfragen, Chatbots oder Online-Rezensionen können Kundenrückmeldungen gesammelt werden. Mithilfe von Text Mining und Stimmungsanalyse können die wichtigsten Problembereiche identifiziert werden.
- Risikomanagement: Text Mining kann durch Stimmungsanalysen und die Extraktion von Informationen aus Analystenberichten und Whitepapers Erkenntnisse über Trends und Finanzmärkte liefern.
- Instandhaltung: Text Mining ermöglicht eine umfassende Bewertung des Betriebszustands und der Funktionalität von Produkten und Maschinen. Im Laufe der Zeit automatisiert das Text Mining die Entscheidungsfindung, indem

es Muster identifiziert, die mit Problemen sowie präventiven und reaktiven Wartungsverfahren korrelieren.

- Gesundheitswesen: Text Mining bietet eine automatisierte Methode zur Extraktion wertvoller Informationen aus medizinischer Fachliteratur, insbesondere durch das Clustering von Informationen.
- Spam-Filter: Spam-E-Mails dienen Hackern häufig als Eingangstor für die Infizierung von Computersystemen mit schädlicher Software. Text Mining bietet eine Methode, um diese E-Mails im Posteingang zu filtern und auszuschließen.

Weitere Anwendungsfelder von Text Mining finden sich laut der Wirtschaftsprüfungs- und Beratungsgesellschaft Deloitte im Erfassen und Bewerten von Stammdaten, im Herausfiltern spezifischer Vertragsbedingungen, im Gruppieren von Dokumenten in unterschiedliche Kategorien, im Generieren von Buchungssätzen auf Basis eingelesener Belege oder im Prüfen aktueller Newsfeeds (*Text Mining: Neue Chance für Unternehmen*, o. D.).
Es wird sichtbar, dass es sich um eine äußerst effektive Technologie handelt, die breite Anwendungsmöglichkeiten bietet und schnell und kostengünstig Einblicke in Trends, Beziehungen und Stimmungen aus großen Datenmengen liefert. Sie wird bereits in verschiedenen Wirtschaftsbereichen erfolgreich eingesetzt.

4.2 Datenstrukturen

Digitale Daten weisen verschiedene Strukturen auf und gemäß Naeem (2023) können folgende Datenstrukturen auftreten:

- Unstrukturierte Daten: Diese Daten haben kein vordefiniertes Datenformat und liegen in ihrer Rohform vor. Dabei kann es sich um Text aus Quellen wie Social-Media-Posts, Chats, Satellitenbilder, IoT-Sensordaten, Produktbewertungen oder beispielsweise um Video- und Audiodateien handeln.
- Strukturierte Daten: Diese Daten sind formatiert und in ein definiertes Datenmodell umgewandelt. Strukturierte Daten besitzen eine normalisierte Form und können unter anderem in relationalen Datenbanken (zeilen- und spaltenorientierten) gespeichert werden. Dies erleichtert die Verwendung von Machine-Learning-Algorithmen oder SQL-Abfragen.

- Halb-, schwach, semi- bzw. teilweise strukturiert Daten: Diese Daten sind eine Mischung aus strukturierten und unstrukturierten Datenformaten. Sie sind zwar relativ geordnet, aber nicht ausreichend strukturiert, um den Anforderungen einer relationalen Datenbank zu genügen. Beispiele sind E-Mails, XML-, JSON- oder HTML-Dateien.

Die heutzutage in Unternehmen anfallenden Daten bestehen größtenteils aus unstrukturierten und halbstrukturierten Daten. Laut einer Studie des IDC werden bis 2025 voraussichtlich fast 80 % aller weltweiten Daten in unstrukturierter Form vorliegen (Schinko, 2021). Die Verarbeitung und Analyse dieser Daten bergen ein enormes Potenzial für Unternehmen, das mithilfe von Text Mining Methoden erschlossen werden kann.

4.3 Topic Modeling

Topic-Modelle basieren auf der Annahme, dass jedem Wort in einem Text ein spezifischer Themenbereich zugeordnet werden kann. Durch die Analyse der Verteilung thematisch zusammengehöriger Wörter lassen sich die zugrunde liegenden Themenbereiche, auch als Topics bezeichnet, ableiten. Dieser Prozess beruht auf statistischen Beobachtungen von Regelmäßigkeiten in der sprachlichen Struktur des Dokuments und ermöglicht die Bildung inhaltlich interpretierbarer Cluster auf der Grundlage des gemeinsamen Vorkommens von Wörtern in Dokumenten (Biemann, 2022, S. 270). Da Topic Modeling ein probabilistisches, unüberwachtes Verfahren ist, hat man keinen direkten Einblick in den automatischen Prozess der Modellierung und selbst die Auswahl der Textsegmente erfolgt zufällig. Obwohl die gegebenen Parameter manuell bestimmt und die Ergebnisse analysiert werden können, basieren die Ergebnisse des Topic Modeling Verfahrens auf komplexen Wahrscheinlichkeitsberechnungen. Daher ist eine exakte Reproduktion eines Topic Modeling Verfahren, selbst bei gleicher Einstellung der Parameter, nicht möglich. Dennoch ist oft eine hohe Ähnlichkeit zwischen den entstehenden Topics zu beobachten (Horstmann, 2018).

Im Kontext dieser Projektarbeit wird Text Mining in Form von Topic Modeling mithilfe der Latent Dirichlet Allocation (LDA) angewandt. Die LDA ist ein generatives, statistisches Wahrscheinlichkeitsmodell, das in Kapitel 5.4.1 näher erläutert wird.

5 Bearbeitung der Geschäftsberichte mit RapidMiner

In diesem Kapitel liegt der Fokus auf der essentiellen Phase der Datenerhebung und -aufbereitung im Rahmen der Analyse von Geschäftsberichten. Zunächst werden geeignete Textdateien, in diesem Kontext Geschäftsbericht, identifiziert. Anschließend wird eine Prozessübersicht dargelegt, um ein besseres Verständnis für den Ablauf und den Mehrwert jeder Phase zu vermitteln. Ein entscheidender Schritt dieses Prozesses ist die Datenverfeinerung, um die Rohdaten in eine geeignete Form für die Analyse zu überführen. Im Mittelpunkt des Kapitels steht zudem die Wissensermittlung mittels Topic Modeling unter Verwendung von Latent Dirichlet Allocation (LDA). Dieser Ansatz ermöglicht es, verborgene Themen und Zusammenhänge in den Geschäftsberichten aufzudecken. Der Prozess der Auswertung wird beleuchtet und gewonnenen Erkenntnisse präsentiert.

Im Anhang A: Ergänzende Informationen“ befindet sich eine Kurzbeschreibung der Umgebung RapidMiner und diesbezüglich Erläuterungen zu fundamentalen Begriffen.

5.1 Identifizierung der Daten

In dieser Ausarbeitung werden die Geschäftsberichte von börsennotierten Unternehmen in Deutschland analysiert. Die Analyse konzentriert sich auf die 40 Konzerne im DAX. Der DAX-Index ist das bekannteste deutsche Börsenbarometer und verfolgt die Wertentwicklung der größten 40 Unternehmen des deutschen Aktienmarktes. Diese Unternehmen repräsentieren etwa 80 Prozent der Marktkapitalisierung börsennotierter Aktiengesellschaften in Deutschland. Der DAX ist eine Marke der Qontigo Index GmbH, die zur Gruppe Deutsche Börse gehört. Die Zusammensetzung des Index wird quartalsweise überprüft, weshalb es möglich ist, dass die betrachteten Unternehmen aus dem Jahr 2021 nicht kontinuierlich Mitglieder des DAX waren. Dennoch bleiben die Unternehmen, auch wenn sie aus dem DAX fallen und ihr Aktienindex sich verschlechtert, weiterhin Teil der DAX-Indexfamilie, zu der auch der MDAX, TecDAX und SDAX gehören, bevor sie die DAX-Indexfamilie vollständig verlassen müssen (*Gruppe Deutsche Börse - DAX-Index – Benchmark und Barometer für die deutsche Wirtschaft*, o. D.).

Seit dem 20. September 2021 umfasst der DAX insgesamt 40 Konzerne. Zuvor waren es nur 30, jedoch wurde der MDAX um zehn Konzerne verkleinert, die in den DAX

überführt wurden. Ziel dieser Neuerung war es unter anderem, die Breite der deutschen Wirtschaft im Leitindex besser abzubilden (tagesschau.de et al., 2022).
Für diese Analyse wurden die Geschäftsberichte der einzelnen DAX 40-Konzerne im Portable Document Format (PDF) von den jeweiligen Unternehmenswebseiten heruntergeladen. Tabelle 4 gibt einen Überblick über die betrachteten Konzerne.

Tabelle 4: DAX 40-Konzerne aus 2021

	Unternehmen	DAX-Einstieg	DAX-Ausstieg	DAX-Wiedereinstieg	Präsentations-sprache
1	Adidas AG	Juni 1998	-	-	Deutsch
2	Airbus SE	Sept. 2021	-	-	Englisch
3	Allianz SE	Juli 1988	-	-	Deutsch
4	BASF SE	Juli 1988	-	-	Deutsch
5	Bayer AG	Juli 1988	-	-	Deutsch
6	Beiersdorf AG	Okt. 2021	März 2022	Juni 2022	Deutsch
7	BMW AG	Juli 1988	-	-	Deutsch
8	Brenntag	Sept. 2021	-	-	Deutsch
9	Continental AG	2003	Dez. 2008	Sep. 2012	Deutsch
10	Covestro AG	März 2018	-	-	Deutsch
11	Delivery Hero SE	Aug. 2020	Juni 2022		Deutsch
12	Deutsche Bank AG	Juli 1988	-	-	Deutsch
13	Deutsche Börse AG	Dez. 2002	-	-	Deutsch
14	Deutsche Post AG	März 2001	-	-	Deutsch
15	Deutsche Telekom AG	Nov. 1996	-	-	Deutsch
16	E.ON SE	Juni 2000	-	-	Deutsch
17	Fresenius SE	März 2009	-	-	Deutsch
18	Fresenius Medical Care AG	Sep. 1999	-	-	Deutsch
19	Hannover Rück	März 2022	-	-	
20	Heidelberg Cement AG	Juni 2010	-	-	Deutsch
21	HelloFresh SE	Sept. 2021	Sept. 2022		Deutsch
22	Henkel AG	Juli 1988	-	-	Deutsch

23	Infineon Technologies AG	Sept. 2009	-	-	Deutsch
24	Mercedes-Benz Group AG	Dez. 1998	-	-	Deutsch
25	Merck KGaA	Juni 2007	-	-	Deutsch
26	MTU Aero Engines AG	Sep. 2019	-	-	Deutsch
27	Münchner Rück AG	Sep. 1996	-	-	Deutsch
28	Linde plc	Okt. 2018	März 2023		Englisch
29	Porsche Automobil Holding SE	Sep. 2021	-	-	Deutsch
30	Puma SE	Sep. 2021	Dez. 2022	-	Deutsch
31	Qiagen N.V.	Sep. 2021	-	-	Englisch
32	RWE AG	Juli 1988	-	-	Deutsch
33	SAP SE	Sep. 1995	-	-	Deutsch
34	Sartorius AG	Sep. 2021	-	-	Deutsch
35	Siemens AG	Juli 1988	-	-	Deutsch
36	Siemens Energy AG	März 2021	März 2022	Sept. 2022	Deutsch
37	Siemens Healthineers AG	Sep. 2021	-	-	Deutsch
38	Symrise AG	Sep. 2021	-	-	Deutsch
39	Volkswagen AG	Juli 1988	-	-	Deutsch
40	Vonovia SE	Sep. 2015	-	-	Deutsch
41	Zalando SE	Sep. 2021	-	-	Deutsch

Legende
10 DAX-Erweiterungen im September 2021
Geschäftsbericht auf Englisch

Quelle: in Anlehnung an DAX 40 Liste (o. D.)

Im Verlauf des Jahres 2021 kam es zu Veränderungen innerhalb der DAX 40 Konzerne, wodurch 41 Konzerne in Tabelle 4 aufgeführt sind. Allerdings werden drei Geschäftsberichte von der Analyse ausgeschlossen, da sie auf Englisch verfasst sind: Airbus SE, Linde plc und Qiagen N.V. Insgesamt liegen der Analyse somit 38

Geschäftsberichte vor. Im Rahmen der Datenverarbeitung erfolgt eine Aufbereitung, bei der die Daten in ein einheitliches Format überführt werden. Um eine Verarbeitung in RapidMiner zu ermöglichen, musste der schreibgeschützte Status bei 15 Geschäftsberichten entfernt werden.

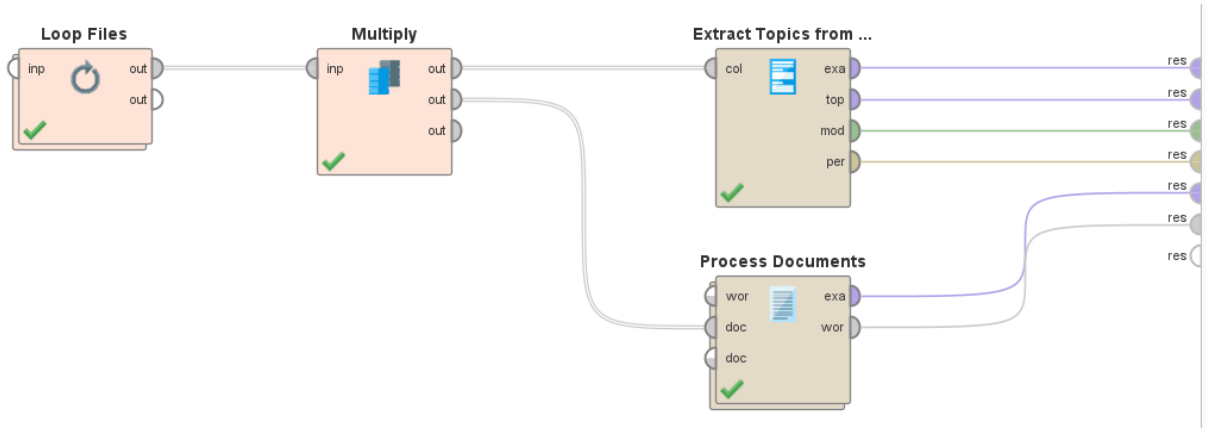
5.2 Prozessübersicht

Für Computerprogramme stellen die Verarbeitung und Analyse unstrukturierter Daten eine große Herausforderung dar. Um diese handhabbar zu machen, muss zunächst eine gewisse Struktur aus den Daten extrahiert werden (Litzel & Luber, 2019b). Bei der Aufbereitung der Daten in RapidMiner können verschiedene Methoden eingesetzt werden. Zum einen gibt es linguistische Methoden, die dem Natural Language Processing (NLP) entsprechen. Beim linguistisch basierten Text Mining werden die Prinzipien der natürlichen Sprachverarbeitung auf die Analyse von Wörtern, Phrasen und Syntax angewendet (IBM Technology, 2022). Hierzu gehören beispielsweise die Identifizierung von Satzstrukturen, die Erkennung von semantischen Beziehungen zwischen Wörtern, die Klassifizierung von Texten nach Themen oder die Extraktion von Entitäten, wie Personen, Orten oder Organisationen.
Zum anderen werden statistische Methoden verwendet, die Frequenzberechnungen nutzen, um verwandte Begriffe abzuleiten (IBM Technology, 2022). Dabei dominieren die linguistischen Methoden, da sie besonders gut für die Arbeit mit halb- oder unstrukturierten Daten geeignet sind (Tiedemann, 2021) und ein zuverlässigeres Verständnis der Sprache ermöglichen, indem Mehrdeutigkeiten aufgelöst werden (IBM Technology, 2022). Es ermöglicht komplexe Informationen aus Textdaten zu gewinnen und inhaltliche Erkenntnisse zu generieren. Gemäß Sumathy und Chidambaram (2013, S. 29) besteht die Analyse aus

- Phase I, der Datenverfeinerung (Kapitel 5.3) und
- Phase II, der Wissensermittlung (Kapitel 5.4).

Die Datenverfeinerung umfasst den Import der Daten und deren Vorverarbeitung (Operator *Loop Files* mit Subprozess). Die Wissensermittlung analysiert den Datensatz anhand bestimmter Kriterien (Operatoren *Extract Topics from Documents (LDA)* und *Process Documents*).

Abbildung 3: externe Prozessübersicht RapidMiner



Quelle: eigene Darstellung in RapidMiner

Der Prozess besteht aus verschiedenen externen Prozessen, die teilweise innen liegende Subprozesse beinhalten. Abbildung 3 veranschaulicht den gesamten externen Prozess in der Entwurfsansicht der grafischen Benutzeroberfläche:

- Loop Files
  - Read Document
  - Tokenize
  - Transform Cases
  - Filter Stopwords (German)
  - Filter Stopwords (Dictionary)
  - Stem (German)
  - Filter Tokens (by Length)
- Multiply
- Extract Topics from Documents (LDA)
- Process Documents

Die Bedeutung und Einstellungen der einzelnen Operatoren werden in den Kapitel 5.3. und 5.4 detailliert erläutert

5.3 Phase I: Datenverfeinerung

Die gesammelten Textdaten aus den Geschäftsberichten werden aufbereitet, um sie für die bevorstehende Analyse gezielt vorzubereiten. Nachfolgende Schritte schaffen eine solide Grundlage, um verlässliche und aussagekräftige Ergebnisse zu erzielen.

5.3.1 Datenimport

Um das separate Hochladen der 38 PDF-Dateien in das Repository des RapidMiners zu vermeiden, wird der *Loop Files* Operator verwendet, der mehrere Dateien gleichzeitig importieren und verarbeiten kann. Im Reiter *Parameters* wird der Dateipfad unter *directory* festgelegt. Der Parameter *filter type* bestimmt, wie die Dateinamen gefiltert werden sollen, und bleibt standardmäßig auf "global" eingestellt. Der Parameter *skip inaccessible* wird aktiviert, um Dateien zu ignorieren, auf die nicht zugegriffen werden kann, und den Operator fortzusetzen. Zusätzlich ist *enable macros* angekreuzt, was die folgenden drei Makros jeder Datei für die Ergebnisdarstellung einbezieht: Dateiname, Dateityp und Ordnername. Zudem wird bei *enable parallel execution* ein Haken gesetzt, um die parallele Ausführung der inneren Prozesse des *Loop Files* Operators zu ermöglichen. Der Operator führt für jede Datei den darin enthaltenen Subprozess durch (siehe Kapitel 5.3.2). Das Endergebnis wird dann an den äußeren Prozess weitergegeben.

An den *Loop Files* Operator schließt sich der Operator *Multiply* an. Es entstehen zwei unabhängige Kopien der eingelesenen und vorverarbeiteten Dokumente. Hierdurch können im weiteren Verlauf des Prozesses zwei unterschiedliche Analyseverfahren auf den Datensatz angewandt werden, ohne dass sich diese gegenseitig beeinflussen.

5.3.2 Datenvorverarbeitung (Subprozess)

Die Vorverarbeitungsphase stellt eine entscheidende Komponente für die Effizienz und den Erfolg des Text Mining Prozesses dar. Sie bildet den inneren Prozess (Subprozess) des *Loop Files* Operators. In dieser Phase werden die Daten mithilfe verschiedener Operatoren aufbereitet, um die Datenbank so zu strukturieren, dass sie weiterverarbeitet werden kann und wertvolle Informationen extrahiert werden können. Die in diesem Kapitel behandelten Operatoren gehören zur Textverarbeitungserweiterung, die auf dem RapidMiner-Marktplatz verfügbar ist. Abbildung 4 verdeutlicht den inneren Prozess des *Loop Files* Operators auf der grafischen Benutzeroberfläche.

Abbildung 4: Innerer Loop Files Prozess (Subprozess)

Quelle: eigene Darstellung in RapidMiner

Die korrekte Platzierung der Operatoren in der Ausführungsreihenfolge spielt eine entscheidende Rolle für die Funktionalität des Prozesses. Folgende Erklärungen wurden mit Hilfe des *Operator Reference Guide* der RapidMiner Documentation erstellt (https://docs.rapidminer.com/10.1/studio/operators/).

Zunächst erfolgt die Datenextraktion durch den Operator *Read Document*. Der Parameter *extract text only* wird aktiviert, damit ausschließlich reine Textdaten aus den Dokumenten extrahiert werden, während andere nicht-textuelle Elemente, wie Formatierungen, Bilder oder Tabellen, nicht berücksichtigt werden. Des Weiteren wird der Parameter *content type* auf “pdf“ festgelegt. Die Codierung wird mit der Option “SYSTEM“ unter *encoding* auf der Standardeinstellung beibehalten. Der Typ der Codierung hängt von der Sprache der Dateien ab. Bei chinesischen oder koreanischen Schriftzeichen muss beispielsweise ein anderer Codierungstyp gewählt werden.
Im zweiten Schritt wird der Operator *Tokenize* angewandt. Dieser teilt den Text der Dokumente in eine Sequenz von Tokens auf. Durch die Auswahl des Parameters *mode* als “non-letters“ dienen Trennzeichen wie Leerzeichen oder Bindestriche, also alle Zeichen, die keine Buchstaben sind, als Trennungsbereiche. Somit werden Grammatikregeln ignoriert und Sätze werden in einzelne Wörter aufgeteilt, die als Tokens bezeichnet werden. Da die Grammatik keine Rolle mehr spielt, können die Tokens in beliebiger Reihenfolge angeordnet werden und gleiche Wörter werden, wie in einer Tabelle, untereinander platziert. Dadurch ergibt sich eine Struktur, in der die Plattform RapidMiner Wörter zählen oder anderweitig verarbeiten kann.
Anschließend folgt der Operator *Transform Cases*, der alle Wörter oder Tokens einheitlich in Klein- oder Großbuchstaben umwandelt. In diesem Fall wird der Parameter *transform to* auf den Wert “lower case“ gesetzt, um sicherzustellen, dass alle Tokens in Kleinbuchstaben geschrieben werden. Dadurch wird gewährleistet, dass ein Wort, das einmal in Kleinbuchstaben und ein anderes Mal in Großbuchstaben vorkommt, gleichwertig behandelt wird und später als dasselbe Wort gezählt wird.

Mittels des Operators *Filter Stopwords (German)* werden häufig vorkommende Wörter wie Artikel, Präpositionen und Pronomen entfernt, die in einer vordefinierten Stoppwortliste für deutsche Sprache enthalten sind. Beispiele für solche Stoppwörter sind "an", "der","und“ und "wir". Das Entfernen dieser Stoppwörter ist notwendig, um eine Verzerrung der Ergebnisse zu vermeiden, da sie zwar häufig vorkommen, aber keine inhaltliche Relevanz für den Text haben. Die Standardeinstellung wird belassen. Im nächsten Schritt wird ein benutzerdefinierter Stoppwortfilter angewandt. Hierbei wird der Operator *Filter Stopwords (Dictionary)* verwendet, um den Textkorpus zu durchlaufen und jedes Wort (Token) mit den Einträgen im Wörterbuch abzugleichen. Wenn ein Wort mit einem Eintrag übereinstimmt, wird es aus dem Text entfernt. Das Ergebnis ist ein bereinigter Textkorpus. Das zusätzliche Stoppwort-Wörterbuch enthält spezifische Wörter, die typischerweise in Geschäftsberichten auftreten oder in der Liste der häufigsten Wörter des Operators *Process Documents* ganz oben stehen, jedoch für das Topic Modeling Verfahren keinen Mehrwert bieten. Hierzu gehören beispielsweise Unternehmensnamen, Monatsnamen aber auch Begriffe wie "Lagebericht", "Konzern" und "Umsatz", die generell in Geschäftsberichten vermehrt auftreten. Das erstellte Wörterbuch kann im Anhang B eingesehen werden.

Der Operator *Stem (German)* repräsentiert das Stemming Verfahren, durch das Wörter auf ihren ursprünglichen Wortstamm reduziert werden, indem Prä- oder Suffixe entfernt werden. Dadurch verringert sich die Bandbreite der Wortvariationen, indem diverse grammatikalische Formen eines Wortes, etwa Pluralformen, Zeitformen, Verbkonjugationen, auf eine gemeinsame Grundform zurückgeführt werden. Hierdurch kann das Topic Modeling die inhaltliche Bedeutung sowie die zugrundeliegenden Themen präziser erfassen, was die Modellkohärenz steigert. Die Wörter werden nach dem "A Fast and Simple Stemming Algorithm for German Words" von Jörg Caumanns (RapidMiner, o. D.) gefiltert. Dies reduziert die Anzahl der Originalwörter und damit auch die anschließend erforderliche Bearbeitungszeit.

Im weiteren Verlauf wird der Operator *Filter Tokens (by Length)* angewandt, um Wörter zu entfernen, die nach der Tokenisierung eine sehr geringe Anzahl von Zeichen aufweisen. Dieser Operator ermöglicht nochmals eine gezielte Entfernung von Tokens, die keine relevante Aussagekraft für die Themenmodellierung haben. Über die Parameter *min chars* und *max chars* wird die minimale und maximale Länge eines Tokens festgelegt, damit es nicht entfernt wird. Für den Parameter *min chars* wird der Wert fünf gewählt, wodurch Wörter, wie AG, SE, AAA (Triple A Bewertung), BMW oder

BASF, Benz oder KGaA herausfallen. Der Parameter *max chars* wird auf eine sehr hohe Zahl (9999) gesetzt, um sicherzustellen, dass alle verbleibenden Tokens eine Länge von mindestens fünf Zeichen haben.

Die Anwendung der genannten Operatoren trägt zur Verbesserung der Qualität des Verfahrens bei, indem sie die Verarbeitungsgeschwindigkeit optimieren, den Speicherbedarf reduzieren und die Effizienz des Prozesses verbessern. Der Fokus liegt dadurch auf semantisch bedeutsamen Informationen und irrelevante Variationen und Rauschen werde reduziert.

5.4 Phase II: Wissensermittlung

Nach Abschluss der Datenverfeinerung der Dokumente erfolgt im nächsten Schritt die Analyse. Durch das Hinzufügen des Operators *Multiply* entstehen zwei unabhängige Datensätze. Einer davon wird mithilfe des Operators *Extract Topics from Documents (LDA)* weiterverarbeitet und durchläuft somit das Topic Modeling Verfahren. Der andere Datensatz wird anahnd des Operators *Process Documents* einer Wörterfrequenzanalyse unterzogen. Am Ende des Prozesses werden die Ergebnisse präsentiert und interpretiert.

5.4.1 Topic Modeling

Gemäß Tomar (2018) ist die Themenmodellierung ein Teilgebiet des Natural Language Processing (NLP), das dazu dient, ein Textdokument mithilfe mehrerer Themen darzustellen, die die zugrunde liegenden Informationen in einem bestimmten Dokument am besten erklären können. Die Themenmodellierung ist ein Lernverfahren im Bereich “unsupervised learning“ (unüberwachtes Lernen), worunter man laut Vajjala et al. (2020, S. 15) eine Sammlung von maschinellen Lernmethoden versteht, die darauf abzielen, versteckte Muster in gegebenen Eingabedaten, ohne jegliche referenzierte Ausgabe zu finden. Unüberwachtes Lernen arbeitet mit großen Sammlungen von nicht beschrifteten Daten. Das Topic Modeling wird durch den Operator *Extract Topics from Documents (LDA)* realisiert.

Tomar (2018) beschreibt LDA (Latent Dirichlet Allocation) als einen probabilistischen Modellierungsansatz, der Themengebiete (Topics) in Dokumenten identifizieren kann. Es handelt sich um eine Erweiterung der Probabilistischen Latenten Semantischen

Analyse (PLSA), die von David M. Blei im Jahr 2003 entwickelt wurde und auf der Arbeit von Thomas Hoffman aus dem Jahr 1999 basiert.

Der Begriff "latent" bezieht sich darauf, dass die Themen, aus denen das Dokument besteht, zu Beginn nicht bekannt sind, da sie nicht unmittelbar sichtbar sind. Es wird jedoch angenommen, dass diese Themen existieren, da der Text auf der Grundlage dieser Themen erstellt wurde. Im Zusammenhang mit der Themenmodellierung bezieht sich "Dirichlet" auf die Verteilung der Themen in den Dokumenten und die Verteilung der Wörter innerhalb der Themen. "Allocation" bedeutet, dass, sobald die Dirichlet-Verteilung festgelegt wurde, die Themen den Dokumenten und die Wörter den Themen zugeordnet werden. LDA besagt also, dass jedes Wort in jedem Dokument aus einem bestimmten Thema stammt und dass das Thema aus einer Verteilung pro Dokument über Themen ausgewählt wird (Tomar, 2018).

Die Themenmodellierung ist ein häufig angewandtes Text Mining Verfahren zur Entdeckung versteckter semantischer Strukturen in einem Textkorpus (Maheta, 2022) und stellt das Hauptziel dieser Arbeit dar. Der Latent Dirichlet Allocation Algorithmus bietet dafür eine effektive Methodik zur Identifikation von Themen in einer gegebenen Sammlung von Textdokumente. Bei der Anwendung des LDA-Algorithmus wird manuell die Anzahl der Themen in der Dokumentensammlung festgelegt und jedem Wort in der Sammlung automatisch ein entsprechendes Thema zugeordnet. Um jedem Dokument ein spezifisches Thema zuzuweisen, werden die Häufigkeiten der Wörter für jedes Thema gezählt, wodurch das vorherrschende Thema für jedes einzelne Dokument ermittelt wird (Burkov, 2019, S. 183). Dies ermöglicht die Untersuchung der Themen, die die DAX 40-Konzerne im Jahr 2021 bewegten. Als Ausgangspunkt dient eine der erzeugten Kopien des Datensatzes.

Zunächst wird der Parameter *number of topics* der Operators *Extract Topics from Documents (LDA)* auf eine beliebige Zahl *n* festgelegt, was bedeutet, dass *n* Themengebiete aus dem Datensatz generiert werden sollen. Die Anzahl der Durchläufe des Lernprozesses des LDA-Algorithmus wird standardmäßig über den Parameter *iterations* auf “1.000“ gesetzt. Durch die Durchführung mehrerer Iterationen kann der LDA-Algorithmus die Qualität der Themenextraktion verbessern, indem er schrittweise die optimalen Zuordnungen von Themen zu Dokumenten und von Wörtern zu Themen findet. Zusätzlich wird der Parameter *top words per topic* auf fünf gesetzt, was bedeutet, dass pro Topic mindestens fünf Wörter enthalten sein sollen. Schließlich wird für den Parameter *stopword language* "german" ausgewählt, da alle eingelesenen

Dokumente in deutscher Sprache verfasst sind. Für die Ausgabe besitzt der Operator verschiedene Output-Ports:

- Der Output-Port *exa* generiert ein ExampleSet mit Attributen wie "documentid", "prediction" (Topic-Zuordnung) oder "confidence" (Vertrauen in die Topic-Zuordnung). Hierbei kann man nachvollziehen, welchem Thema das jeweilige Dokument zugeordnet wurde und mit welchem Prozentsatz. Zusätzlich werden Metadaten, wie der Dateityp und der Dateipfad angezeigt.
- Der Output-Port *top* erzeugt ebenfalls ein ExampleSet mit Details zu den einzelnen Themen. Der Operator gibt die fünf am häufigsten gezählten Wörter für jedes Thema zurück.
- Über den Output-Port *mod* erhält man das Topic-Modell. Es werden detaillierte Informationen zu jedem Topic und den darin enthaltenen Tokens angezeigt, beispielsweise die Anzahl der Tokens pro Topic oder die *exclusivity*.
- Der Output-Port *per* stellt die Performance dar. Es werden verschiedene Kriterien geliefert, wie die durchschnittliche Tokenlänge oder die Perplexität.

All diese Output-Ports werden mit dem Port Result (*res*) des Process-Views verbunden und präsentieren ihre Ergebnisse im Result-View, der grafischen Benutzeroberfläche.

5.4.2 Wörterfrequenzanalyse

Der zweite Datensatz, der durch den Operator *Multiply* erzeugt wird, wird mit dem Operator *Process Documents* verknüpft. Dieser Operator generiert Wortvektoren in einem mehrdimensionalen Vektorraum aus einem Textobjekt, um jedem Wort einen eindeutigen Vektor zuzuordnen. Diese Vektoren spiegeln die Bedeutung und die Beziehung der Wörter zueinander im Text wider (*Process documents - RapidMiner Documentation*, o. D.). Ein Wortvektor ist eine numerische Repräsentation von Wörtern, die es Computerprogrammen ermöglichen natürlichere Interaktionen mit menschlicher Sprache zu führen. Anstatt Wörter als Zeichenketten darzustellen, werden sie als Vektoren reeller Zahlen behandelt. Diese Vektoren kodieren die semantische Bedeutung der Wörter, wodurch das Computerprogramm Strukturen innerhalb der Sprache erkennen kann. Auf diese Weise kann dem System natürliche Sprache verständlich gemacht werden (Laasch, B. M., 2018).

Der Operator *Process Documents* verfügt über einen Subprozess, in dem die Daten nicht weiterbearbeitet werden. Daher wird der Input-Port des Subprozesses direkt mit dem Output-Port verbunden.

Der äußere Prozess bietet konfigurierbare Parameter. Im Feld *create word vector* wird ein Haken gesetzt. Für den Parameter *vector creation* wird der Wert “TF-IDF“ ausgewählt. TF-IDF steht für "Term Frequency Inverse Document Frequency" und gibt die relative Wichtigkeit eines Wortes in einem bestimmten Dokument im Verhältnis zur Wichtigkeit des Wortes in allen Dokumenten im Korpus an. Eine detailliertere Betrachtung dazu findet sich im Exkurs zur Berechnung von TF-IDF in **Fehler! Verweisquelle konnte nicht gefunden werden.**5. Neben dem Parameter *vector creation* besteht die Möglichkeit, durch Aktivieren der Option *add meta information*, Metadaten im Result-View anzeigen zu lassen. Der Parameter *prune method* gibt an, auf welche Weise der Datensatz beschnitten werden soll. Es wird festgelegt, ob zu häufige oder zu seltene Wörter bei der Erstellung der Wortliste ignoriert werden sollen und wie die Häufigkeiten angegeben werden. Dies fungiert als ein Filter, der Wörter entsprechend der gewählten Methode aussortiert. In diesem Fall wird "absolute" ausgewählt. Die Felder *prune below absolute* und *prune above absolute* werden verwendet, um den Bereich der Wortliste einzuschränken. Für *prune below absolute* wird der Wert zwei angegeben, was bedeutet, dass alle Wörter, die in weniger als zwei Dokumenten vorkommen, ignoriert werden. Für *prune above absolute* wird eine sehr hohe Zahl (9999) gewählt, um sicherzustellen, dass alle Wörter, die in mehr als zwei Dokumenten vorkommen, erhalten bleiben. Der Parameter *data management* bleibt auf der Standardoption "auto".

Der Operator *Process Documents* hat zwei Output-Ports. Durch den Output-Port *exa* erhält man ein ExampleSet, das alle verbleibenden Tokens auflistet. Jede Zeile repräsentiert ein Dokument, und die Spalten stellen die einzelnen Tokens/Wörter dar. Unterhalb der Tabelle wird jeweils die relative Häufigkeit (Term Frequency) jedes Wortes für das entsprechende Dokument angezeigt. Der Output-Port *wor* erzeugt eine Wortliste, die anzeigt, wie häufig jedes Wort aufgetreten ist und in wie vielen Dokumenten es vorkommt.

Abbildung 5: Exkurs TF-IDF Berechnung

***Exkurs:** TF-IDF Berechnung*

Definition

Die Term Frequency-Inverse Document Frequency (TF-IDF) ist eine statistische Maßnahme zur Bewertung der Relevanz eines Begriffs (Terms) in einem Dokument oder einer Sammlung von Dokumenten. Die Berechnung erfolgt in zwei Schritten: die Berechnung der Term Frequency (TF) und die Berechnung der Inverse Document Frequency (IDF). Das Produkt dieser beiden Werte ergibt den TF-IDF-Wert eines Terms.

Beschreibung

Die Termfrequenz (Term Frequency, TF) TF(t,d) bestimmt die Häufigkeit des Auftretens eines Begriffs t in einem bestimmten Dokument d und setzt diese ins Verhältnis zur Gesamtzahl aller Begriffe in diesem Dokument.

Die inverse Dokumentenhäufigkeit (Inverse Document Frequency, IDF) ermittelt die Seltenheit eines Begriffs in der gesamten Sammlung von Dokumenten, um zu bestimmen, wie wichtig oder selten der Begriff in der gesamten Sammlung ist.

Berechnung

t = Term, Begriff *D* = Gesamtanzahl der Dokumente; Korpus *d* = ein Dokument

TF

Die Termfrequenz beschreibt, wie oft ein Term auftritt. Wenn ein Wort häufig vorkommt, hat es eine höhere TF.

$$\mathrm{TF}(t,d) = \frac{(\text{Gesamtanzahl der Vorkommen des Tokens } t \text{ in Dokument } d)}{(\text{Gesamtanzahl der Tokens in Dokument } d)}$$

IDF

Die inverse Dokumentenhäufigkeit ist ein Maß dafür, wie selten ein Wort in einer Sammlung von Dokumenten vorkommt. Wenn ein Wort in vielen Dokumenten vorkommt, hat es eine niedrigere IDF. Dieser Wert wird anhand des gesamten Korpus berechnet und ist daher konstant für den gesamten Korpus. Die IDF wird

verwendet, um dem Termfrequenz-Wert (TF) eine Gewichtung zu geben. Zur Berechnung der IDF wird zunächst die Dokumentenhäufigkeit (Document Frequency, DF) eines Begriffs benötigt.

$$\text{DF}(t) = \text{(Anzahl der Dokumente } d\text{, in denen der Term } t \text{ auftritt)} + 1$$

Falls der Wert der Dokumentenhäufigkeit eines Begriffs den Wert Null aufweisen sollte, wird zur Vermeidung einer Division durch Null immer der Wert eins addiert. Als nächstes wird das Inverse der Dokumentenhäufigkeit DF(t) gebildet und anschließend der Logarithmus davon genommen wird.

$$\text{IDF}(t) = \log \frac{\text{(Gesamtanzahl der Dokumente)}}{\text{(Anzahl der Dokumente } d\text{,in denen der Term } t \text{ auftritt)}+1} = \log \frac{(|D|)}{\text{DF}(t)}$$

Warum das Inverse?

Das Inverse der Dokumentenhäufigkeit (Document Frequency, DF(t)) wird verwendet, um die Gewichtung der Wörter in einem Dokument zu skalieren. Durch diese Skalierung wird die Bedeutung seltener Wörter im Vergleich zu häufig vorkommenden Wörtern erhöht. Ein Wort mit einer geringen Dokumentenhäufigkeit hat einen höheren IDF-Wert, da es möglicherweise eine größere Bedeutung für den Kontext des Dokuments aufweist. Dem entsprechend haben allgemeine Wörter, die in vielen Dokumenten auftreten, weniger Einfluss auf die Gewichtung.

Warum der Logarithmus?

Wenn der Wert der Begriffshäufigkeit (TF(t,d) eines Wortes in einem Dokument sehr hoch ist, bedeutet das, dass der Begriff häufig in diesem Dokument vorkommt. Ohne die Anwendung des Logarithmus würde die Gewichtung dieses häufig auftretenden Wortes stark ansteigen. Dies kann auf generelles (Fach-) Vokabular zurückzuführen sein, das in allen Dokumenten wiederholt erscheint und nicht durch einfache Filter entfernt wurde. Beispiel für die Arbeit wären Begriffe wie "Konzern" oder Markennamen wie "adidas". Um eine bessere Vergleichbarkeit der Gewichtungen zu ermöglichen und die tatsächliche Relevanz der Wörter im Dokument widerzuspiegeln, wird der Logarithmus angewandt. Durch diese Anwendung werden extreme Unterschiede der TF-Werte gedämpft, die häufig vorkommen und die

Gewichtung besser ausbalanciert. Gleichzeitig erhöht der Logarithmus die Gewichtung seltener Wörter.

TF-IDF

Jeds relevante Wort im Korpus *D* erhält einen TF-IDF-Wert. Dieser sagt aus, wie wichtig das Wort *t* für die Dokumente im Korpus *D* ist.

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) * \text{IDF}(t)$$

Als Informationsquelle dienten folgende Videos:

- codebasics, 2022
- Data Science Garage, 2021
- dbislab, 2020
- *Machine Learning and RapidMiner Tutorials | RapidMiner Academy*, o. D.

Quelle: eigene Darstellung

5.5 Auswertung Topic Modeling Verfahren

Die Auswertung des Topic Modeling Verfahrens bildet den Abschluss der computergestützten Analyse. Zuerst werden die angewendeten Bewertungskriterien beleuchtet, die Aufschluss über die Qualität und Relevanz der identifizierten Themen geben. Anschließend werden die erzielten Ergebnisse präsentiert, die Erkenntnisse über verborgene Strukturen und Zusammenhänge aufdecken.

5.5.1 Bewertungskriterien

Nach der Optimierung des Prozesses anhand verschiedener Operatoren, erfolgt die Durchführung einer Reihe von Iterationen, welche sich in der Anzahl der Topics unterscheiden. Begonnen wird mit fünf Topics, wobei diese Anzahl schrittweise bis zu 20 gesteigert wird. Hieraufhin setzt sich die Steigerung in Intervallen von fünf fort, bis schließlich eine Anzahl von 65 Topics erreicht wird.

Die Ergebnisse des Topic Modeling Verfahrens werden mittels drei Metriken evaluiert. Zum einen anhand der LogLikelihood sowie der Perplexity des Modells. Der LogLikelihood-Wert ist ein Maß für die Anpassungsfähigkeit eines Topic Modeling-Modells an die gegebenen Daten. Grundsätzlich wird eine Erhöhung des LogLikelihood-Werts angestrebt, da dieser darauf hinweist, dass das Modell die

beobachteten Daten besser erklärt. Jener Wert misst die Wahrscheinlichkeit, dass die gegebenen Daten unter Verwendung des Modells generiert wurden. Ein hoher LogLikelihood-Wert indiziert eine Übereinstimmung des Modells mit den Daten und eine effiziente Reproduktion der beobachteten Informationen (Tijare & Rani, 2020). Dadurch wird unter anderem das Kriterium der Gültigkeit, welches Simoudis et al. (1996) bezüglich der Wissensentdeckung definierten. Die Perplexität fungiert als Indikator für die Verallgemeinerungsfähigkeit des Modells, sprich wie gut das Modell die gegebenen Daten beschreibt. Es wird ein niedriger Wert erstrebt, was darauf deutet, dass das Modell eine minimale Unsicherheit in seinen Prognosen aufweist und es effektiv Themen aus den Daten extrahiert. Das wiederum führt zu einem besseren Vorhersagen für nicht gesehene Daten (Tijare & Rani, 2020).

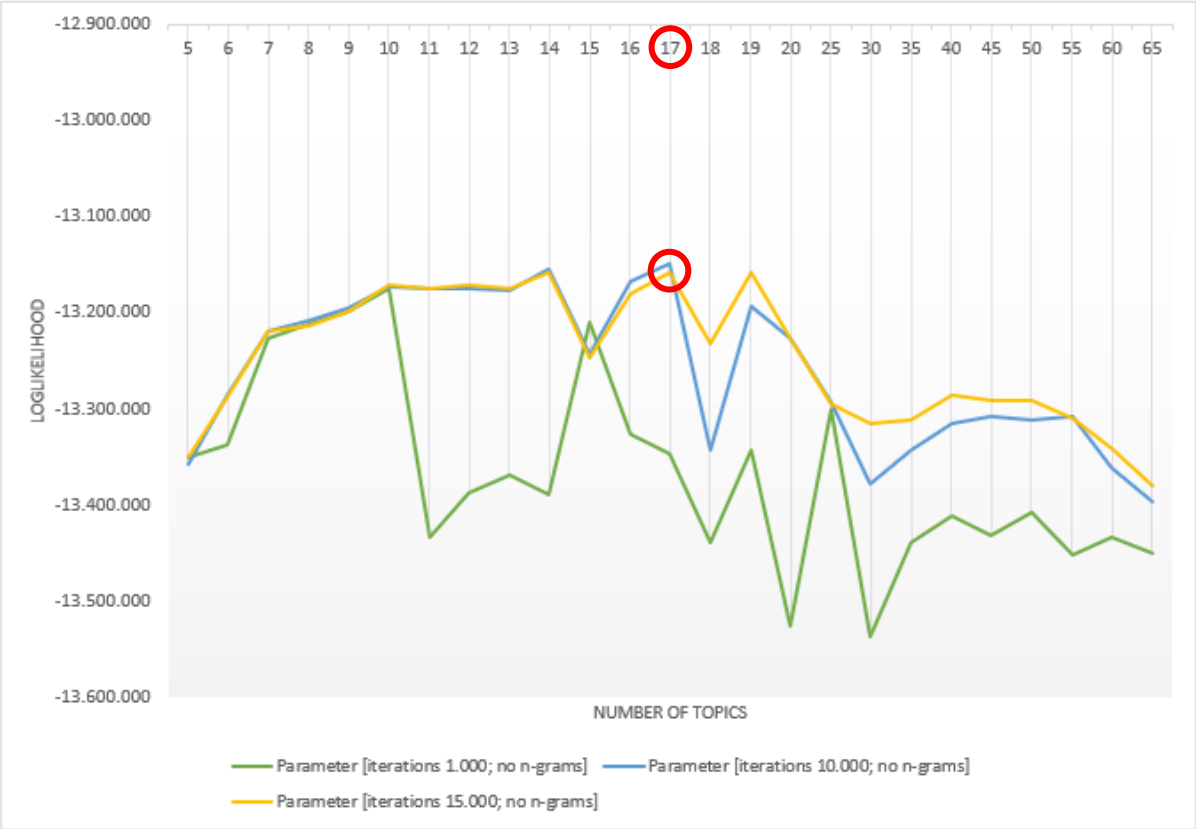
Zum anderen wird die Qualität der ermittelten Topics anhand ihres Coherence-Werts bewertet. Bevorzugt wird eine höhere Kohärenz, da sie darauf hinweist, dass die Wörter innerhalb eines Topics semantisch zusammenhängender und aussagekräftiger sind. Diese Metrik misst die Konsistenz der Wörter innerhalb eines Topics durch die Beurteilung der Ähnlichkeit der Worte im gegebenen Topic. Ein erhöhter Coherence-Wert wird erreicht, wenn die Worte innerhalb eines Topics eine klare semantische Verknüpfung haben (Pedro, 2022).

5.5.2 Ergebnisse

Im anschließenden Abschnitt werden die Resultate der durchgeführten Datenanalyse präsentiert. Hierbei finden insbesondere die erzielten Werte der Gütemaße LogLikelihood und Perplexity Beachtung. Darüber hinaus erfolgt eine Darstellung und Zusammenfassung der erzeugten Latent Dirichlet Allocation (LDA)-Modelle.

Gemäß Abbildung 6 wurden diverse Parametereinstellungen hinsichtlich der Anzahl der generierten Topics sowie der Anzahl der Iterationen getestet, um einen optimalen LogLikelihood-Wert zu erzielen. Die grüne Linie in der Abbildung illustriert die LogLikelihood-Werte bei einer Iteration von 1.000 ohne Einbeziehung von n-grams. Die Blaue hingegen repräsentiert den LogLikelihood-Wert bei einer Iteration von 10.000 ohne Berücksichtigung von n-grams. Die gelbe Linie veranschaulicht die Werte bei einer Iteration von 15.000 erneut ohne Einbeziehung von n-grams.

Abbildung 6: LogLikelihood Werte



Quelle: eigene Darstellung zu Anhang C

Der maximale LogLikelihood-Wert von -13.158.294,263 wird in Verwendung der nachfolgend genannten Konfiguration **LL1** erreicht:

LL1	
	▪ *number of topics*: 17 ▪ *iterations*: 15.000 ▪ Keine *n-grams*

In Tabelle 5 werden sämtliche 17 Topics visualisiert, wobei für jedes dieser Topics die fünf häufigsten auftretenden Wörter aufgeführt sind. Darüber hinaus erfolgt eine manuelle Zuordnung einer passenden Kategorie. Es ist anzumerken, dass die Interpretierbarkeit der fünf aufgeführten Begriffe in jedem Topic, wenn auch nicht unmittelbar offensichtlich, dennoch eine Identifizierung der thematischen Ausrichtung problemlos ermöglicht. Die Begriffe bieten Anhaltspunkte, die bei der Analyse helfen, auch wenn sie nicht in allen Fällen eine klare und eindeutige Bedeutung vermitteln.

Die 38 Geschäftsberichte werden hinsichtlich der höchsten inhaltlichen Übereinstimmung automatisch einem identifizierten Topic zugeteilt. Die Analyse gestattet eine detaillierte Einsicht in die behandelten Themengebiete der Geschäftsberichte und ermöglicht eine Gruppierung der Unternehmen anhand ihrer Schwerpunktsetzung und gemeinsamer Themen.

Tabelle 5: LDA-Model Ergebnisse in LL1

Topic Nr.	Begriff 1	Begriff 2	Begriff 3	Begriff 4	Begriff 5	Kategorie
0	mitarbei	global	produktio	rohstoff	standor	Unternehmensstrategie und globale Produktion
1	optio	finanzjah	serivc	anzahl	gewahr	Finanz- und Geschäftsanalysen
2	kontak	varia	finanzkal	limited	finanzschuld	Finanzenwesen und Unternehmensstruktur
3	erwar	leistung	enthal	veranderung	prufung	Entwicklung und Leistung
4	investor	vertr	versicherung	clearing	nettoerlo	Finanzwesen und Investitionen
5	medical	patie	nichtfinanziell	angab	covid	Gesundheitswesen und Medizin
6	energie	taxonomie	strom	aspek	pensio	Energie und Umwelt
7	variabl	schaf	schen	unterneh	grund	Unternehmensstrukturen und Finanzierung
8	limited	kompo	tranch	holding	produc	Unternehmensstrukturen und Finanzierung
9	fahrzeug	automobil	informationenzusammengefass	equity	truck	Transport und Fahrzeuge
10	gefass	immobilie	developm	victoriah	nanziell	Immobilien und Finanzierung
11	schad	ruckversicherung	versicherungstechnsich	versicherung	insuranc	Versicherungs- und Schadensmanagement
12	zweitw	finanzinstitu	risiko	bestimm	verpflichtung	Finanz- und Risikomanagement
13	mobil	technologie	deutsch	mobilfunk	servic	Mobile Technologie und Dienstleistungen
14	integrier	mitarbei	nachhaltigkei	emissio	servic	Nachhaltigkeit und Integration
15	vertra	versicherung	technologie	fortgefuhr	geschaftsfeld	Versicherung und Technologie
16	scienc	foschung	organisch	healthcar	solutio	Forschung und Gesundheitswesen

Quelle: eigene Darstellung, Details siehe Anhang D

Bei der Analyse der Ergebnisse zeigt sich überraschenderweise, dass mit einer einzigen Ausnahme alle 38 Geschäftsberichte einem identischen Topic zugeordnet werden. Bemerkenswerterweise sind 37 der Geschäftsberichte dem Topic 3 zugewiesen, das sich überwiegend der Kategorie “Entwicklung und Leistung“ widmet. Die dargelegte Zuordnung scheint ungewöhnlich und bedarf einer genaueren Untersuchung, um ihre Hintergründe näher zu beleuchten.

Abbildung 7: Details zu Topic 3

Topic 3
tokens=864579.0000
document_entropy=3.5995
word-length=7.4000
coherence=0.0000
uniform_dist=3.4533
corpus_dist=0.2783
eff_num_words=1158.1065
token-doc-diff=0.0002
rank_1_docs=0.9737
allocation_ratio=0.9474
allocation_count=1.0000
exclusivity=0.4675

erwar	word-length=5.0000	coherence=0.0000	uniform_dist=0.0222	corpus_dist=0.0018	token-doc-diff=0.0000	exclusivity=0.4242
leistung	word-length=8.0000	coherence=0.0000	uniform_dist=0.0221	corpus_dist=0.0020	token-doc-diff=0.0000	exclusivity=0.8990
enthal	word-length=6.0000	coherence=0.0000	uniform_dist=0.0217	corpus_dist=0.0017	token-doc-diff=0.0000	exclusivity=0.3972
veranderung	word-length=11.0000	coherence=0.0000	uniform_dist=0.0200	corpus_dist=0.0014	token-doc-diff=0.0001	exclusivity=0.2768
prufung	word-length=7.0000	coherence=0.0000	uniform_dist=0.0197	corpus_dist=0.0016	token-doc-diff=0.0001	exclusivity=0.3404

Quelle: eigene Darstellung

Gemäß Abbildung 7 zeigt Topic 3 einen Kohärenzwert von 0,000 auf, da sämtliche fünf zugehörigen Begriffe (*erwar, leistung, enthal, veranderung, prufung)* ebenso einen Kohärenzwert von 0,000 aufweise. Das impliziert, dass diese Begriffe untereinander keinerlei semantische Verknüpfungen aufweisen, sondern vielmehr in einer lockeren Zusammensetzung erscheinen. Abbildung 8 veranschaulicht zusätzlich, dass diese fünf Begriffe unter den acht am häufigsten vorkommenden Begriffen im Textkorpus zu finden sind und ebenfalls in allen 38 Dokumenten präsentiert sind.

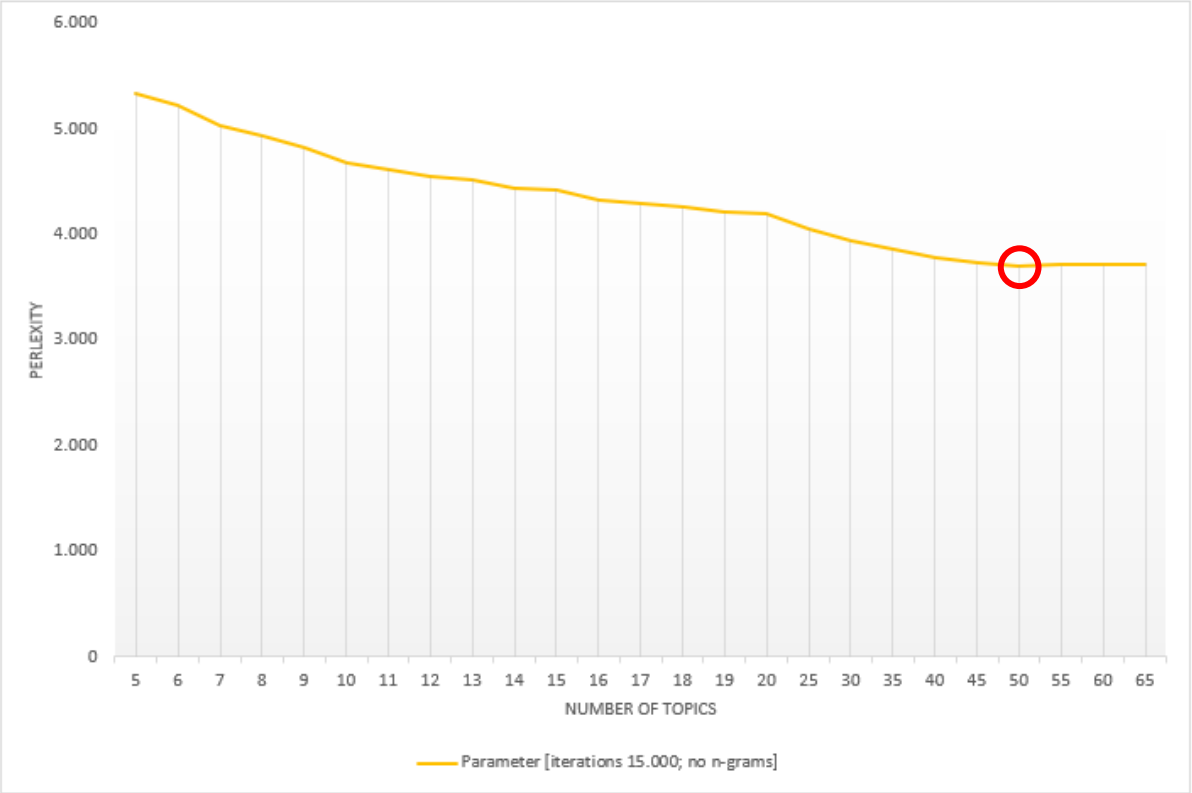
Abbildung 8: Ranking Tokens nach totalem Aufkommen

Word	Attribute Name	Total Occurences ↓	Document Occurences
zeitw	zeitw	4438	38
erwar	erwar	3643	38
enthal	enthal	3602	38
veranderung	veranderung	3513	38
angab	angab	3445	38
leistung	leistung	3378	38
operativ	operativ	3341	38
prufung	prufung	3316	38
risiko	risiko	3070	38
global	global	3045	38

Quelle: eigene Darstellung aus RapidMiner

Trotz der mangelnden Kohärenz verbucht Topic 3 den besten Kohärenzwert, da alle übrigen Themen in Bezug auf diese Metrik negative Werte aufweisen. Bezüglich des Exklusivitätswerts (*exclusivity* = 0,4675), welcher ein Maß für die Eindeutigkeit des Themas in Bezug auf die Verteilung der Wörter in anderen Themen ist, erreicht Topic 3 keinen gewünschten hohen Wert. Ein höherer Wert signalisiert, dass die Wörter innerhalb eines Themas eindeutig sind und keine Überschneidungen mit anderen Themen aufweisen. Diese Kennzahl misst somit, inwiefern die Wörter in einem Thema charakteristisch für dieses Thema sind und sich nicht auf andere Themen erstrecken. Dies verdeutlicht, dass Topic 3 die meist auftretenden Wörter im Korpus präsentiert, die generell dem Themengebiet der Geschäftsberichte zugeordnet werden können.
Der zweitbeste LogLikelihood-Wert wurde mit einer Anzahl von 14 Topics erreicht. Hier zeigt sich erneut, dass alle 38 Geschäftsberichte lediglich zwei Topics zugeordnet werden. Durch Erhöhung der Topic-Anzahl auf 19 wird der drittbeste Wert erzielt und die Geschäftsberichte werden insgesamt drei verschiedenen Topics zugewiesen.
Neben den LogLikelihood-Werten werden auch die Perplexity-Werte ermittelt. Nachdem festgestellt wurde, dass eine Iteration von 15.000 und keine n-grams den höchsten LogLikelihood-Wert erzielt, wird diese Einstellungen beibehalten. Für jede Anzahl von Themen ergibt sich ein eigener Perplexity-Wert, wie in Abbildung 9 dargestellt ist.

Abbildung 9: Perplexity-Werte



Quelle: eigene Darstellung

Der beste Wert von 3.703,611 wird bei einer Anzahl von 50 Topics und der hierin angegebenen Konfiguration **PX1** erreicht:

PX1	
	▪ *number of topics*: 50 ▪ *iterations*: 15.000 ▪ Keine *n-grams*

Die Perplexity ist eine Metrik, die die Qualität des LDA-Modells bewertet. Ein niedriger Perplexitätswert deutet darauf hin, dass das Modell voraussichtlich in der Lage ist, genaue Vorhersagen für neue Daten zu generieren. Unter der Verwendung der Einstellung **PX1** zeigen insgesamt elf der 50 Topics einen Coherence-Wert von 0,000. Dadurch erreichen sie, ähnlich wie in der Einstellung **LL1**, die höchstmöglichen Coherence-Werte, da die übrigen Topics negative Werte aufweisen. In Tabelle 6 sind die Begriffe dieser elf Topics sowie die zugeordneten Kategorien aufgeführt:

Tabelle 6: LDA-Model ausgewählte Ergebnisse in PX1

Topic Nr.	Begriff 1	Begriff 2	Begriff 3	Begriff 4	Begriff 5	Kategorie
12	leistung	prufung	hauptversammlung	vermogensw	vorsitx	Unternehmensmanagement
16	informationenzusammengefass	fahrzeug	kompo	emissio	brillianc	Fahrzeug und Innovation
25	stand	verkauf	verausserung	buchw	investitio	Finanztransaktion
30	medical	personlich	patie	gesellschafteri	grundlag	Gesundheitswesen
33	immobilie	developm	victoriah	osterreich	propertie	Immobilien
35	gefass	nanziell	tranch	lokal	transaktio	Finanztransaktion
38	clearing	nettoerlo	performanc	mitarbei	erlauterung	Unternehmensleistung
41	limited	kompo	holding	tranch	langfristbonu	Unternehmensstruktur
42	erwar	enthal	auswirkung	veranderung	geschaf	Entwicklung
45	vertr	versicherung	zivil	triebwek	erlauterung	Versicherungswesen
47	bewertung	positiv	stark	anderung	handl	Wahrnehmung

Quelle: eigene Darstellung, Details siehe Anhang E

Tabelle 7 präsentiert in einer zusammenfassenden Darstellung die drei besten Ergebnisse hinsichtlich der Qualitätskriterien LogLikelihood und Perplexity. Diese Ergebnisse wurden unter der Bedingung einer Iteration von 15.000 und keiner Verwendung von n-grams erzielt. Die Tabelle führt die jeweiligen Topic-Nummern sowie die dazugehörigen Ergebniswerte auf.

Tabelle 7: Bewertungskriterien Top 3

	LogLikelihood	Perplexity
Top 1	Topic # 17 (-13.158.294,263)	Topic # 50 (3703,611)
Top 2	Topic # 14 (-13.158.748,781)	Topic # 55 (3708,097)
Top 3	Topic # 19 (-13.159.392,921)	Topic # 60 (3712,729)

Quelle: eigenen Darstellung

Wie ersichtlich wird, bestehen erhebliche Unterschiede in den Höchstwerten dieser beiden Metriken in Bezug auf die Anzahl der Topics. Um die optimale Parametereinstellung zu ermitteln, bedarf es eines Kompromisses zwischen diesen Gütemaßen. Dieser Abwägungsprozess erfolgt mittels eines simplen Ausschlussverfahrens. Der zur Auswahl stehende Bereich erstreckt sich zwischen 17 und 50 Topics, da hier am oberen Ende der beste LogLikelihood-Wert und am unteren Ende der beste Perplexity-Wert zu finden sind. Um den Perplexity-Wert unter 4000 zu halten, wird der Bereich weiter auf die Anzahl der Topics zwischen 30 und 50 eingeschränkt. Innerhalb dieses definierten Intervalls ergeben sich fünf geeignete Werte für die Anzahl der Topics: 30, 35, 40, 45 und 50. Zur weiteren Eingrenzung werden die dazugehörigen LogLikelihood-Werte betrachtet. Die beiden besten Werte

resultieren aus der Anzahl der Topics 40 und 45. Da sowohl 40 als auch 45 eine hohe Anzahl an Topics darstellen, wird die Entscheidung zugunsten des kleineren Werts getroffen. Folglich werden die Ergebnisse für die nachfolgend genannte Konfiguration **Compro** genauer untersucht:

Compro	
	▪ *number of topics*: 40 ▪ *iterations*: 15.000 ▪ Keine *n-grams*

Innerhalb der 40 Topics weisen sechs Topics eine Kohärenz von 0,000 auf. 55 % der Geschäftsberichte werden diesen sechs Topics zugeordnet, was 21 von insgesamt 38 entspricht. Demzufolge fällt mehr als die Hälfte der Geschäftsberichte in ein Topic, das keine erkennbaren inhaltlichen Zusammenhänge aufweist. Tabelle 8 präsentiert diese sechs Topics sowie die zugehörigen ermittelten Kategorien:

Tabelle 8: LDA-Modell ausgewählte Ergebnisse in Compro

Topic Nr.	Begriff 1	Begriff 2	Begriff 3	Begriff 4	Begriff 5	Kategorie
0	ruckversicherung	schad	festverzinslich	ruckversich	perso	Finanzwesen und Versicherung
3	anzahl	zweitw	optio	servic	gewahr	Dienstleistung und Qualität
16	limited	kompo	holding	tranch	langfristbonu	Unternhemensstruktur
27	complianc	global	entwickel	standard	weltweit	Internationale Standards
33	lieferung	erweb	leistung	schuld	betrieblich	Verpflichtung
35	erwar	prufung	enthal	veranderung	leistung	Entwicklung und Leistung

Quelle: eigene Darstellung, Details siehe Anhang F

Überraschenderweise definiert sich Topic 35 aus denselben Tokens, wie Topic 3 der Einstellung **LL1**. Auch hier sind die fünf am häufigsten auftretenden Tokens *erwar, prufung, enthal, veranderung, leistung*. Diesem Topic werden wiederum die meisten der 21 Geschäftsberichte, die einem Topic mit Coherence-Wert 0,000 zugewiesen sind, zugeordnet. Es zeigt sich erneut, dass der Coherence-Wert 0,000 der höchste ist, der von allen Topics erreicht wird. Alle übrigen Werte befinden sich im negativen Bereich. Obwohl die Anzahl der Topics mit 40 recht groß ist und die Coherence-Werte aller Topics enttäuschend sind, weisen dennoch alle eine gute Interpretierbarkeit auf. Jedem Topic kann eine übergeordnete Kategorie vergeben werden, die erahnen lässt, welche Angelegenheiten oder Problematiken adressiert werden.

Nachdem zunächst der Fokus auf der Iterationsgröße und der Topic-Anzahl lag, lässt sich zusammenfassend festhalten, dass die Gütemetriken LogLikelihood, Perplexity und Coherence hilfreiche Anhaltspunkte für die Leistung der Modelle bieten. Die Orientierung an diesen und weiteren Metriken verschafft ein verständliches Bild der Modelle und ihrer Topics. Dennoch bleibt es herausfordernd, alle relevanten Maße angemessen abzuwägen und Kompromisse zu finden, um schließlich ein Modell mit guten Ergebnissen zu erhalte. In dieser Arbeit ergibt sich bedauerlicherweise, dass unabhängig von den gewählten Parametereinstellungen immer mindestens ein Topic aus zusammenhangslosen Tokens besteht und automatisch der Mehrheit der Geschäftsberichte zugeordnet wird. Trotzdem bieten die Topics eine hohe Diversität und sind gut interpretierbar. Dies bedeutet, dass die Modelle eine breite Palette von Themen abdecken und die Inhalte der Tpoics anhand der zugehörigen Tokens leicht verständlich sind. Drei nützliche Modelle mit ihren Parametereinstellungen (**LL1**, **PX1** und **Compro**) wurden genauer untersucht, wovon die Kompromissvariante für den anschließenden Schritt der Zuordnung der Unternehmen angewandt wird.

6 Korrelation zwischen Unternehmenserfolg und Topics

Im nachfolgenden Abschnitt wird versucht, einen Zusammenhang zwischen den erzeugten Topics der Einstellung **Compro** und dem wirtschaftlichen Erfolg von Unternehmen herzustellen. Vor diesem Hintergrund erfolgt zunächst eine Definition des wirtschaftlichen Erfolgs und eine Gruppierung der Unternehmen in „wirtschaftlich erfolgreich“ und „wirtschaftlich nicht erfolgreich“ für das Geschäftsjahr 2021. Im Anschluss wird untersucht, welche Topics vermehrt bei den erfolgreichen Unternehmen auftreten. Das Ziel besteht darin herauszufinden, ob es Gemeinsamkeiten zwischen den Topics der erfolgreichen Unternehmen gibt und ob sich diese von den Topics der nicht erfolgreichen Unternehmen unterscheiden. Hierbei wird versucht, festzustellen, ob ein unmittelbarer Zusammenhang zwischen dem Erfolg eines Unternehmens und den behandelten Topics besteht.

6.1 Unternehmenseinteilung

Ein zentraler Indikator, der in dieser Studie zur Beurteilung der wirtschaftlichen Performance von Unternehmen im Geschäftsjahr 2021 verwendet wird, ist die Profitabilität. Es ist jedoch zu beachten, dass diese Kennzahl allein begrenzte Aussagekraft besitzt, da sie lediglich eine Momentaufnahme des Geschäftsjahres 2021 darstellt und keine umfassendere historische Perspektive berücksichtigt. Zudem werden Unternehmen aus verschiedenen Branchen und Märkten miteinander verglichen, was unterschiedliche Geschäftsmodelle und Rahmenbedingungen mit sich bringt, einschließlich zyklischer Schwankungen, Personalkosten und Investitionsvolumen (Finanzfluss, 2017).

Gemäß dem Online-Wirtschaftslexikon der Bundeszentrale für politische Bildung wird wirtschaftlicher Erfolg als das Ergebnis der wirtschaftlichen Tätigkeit eines Unternehmens innerhalb eines bestimmten Betrachtungszeitraums definiert. Dieser Erfolg wird positiv (Gewinn, Überschuss) verbucht, wenn der Wertzuwachs größer ist als der Wertverlust, und negativ (Verlust), wenn der Wertverlust überwiegt („Erfolg“, 2021). Die Profitabilität ermöglicht die Messung dieses Erfolgs, da sie die Ertragskraft eines Unternehmens verdeutlicht. Ein Unternehmen gilt nur dann als profitabel, wenn es in der Lage ist, langfristig Gewinne zu erzielen (Profitabilität, o. D.). Die Kennzahl "Gewinn" bewertet die Höhe und den Zustand des Erfolgs möglichst unabhängig von bilanzpolitischen, steuerlichen und handelsrechtlichen Einflüssen („Erfolg“, 2021).

Um die Berechnung des Gewinns zu veranschaulichen und diese Kennzahl von anderen abzugrenzen, wird dies anhand eines Beispiels in Tabelle 99 erläutert. Der Umsatz bildet die Grundlage für die Gewinnermittlung. Davon werden zunächst die operativen Kosten wie Wasser, Strom, Personalkosten, Rohstoff- und Materialkosten, Treibstoff oder Miete abgezogen. Dadurch ergibt sich das EBITDA. Wenn von diesem Wert noch die Abschreibungen subtrahiert werden, erhält man das EBIT. Schließlich wird dieser Wert um Steuern und Zinsen bereinigt, um den Gewinn zu ermitteln.

Tabelle 9: Beispiel zur Gewinnermittlung

			1000 € Umsatz
400 € Ausgaben			600 € EBITDA
100 € Abschreibung	400 € Ausgaben		500 € EBIT
100 € Steuern & Zinsen	100 € Abschreibung	400 € Ausgaben	400 € Gewinn

Quelle: in Anlehnung an Martin Schengel [sevDesk], 2020

Die Finanzberichte der im DAX gelisteten Unternehmen enthalten in ihren Gewinn- und Verlustrechnungen für das Geschäftsjahr 2021 Angaben zum Jahresüberschuss bzw. zum Jahresfehlbetrag. Die entsprechenden Werte sind in Tabelle 10 dargestellt.

Tabelle 10: Gewinne der DAX-Unternehmen 2021

#	Unternehmen	Gewinn 2020*	Gewinn 2021*	Delta*	Prozent-punkte
1	Adidas AG	443	2.158	+1.715	+387,13
2	Allianz SE	7.133	7.105	-28	-0,39
3	BASF SE	-1.075	5.982	+7.057	+656,47
4	Bayer AG	-10.495	1.000	+9.495	+90,47
5	Beiersdorf AG	577	655	+78	+13,52
6	BMW AG	3.857	12.463	+8.606	+223,13

7	Brenntag	473,8	461,4	-12.4	-2,62
8	Continental AG	-918,8	1.506,9	+2.425,7	+264
9	Covestro AG	454	1.619	+1.165	+256,61
10	Delivery Hero SE	-1.407,2	-1.096,5	+310,7	+22,08
11	Deutsche Bank AG	600	2.500	+1.900	+316,67
12	Deutsche Börse AG	1.125,1	1.264,9	+139,8	+12,43
13	Deutsche Post AG	3.176	5.423	+2.247	+70,75
14	Deutsche Telekom AG	6.747	6.103	-644	-9,54
15	E.ON SE	1.270	5.305	+4.035	+317,72
16	Fresenius SE	1.796	1.867	+71	+3,95
17	Fresenius Medical Care AG	1.435,8	1.219	-216,8	-15,1
18	Hannover Rück	883,1	1.231,3	+348,2	+39,43
19	Heidelberg Cement AG	-2.009	1.902	+3.911	+194,67
20	HelloFresh SE	369,1	256,3	-112,8	-30,56
21	Henkel AG	1.424	1.629	+205	+14,4
22	Infineon Technologies AG	368	1.169	+801	+217,66
23	Mercedes-Benz Group AG	4.009	23.396	+19.387	+483,59
24	Merck KGaA	1.994	3.065	+1.071	+53,71
25	MTU Aero Engines AG	294	342	+48	+16,33
26	Münchner Rück AG	1.211	2.932	+1.721	+142,11
27	Porsche Automobil Holding SE	703	824	+121	+17,21
28	Puma SE	123,1	376,8	+253,7	+206,09
29	RWE AG	1.110	832	-278	-25,05
30	SAP SE	5.283	5.376	+93	+1,76
31	Sartorius AG	299	426	+127	+42,47
32	Siemens AG	4.200	6.697	+2.497	+59,45
33	Siemens Energy AG	-1.859	-560	+1299	+69,88
34	Siemens Healthineers AG	1.423	1.746	+323	+22,7
35	Symrise AG	314	385	+71	+22,61
36	Volkswagen AG	8.824	15.428	+6.604	+74,84
37	Vonovia SE	3.340,0	2.830,9	-509,1	-15,24
38	Zalando SE	226,1	234,5	+8,4	+3,72

Quelle: entsprechender Geschäftsbericht *in Millionen

Die Auswertung der Geschäftsberichte der im DAX gelisteten Unternehmen ergab, dass 81,6 % der insgesamt 38 betrachteten Unternehmen einen Jahresüberschuss erwirtschafteten, während 18,4 % einen Jahresverlust verbuchten (in der Tabelle rot markiert). Diese Verteilung der Ergebnisse entspricht weitgehend dem Verhältnis, das dem Pareto-Prinzip von Vilfredo Pareto zugrunde liegt, wonach etwa 80 % der Ergebnisse durch etwa 20 % der Ursachen oder Anstrengungen erzielt werden. Obwohl das Ursache-Wirkungsprinzip hier nicht direkt angewandt werden kann, wurde in Anlehnung an das Pareto-Prinzip eine Aufteilung der Unternehmen in zwei Gruppen vorgenommen. Die erste Gruppe umfasst die Unternehmen, die etwa 80 % der Gesamtsumme erwirtschaftet haben und somit den höchsten Jahresüberschuss erzielten. Die zweite Gruppe deckt die restlichen 20 % der erwirtschafteten Gesamtsumme ab und umfasst Unternehmen, die entweder einen niedrigeren Jahresüberschuss erzielten oder einen Jahresverlust verbuchten. Diese Aufteilung wird in Tabelle 11 dargestellt. Als Berechnungshilfe wird die ABC-Analyse herangezogen. Diese Analyse ermöglicht die Klassifizierung von Elementen basierend auf ihrer relativen Bedeutung oder Wertigkeit und teilt sie in die Kategorien A, B und C ein. In dieser Arbeit wurde eine Beschränkung auf zwei Kategorien im Verhältnis 80/20 vorgenommen. Durch die Anwendung dieser Verteilung ist es möglich, die Gruppe der größten Gewinner zu identifizieren und zu untersuchen, welche Topics vermehrt in deren Geschäftsberichten auftreten. Gleichzeitig wird analysiert, ob diese Topics auch in den Geschäftsberichten der unteren 20 % der Unternehmen zu finden sind.

Tabelle 11: Gruppenbildung nach Gewinndelta

Jahresgewinne

ID	Delta Gewinn 2020 zu 2021*	prozentualer Anteil der Gesamtsumme	prozentualer Anteil kummuliert	Gruppe
3	656,47	14,87	14,87	A
23	483,59	10,95	25,82	A
1	387,13	8,77	34,58	A
15	317,72	7,19	41,78	A
11	316,67	7,17	48,95	A
8	264	5,98	54,93	A
9	256,61	5,81	60,74	A
6	223,13	5,05	65,79	A
22	217,66	4,93	70,72	A
28	206,09	4,67	75,39	A
19	194,67	4,41	79,79	A
26	142,11	3,22	83,01	A
4	90,47	2,05	85,06	B
36	74,84	1,69	86,76	B
13	70,75	1,60	88,36	B
33	69,88	1,58	89,94	B
32	59,45	1,35	91,29	B
24	53,71	1,22	92,50	B
31	42,47	0,96	93,46	B
18	39,43	0,89	94,36	B
34	22,7	0,51	94,87	B
35	22,61	0,51	95,38	B
10	22,08	0,50	95,88	B
27	17,21	0,39	96,27	B
25	16,33	0,37	96,64	B
21	14,4	0,33	96,97	B
5	13,52	0,31	97,27	B
12	12,43	0,28	97,56	B
16	3,95	0,09	97,65	B
38	3,72	0,08	97,73	B
30	1,76	0,04	**97,77**	B
Summe	**4317,56**			

Jahresverluste

ID	Delta Gewinn 2020 zu 2021*	Betragswerte	prozentualer Anteil der Gesamtsumme	prozentualer Anteil kummuliert	Gruppe
2	-0,39	0,39	0,01	97,78	B
7	-2,62	2,62	0,06	97,84	B
14	-9,54	9,54	0,22	98,05	B
17	-15,1	15,1	0,34	98,40	B
37	-15,24	15,24	0,35	98,74	B
29	-25,05	25,05	0,57	99,31	B
20	-30,56	30,56	0,69	**100,00**	B
Summe	-98,5	**98,5**			
Gesamtsumme		**4416,06**			

*in Millionen

Quelle: eigene Darstellung

Die Unternehmen werden anhand des Gewinndeltas zwischen den Jahren 2020 und 2021 in absteigender Reihenfolge sortiert. In der Gruppe A befinden sich diejenigen Unternehmen, die im Geschäftsjahr 2021 wirtschaftlich erfolgreich waren, gemessen am definierten Indikator "Gewinn". Die oberen zwölf Unternehmen, die zusammen 83,01 % der Gesamtsumme erwirtschafteten und etwa ein Drittel der Stichprobe ausmachen, weisen jeweils ein dreistelliges Gewinndelta auf. Sie bilden die Gruppe A. Die verbleibenden 16,99 % der erwirtschafteten Gesamtsumme werden von den 26 Unternehmen in Gruppe B generiert. Da sich in Gruppe B Unternehmen befinden, die Verluste verzeichneten, wird zur Berechnung der Gesamtsumme dieser negative Wert als Betragswert verwandet. Um sicherzustellen, dass diese Betragswerte nicht in Gruppe A platziert werden, werden sie gesondert am Ende der Tabelle 11 platziert.

6.2 Unternehmensvergleich

Für einen Vergleich der Topic-Verteilung innerhalb der Gruppen A und B wird die Einstellung **Compro** gewählt. Insgesamt stehen 40 Topics zur Verfügung, von denen sechs einen Coherence-Wert von 0,000 aufweisen. Tabelle 12 illustriert, welche Topics den zwölf Unternehmen der Gruppe A zugewiesen werden. Die Markierung mit einem Stern (*) weist auf einen Coherence-Wert von 0,000 hin:

Tabelle 12: Topic-Verteilung Gruppe A in Compro

ID	Unternehmen	Compro	Kategorie
3	BASF SE	Topic 35*	Entwicklung und Leistung
23	Mercedes-Benz Group AG	Topic 17	Gesundheit und Technologie
1	Adidas AG	Topic 25	Versicherung
15	E.ON SE	Topic 20	Gesundheit und Organisation
11	Deutsche Bank AG	Topic 18	Unternehmensleistung
8	Continental AG	Topic 35*	Entwicklung und Leistung
9	Covestro AG	Topic 35*	Entwicklung und Leistung
6	BMW AG	Topic 35*	Entwicklung und Leistung
22	Infineon Technologies AG	Topic 35*	Entwicklung und Leistung
28	Puma AG	Topic 6	Energie, Investition und Versicherung
19	Heidelberg Cement AG	Topic 35*	Entwicklung und Leistung
26	Münchner Rück AG	Topic 35*	Entwicklung und Leistung
	Anzahl *	**7**	

Quelle: eigene Darstellung, Details siehe Anhang F

Sieben Unternehmen der Gruppe A werden dem Topic 35 zugeteilt, das eine Coherence von 0,000 aufweist und die Kategorie “Entwicklung und Leistung“ präsentiert. Andere Unternehmen der Gruppe A fokussieren sich in ihren Geschäftsberichten hauptsächlich auf Themen wie “Gesundheit und Technologie“, “Versicherung“, “Gesundheitswesen und Organisation“, “Unternehmensleistung“ und “Energie, Investition und Versicherung“. Daraus lässt sich schließen, dass der Fokus wirtschaftlich erfolgreicher Unternehmen im Geschäftsjahr 2021 in den Bereichen Entwicklung, Energie, Gesundheit und Versicherung lag. Wobei festzuhalten ist, dass sich 60 % dieser Unternehmen in einer inkohärenten Gruppe befinden.

In der Gruppe B werden knapp 75 % der Unternehmen einem Topic zugeteilt, welches eine Coherence von 0,000 aufweist. Das entspricht 16 Unternehmen von insgesamt 26 aus der Gruppe B, wie in Tabelle 13 deutlich wird. Diese 16 Unternehmen beschäftigen sich mit den Inhalten “Entwicklung und Leistung“, “Finanzwesen und Versicherung“, “Internationale Standards“ sowie “Dienstleistung und Qualität“. Die übrigen Unternehmen der Gruppe B setzen sich in ihren Geschäftsberichten aus dem Jahr 2021 mit den Themen “Automobil und Dieselkrise“, “Gesundheit und Technologie“ beziehungsweise “Gesundheit und Organisation“ oder “Gesundheitsmanagement“, “Unternehmensleistung“, “Integration, Nachhaltigkeit und Cloud-Service“, “Versicherung“, “Mobilfunk“, “Immobilienentwicklung“ und “Energie, Investition und Versicherung“ auseinander. Zusammenfassend lässt sich festhalten, dass Gruppe B eine wesentlich vielfältigere Themenauswahl aufweist als Gruppe A. Dies schließt die Themen Entwicklung, Energie, Gesundheit und Versicherung ein, die ebenfalls in Gruppe A präsent sind, sowie die zusätzliche Bereiche Automobil, Dieselkrise, Finanzwesen, Immobilien, Mobilfunk, Nachhaltigkeit und auch Qualität und Standards. Ebenso sei angemerkt, dass sich 75 % dieser Unternehmen in einer inkohärenten Gruppe befinden.

Tabelle 13: Topic-Verteilung Gruppe B in Compro

ID	Unternehmen	Topic Nr.	Kategorie
4	Bayer AG	35*	Entwicklung und Leistung
36	Volkswagen AG	9	Automobile und Dieselkrise
13	Deutsche Post AG	35*	Entwicklung und Leistung
33	Siemens Energy AG	35*	Automobile und Dieselkrise
32	Siemens AG	35*	Entwicklung und Leistung
24	Merck KGaA	17	Gesundheit und Technologie
31	Sartorius AG	35*	Entwicklung und Leistung
18	Hannover Rück	0*	Finanzwesen und Versicherung
34	Siemens Healthineers AG	35*	Entwicklung und Leistung
35	Symrise AG	35*	Entwicklung und Leistung
10	Delivery Hero SE	35*	Entwicklung und Leistung
27	Prosche Automobile Holding SE	35*	Entwicklung und Leistung
25	MTU Aero Engines AG	35*	Entwicklung und Leistung
21	Henkel AG	35*	Entwicklung und Leistung
5	Beiersdorf AG	27*	Internationale Standards
12	Deutsche Börse AG	18	Unternehmensleistung
16	Fresenius SE	20	Gesundheitswesen und Organisation
38	Zalando SE	3*	Dienstleistung und Qualität
30	SAP SE	39	Integration, Nachhaltigkeit und Cloud-Service
2	Allianz SE	25	Versicherung
7	Brenntag	35*	Entwicklung und Leistung
14	Deutsche Telekom AG	31	Mobilfunk
17	Fresenius Medical Care AG	37	Gesundheitsmanagement
37	Vonovia AG	8	Immobilienentwicklung
29	RWE AG	6	Energie, Investition und Versicherung
20	HelloFresh SE	35*	Entwicklung und Leistung
	Anzahl *	**16**	

Quelle: eigene Darstellung, Details siehe Anhang F

Abschließend ist zusagen, dass sowohl Unternehmen der Gruppe A als auch der Gruppe B im Verlauf des Geschäftsjahres 2021 intensive Auseinandersetzung mit den Themenfeldern

- Entwicklung,
- Energie,
- Gesundheit und
- Versicherung

zeigten. Diese Bereiche wurden von sämtlichen Unternehmen der Gruppe A und von etwa 80 % der Unternehmen der Gruppe B adressiert. Unter den adressierenden Unternehmen befinden sich auch solche, die im Geschäftsjahr 2021 vergleichsweise

schwache wirtschaftliche Ergebnisse erzielten. Folglich beschäftigten sich diese Firmen mit denselben Themen wie die erfolgreicheren Unternehmen in der Gruppe A, jedoch mit unterschiedlicher oder weniger effektiver Umsetzung. Hingegen fokussierten sich die verbleibenden 20 % der Unternehmen in Gruppe B im genannten Geschäftsjahr auf abweichende Themen.
In der Gesamtheit ergibt sich, dass wirtschaftlich erfolgreiche Unternehmen hauptsächlich ein bestimmtes Thema fokussierten. Es zeigt sich, dass auch weniger erfolgreiche Unternehmen ähnliche Themengebiete in Betracht ziehen. Was daraufhin deutet, dass es keine direkte Korrelation zwischen wirtschaftlichem Erfolg und dem thematischen Fokus im Geschäftsjahr gibt. Diese Beobachtung lässt Raum für die Hypothese, dass, obwohl beide Gruppen selbe Schwerpunkte in Bezug auf Themen vorweisen, unterschiedliche Schlussfolgerungen, Entscheidungen und Umsetzungen getroffen werden, die den Unterschied zwischen Erfolg und Misserfolg ausmachen könnten. Es ist daher abschließend festzustellen, dass der Vergleich von Unternehmen verschiedener Branchen unter Berücksichtigung einer geringen Anzahl von Gütemetriken eine gewisse Begrenztheit aufweist.

6.3 Ursachen für Topic-Modeling Ergebnisse

Zusammenfassend lässt sich festhalten, dass eine klare Differenzierung der behandelten Themen zwischen Gruppe A und Gruppe B nicht ermittelt werden konnte. Sowohl Unternehmen, die im Geschäftsjahr 2021 Erfolg verzeichneten, als auch jene mit geringerem Erfolg beschäftigten sich im Wesentlichen mit vergleichbaren Themen.

Ein möglicher Grund für die fehlende Trennung der behandelten Themen könnte in gemeinsamen Marktentwicklungen begründet liegen. Unternehmen unterschiedlicher Branchen sehen sich mit ähnlichen Marktbedingungen und Herausforderungen konfrontiert, was zu einer natürlichen Ähnlichkeit der thematischen Schwerpunkte führen könnte.
Ein weiterer Aspekt könnte darin bestehen, dass Unternehmen trotz unterschiedlicher Performance ähnliche Reaktionen auf externe Einflüsse zeigen und daher ähnliche Themen adressieren. Dies könnte insbesondere im Bereich Technologie, Innovation und Wettbewerbsstrategien der Fall sein.
Darüber hinaus könnte die Ursache für diese ungleiche Zuordnung in unausgewogenen Daten liegen. Es ist möglich, dass die analysierten Texte nicht

hinreichend variieren, um unterschiedliche Themen korrekt zu identifizieren. Wenn sämtliche Berichte einem einzigen Thema zugeordnet werden, könnte dies darauf hinweisen, dass das Modell Schwierigkeiten hat, die subtilen Nuancen und Unterschiede zwischen den Texten zu erfassen.

Zudem könnte es zu einer Themenüberschneidung kommen. Es ist denkbar, dass Geschäftsberichte vergleichbare oder sich überschneidende Themenfelder behandeln, was wiederum zu einer starken Konzentration in einem einzigen Hauptthema führen könnte.

Nicht zuletzt spielen auch Datenqualität und -vielfalt eine wesentliche Rolle. Wenn die Daten ähnliche Schreibweisen, Begrifflichkeiten oder Strukturen aufweisen, könnte dies dazu führen, dass das Modell Schwierigkeiten hat, verschiedene Themen differenziert zu erfassen. Dies könnte ebenfalls erklären, warum die Mehrheit der Berichte im selben Hauptthema verortet ist.

7 Fazit

Nach einer Reihe von Datenverfeinerungsschritten wurde die Analyse initiier. Dabei wurden sowohl Topic Modeling als auch eine Analyse der Wortfrequenzen durchgeführt, um reichhaltige Bewertungsinformationen zur Verfügung zu haben. Dies erforderte eine beträchtliche Anzahl von Iterationen, bei denen kontinuierliche und teilweise fein nuancierte Anpassungen der Parameter vorgenommen wurden. Es war notwendig, sorgfältige Abwägungen zwischen den Qualitätsmaßen zu treffen, um schließlich einen Kompromiss und ein ausgewogenes Modell zu erreichen. Zusätzlich musste noch wirtschaftlicher Erfolg greifbar gemacht werden, um die untersuchten Unternehmen in zwei Kategorien zu unterteilen: solche mit nachweislich wirtschaftlichem Erfolg und solche mit geringerem Erfolg. Letztlich wurden diese beiden Gruppen anhand der ihnen zugeordneten Topics verglichen.

Durch die Evaluierung der Gütemaße LogLikelihood, Perplexity und Coherence wurde angestrebt, einen optimalen Prozess zu entwickeln, der aussagekräftige Resultate liefert und gleichzeitig den Kriterien der Wissensentdeckung gemäß Simoudis et al. (1996) entspricht. Mithilfe des LogLikelihood-Wertes wurde die Gültigkeit des Modells sichergestellt, indem geprüft wurde, inwieweit es die zugrundeliegenden Daten repräsentiert. Der Perplexity-Wert wurde einbezogen, um ein Modell zu generieren, das potenziell nützliche Vorhersagen liefert und robust mit neuen Daten umgehen kann. Die Berücksichtigung der Coherence-Werte zielte darauf ab, die Verständlichkeit der Topics zu verbessern und deren Interpretierbarkeit zu ermöglichen.

Im Verlauf dieser Untersuchung wurde ein Muster deutlich, bei dem wirtschaftlich erfolgreiche Unternehmen in ihren Geschäftsberichten für das Jahr 2021 überwiegend den Schwerpunkt auf die Kategorie "Entwicklung und Leistung" legten. Es zeigten sich auch Unterschiede in den behandelten Themen im Vergleich zu weniger erfolgreichen Unternehmen. Letztere wiesen eine größere thematische Vielfalt auf. Sowohl erfolgreiche als auch weniger erfolgreiche Unternehmen behandelten im Geschäftsjahr 2021 Aspekte der umfassenden Kategorie "Entwicklung und Leistung". Dennoch scheinen verschiedene Resultate und Maßnahmen zu unterschiedlicher finanzieller Performance geführt zu haben, die sowohl positiv als auch negativ waren. Daher lässt sich keine klare Trennung zwischen den Themen beider Gruppen ziehen. Ebenso konnte kein bestimmtes Thema als zuverlässiger Indikator für die wirtschaftliche Leistungsfähigkeit identifiziert werden.

Alles betrachtend lässt sich feststellen, dass keine klare Korrelation zwischen den behandelten Themen in den Geschäftsberichten und dem unternehmerischen Erfolg besteht. Die identifizierten Topics weisen eine moderate thematische Vielfalt auf, wobei der Schwerpunkt naturgemäß auf finanziellen Aspekten liegt. Trotz mehrerer Anpassungsversuche zeigte sich immer wieder mindestens ein Topic mit einem Coherence-Wert von 0,000. Dieses Topic wird den meisten Geschäftsberichten zugeordnet, was die Unterscheidung zwischen den behandelten Themen von wirtschaftlich erfolgreichen und weniger erfolgreichen Unternehmen im Geschäftsjahr 2021 erschwert.

8 Ausblick

Die Ergebnisse werden zweifelsohne von den Modellparametern und -einstellungen beeinflusst. Eine mögliche Optimierung dieser Parameter könnte zu einer effektiveren Verteilung der Geschäftsberichte auf verschiedene Topics führen.

Eine weitere Möglichkeit zur Feinjustierung besteht in der Anpassung des maßgeschneiderten Stoppwörterbuchs, um eine noch differenziertere Themendarstellung zu erzielen. Dieses Stoppwörterbuch ermöglicht die gezielte Ausschließung bestimmter Wörter oder Ausdrücke, die möglicherweise die Analyse oder den Prozess beeinflussen könnten. Eine Erweiterung des Wörterbuchs könnte in Betracht gezogen werden, indem beispielsweise sämtliche Ausdrücke mit einem Kohärenzwert von 0,000 oder Termini und Wörter, die in allen 38 Berichten vorkommen, hinzugefügt werden. Bei der Zusammenstellung des Stoppwörterbuchs ist jedoch höchste Genauigkeit und Abwägung geboten, um wichtige Informationen nicht unabsichtlich auszufiltern.

Der Prozess könnte auch dahingehen angepasst werden, indem der Prozessschritt des Stemming entfernt wird. Hierdurch könnte die ursprüngliche Bedeutung der Wörter beibehalten werden und gleichzeitig die Klarheit und Interpretierbarkeit der generierten Topics gesteigert werden. Diese Anpassung könnte aber Einbußen hinsichtlich des LogLikelihood- und Perplexity-Werts mit sich bringen.

Darüber hinaus könnte die Einführung von n-grams eine Steigerung der Coherence-Werte und Verbesserung der Topic-Einteilung bewirken. Hierbei könnte der Operator *Generate n-Grams (Terms*) Verwendung finden. Dies ermöglicht die Analyse ganzer Ausdrücke, die als n-Grams bezeichnet werden. Ein n-Gram ist eine Abfolge aufeinanderfolgender Tokens mit einer festgelegten Länge von n. Dadurch entstehen semantische Beziehungen zwischen den Tokens. Die Variable n kann flexibel gewählt werden und repräsentiert die maximale Anzahl von Tokens in einer Sequenz, nach der der Prozess die Dokumente durchsucht.

Des Weiteren könnte für eine repräsentative Stichprobe von Geschäftsberichten eine manuelle Überprüfung der zugewiesenen Topics in Betracht gezogen werden. Dadurch kann ein Vergleich der manuellen Zuordnungen mit den automatisierten Zuordnungen des LDA-Modells erfolgen, um mögliche Diskrepanzen zu aufzudecken. Obwohl dieses Verfahren zeitaufwendig ist, könnten die gewonnenen Erkenntnisse äußerst wertvoll sein. Die Untersuchung könnte aufzeigen, ob tatsächlich eine starke

thematische Überschneidung vorliegt oder ob bestimmte Themen in den Texten eventuell unzureichend erfasst wurden.
Abschließend wären zusätzliche Analysen denkbar, um die Gründe für die starke Zuordnung zu einem Topic genauer zu untersuchen und gegebenenfalls alternative Modelle oder Methoden in Erwägung zu ziehen:

- Sentiment Analyse
- Cluster Analyse
- andere Analyseverfahren

Eine vielversprechende Option für eine alternative Analysemethode wäre die Non-Negative Matrix Factorization (NMF), ein statistischer Algorithmus zur Reduzierung der Dimensionalität eines gegebenen Korpus. Der Begriff "Nicht-Negativ" verdeutlicht, dass dieser Algorithmus auf Matrizen angewendet wird, bei denen keine negativen Werte vorkommen. Das macht ihn besonders geeignet für Datensätze wie Bilder, Texte und andere nicht-negativ beschränkte Wertebereiche. Die NMF-Technik nutzt Prinzipien der Faktorenanalyse, um Wörtern mit geringer Kohärenz vergleichsweise weniger Gewichtung zu verleihen (Rocky Suven Datascience, 2023).
Ebenfalls bietet sich die Anwendung des Latent Semantic Analysis (LSA) Algorithmus an. LSA ist eine Methode zur Dimensionsreduktion von Textdaten, bei der aus einem Textkorpus ein semantischer Raum erstellt wird. Dieser semantische Raum wird verwendet, um die Ähnlichkeiten zwischen Wörtern, Sätzen, Absätzen oder ganzen Dokumenten für diverse Zwecke zu quantifizieren (Cvitanic et al., 2016).
Da LDA nicht in der Lage ist, Themenkorrelationen zu modellieren, könnte der Correlated Topic Model (CTM) eine sinnvolle Alternative sein. Dieser stellt eine Erweiterung des LDA-Algorithmus dar. Trotz seiner erhöhten Komplexität gegenüber LDA und möglicher Anforderungen an Daten und Rechenressourcen kann der CTM eine genauere Darstellung der tatsächlichen Beziehungen zwischen Themen bieten, sofern solche Korrelationen in den Daten vorhanden sind (Lafferty & Blei, 2005).

Literaturverzeichnis

Alexander Meneikis [Alexander Meneikis]. (2014, 24. Januar). *Grundlagen IFRS Kurzpräsentation* [Video]. YouTube. https://www.youtube.com/watch?v=UdEYjmkmrq8

Best Data Science Platform for Your Enterprise | RapidMiner. (2022, 16. September). RapidMiner. https://rapidminer.com/platform/

BfJ - Bestandteile des Jahresabschlusses. (o. D.). Bundesamt für Justiz. Abgerufen am 14. August 2023, von https://www.bundesjustizamt.de/DE/Themen/OrdnungsgeldVollstreckung/Jahresabschluesse/Verstoesse/InhaltJahresabschluss/Bestandteile/Bestandteile_node.html#AnkerDokument41598

Biemann, C. (2022). *Wissensrohstoff text - konzepte, algorithmen, ergebnisse: eine einfhrung in das text mining*.

Bilanz nach IAS / IFRS BETROFFENE UNTERNEHMEN. (o. D.). bibukurse.de. Abgerufen am 18. April 2022, von https://www.bibukurse.de/internationale-rechnungslegung/bedeutung-und-entwicklung-der-internationalen-rechnungslegung/betroffene-unternehmen.html

Büchel, J. & Engels, B. (2023). Digitalisierung der Wirtschaft in Deutschland: Digitalisierungsindex 2022. *Bundesministerium für Wirtschaft und Klimaschutz*. https://www.de.digital/DIGITAL/Redaktion/DE/Digitalisierungsindex/Publikationen/publikation-digitalisierungsindex-2022-kurzfassung.pdf?__blob=publicationFile&v=4

codebasics. (2022, 17. August). *Text Representation Using TF-IDF: NLP Tutorial For Beginners - 18* [Video]. YouTube. https://www.youtube.com/watch?v=ATK6fm3cYfI

Cvitanic, T., Lee, B., Song, H. I., Fu, K. & Rosen, D. B. (2016). LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents. *ICCBR Workshops*, 41–50. http://ceur-ws.org/Vol-1815/paper4.pdf

Data Science Garage. (2021, 23. Februar). *Calculate TF-IDF in NLP (Simple Example)* [Video]. YouTube. https://www.youtube.com/watch?v=vZAXpvHhQow

DAX 40 Liste. (o. D.). boerse.de. Abgerufen am 15. April 2023, von https://www.boerse.de/kurse/Dax-Aktien/DE0008469008

dbislab. (2020, 6. Mai). *005 tfidf* [Video]. YouTube. https://www.youtube.com/watch?v=Quclt1YkqF0

Erfolg. (2021). In *bpb.de*. Abgerufen am 8. Juni 2023, von https://www.bpb.de/kurz-knapp/lexika/lexikon-der-

wirtschaft/19222/erfolg/#:~:text=das%20Ergebnis%20der%20wirtschaftlichen%20T%C3%A4tigkeit,oder%20negativ%20(Verlust)%20sein.

Feldman, R. & Dagan, I. (1995, 28. Juni). *Knowledge Discovery in Textual Databases (KDT)*. ResearchGate. https://www.researchgate.net/publication/2781984_Knowledge_Discovery_in_Textual_Databases_KDT/link/09e4150e18fb654f9b000000/download

Finanzfluss. (2017, 11. November). *In Aktien investieren: 12 wichtige Aktienkennzahlen!* [Video]. YouTube. Abgerufen am 8. Juni 2023, von https://www.youtube.com/watch?v=qie9sxCIhHM

Geschäftsbericht Definition. (o. D.). finanzen.net. Abgerufen am 29. April 2023, von https://www.finanzen.net/wirtschaftslexikon/geschaeftsbericht

Horstmann, J. (2018). *Topic Modeling*. forTEXT. Literatur digital erforschen. Abgerufen am 6. Juli 2023, von https://fortext.net/routinen/methoden/topic-modeling

IBM Technology. (2022, 27. April). *What is Text Mining?* [Video]. YouTube. https://www.youtube.com/watch?v=BeDeHntF68M

IDL Wissenswert: Wer muss nach IFRS bilanzieren? (2021, 20. Juni). Insight Software. Abgerufen am 19. April 2022, von https://insightsoftware.com/de/blog/idl-wissenswert-wer-muss-nach-ifrs-bilanzieren/

IFRS - Who uses IFRS Accounting Standards? (o. D.). https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/

IWD. (2019, 7. Juni). *Datenmenge explodiert*. iwd. https://www.iwd.de/artikel/datenmenge-explodiert-431851/

KMPG - Geschäftsberichte lesen und verstehen. (2014). In *Assets KMPG*.

Lafferty, J. & Blei, D. M. (2005). Correlated topic models. *Neural Information Processing Systems*, *18*, 147–154. http://ece.duke.edu/~lcarin/Blei2005CTM.pdf

Litzel, N. & Luber, S. (2019a, März 19). *Was ist Data Mining?* BigData-Insider. https://www.bigdata-insider.de/was-ist-data-mining-a-593421/

Litzel, N. & Luber, S. (2019b, März 19). *Was sind unstrukturierte Daten?* BigData-Insider. https://www.bigdata-insider.de/was-sind-unstrukturierte-daten-a-666378/

Machine Learning and RapidMiner Tutorials | RapidMiner Academy. (o. D.). [Video]. https://academy.rapidminer.com/learn/course/text-and-web-mining-with-rapidminer/text-and-web-mining/comparison-classification-and-clustering?page=2

Maheta, D. M. (2022, 14. Juni). *64. Topic Modelling in Rapidminer || Dr. Dhaval Maheta* [Video]. YouTube. Abgerufen am 21. August 2023, von https://www.youtube.com/watch?v=Q4-Ve1bBmDw

Martin Schengel [sevDesk]. (2020, 19. August). *Unterschied zwischen #EBIT und #EBITDA? | Einfach erklärt!* [Video]. YouTube. Abgerufen am 8. Juni 2023, von https://www.youtube.com/watch?v=dBcYgoyts0E

Naeem, T. (2023, 20. Januar). *Grundlegendes zu strukturierten, halbstrukturierten und unstrukturierten Daten*. Astera. https://www.astera.com/de/type/blog/structured-semi-structured-and-unstructured-data/

Process documents - RapidMiner Documentation. (o. D.). RapidMiner Documentation. Abgerufen am 14. August 2023, von https://docs.rapidminer.com/10.0/studio/operators/extensions/Text%20Processing/process_documents.html

Profitabilität. (o. D.). BWL-Lexikon. Abgerufen am 8. Juni 2023, von https://www.bwl-lexikon.de/wiki/profitabilitaet/#abgrenzung-profitabilitaet-und-rentabilitaet

RapidMiner. (o. D.). *Stem (German) - RapidMiner Documentation*. https://docs.rapidminer.com/latest/studio/operators/extensions/Text%20Processing/stemming/stem_german.html

Redaktion RWP. (2023, 12. April). *IFRS - International Financial Reporting Standards*. Rechnungswesen-Portal. Abgerufen am 26. April 2023, von https://www.rechnungswesen-portal.de/Fachinfo/IAS--IFRS--US-GAAP/IFRS.html

Rega, I., Laue, J. C., Böckem, H. & Beyhs, O. (2014). Geschäftsberichte lesen und verstehen. *KPMG*, *7*. https://assets.kpmg.com/content/dam/kpmg/pdf/2014/10/geschaeftsberichte-lesen-und-verstehen-compressed.pdf

Rocky Suven Datascience. (2023, 1. April). *Topic modelling using NMF*. Kaggle. Abgerufen am 21. August 2023, von https://www.kaggle.com/code/rockystats/topic-modelling-using-nmf

Schinko, C. (2021, 6. Dezember). *Unstrukturierte Daten: Wie Unternehmen diese Herausforderung meistern können - CANCOM.info*. CANCOM.info. Abgerufen am 29. April 2023, von https://www.cancom.info/2021/12/unstrukturierte-daten-wie-unternehmen-diese-herausforderung-meistern-koennen/#:~:text=%E2%80%9C,unstrukturierten%20Daten%20im%20Betrieb%20%C3%BCberfordert.

Simoudis, E., Han, J. & Fayyad, U. M. (1996). *KDD-96: Proceedings*. Amer Assn for Artificial. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230/1131

Sumathy, K. L. & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues - An Overview. *International journal of computer applications*, *80*(4), 29–32. https://doi.org/10.5120/13851-1685

tagesschau.de, ARD-Börsenstudio & Mannweiler, A. (2022, 2. September). Ein Jahr DAX 40: Haben sich die Erwartungen erfüllt? *tagesschau.de*. Abgerufen am 7. Juli 2023, von https://www.tagesschau.de/wirtschaft/finanzen/dax-40-erweiterung-reform-airbus-zalando-101.html

Text Mining: Neue Chance für Unternehmen. (o. D.). Deloitte Deutschland. Abgerufen am 5. Juli 2023, von https://www2.deloitte.com/de/de/pages/risk/articles/text-mining.html

Thiele, S. (o. D.). *Grundlagen der Bilanzanalyse* [Vorlesungsfolien]. Windersemester 2009/2010, Wuppertal, Nordrhein-Westfalen, Deutschland. Bergische Universität Wuppertal. https://www.wp.uni-wuppertal.de/fileadmin/wp/200910_WS/Bilanzanalyse/Bilanzanalyse_WS0910_Teil01_ohne_Kennwort.pdf

Tiedemann, M. (2021, 25. Januar). *Text Mining – Grundlagen, Methoden und Anwendungsfälle*. Alexander Thamm GmbH. https://www.alexanderthamm.com/de/blog/text-mining-grundlagen-methoden-und-anwendungsfaelle/

Tijare, P. & Rani, P. (2020). Exploring popular topic models. *Journal of physics*, *1706*(1), 012171. https://doi.org/10.1088/1742-6596/1706/1/012171

Tomar, A. (2018, 25. November). *Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained!* Medium. Abgerufen am 7. Juli 2023, von https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045

Vajjala, S., Majumder, B., Surana, H. & Gupta, A. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.

Was ist Digitalisierung? (o. D.). https://www.de.digital/DIGITAL/Navigation/DE/Lagebild/Was-ist-Digitalisierung/was-ist-digitalisierung.html

Was ist Text-Mining? | *IBM*. (o. D.). https://www.ibm.com/de-de/topics/text-mining

Why global accounting standards? (o. D.). IFRS. Abgerufen am 26. April 2023, von https://www.ifrs.org/use-around-the-world/why-global-accounting-standards/

Winter, A. (2023, 24. Januar). *Was ist automatisierte Textanalyse I Marktforschung*. Cogitaris. https://cogitaris.de/was-ist-automatisierte-textanalyse/

Anhang A: Ergänzende Informationen

Anhang A1: Umgebung RapidMiner Studio

RapidMiner ist eine Open-Source-Plattform für Data-Mining und maschinelles Lernen. Die RapidMiner Data Science Plattform umfasst mehrere Produkte, von denen für diese Arbeit RapidMiner Studio genutzt wird. Die Plattform bietet eine grafische, benutzerfreundliche Drag-and-Drop-Oberfläche für die Erstellung von Workflow-basierten Prozessen, ohne Programmierkenntnisse zu benötigen. RapidMiner bietet eine Vielzahl von Funktionen, desgleichen der Möglichkeiten, Daten aus verschiedenen Quellen zu importieren und zu integrieren, Daten zu transformieren und zu bereinigen, Modelle zu trainieren und zu validieren, Vorhersagen zu treffen und Ergebnisse zu visualisieren und bildet damit den gesamten Data-Science Lebenszyklus ab. RapidMiner unterstützt eine Vielzahl von Datenquellen, einschließlich Datenbanken, Dateien und Webservices, und bietet eine umfangreiche Sammlung von Algorithmen für maschinelles Lernen, wie Klassifikation, Clustering, Regression, Textanalyse (*Best Data Science Platform for Your Enterprise | RapidMiner*, 2022). Für deren Realisierung bietet der RapidMiner Marketplace verschiedene Erweiterungen an, von denen in dieser Arbeit die Erweiterungen *Text Processing* genutzt wird.

RapidMiner vereinfacht den Prozess der Text Mining Analyse durch die Bereitstellung von Tools und Funktionen, die es Benutzern ermöglichen, schnell und einfach Erkenntnisse aus ihren Textdaten zu gewinnen. Die Plattform wird von Unternehmen und Organisationen in verschiedenen Branchen eingesetzt, darunter Finanzdienstleistungen, Gesundheitswesen, Einzelhandel, Marketing (*Best Data Science Platform for Your Enterprise | RapidMiner*, 2022).

Anhang A2: Fundamentale Begriffe

Operatoren

Operatoren sind die Bausteine eines RapidMiner-Prozesses. Jeder Operator hat einen Eingangs- und Ausgangsport, zwischen denen die Aktion des Operators stattfindet. Es gibt mehr als 1500 Operatoren, die in unterschiedliche Kategorien unterteilt werden, wie Data Access, Blending, Cleansing oder Modeling (RapidMiner,2014). Sie können ausgewählt und per Drag & Drop an gewünschte Stelle im Prozessfenster platziert werden. Durch das Ziehen einer Linie zwischen den Eingangs- und Ausgangsports werden die Operatoren verbunden und ein Prozess entsteht.

Prozess

Ein Prozess ist das Produkt mehrerer miteinander verbundener Operatoren. Sehr wichtig ist die Reihenfolge der Operatoren, damit die Ausgaben zu den Eingangports passen und sich die Bearbeitungsschritte innerhalb des Prozrsses nicht im Weg stehen. Der Prozess wird in der Entwurfsansicht erstellt und die Ergebnisse werden in der Ergebnisansicht dargestellt.

Subprocess Level

Operatoren können Unter- bzw. Subprozesse haben, die durch einen Doppelklick auf den Operator angezeigt werden. Hier können weitere Operatoren eingefügt werden, welche im Hintergrund ausgeführt werden.

Parameter

Parameter sind die Einstellungen eines jeden Operators. Sie bestimmen und steuern dessen Ausführung. Sie werden mit einem Klick auf den Operator auf der rechten Seite der Benutzeroberfläche angezeigt. Die Parameter sind von Operator zu Operator unterschiedlich und besitzen stets eine Standardeinstellung, wie auch Empfehlungen, die über einen grünen Pfeil eingesehen werden können.

Ports

Ein- und Ausgangports verbinden die Operatoren und formen sie so zu einem funktionierenden Prozess. Die Hilfeansicht auf der rechten Seite der Benutzeroberfläche gibt genau an, welche Art von Eingabe ein bestimmter Operator benötigt und welche Art von Ausgabe er erzeugt. Diese Eingabe bzw. Ausgabe kann zum Beispiel ein Dokument oder eine Sammlung von Dokumenten sein.

Anhang B: Stopwords-Dictionary

Adidas
Adjusted
Aktie
Aktien
aktionär
Aktionär
Aktionäre
Aktionären
Aktionärinnen
Allianz
Anhang
anleihen
Anteil
April
Aufsichtsrat
Aufsichtsrats
Aufwendungen
August
BASF
Bayer
Beiersdorf
Beizulegenden
Bereinigt
bereinigte
Bericht
Berichtsjahr
Berlin
Bestätigungsvermerk
Bestätigungsvermerke
beziehungsweise
Bilanz
BMW
Brenntag
buwog
BUWOG
Cashflow
Cement
clearstream
Continental
Corporate
Covestro
Daimler
Delivery
DeliveryHero
deutsch
Deutsch
Deutsche
deutscher
Deutschland
Dezember
Ebit
Ebitda
Eigenkapital
Energy
EON
Erfasst
erfolgt
Ergebnis
Ertrag
Erträge
eurex
Februar
Finanzielle
finanziellen
Forderungen
Fresenius
Gegenüber
Gemäß
Gesamt
Gesamte
Gesamten
Geschäfts
Geschäftsbereiche
Geschäftsbericht
Geschäftsjahr
Geschäftsleitung
Gesellschaft
Gesetzlich
Gesetzliche
Gesetzlichen
Gesetzlicher
Gewinn
Governance
Group
gruppe
Gruppe
Hannover
Hanson
Healthineers
HeidelberAbschluss
Heidelberg
HeidelbergCement
Helio
Helios
Hellofresh
Henkel
Hero
Hinaus
immaterielle

Impressum
Infineon
Information
Informationen
informationenzusammengefass
Insbesondere
Insgesamt
Jahre
Januar
Jeweiligen
Juli
Juni
Kapitalanlage
Kapitalanlagen
Konzern
Konzernabschluss
Konzernanhang
Konzernbereich
Konzernbereiche
Konzernlagebericht
Konzerns
Kunden
Kundinnen
Kurzfristig
Lagebericht
lageberichtbmw
Langfristig
Langfristigen
Mai
Maidenhead
Management
März
Mercedes
merck
Millionen
Mitarbeiter
Mitarbeiterinnen
Mitarbeitern
Mitglieder
München
Münchner
Munich
nicht
nichtfinanziell
November
Oktober
Porsche
Produkte
Prozent
Puma
Rahmen
Reebok
Risiken
Rück
Rückstellungen
RWE
SAP
Sartorius
September
Siemens
Sonstige
Sonstige
Sonstigen
sonstiges
Stakeholder
Stakeholders
Symrise
Telekom
Übrig
Übrige
Übrigen
übriges
Umfasst
Umsatz
umsatzerlöse
Unternehmen
Unternehmensbereich
Unternehmensbereiche
Verbindlichkeiten
Vergütung
Vergütungsbericht
Vermögenswert
Vermögenswerte
Volkswagen
Vonovia
Vorjahr
Vorstand
Vorstands
Vorzugsaktie
Vorzugsaktien
Wertpapiere
wesentlich
Wesentlichen
Woowa
Zalando
Zusammengefasst
Zusammengefasste
zusammengefassten
Zusammengefasster
Zusammenhang
Zusätzlich
Zusätzliche
Zusätzlichen

Anhang C: LogLikelihood Werte

Number of topics	LogLikelihood [iterations 1.000; no n-grams]	LogLikelihood [iterations 10.000; no n-grams]	LogLikelihood [iterations 15.000; no n-grams]
5	-13350052,755	-13357870,421	-13351133,881
6	-13337880,347	-13285471,713	-13287757,678
7	-13227805,275	-13219872,542	-13220334,729
8	-13212315,195	-13208414,821	-13213258,726
9	-13198604,382	-13195249,660	-13198433,489
10	-13175839,871	-13174152,541	-13171183,483
11	-13433969,057	-13175414,605	-13175404,572
12	-13387482,271	-13175414,605	-13171863,251
13	-13368205,992	-13176701,639	-13175923,060
14	-13390266,784	-13155047,969	-13158748,781
15	-13209409,622	-13243186,357	-13247438,594
16	-13326719,780	-13168643,512	-13179862,578
17	-13347486,038	-13149191,769	-13158294,263
18	-13438567,242	-13342894,938	-13231926,456
19	-13342443,040	-13192927,465	-13159392,921
20	-13526036,165	-13227065,978	-13226975,046
25	-13301606,341	-13294296,263	-13294885,791
30	-13537051,425	-13379097,916	-13316273,382
35	-13439906,834	-13342704,313	-13312696,895
40	-13412231,755	-13314672,985	-13285236,584
45	-13432436,249	-13307280,196	-13291683,359
50	-13407446,007	-13312472,180	-13290690,795
55	-13452758,690	-13307902,937	-13309091,528
60	-13433647,623	-13361836,619	-13341598,153
65	-13450428,575	-13396268,384	-13380157,432

Anhang D: Ergebnisse LDA Modell LL1

LDA Model with 17 topics
alphaSum = 5.038531270494173
beta = 0.033049438207988405

```
Topic 0 tokens=53519.0000      document_entropy=1.7800        word-
length=7.6000   coherence=-1.4695      uniform_dist=3.5082
        corpus_dist=1.9114     eff_num_words=734.8546 token-doc-diff=0.0124
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0690        exclusivity=0.3316
  mitarbei      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0989    corpus_dist=0.0356     token-doc-diff=0.0079
        exclusivity=0.3353
  global        word-length=6.0000     coherence=-0.2227
        uniform_dist=0.0499    corpus_dist=0.0104     token-doc-diff=0.0000
        exclusivity=0.2963
  produktio     word-length=9.0000     coherence=-0.2227
        uniform_dist=0.0432    corpus_dist=0.0155     token-doc-diff=0.0008
        exclusivity=0.2935
  rohstoff      word-length=8.0000     coherence=-0.2227
        uniform_dist=0.0407    corpus_dist=0.0176     token-doc-diff=0.0007
        exclusivity=0.4581
  standor       word-length=7.0000     coherence=-0.1820
        uniform_dist=0.0368    corpus_dist=0.0113     token-doc-
diff=0.0030    exclusivity=0.2750
```

```
Topic 1 tokens=37757.0000      document_entropy=1.8768        word-
length=6.4000   coherence=-5.3360      uniform_dist=3.6323
        corpus_dist=2.1710     eff_num_words=699.5231 token-doc-diff=0.0687
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.1000        exclusivity=0.5647
  optio word-length=5.0000     coherence=0.0000       uniform_dist=0.1137
        corpus_dist=0.0451     token-doc-diff=0.0017  exclusivity=0.7452
  finanzjah     word-length=9.0000     coherence=-2.9137
        uniform_dist=0.0625    corpus_dist=0.0348     token-doc-diff=0.0508
        exclusivity=0.9985
  servic        word-length=6.0000     coherence=-0.2359
        uniform_dist=0.0514    corpus_dist=0.0139     token-doc-diff=0.0053
        exclusivity=0.1978
  anzahl        word-length=6.0000     coherence=-0.4913
        uniform_dist=0.0381    corpus_dist=0.0098     token-doc-diff=0.0028
        exclusivity=0.4730
  gewahr        word-length=6.0000     coherence=-0.4913
        uniform_dist=0.0299    corpus_dist=0.0072     token-doc-diff=0.0082
        exclusivity=0.4091
```

```
Topic 2 tokens=39408.0000      document_entropy=1.6398        word-
length=7.8000   coherence=-6.5838      uniform_dist=3.6593
        corpus_dist=2.0078     eff_num_words=731.4703 token-doc-diff=0.0044
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0741        exclusivity=0.5936
  kontak        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0545    corpus_dist=0.0268     token-doc-diff=0.0000
        exclusivity=0.8595
```

```
  varia word-length=5.0000     coherence=-0.2226      uniform_dist=0.0540
        corpus_dist=0.0279     token-doc-diff=0.0010  exclusivity=0.7991
  finanzkal     word-length=9.0000     coherence=-0.9130
        uniform_dist=0.0521    corpus_dist=0.0263     token-doc-diff=0.0031
        exclusivity=0.8380
  limited       word-length=7.0000     coherence=-0.5086
        uniform_dist=0.0446    corpus_dist=0.0136     token-doc-diff=0.0002
        exclusivity=0.1437
  finanzschuld word-length=12.0000    coherence=-1.0254
        uniform_dist=0.0323    corpus_dist=0.0138     token-doc-diff=0.0000
        exclusivity=0.3276
```

```
Topic 3 tokens=864579.0000     document_entropy=3.5995        word-
length=7.4000  coherence=0.0000       uniform_dist=3.4533
        corpus_dist=0.2783     eff_num_words=1158.1065        token-doc-
diff=0.0002    rank_1_docs=0.9737     allocation_ratio=0.9474
        allocation_count=1.0000        exclusivity=0.4675
  erwar word-length=5.0000     coherence=0.0000       uniform_dist=0.0222
        corpus_dist=0.0018     token-doc-diff=0.0000  exclusivity=0.4242
  leistung      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0221    corpus_dist=0.0020     token-doc-diff=0.0000
        exclusivity=0.8990
  enthal        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0217    corpus_dist=0.0017     token-doc-diff=0.0000
        exclusivity=0.3972
  veranderung   word-length=11.0000    coherence=0.0000
        uniform_dist=0.0200    corpus_dist=0.0014     token-doc-diff=0.0001
        exclusivity=0.2768
  prufung       word-length=7.0000     coherence=0.0000
        uniform_dist=0.0197    corpus_dist=0.0016     token-doc-diff=0.0001
        exclusivity=0.3404
```

```
Topic 4 tokens=24263.0000      document_entropy=1.0944        word-
length=8.4000  coherence=-9.4600      uniform_dist=3.8062
        corpus_dist=2.5435     eff_num_words=573.2347 token-doc-diff=0.0053
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0714        exclusivity=0.6682
  investor      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0826    corpus_dist=0.0330     token-doc-diff=0.0007
        exclusivity=0.7060
  vertr word-length=5.0000     coherence=-0.4036      uniform_dist=0.0826
        corpus_dist=0.0248     token-doc-diff=0.0000  exclusivity=0.3766
  versicherung word-length=12.0000    coherence=-0.6890
        uniform_dist=0.0724    corpus_dist=0.0271     token-doc-diff=0.0006
        exclusivity=0.3180
  clearing      word-length=8.0000     coherence=-1.3740
        uniform_dist=0.0481    corpus_dist=0.0304     token-doc-diff=0.0006
        exclusivity=0.9561
  nettoerlo     word-length=9.0000     coherence=-1.3740
        uniform_dist=0.0475    corpus_dist=0.0306     token-doc-diff=0.0032
        exclusivity=0.9843
```

```
Topic 5 tokens=51834.0000      document_entropy=0.8526        word-
length=7.4000  coherence=-1.7679      uniform_dist=3.8415
        corpus_dist=1.8935     eff_num_words=490.1903 token-doc-diff=0.0155
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.1250        exclusivity=0.5027
```

```
  medical      word-length=7.0000     coherence=0.0000
       uniform_dist=0.1522    corpus_dist=0.0673     token-doc-diff=0.0047
       exclusivity=0.8856
  patie word-length=5.0000     coherence=-0.2503      uniform_dist=0.1057
       corpus_dist=0.0475     token-doc-diff=0.0023  exclusivity=0.7541
  nichtfinanziell      word-length=15.0000    coherence=-0.2503
       uniform_dist=0.0576    corpus_dist=0.0160     token-doc-diff=0.0012
       exclusivity=0.3033
  angab word-length=5.0000     coherence=-0.2503      uniform_dist=0.0524
       corpus_dist=0.0102     token-doc-diff=0.0020  exclusivity=0.3603
  covid word-length=5.0000     coherence=-0.2503      uniform_dist=0.0408
       corpus_dist=0.0109     token-doc-diff=0.0052  exclusivity=0.2102

---

Topic 6 tokens=26576.0000      document_entropy=1.0000        word-
length=6.4000  coherence=-2.3436      uniform_dist=3.6664
       corpus_dist=2.5301     eff_num_words=753.4771 token-doc-diff=0.0116
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0909        exclusivity=0.4172
  energie      word-length=7.0000     coherence=0.0000
       uniform_dist=0.0597    corpus_dist=0.0223     token-doc-diff=0.0029
       exclusivity=0.2998
  taxonomie    word-length=9.0000     coherence=-0.4027
       uniform_dist=0.0406    corpus_dist=0.0174     token-doc-diff=0.0031
       exclusivity=0.3691
  strom word-length=5.0000     coherence=-0.1812      uniform_dist=0.0318
       corpus_dist=0.0137     token-doc-diff=0.0040  exclusivity=0.3376
  aspek word-length=5.0000     coherence=-0.3346      uniform_dist=0.0313
       corpus_dist=0.0128     token-doc-diff=0.0001  exclusivity=0.6731
  pensio       word-length=6.0000     coherence=-0.5561
       uniform_dist=0.0310    corpus_dist=0.0119     token-doc-diff=0.0016
       exclusivity=0.4063

---

Topic 7 tokens=55246.0000      document_entropy=3.2734        word-
length=6.0000  coherence=-0.4985      uniform_dist=3.3463
       corpus_dist=2.5491     eff_num_words=928.7887 token-doc-diff=0.0027
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.7308
  variabl      word-length=7.0000     coherence=0.0000
       uniform_dist=0.0570    corpus_dist=0.0158     token-doc-diff=0.0008
       exclusivity=0.3807
  schaf word-length=5.0000     coherence=-0.0317      uniform_dist=0.0508
       corpus_dist=0.0203     token-doc-diff=0.0003  exclusivity=0.8091
  schen word-length=5.0000     coherence=-0.0983      uniform_dist=0.0435
       corpus_dist=0.0214     token-doc-diff=0.0000  exclusivity=0.7729
  unterneh     word-length=8.0000     coherence=-0.1334
       uniform_dist=0.0374    corpus_dist=0.0193     token-doc-diff=0.0002
       exclusivity=0.8488
  grund word-length=5.0000     coherence=-0.0317      uniform_dist=0.0369
       corpus_dist=0.0120     token-doc-diff=0.0014  exclusivity=0.8428

---

Topic 8 tokens=18231.0000      document_entropy=0.6486        word-
length=6.2000  coherence=-3.2003      uniform_dist=3.6784
       corpus_dist=2.8575     eff_num_words=671.2024 token-doc-diff=0.0218
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0833        exclusivity=0.4089
  limited      word-length=7.0000     coherence=0.0000
       uniform_dist=0.1365    corpus_dist=0.0543     token-doc-diff=0.0129
       exclusivity=0.3559
```

```
  kompo word-length=5.0000     coherence=-0.4000      uniform_dist=0.0603
        corpus_dist=0.0243     token-doc-diff=0.0005  exclusivity=0.5161
  tranch       word-length=6.0000     coherence=-0.4000
        uniform_dist=0.0389    corpus_dist=0.0132     token-doc-diff=0.0002
        exclusivity=0.2082
  holding      word-length=7.0000     coherence=-0.4000
        uniform_dist=0.0361    corpus_dist=0.0117     token-doc-diff=0.0072
        exclusivity=0.2507
  produc       word-length=6.0000     coherence=-0.4000
        uniform_dist=0.0330    corpus_dist=0.0200     token-doc-diff=0.0011
        exclusivity=0.7137
```

```
Topic 9 tokens=51534.0000      document_entropy=1.4407       word-
length=11.0000 coherence=-6.1538      uniform_dist=3.4581
        corpus_dist=2.0957     eff_num_words=727.5223 token-doc-diff=0.0168
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0526        exclusivity=0.7483
  fahrzeug     word-length=8.0000     coherence=0.0000
        uniform_dist=0.1278    corpus_dist=0.0572     token-doc-diff=0.0016
        exclusivity=0.8821
  automobil    word-length=9.0000     coherence=-0.4503
        uniform_dist=0.0610    corpus_dist=0.0301     token-doc-diff=0.0005
        exclusivity=0.8825
  informationenzusammengefass  word-length=27.0000    coherence=-1.6914
        uniform_dist=0.0401    corpus_dist=0.0204     token-doc-diff=0.0090
        exclusivity=0.7685
  equity       word-length=6.0000     coherence=-0.4503
        uniform_dist=0.0372    corpus_dist=0.0116     token-doc-diff=0.0039
        exclusivity=0.2496
  truck word-length=5.0000     coherence=-0.6770      uniform_dist=0.0355
        corpus_dist=0.0194     token-doc-diff=0.0019  exclusivity=0.9588
```

```
Topic 10       tokens=21528.0000      document_entropy=0.6578       word-
length=8.0000  coherence=-11.8698     uniform_dist=3.6134
        corpus_dist=2.8256     eff_num_words=764.5800 token-doc-diff=0.0161
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.2500        exclusivity=0.8203
  gefass       word-length=6.0000     coherence=0.0000
        uniform_dist=0.0647    corpus_dist=0.0348     token-doc-diff=0.0054
        exclusivity=0.9481
  immobilie    word-length=9.0000     coherence=-1.0771
        uniform_dist=0.0549    corpus_dist=0.0267     token-doc-diff=0.0072
        exclusivity=0.6708
  developm     word-length=8.0000     coherence=-0.4000
        uniform_dist=0.0477    corpus_dist=0.0244     token-doc-diff=0.0015
        exclusivity=0.6204
  victoriah    word-length=9.0000     coherence=-1.0771
        uniform_dist=0.0370    corpus_dist=0.0255     token-doc-diff=0.0010
        exclusivity=0.9976
  nanziell     word-length=8.0000     coherence=-3.4423
        uniform_dist=0.0370    corpus_dist=0.0237     token-doc-diff=0.0010
        exclusivity=0.8644
```

```
Topic 11       tokens=43024.0000      document_entropy=1.1145       word-
length=12.6000 coherence=-1.9624      uniform_dist=3.5327
        corpus_dist=2.2292     eff_num_words=654.7097 token-doc-diff=0.0200
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.3333        exclusivity=0.8070
```

```
  schad word-length=5.0000     coherence=0.0000       uniform_dist=0.1196
        corpus_dist=0.0538     token-doc-diff=0.0110  exclusivity=0.9377
  ruckversicherung     word-length=16.0000    coherence=0.0000
        uniform_dist=0.0917    corpus_dist=0.0467     token-doc-diff=0.0006
        exclusivity=0.9874
  versicherungstechnisch       word-length=22.0000    coherence=-0.2215
        uniform_dist=0.0501    corpus_dist=0.0277     token-doc-diff=0.0034
        exclusivity=0.9934
  versicherung word-length=12.0000    coherence=-0.2215
        uniform_dist=0.0400    corpus_dist=0.0129     token-doc-diff=0.0024
        exclusivity=0.1983
  insuranc     word-length=8.0000     coherence=-0.5065
        uniform_dist=0.0391    corpus_dist=0.0208     token-doc-diff=0.0026
        exclusivity=0.9183
```

```
Topic 12       tokens=53638.0000      document_entropy=0.3468        word-
length=8.8000  coherence=-5.7942      uniform_dist=3.6144
        corpus_dist=1.7048     eff_num_words=761.1788 token-doc-diff=0.0132
        rank_1_docs=0.0455     allocation_ratio=0.3333
        allocation_count=0.0455        exclusivity=0.4657
  zeitw word-length=5.0000     coherence=0.0000       uniform_dist=0.0954
        corpus_dist=0.0208     token-doc-diff=0.0034  exclusivity=0.3472
  finanzinstitu word-length=13.0000    coherence=-0.5086
        uniform_dist=0.0587    corpus_dist=0.0291     token-doc-diff=0.0003
        exclusivity=0.9394
  risiko       word-length=6.0000     coherence=-0.5086
        uniform_dist=0.0396    corpus_dist=0.0072     token-doc-diff=0.0053
        exclusivity=0.3046
  bestimm      word-length=7.0000     coherence=-1.1963
        uniform_dist=0.0334    corpus_dist=0.0056     token-doc-diff=0.0015
        exclusivity=0.3844
  verpflichtung word-length=13.0000    coherence=-0.6899
        uniform_dist=0.0286    corpus_dist=0.0052     token-doc-diff=0.0027
        exclusivity=0.3531
```

```
Topic 13       tokens=36460.0000      document_entropy=0.8289        word-
length=7.6000  coherence=-6.5629      uniform_dist=3.7138
        corpus_dist=2.1342     eff_num_words=523.9113 token-doc-diff=0.0101
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.1333        exclusivity=0.5805
  mobil word-length=5.0000     coherence=0.0000       uniform_dist=0.1388
        corpus_dist=0.0679     token-doc-diff=0.0013  exclusivity=0.8947
  technologie  word-length=11.0000    coherence=0.0000
        uniform_dist=0.1179    corpus_dist=0.0394     token-doc-diff=0.0001
        exclusivity=0.3829
  deutsch      word-length=7.0000     coherence=-0.4036
        uniform_dist=0.0599    corpus_dist=0.0164     token-doc-diff=0.0041
        exclusivity=0.4620
  mobilfunk    word-length=9.0000     coherence=-1.3740
        uniform_dist=0.0586    corpus_dist=0.0332     token-doc-diff=0.0041
        exclusivity=0.9815
  servic       word-length=6.0000     coherence=-0.6899
        uniform_dist=0.0466    corpus_dist=0.0121     token-doc-diff=0.0005
        exclusivity=0.1815
```

```
Topic 14       tokens=33790.0000      document_entropy=1.7537        word-
length=8.6000  coherence=-2.7918      uniform_dist=3.9198
        corpus_dist=2.1113     eff_num_words=433.5924 token-doc-diff=0.0110
```

```
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0278        exclusivity=0.4817
  integrier     word-length=9.0000     coherence=0.0000
        uniform_dist=0.1452    corpus_dist=0.0641     token-doc-diff=0.0014
        exclusivity=0.8379
  mitarbei      word-length=8.0000     coherence=-0.1757
        uniform_dist=0.1371    corpus_dist=0.0528     token-doc-diff=0.0011
        exclusivity=0.4352
  nachhaltigkei word-length=13.0000    coherence=-0.2556
        uniform_dist=0.1107    corpus_dist=0.0415     token-doc-diff=0.0006
        exclusivity=0.6345
  emissio       word-length=7.0000     coherence=-0.4049
        uniform_dist=0.0487    corpus_dist=0.0146     token-doc-diff=0.0027
        exclusivity=0.3157
  servic        word-length=6.0000     coherence=-0.3132
        uniform_dist=0.0481    corpus_dist=0.0126     token-doc-diff=0.0052
        exclusivity=0.1854

---

Topic 15        tokens=25967.0000      document_entropy=1.2969        word-
length=10.2000  coherence=-3.1275      uniform_dist=3.5870
        corpus_dist=2.5286     eff_num_words=716.4412 token-doc-diff=0.0014
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0556        exclusivity=0.3195
  vertr word-length=5.0000     coherence=0.0000       uniform_dist=0.0693
        corpus_dist=0.0197     token-doc-diff=0.0003  exclusivity=0.3252
  versicherung  word-length=12.0000    coherence=-0.2503
        uniform_dist=0.0622    corpus_dist=0.0225     token-doc-diff=0.0002
        exclusivity=0.2807
  technologie   word-length=11.0000    coherence=-0.3553
        uniform_dist=0.0617    corpus_dist=0.0170     token-doc-diff=0.0004
        exclusivity=0.2121
  fortgefuhr    word-length=10.0000    coherence=-0.3553
        uniform_dist=0.0513    corpus_dist=0.0142     token-doc-diff=0.0000
        exclusivity=0.2545
  geschaftsfeld word-length=13.0000    coherence=-0.5849
        uniform_dist=0.0389    corpus_dist=0.0177     token-doc-diff=0.0004
        exclusivity=0.5252

---

Topic 16        tokens=31146.0000      document_entropy=1.0352        word-
length=8.0000   coherence=-5.7765      uniform_dist=3.5989
        corpus_dist=2.3758     eff_num_words=752.3879 token-doc-diff=0.0052
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0714        exclusivity=0.5316
  scienc        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0903    corpus_dist=0.0486     token-doc-diff=0.0022
        exclusivity=0.9717
  forschung     word-length=9.0000     coherence=-0.5849
        uniform_dist=0.0488    corpus_dist=0.0167     token-doc-diff=0.0003
        exclusivity=0.2511
  organisch     word-length=9.0000     coherence=-0.4036
        uniform_dist=0.0429    corpus_dist=0.0190     token-doc-diff=0.0005
        exclusivity=0.5198
  healthcar     word-length=9.0000     coherence=-0.8411
        uniform_dist=0.0392    corpus_dist=0.0197     token-doc-diff=0.0001
        exclusivity=0.5271
  solutio       word-length=7.0000     coherence=-0.8411
        uniform_dist=0.0353    corpus_dist=0.0138     token-doc-diff=0.0021
        exclusivity=0.3886
```

Anhang E: Ausgewählte Ergebnisse LDA Modell PX1

```
LDA Model with 50 topics
alphaSum = 13.098772274585459
beta = 0.011162723925289228
```

```
Topic 12        tokens=232741.0000      document_entropy=3.6214         word-
length=9.6000   coherence=0.0000        uniform_dist=4.5667
        corpus_dist=1.3341      eff_num_words=422.5773 token-doc-diff=0.0072
        rank_1_docs=0.1053      allocation_ratio=0.0000
        allocation_count=0.0000         exclusivity=0.3775
  leistung      word-length=8.0000      coherence=0.0000
        uniform_dist=0.0866     corpus_dist=0.0215      token-doc-diff=0.0032
        exclusivity=0.3932
  prufung       word-length=7.0000      coherence=0.0000
        uniform_dist=0.0705     corpus_dist=0.0163      token-doc-diff=0.0006
        exclusivity=0.3065
  hauptversammlung      word-length=16.0000     coherence=0.0000
        uniform_dist=0.0534     corpus_dist=0.0151      token-doc-diff=0.0002
        exclusivity=0.6713
  vermogensw    word-length=10.0000     coherence=0.0000
        uniform_dist=0.0435     corpus_dist=0.0114      token-doc-diff=0.0016
        exclusivity=0.3113
  vorsitx       word-length=7.0000      coherence=0.0000
        uniform_dist=0.0433     corpus_dist=0.0100      token-doc-diff=0.0016
        exclusivity=0.2052
```

```
Topic 16        tokens=10892.0000       document_entropy=0.0018         word-
length=11.2000  coherence=0.0000        uniform_dist=4.5831
        corpus_dist=3.4390      eff_num_words=267.0651 token-doc-diff=0.0176
        rank_1_docs=0.5000      allocation_ratio=0.0000
        allocation_count=0.0000         exclusivity=0.5103
  informationenzusammengefass  word-length=27.0000     coherence=0.0000
        uniform_dist=0.2311     corpus_dist=0.1402      token-doc-diff=0.0108
        exclusivity=0.8212
  fahrzeug      word-length=8.0000      coherence=0.0000
        uniform_dist=0.1193     corpus_dist=0.0528      token-doc-diff=0.0000
        exclusivity=0.2070
  kompo word-length=5.0000      coherence=0.0000        uniform_dist=0.1059
        corpus_dist=0.0471      token-doc-diff=0.0003   exclusivity=0.3159
  emissio       word-length=7.0000      coherence=0.0000
        uniform_dist=0.0826     corpus_dist=0.0287      token-doc-diff=0.0023
        exclusivity=0.2117
  brillianc     word-length=9.0000      coherence=0.0000
        uniform_dist=0.0719     corpus_dist=0.0527      token-doc-diff=0.0041
        exclusivity=0.9958
```

```
Topic 25        tokens=63138.0000       document_entropy=3.4025         word-
length=7.8000   coherence=0.0000        uniform_dist=4.7738
        corpus_dist=2.1993      eff_num_words=274.1439 token-doc-diff=0.0011
        rank_1_docs=0.0000      allocation_ratio=0.0000
        allocation_count=0.0000         exclusivity=0.3777
  stand word-length=5.0000      coherence=0.0000        uniform_dist=0.1142
        corpus_dist=0.0369      token-doc-diff=0.0008   exclusivity=0.4219
  verkauf       word-length=7.0000      coherence=0.0000
        uniform_dist=0.0966     corpus_dist=0.0345      token-doc-diff=0.0000
        exclusivity=0.4190
```

```
  verausserung word-length=12.0000    coherence=0.0000
        uniform_dist=0.0893    corpus_dist=0.0326     token-doc-diff=0.0001
        exclusivity=0.2396
  buchw word-length=5.0000     coherence=0.0000       uniform_dist=0.0868
        corpus_dist=0.0295     token-doc-diff=0.0001  exclusivity=0.5002
  investitio    word-length=10.0000    coherence=0.0000
        uniform_dist=0.0858    corpus_dist=0.0282     token-doc-diff=0.0002
        exclusivity=0.3078
```

```
Topic 30        tokens=17050.0000      document_entropy=0.4993        word-
length=9.0000  coherence=0.0000        uniform_dist=4.7512
        corpus_dist=2.8089     eff_num_words=270.0898 token-doc-diff=0.0106
        rank_1_docs=0.2500     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.4589
  medical       word-length=7.0000     coherence=0.0000
        uniform_dist=0.2004    corpus_dist=0.0922     token-doc-diff=0.0072
        exclusivity=0.3878
  personlich    word-length=10.0000    coherence=0.0000
        uniform_dist=0.1046    corpus_dist=0.0512     token-doc-diff=0.0002
        exclusivity=0.5504
  patie word-length=5.0000     coherence=0.0000       uniform_dist=0.1037
        corpus_dist=0.0464     token-doc-diff=0.0003  exclusivity=0.2895
  gesellschafteri       word-length=15.0000    coherence=0.0000
        uniform_dist=0.1004    corpus_dist=0.0603     token-doc-diff=0.0004
        exclusivity=0.7194
  grundlag      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0782    corpus_dist=0.0234     token-doc-diff=0.0026
        exclusivity=0.3472
```

```
Topic 33        tokens=12626.0000      document_entropy=0.0000        word-
length=9.0000  coherence=0.0000        uniform_dist=4.2727
        corpus_dist=3.2924     eff_num_words=422.7461 token-doc-diff=0.0055
        rank_1_docs=1.0000     allocation_ratio=0.0000
        allocation_count=1.0000        exclusivity=0.6860
  immobilie     word-length=9.0000     coherence=0.0000
        uniform_dist=0.1046    corpus_dist=0.0551     token-doc-diff=0.0024
        exclusivity=0.4562
  developm      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0888    corpus_dist=0.0488     token-doc-diff=0.0006
        exclusivity=0.6514
  victoriah     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0685    corpus_dist=0.0490     token-doc-diff=0.0002
        exclusivity=0.9957
  osterreich    word-length=10.0000    coherence=0.0000
        uniform_dist=0.0583    corpus_dist=0.0326     token-doc-diff=0.0011
        exclusivity=0.4712
  propertie     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0571    corpus_dist=0.0398     token-doc-diff=0.0013
        exclusivity=0.8558
        corpus_dist=0.0151     token-doc-diff=0.0072  exclusivity=0.0954
```

```
Topic 35        tokens=8722.0000       document_entropy=0.0022        word-
length=7.0000  coherence=0.0000        uniform_dist=4.4514
        corpus_dist=3.5217     eff_num_words=354.7698 token-doc-diff=0.0175
        rank_1_docs=0.5000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.4772
  gefass        word-length=6.0000     coherence=0.0000
        uniform_dist=0.1797    corpus_dist=0.1064     token-doc-diff=0.0088
        exclusivity=0.8983
```

```
  nanziell      word-length=8.0000     coherence=0.0000
        uniform_dist=0.1056    corpus_dist=0.0726     token-doc-diff=0.0001
        exclusivity=0.8977
  tranch        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0983    corpus_dist=0.0408     token-doc-diff=0.0000
        exclusivity=0.1477
  lokal word-length=5.0000     coherence=0.0000       uniform_dist=0.0669
        corpus_dist=0.0255     token-doc-diff=0.0025  exclusivity=0.2713
  transaktio    word-length=10.0000    coherence=0.0000
        uniform_dist=0.0514    corpus_dist=0.0170     token-doc-diff=0.0062
        exclusivity=0.1708
  holzmi        word-length=6.0000     coherence=-1.0912
        uniform_dist=0.0556    corpus_dist=0.0483     token-doc-diff=0.0042
        exclusivity=0.9706

---

Topic 38        tokens=10610.0000      document_entropy=0.2295        word-
length=9.2000  coherence=0.0000       uniform_dist=4.5667
        corpus_dist=3.1368     eff_num_words=324.3013 token-doc-diff=0.0126
        rank_1_docs=0.3333     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.4859
  clearing      word-length=8.0000     coherence=0.0000
        uniform_dist=0.1360    corpus_dist=0.0923     token-doc-diff=0.0046
        exclusivity=0.9958
  nettoerlo     word-length=9.0000     coherence=0.0000
        uniform_dist=0.1221    corpus_dist=0.0836     token-doc-diff=0.0027
        exclusivity=0.9822
  performanc    word-length=10.0000    coherence=0.0000
        uniform_dist=0.1175    corpus_dist=0.0440     token-doc-diff=0.0003
        exclusivity=0.1655
  mitarbei      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0925    corpus_dist=0.0328     token-doc-diff=0.0023
        exclusivity=0.0960
  erlauterung   word-length=11.0000    coherence=0.0000
        uniform_dist=0.0896    corpus_dist=0.0370     token-doc-diff=0.0027
        exclusivity=0.1902

---

Topic 41        tokens=11878.0000      document_entropy=0.0000        word-
length=7.6000  coherence=0.0000       uniform_dist=4.2845
        corpus_dist=3.2073     eff_num_words=377.1247 token-doc-diff=0.0374
        rank_1_docs=1.0000     allocation_ratio=0.0000
        allocation_count=1.0000        exclusivity=0.3516
  limited       word-length=7.0000     coherence=0.0000
        uniform_dist=0.2190    corpus_dist=0.0943     token-doc-diff=0.0231
        exclusivity=0.2809
  kompo word-length=5.0000     coherence=0.0000       uniform_dist=0.0853
        corpus_dist=0.0366     token-doc-diff=0.0000  exclusivity=0.2630
  holding       word-length=7.0000     coherence=0.0000
        uniform_dist=0.0537    corpus_dist=0.0193     token-doc-diff=0.0039
        exclusivity=0.1336
  tranch        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0518    corpus_dist=0.0189     token-doc-diff=0.0044
        exclusivity=0.0867
  langfristbonu word-length=13.0000    coherence=0.0000
        uniform_dist=0.0468    corpus_dist=0.0357     token-doc-diff=0.0059
        exclusivity=0.9940

---

Topic 42        tokens=262740.0000     document_entropy=3.5498        word-
length=7.8000  coherence=0.0000       uniform_dist=4.6710
        corpus_dist=1.3081     eff_num_words=385.7643 token-doc-diff=0.0019
```

```
        rank_1_docs=0.0263     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.3995
  erwar word-length=5.0000     coherence=0.0000       uniform_dist=0.0884
        corpus_dist=0.0211     token-doc-diff=0.0012  exclusivity=0.4452
  enthal       word-length=6.0000     coherence=0.0000
        uniform_dist=0.0696    corpus_dist=0.0151     token-doc-diff=0.0000
        exclusivity=0.4968
  auswirkung   word-length=10.0000    coherence=0.0000
        uniform_dist=0.0685    corpus_dist=0.0172     token-doc-diff=0.0000
        exclusivity=0.5651
  veranderung  word-length=11.0000    coherence=0.0000
        uniform_dist=0.0634    corpus_dist=0.0134     token-doc-diff=0.0002
        exclusivity=0.2349
  geschaf      word-length=7.0000     coherence=0.0000
        uniform_dist=0.0587    corpus_dist=0.0140     token-doc-diff=0.0005
        exclusivity=0.2554
```

```
Topic 45       tokens=9260.0000       document_entropy=0.0040       word-
length=8.4000  coherence=0.0000       uniform_dist=4.3803
        corpus_dist=3.3716     eff_num_words=323.0565 token-doc-diff=0.0224
        rank_1_docs=0.3333     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.5450
  vertr word-length=5.0000     coherence=0.0000       uniform_dist=0.1899
        corpus_dist=0.0699     token-doc-diff=0.0052  exclusivity=0.3660
  versicherung word-length=12.0000    coherence=0.0000
        uniform_dist=0.1834    corpus_dist=0.0810     token-doc-diff=0.0044
        exclusivity=0.3038
  zivil word-length=5.0000     coherence=0.0000       uniform_dist=0.1021
        corpus_dist=0.0719     token-doc-diff=0.0004  exclusivity=0.9438
  triebwerk    word-length=9.0000     coherence=0.0000
        uniform_dist=0.0765    corpus_dist=0.0574     token-doc-diff=0.0031
        exclusivity=0.9960
  erlauterung  word-length=11.0000    coherence=0.0000
        uniform_dist=0.0511    corpus_dist=0.0189     token-doc-diff=0.0094
        exclusivity=0.1156
```

```
Topic 47       tokens=159970.0000     document_entropy=3.5713       word-
length=7.0000  coherence=0.0000       uniform_dist=4.7192
        corpus_dist=1.7283     eff_num_words=343.1273 token-doc-diff=0.0016
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.3261
  bewertung    word-length=9.0000     coherence=0.0000
        uniform_dist=0.0832    corpus_dist=0.0235     token-doc-diff=0.0011
        exclusivity=0.3061
  positiv      word-length=7.0000     coherence=0.0000
        uniform_dist=0.0659    corpus_dist=0.0207     token-doc-diff=0.0000
        exclusivity=0.4618
  stark word-length=5.0000     coherence=0.0000       uniform_dist=0.0636
        corpus_dist=0.0165     token-doc-diff=0.0000  exclusivity=0.1485
  anderung     word-length=8.0000     coherence=0.0000
        uniform_dist=0.0620    corpus_dist=0.0147     token-doc-diff=0.0001
        exclusivity=0.1968
  handel       word-length=6.0000     coherence=0.0000
        uniform_dist=0.0571    corpus_dist=0.0187     token-doc-diff=0.0004
        exclusivity=0.5174
```

Anhang F: Ergebnisse LDA Modell Compro

```
LDA Model with 40 topics
alphaSum = 11.136156840168322
beta = 0.014005539683537226

------------------------------------------------------------------------------

Topic 0 tokens=14826.0000      document_entropy=0.4286        word-
length=10.4000  coherence=0.0000       uniform_dist=4.1595
        corpus_dist=3.0850     eff_num_words=394.5588 token-doc-diff=0.0264
        rank_1_docs=0.2000     allocation_ratio=0.0000
        allocation_count=0.2000        exclusivity=0.5235
  ruckversicherung     word-length=16.0000    coherence=0.0000
        uniform_dist=0.1852    corpus_dist=0.1019     token-doc-diff=0.0114
        exclusivity=0.6042
  schad word-length=5.0000     coherence=0.0000       uniform_dist=0.1310
        corpus_dist=0.0597     token-doc-diff=0.0022  exclusivity=0.3171
  festverzinslich      word-length=15.0000    coherence=0.0000
        uniform_dist=0.0595    corpus_dist=0.0306     token-doc-diff=0.0035
        exclusivity=0.4685
  ruckversich  word-length=11.0000    coherence=0.0000
        uniform_dist=0.0564    corpus_dist=0.0358     token-doc-diff=0.0042
        exclusivity=0.6047
  perso word-length=5.0000     coherence=0.0000       uniform_dist=0.0524
        corpus_dist=0.0206     token-doc-diff=0.0052  exclusivity=0.6228

------------------------------------------------------------------------------

Topic 1 tokens=13681.0000      document_entropy=0.5971        word-
length=9.4000   coherence=-7.3957      uniform_dist=4.0104
        corpus_dist=3.2352     eff_num_words=592.5850 token-doc-diff=0.0324
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.5560
  health       word-length=6.0000     coherence=0.0000
        uniform_dist=0.0896    corpus_dist=0.0505     token-doc-diff=0.0007
        exclusivity=0.5870
  scienc       word-length=6.0000     coherence=0.0000
        uniform_dist=0.0714    corpus_dist=0.0374     token-doc-diff=0.0145
        exclusivity=0.3626
  pharmaceutical       word-length=14.0000    coherence=-1.3759
        uniform_dist=0.0669    corpus_dist=0.0454     token-doc-diff=0.0087
        exclusivity=0.8649
  sondereinfluss       word-length=14.0000    coherence=-1.3759
        uniform_dist=0.0625    corpus_dist=0.0301     token-doc-diff=0.0070
        exclusivity=0.2651
  divisio      word-length=7.0000     coherence=-1.3759
        uniform_dist=0.0429    corpus_dist=0.0281     token-doc-diff=0.0015
        exclusivity=0.7003

------------------------------------------------------------------------------

Topic 2 tokens=14279.0000      document_entropy=0.0929        word-
length=7.4000   coherence=-2.4618      uniform_dist=4.2945
        corpus_dist=2.9404     eff_num_words=398.0070 token-doc-diff=0.0358
        rank_1_docs=0.3333     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.2710
  global       word-length=6.0000     coherence=0.0000
        uniform_dist=0.1065    corpus_dist=0.0301     token-doc-diff=0.0010
        exclusivity=0.1823
  mitarbei     word-length=8.0000     coherence=-0.6862
        uniform_dist=0.1036    corpus_dist=0.0377     token-doc-diff=0.0142
        exclusivity=0.1290
```

```
  rohstoff     word-length=8.0000     coherence=0.0000
       uniform_dist=0.0798    corpus_dist=0.0385     token-doc-diff=0.0065
       exclusivity=0.2831
  solutio      word-length=7.0000     coherence=-1.0894
       uniform_dist=0.0701    corpus_dist=0.0312     token-doc-diff=0.0044
       exclusivity=0.1886
  chemical     word-length=8.0000     coherence=0.0000
       uniform_dist=0.0690    corpus_dist=0.0420     token-doc-diff=0.0098
       exclusivity=0.5720

---

Topic 3 tokens=83581.0000      document_entropy=3.4328        word-
length=5.6000  coherence=0.0000       uniform_dist=4.6236
       corpus_dist=1.9590     eff_num_words=325.6186 token-doc-diff=0.0035
       rank_1_docs=0.0263     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.3420
  anzahl       word-length=6.0000     coherence=0.0000
       uniform_dist=0.1151    corpus_dist=0.0412     token-doc-diff=0.0012
       exclusivity=0.5060
  zeitw word-length=5.0000     coherence=0.0000       uniform_dist=0.1031
       corpus_dist=0.0233     token-doc-diff=0.0003  exclusivity=0.1898
  optio word-length=5.0000     coherence=0.0000       uniform_dist=0.0921
       corpus_dist=0.0351     token-doc-diff=0.0000  exclusivity=0.4429
  servic       word-length=6.0000     coherence=0.0000
       uniform_dist=0.0748    corpus_dist=0.0228     token-doc-diff=0.0006
       exclusivity=0.1354
  gewahr       word-length=6.0000     coherence=0.0000
       uniform_dist=0.0685    corpus_dist=0.0222     token-doc-diff=0.0013
       exclusivity=0.4358

---

Topic 4 tokens=43056.0000      document_entropy=3.0480        word-
length=8.6000  coherence=-0.6660      uniform_dist=4.6828
       corpus_dist=2.5134     eff_num_words=242.6156 token-doc-diff=0.0014
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.3852
  standor      word-length=7.0000     coherence=0.0000
       uniform_dist=0.1562    corpus_dist=0.0663     token-doc-diff=0.0009
       exclusivity=0.6043
  produktio    word-length=9.0000     coherence=-0.0625
       uniform_dist=0.1209    corpus_dist=0.0531     token-doc-diff=0.0000
       exclusivity=0.4480
  emissio      word-length=7.0000     coherence=-0.1292
       uniform_dist=0.1178    corpus_dist=0.0443     token-doc-diff=0.0000
       exclusivity=0.3799
  forschung    word-length=9.0000     coherence=-0.0625
       uniform_dist=0.1173    corpus_dist=0.0484     token-doc-diff=0.0001
       exclusivity=0.3167
  technologie  word-length=11.0000    coherence=-0.0953
       uniform_dist=0.1154    corpus_dist=0.0383     token-doc-diff=0.0003
       exclusivity=0.1773

---

Topic 5 tokens=30670.0000      document_entropy=0.0084        word-
length=8.4000  coherence=-2.7449      uniform_dist=4.2081
       corpus_dist=2.4656     eff_num_words=429.5161 token-doc-diff=0.0340
       rank_1_docs=0.5000     allocation_ratio=0.0000
       allocation_count=0.5000        exclusivity=0.4049
  zeitw word-length=5.0000     coherence=0.0000       uniform_dist=0.1408
       corpus_dist=0.0359     token-doc-diff=0.0200  exclusivity=0.2444
```

```
  finanzinstitu word-length=13.0000    coherence=0.0000
        uniform_dist=0.1073    corpus_dist=0.0572     token-doc-diff=0.0002
        exclusivity=0.9406
  vermogensw    word-length=10.0000    coherence=0.0000
        uniform_dist=0.0499    corpus_dist=0.0138     token-doc-diff=0.0136
        exclusivity=0.3473
  bestimm       word-length=7.0000     coherence=-0.6862
        uniform_dist=0.0487    corpus_dist=0.0102     token-doc-diff=0.0001
        exclusivity=0.2517
  deutsch       word-length=7.0000     coherence=-0.6862
        uniform_dist=0.0478    corpus_dist=0.0120     token-doc-diff=0.0001
        exclusivity=0.2407
```

```
Topic 6 tokens=13862.0000      document_entropy=0.4365        word-
length=8.2000  coherence=-1.3542      uniform_dist=4.3752
        corpus_dist=3.0292     eff_num_words=323.9930 token-doc-diff=0.0131
        rank_1_docs=0.0909     allocation_ratio=0.0000
        allocation_count=0.0909        exclusivity=0.3059
  investor      word-length=8.0000     coherence=0.0000
        uniform_dist=0.1360    corpus_dist=0.0593     token-doc-diff=0.0000
        exclusivity=0.4448
  versicherung word-length=12.0000    coherence=-0.1819
        uniform_dist=0.1354    corpus_dist=0.0571     token-doc-diff=0.0013
        exclusivity=0.2806
  vertr word-length=5.0000     coherence=-0.4043      uniform_dist=0.1348
        corpus_dist=0.0461     token-doc-diff=0.0046  exclusivity=0.3656
  wilmingto     word-length=9.0000     coherence=-0.1819
        uniform_dist=0.0781    corpus_dist=0.0346     token-doc-diff=0.0046
        exclusivity=0.2156
  energie       word-length=7.0000     coherence=-0.1819
        uniform_dist=0.0698    corpus_dist=0.0269     token-doc-diff=0.0026
        exclusivity=0.2229
```

```
Topic 7 tokens=35052.0000      document_entropy=3.2481        word-
length=6.2000  coherence=-0.8826      uniform_dist=4.1234
        corpus_dist=3.1147     eff_num_words=464.6382 token-doc-diff=0.0064
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.6988
  schaf word-length=5.0000     coherence=0.0000       uniform_dist=0.1133
        corpus_dist=0.0519     token-doc-diff=0.0036  exclusivity=0.6358
  schen word-length=5.0000     coherence=-0.0645      uniform_dist=0.0782
        corpus_dist=0.0414     token-doc-diff=0.0002  exclusivity=0.8113
  grund word-length=5.0000     coherence=-0.0317      uniform_dist=0.0624
        corpus_dist=0.0232     token-doc-diff=0.0008  exclusivity=0.6232
  unterneh      word-length=8.0000     coherence=-0.1335
        uniform_dist=0.0594    corpus_dist=0.0325     token-doc-diff=0.0003
        exclusivity=0.7677
  sichtsra      word-length=8.0000     coherence=-0.1483
        uniform_dist=0.0503    corpus_dist=0.0273     token-doc-diff=0.0015
        exclusivity=0.6561
```

```
Topic 8 tokens=13907.0000      document_entropy=0.0427        word-
length=9.0000  coherence=-2.7449      uniform_dist=4.0674
        corpus_dist=3.1880     eff_num_words=497.3440 token-doc-diff=0.0043
        rank_1_docs=0.5000     allocation_ratio=0.0000
        allocation_count=0.5000        exclusivity=0.6814
  immobilie     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0942    corpus_dist=0.0490     token-doc-diff=0.0017
        exclusivity=0.4433
```

```
  developm     word-length=8.0000     coherence=-0.6862
       uniform_dist=0.0773    corpus_dist=0.0418     token-doc-diff=0.0023
       exclusivity=0.6177
  victoriah    word-length=9.0000     coherence=-0.6862
       uniform_dist=0.0613    corpus_dist=0.0436     token-doc-diff=0.0002
       exclusivity=0.9962
  osterreich   word-length=10.0000    coherence=-0.6862
       uniform_dist=0.0521    corpus_dist=0.0288     token-doc-diff=0.0000
       exclusivity=0.4891
  propertie    word-length=9.0000     coherence=-0.6862
       uniform_dist=0.0511    corpus_dist=0.0353     token-doc-diff=0.0001
       exclusivity=0.8608
```

```
Topic 9 tokens=23087.0000      document_entropy=0.7227        word-
length=10.2000 coherence=-3.7017      uniform_dist=3.9462
       corpus_dist=2.6493     eff_num_words=594.3388 token-doc-diff=0.0111
       rank_1_docs=0.1250     allocation_ratio=0.0000
       allocation_count=0.1250        exclusivity=0.5008
  fahrzeug     word-length=8.0000     coherence=0.0000
       uniform_dist=0.1204    corpus_dist=0.0534     token-doc-diff=0.0037
       exclusivity=0.2866
  automobil    word-length=9.0000     coherence=-0.2224
       uniform_dist=0.0500    corpus_dist=0.0240     token-doc-diff=0.0014
       exclusivity=0.2406
  dieselthematik       word-length=14.0000    coherence=-0.9121
       uniform_dist=0.0443    corpus_dist=0.0293     token-doc-diff=0.0020
       exclusivity=0.9951
  nutzfahrzeug word-length=12.0000    coherence=-0.5090
       uniform_dist=0.0418    corpus_dist=0.0255     token-doc-diff=0.0036
       exclusivity=0.5912
  stuttgar     word-length=8.0000     coherence=-0.5090
       uniform_dist=0.0396    corpus_dist=0.0216     token-doc-diff=0.0005
       exclusivity=0.3905
```

```
Topic 10       tokens=8989.0000       document_entropy=0.0558        word-
length=6.2000  coherence=-1.3724      uniform_dist=4.2406
       corpus_dist=3.3702     eff_num_words=405.8715 token-doc-diff=0.0277
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.4226
  mitarbei     word-length=8.0000     coherence=0.0000
       uniform_dist=0.1346    corpus_dist=0.0517     token-doc-diff=0.0136
       exclusivity=0.1559
  consum       word-length=6.0000     coherence=0.0000
       uniform_dist=0.1282    corpus_dist=0.0785     token-doc-diff=0.0000
       exclusivity=0.6429
  variabl      word-length=7.0000     coherence=0.0000
       uniform_dist=0.0870    corpus_dist=0.0275     token-doc-diff=0.0025
       exclusivity=0.1415
  nivea word-length=5.0000     coherence=-0.6862      uniform_dist=0.0749
       corpus_dist=0.0567     token-doc-diff=0.0012  exclusivity=0.9968
  stark word-length=5.0000     coherence=0.0000       uniform_dist=0.0562
       corpus_dist=0.0139     token-doc-diff=0.0104  exclusivity=0.1761
```

```
Topic 11       tokens=100828.0000     document_entropy=3.4742        word-
length=7.6000  coherence=-0.0533      uniform_dist=4.4222
       corpus_dist=1.8741     eff_num_words=427.8232 token-doc-diff=0.0016
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.3259
```

```
  bewertung     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0676    corpus_dist=0.0178     token-doc-diff=0.0006
        exclusivity=0.2773
  stand word-length=5.0000     coherence=0.0000       uniform_dist=0.0625
        corpus_dist=0.0168     token-doc-diff=0.0002  exclusivity=0.4061
  gesellschaf   word-length=11.0000    coherence=-0.0267
        uniform_dist=0.0520    corpus_dist=0.0161     token-doc-diff=0.0001
        exclusivity=0.4172
  stell word-length=5.0000     coherence=0.0000       uniform_dist=0.0518
        corpus_dist=0.0119     token-doc-diff=0.0001  exclusivity=0.2558
  anderung      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0475    corpus_dist=0.0100     token-doc-diff=0.0005
        exclusivity=0.2728

---

Topic 12        tokens=14939.0000      document_entropy=0.2512        word-
length=11.2000 coherence=-0.4031      uniform_dist=4.2853
        corpus_dist=3.0248     eff_num_words=354.4606 token-doc-diff=0.0617
        rank_1_docs=0.2000     allocation_ratio=0.0000
        allocation_count=0.2000        exclusivity=0.4096
  informationenzusammengefass  word-length=27.0000    coherence=0.0000
        uniform_dist=0.1616    corpus_dist=0.0953     token-doc-diff=0.0322
        exclusivity=0.7866
  fahrzeug      word-length=8.0000     coherence=0.0000
        uniform_dist=0.1448    corpus_dist=0.0660     token-doc-diff=0.0004
        exclusivity=0.3292
  automobil     word-length=9.0000     coherence=-0.4031
        uniform_dist=0.1027    corpus_dist=0.0540     token-doc-diff=0.0011
        exclusivity=0.4394
  kompo word-length=5.0000     coherence=0.0000       uniform_dist=0.0763
        corpus_dist=0.0321     token-doc-diff=0.0045  exclusivity=0.3043
  emissio       word-length=7.0000     coherence=0.0000
        uniform_dist=0.0554    corpus_dist=0.0173     token-doc-diff=0.0236
        exclusivity=0.1884

---

Topic 13        tokens=9570.0000       document_entropy=0.3982        word-
length=10.6000 coherence=-8.6298      uniform_dist=4.7481
        corpus_dist=3.0014     eff_num_words=214.3019 token-doc-diff=0.0242
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.7674
  kontak        word-length=6.0000     coherence=0.0000
        uniform_dist=0.2794    corpus_dist=0.1632     token-doc-diff=0.0005
        exclusivity=0.9332
  finanzkal     word-length=9.0000     coherence=-1.4987
        uniform_dist=0.2665    corpus_dist=0.1586     token-doc-diff=0.0135
        exclusivity=0.8953
  beauty        word-length=6.0000     coherence=-1.4987
        uniform_dist=0.0705    corpus_dist=0.0498     token-doc-diff=0.0004
        exclusivity=0.7810
  organisch     word-length=9.0000     coherence=-1.0955
        uniform_dist=0.0681    corpus_dist=0.0325     token-doc-diff=0.0047
        exclusivity=0.2752
  gesellschafterausschuss       word-length=23.0000    coherence=-1.3759
        uniform_dist=0.0665    corpus_dist=0.0498     token-doc-diff=0.0050
        exclusivity=0.9521

---

Topic 14        tokens=11115.0000      document_entropy=0.5561        word-
length=11.0000 coherence=-5.0141      uniform_dist=4.3589
        corpus_dist=3.3129     eff_num_words=360.2547 token-doc-diff=0.0330
```

```
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.3611
  mitarbei      word-length=8.0000     coherence=0.0000
        uniform_dist=0.2250    corpus_dist=0.0946     token-doc-diff=0.0203
        exclusivity=0.2503
  kreislaufwirtschaf   word-length=18.0000    coherence=-0.6908
        uniform_dist=0.0703    corpus_dist=0.0437     token-doc-diff=0.0003
        exclusivity=0.6152
  nachhaltigkei word-length=13.0000    coherence=-0.5090
        uniform_dist=0.0600    corpus_dist=0.0194     token-doc-diff=0.0086
        exclusivity=0.1953
  material      word-length=8.0000     coherence=-0.5090
        uniform_dist=0.0553    corpus_dist=0.0274     token-doc-diff=0.0022
        exclusivity=0.2757
  chemisch      word-length=8.0000     coherence=-0.9121
        uniform_dist=0.0433    corpus_dist=0.0268     token-doc-diff=0.0016
        exclusivity=0.4689
```

```
Topic 15        tokens=12740.0000      document_entropy=0.3768        word-
length=10.2000  coherence=-11.2812     uniform_dist=4.4738
        corpus_dist=2.7693     eff_num_words=344.2878 token-doc-diff=0.0285
        rank_1_docs=0.0500     allocation_ratio=0.0000
        allocation_count=0.0500        exclusivity=0.3729
  fortgefuhr    word-length=10.0000    coherence=0.0000
        uniform_dist=0.1073    corpus_dist=0.0367     token-doc-diff=0.0058
        exclusivity=0.2516
  technologie   word-length=11.0000    coherence=-0.9779
        uniform_dist=0.0985    corpus_dist=0.0314     token-doc-diff=0.0000
        exclusivity=0.1460
  geschaftsfeld word-length=13.0000    coherence=-0.9779
        uniform_dist=0.0854    corpus_dist=0.0434     token-doc-diff=0.0002
        exclusivity=0.3523
  operativ      word-length=8.0000     coherence=-0.9779
        uniform_dist=0.0817    corpus_dist=0.0200     token-doc-diff=0.0007
        exclusivity=0.1176
  contitech     word-length=9.0000     coherence=-2.0673
        uniform_dist=0.0787    corpus_dist=0.0551     token-doc-diff=0.0218
        exclusivity=0.9970
```

```
Topic 16        tokens=12559.0000      document_entropy=0.0000        word-
length=7.6000   coherence=0.0000       uniform_dist=4.2175
        corpus_dist=3.0821     eff_num_words=411.8613 token-doc-diff=0.0380
        rank_1_docs=1.0000     allocation_ratio=0.0000
        allocation_count=1.0000        exclusivity=0.3786
  limited       word-length=7.0000     coherence=0.0000
        uniform_dist=0.2015    corpus_dist=0.0857     token-doc-diff=0.0233
        exclusivity=0.2990
  kompo word-length=5.0000     coherence=0.0000       uniform_dist=0.0787
        corpus_dist=0.0333     token-doc-diff=0.0000  exclusivity=0.3089
  holding       word-length=7.0000     coherence=0.0000
        uniform_dist=0.0497    corpus_dist=0.0176     token-doc-diff=0.0037
        exclusivity=0.1912
  tranch        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0445    corpus_dist=0.0156     token-doc-diff=0.0053
        exclusivity=0.0992
  langfristbonu word-length=13.0000    coherence=0.0000
        uniform_dist=0.0439    corpus_dist=0.0334     token-doc-diff=0.0055
        exclusivity=0.9950
```

```
Topic 17        tokens=17164.0000       document_entropy=0.1442         word-
length=8.2000   coherence=-5.2067       uniform_dist=4.2265
        corpus_dist=2.7336      eff_num_words=459.9390 token-doc-diff=0.0167
        rank_1_docs=0.1250      allocation_ratio=0.0000
        allocation_count=0.1250         exclusivity=0.4137
  scienc        word-length=6.0000      coherence=0.0000
        uniform_dist=0.0908     corpus_dist=0.0489      token-doc-diff=0.0009
        exclusivity=0.4539
  organisch     word-length=9.0000      coherence=-0.6862
        uniform_dist=0.0785     corpus_dist=0.0382      token-doc-diff=0.0001
        exclusivity=0.3259
  healthcar     word-length=9.0000      coherence=-0.6862
        uniform_dist=0.0709     corpus_dist=0.0383      token-doc-diff=0.0075
        exclusivity=0.2916
  electronic    word-length=10.0000     coherence=-0.6862
        uniform_dist=0.0646     corpus_dist=0.0413      token-doc-diff=0.0078
        exclusivity=0.8205
  solutio       word-length=7.0000      coherence=-1.0894
        uniform_dist=0.0642     corpus_dist=0.0281      token-doc-diff=0.0003
        exclusivity=0.1767
```

```
Topic 18        tokens=11513.0000       document_entropy=0.4314         word-
length=9.2000   coherence=-4.6630       uniform_dist=4.4939
        corpus_dist=3.0667      eff_num_words=345.7753 token-doc-diff=0.0102
        rank_1_docs=0.0000      allocation_ratio=0.0000
        allocation_count=0.0588         exclusivity=0.5161
  clearing      word-length=8.0000      coherence=0.0000
        uniform_dist=0.1254     corpus_dist=0.0846      token-doc-diff=0.0044
        exclusivity=0.9980
  nettoerlo     word-length=9.0000      coherence=-0.6897
        uniform_dist=0.1134     corpus_dist=0.0773      token-doc-diff=0.0003
        exclusivity=0.9954
  performanc    word-length=10.0000     coherence=-0.6897
        uniform_dist=0.1077     corpus_dist=0.0396      token-doc-diff=0.0001
        exclusivity=0.2186
  erlauterung   word-length=11.0000     coherence=-0.5090
        uniform_dist=0.0897     corpus_dist=0.0371      token-doc-diff=0.0049
        exclusivity=0.2631
  mitarbei      word-length=8.0000      coherence=-0.6897
        uniform_dist=0.0836     corpus_dist=0.0290      token-doc-diff=0.0005
        exclusivity=0.1052
```

```
Topic 19        tokens=8672.0000        document_entropy=1.2174         word-
length=7.2000   coherence=-5.2018       uniform_dist=4.6136
        corpus_dist=3.4249      eff_num_words=220.2438 token-doc-diff=0.0132
        rank_1_docs=0.0000      allocation_ratio=0.0000
        allocation_count=0.0000         exclusivity=0.5954
  varia word-length=5.0000      coherence=0.0000        uniform_dist=0.3060
        corpus_dist=0.1866      token-doc-diff=0.0057   exclusivity=0.9621
  healthcar     word-length=9.0000      coherence=-0.2228
        uniform_dist=0.1907     corpus_dist=0.1131      token-doc-diff=0.0000
        exclusivity=0.6572
  medical       word-length=7.0000      coherence=-0.6917
        uniform_dist=0.1137     corpus_dist=0.0480      token-doc-diff=0.0000
        exclusivity=0.2590
  covid word-length=5.0000      coherence=-0.3561       uniform_dist=0.0898
        corpus_dist=0.0305      token-doc-diff=0.0073   exclusivity=0.1101
  diagnostic    word-length=10.0000     coherence=-0.8090
        uniform_dist=0.0789     corpus_dist=0.0595      token-doc-diff=0.0002
        exclusivity=0.9884
```

```
Topic 20        tokens=27202.0000       document_entropy=0.7006         word-
length=9.2000   coherence=-4.7717       uniform_dist=4.3418
        corpus_dist=2.3252      eff_num_words=369.8590 token-doc-diff=0.0079
        rank_1_docs=0.0909      allocation_ratio=0.0000
        allocation_count=0.0909         exclusivity=0.5167
  medical       word-length=7.0000      coherence=0.0000
        uniform_dist=0.1614     corpus_dist=0.0720      token-doc-diff=0.0017
        exclusivity=0.3803
  patie word-length=5.0000      coherence=-0.1819       uniform_dist=0.1469
        corpus_dist=0.0692      token-doc-diff=0.0007   exclusivity=0.4809
  gesellschafteri       word-length=15.0000     coherence=-0.6908
        uniform_dist=0.0659     corpus_dist=0.0382      token-doc-diff=0.0031
        exclusivity=0.7385
  personlich    word-length=10.0000     coherence=-0.6908
        uniform_dist=0.0648     corpus_dist=0.0296      token-doc-diff=0.0002
        exclusivity=0.5357
  zuteilung     word-length=9.0000      coherence=-0.5090
        uniform_dist=0.0521     corpus_dist=0.0269      token-doc-diff=0.0022
        exclusivity=0.4481
```

```
Topic 21        tokens=6495.0000        document_entropy=0.0823         word-
length=8.6000   coherence=-2.0587       uniform_dist=4.4201
        corpus_dist=3.6924      eff_num_words=422.2514 token-doc-diff=0.0127
        rank_1_docs=0.0000      allocation_ratio=0.0000
        allocation_count=0.0000         exclusivity=0.6372
  stedim        word-length=6.0000      coherence=0.0000
        uniform_dist=0.0914     corpus_dist=0.0718      token-doc-diff=0.0036
        exclusivity=0.9972
  solutio       word-length=7.0000      coherence=0.0000
        uniform_dist=0.0734     corpus_dist=0.0329      token-doc-diff=0.0083
        exclusivity=0.1814
  bioprocess    word-length=10.0000     coherence=-0.6862
        uniform_dist=0.0686     corpus_dist=0.0559      token-doc-diff=0.0004
        exclusivity=0.9964
  underlying    word-length=10.0000     coherence=-0.6862
        uniform_dist=0.0675     corpus_dist=0.0530      token-doc-diff=0.0003
        exclusivity=0.8878
  akquisitio    word-length=10.0000     coherence=-0.6862
        uniform_dist=0.0640     corpus_dist=0.0243      token-doc-diff=0.0001
        exclusivity=0.1233
```

```
Topic 22        tokens=15962.0000       document_entropy=0.7308         word-
length=7.8000   coherence=-4.1121       uniform_dist=4.3403
        corpus_dist=2.7239      eff_num_words=371.2796 token-doc-diff=0.0140
        rank_1_docs=0.0000      allocation_ratio=0.0000
        allocation_count=0.0000         exclusivity=0.4938
  truck word-length=5.0000      coherence=0.0000        uniform_dist=0.1420
        corpus_dist=0.0881      token-doc-diff=0.0082   exclusivity=0.9982
  mobility      word-length=8.0000      coherence=-0.1333
        uniform_dist=0.1002     corpus_dist=0.0555      token-doc-diff=0.0000
        exclusivity=0.5315
  fahrzeug      word-length=8.0000      coherence=-0.4690
        uniform_dist=0.0819     corpus_dist=0.0341      token-doc-diff=0.0006
        exclusivity=0.2007
  aktivita      word-length=8.0000      coherence=-0.2228
        uniform_dist=0.0795     corpus_dist=0.0258      token-doc-diff=0.0052
        exclusivity=0.2013
```

```
  abspaltung    word-length=10.0000    coherence=-0.9779
        uniform_dist=0.0756    corpus_dist=0.0440     token-doc-diff=0.0000
        exclusivity=0.5374
```

```
Topic 23        tokens=17219.0000      document_entropy=0.7796        word-
length=9.8000  coherence=-2.6420      uniform_dist=4.2896
        corpus_dist=2.8522     eff_num_words=409.1722 token-doc-diff=0.0674
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.3196
  finanzjah     word-length=9.0000     coherence=0.0000
        uniform_dist=0.1534    corpus_dist=0.0927     token-doc-diff=0.0515
        exclusivity=0.9983
  global        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0706    corpus_dist=0.0172     token-doc-diff=0.0060
        exclusivity=0.1289
  geschaftsbereich     word-length=16.0000    coherence=-0.2865
        uniform_dist=0.0693    corpus_dist=0.0305     token-doc-diff=0.0064
        exclusivity=0.2537
  fortgefuhr    word-length=10.0000    coherence=-0.2865
        uniform_dist=0.0548    corpus_dist=0.0155     token-doc-diff=0.0034
        exclusivity=0.1429
  mitarbei      word-length=8.0000     coherence=-0.6897
        uniform_dist=0.0543    corpus_dist=0.0168     token-doc-diff=0.0000
        exclusivity=0.0744
```

```
Topic 24        tokens=13171.0000      document_entropy=0.4193        word-
length=14.0000 coherence=-3.4483      uniform_dist=4.4399
        corpus_dist=3.0701     eff_num_words=316.0261 token-doc-diff=0.0038
        rank_1_docs=0.1429     allocation_ratio=0.0000
        allocation_count=0.1429        exclusivity=0.6492
  schad word-length=5.0000     coherence=0.0000       uniform_dist=0.1696
        corpus_dist=0.0801     token-doc-diff=0.0012  exclusivity=0.3948
  konzerngeschaftsberich       word-length=22.0000    coherence=-0.6897
        uniform_dist=0.1204    corpus_dist=0.0795     token-doc-diff=0.0008
        exclusivity=0.9979
  ruckversicherung     word-length=16.0000    coherence=-0.4031
        uniform_dist=0.1100    corpus_dist=0.0573     token-doc-diff=0.0009
        exclusivity=0.3788
  versicherungstechnisch       word-length=22.0000    coherence=-0.4031
        uniform_dist=0.0840    corpus_dist=0.0489     token-doc-diff=0.0004
        exclusivity=0.4891
  munch word-length=5.0000     coherence=-0.4031      uniform_dist=0.0822
        corpus_dist=0.0565     token-doc-diff=0.0005  exclusivity=0.9856
```

```
Topic 25        tokens=13708.0000      document_entropy=0.2394        word-
length=7.0000  coherence=-0.6897      uniform_dist=4.2621
        corpus_dist=3.0562     eff_num_words=454.8275 token-doc-diff=0.0109
        rank_1_docs=0.2000     allocation_ratio=0.0000
        allocation_count=0.2000        exclusivity=0.2426
  operativ      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0877    corpus_dist=0.0220     token-doc-diff=0.0013
        exclusivity=0.1257
  schad word-length=5.0000     coherence=-0.4031      uniform_dist=0.0752
        corpus_dist=0.0313     token-doc-diff=0.0048  exclusivity=0.1940
  insuranc      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0746    corpus_dist=0.0425     token-doc-diff=0.0012
        exclusivity=0.4935
```

```
  wilmingto     word-length=9.0000     coherence=-0.2865
        uniform_dist=0.0679    corpus_dist=0.0294     token-doc-diff=0.0000
        exclusivity=0.1906
  netto word-length=5.0000     coherence=0.0000       uniform_dist=0.0618
        corpus_dist=0.0230     token-doc-diff=0.0036  exclusivity=0.2089
```

```
Topic 26       tokens=17415.0000      document_entropy=0.1622        word-
length=6.4000  coherence=-2.4721      uniform_dist=4.2093
        corpus_dist=2.7977     eff_num_words=463.8329 token-doc-diff=0.0151
        rank_1_docs=0.2000     allocation_ratio=0.0000
        allocation_count=0.2000        exclusivity=0.5226
  energie       word-length=7.0000     coherence=0.0000
        uniform_dist=0.0971    corpus_dist=0.0400     token-doc-diff=0.0077
        exclusivity=0.3017
  taxonomie     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0588    corpus_dist=0.0268     token-doc-diff=0.0011
        exclusivity=0.4808
  aspek word-length=5.0000     coherence=0.0000       uniform_dist=0.0549
        corpus_dist=0.0249     token-doc-diff=0.0018  exclusivity=0.6012
  strom word-length=5.0000     coherence=0.0000       uniform_dist=0.0511
        corpus_dist=0.0239     token-doc-diff=0.0027  exclusivity=0.2866
  innogy        word-length=6.0000     coherence=-0.6897
        uniform_dist=0.0485    corpus_dist=0.0335     token-doc-diff=0.0018
        exclusivity=0.9430
```

```
Topic 27       tokens=237635.0000     document_entropy=3.5127        word-
length=7.8000  coherence=0.0000       uniform_dist=4.3837
        corpus_dist=1.3161     eff_num_words=482.6495 token-doc-diff=0.0023
        rank_1_docs=0.0263     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.3706
  complianc     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0648    corpus_dist=0.0165     token-doc-diff=0.0014
        exclusivity=0.3433
  global        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0525    corpus_dist=0.0112     token-doc-diff=0.0000
        exclusivity=0.1054
  entwickel     word-length=9.0000     coherence=0.0000
        uniform_dist=0.0475    corpus_dist=0.0135     token-doc-diff=0.0001
        exclusivity=0.7076
  standard      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0438    corpus_dist=0.0122     token-doc-diff=0.0003
        exclusivity=0.5000
  weltwei       word-length=7.0000     coherence=0.0000
        uniform_dist=0.0421    corpus_dist=0.0096     token-doc-diff=0.0005
        exclusivity=0.1968
```

```
Topic 28       tokens=13500.0000      document_entropy=0.1774        word-
length=13.6000 coherence=-5.9067      uniform_dist=4.4322
        corpus_dist=2.9205     eff_num_words=270.1769 token-doc-diff=0.0253
        rank_1_docs=0.1111     allocation_ratio=0.0000
        allocation_count=0.1111        exclusivity=0.5915
  technologie   word-length=11.0000    coherence=0.0000
        uniform_dist=0.3057    corpus_dist=0.1236     token-doc-diff=0.0050
        exclusivity=0.3982
  strategie     word-length=9.0000     coherence=-0.2865
        uniform_dist=0.1341    corpus_dist=0.0470     token-doc-diff=0.0004
        exclusivity=0.4331
```

```
  cypress      word-length=7.0000     coherence=-1.3759
       uniform_dist=0.0679    corpus_dist=0.0479     token-doc-diff=0.0018
       exclusivity=0.9965
  lageberichtgeschaftsausrichtung     word-length=31.0000    coherence=-
1.3759  uniform_dist=0.0551    corpus_dist=0.0399     token-doc-diff=0.0005
       exclusivity=0.9959
  fortgefuhr   word-length=10.0000    coherence=-0.4031
       uniform_dist=0.0518    corpus_dist=0.0144     token-doc-diff=0.0176
       exclusivity=0.1340
```

```
Topic 29       tokens=20923.0000      document_entropy=1.0288        word-
length=7.6000  coherence=-5.1798      uniform_dist=4.2118
       corpus_dist=2.5453     eff_num_words=499.1761 token-doc-diff=0.0012
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.3245
  limited      word-length=7.0000     coherence=0.0000
       uniform_dist=0.0772    corpus_dist=0.0273     token-doc-diff=0.0000
       exclusivity=0.1329
  geschaftsjahr word-length=13.0000    coherence=-0.2509
       uniform_dist=0.0582    corpus_dist=0.0191     token-doc-diff=0.0005
       exclusivity=0.2450
  renewabl     word-length=8.0000     coherence=-1.0955
       uniform_dist=0.0568    corpus_dist=0.0317     token-doc-diff=0.0003
       exclusivity=0.4561
  award word-length=5.0000     coherence=-0.5865      uniform_dist=0.0560
       corpus_dist=0.0278     token-doc-diff=0.0003  exclusivity=0.3977
  stock word-length=5.0000     coherence=-0.5865      uniform_dist=0.0557
       corpus_dist=0.0288     token-doc-diff=0.0001  exclusivity=0.3907
```

```
Topic 30       tokens=9164.0000       document_entropy=0.0893        word-
length=7.0000  coherence=-2.8649      uniform_dist=4.3549
       corpus_dist=3.5003     eff_num_words=375.2140 token-doc-diff=0.0530
       rank_1_docs=0.0000     allocation_ratio=0.0000
       allocation_count=0.0000        exclusivity=0.5056
  gefass       word-length=6.0000     coherence=0.0000
       uniform_dist=0.1736    corpus_dist=0.1025     token-doc-diff=0.0244
       exclusivity=0.9236
  nanziell     word-length=8.0000     coherence=0.0000
       uniform_dist=0.0998    corpus_dist=0.0684     token-doc-diff=0.0048
       exclusivity=0.8992
  tranch       word-length=6.0000     coherence=0.0000
       uniform_dist=0.0911    corpus_dist=0.0374     token-doc-diff=0.0157
       exclusivity=0.1796
  transaktio   word-length=10.0000    coherence=-1.0894
       uniform_dist=0.0640    corpus_dist=0.0224     token-doc-diff=0.0081
       exclusivity=0.2425
  lokal word-length=5.0000     coherence=-1.0894      uniform_dist=0.0624
       corpus_dist=0.0234     token-doc-diff=0.0001  exclusivity=0.2830
```

```
Topic 31       tokens=22285.0000      document_entropy=0.0536        word-
length=7.0000  coherence=-4.1311      uniform_dist=4.3293
       corpus_dist=2.5056     eff_num_words=300.9404 token-doc-diff=0.0049
       rank_1_docs=0.2000     allocation_ratio=0.0000
       allocation_count=0.2000        exclusivity=0.5187
  mobil word-length=5.0000     coherence=0.0000       uniform_dist=0.2271
       corpus_dist=0.1176     token-doc-diff=0.0006  exclusivity=0.8689
  deutsch      word-length=7.0000     coherence=-0.6897
       uniform_dist=0.1097    corpus_dist=0.0359     token-doc-diff=0.0006
       exclusivity=0.4900
```

```
  mobilfunk     word-length=9.0000     coherence=-0.6897
        uniform_dist=0.1014    corpus_dist=0.0603     token-doc-diff=0.0002
        exclusivity=0.9564
  operativ      word-length=8.0000     coherence=-0.2865
        uniform_dist=0.0935    corpus_dist=0.0241     token-doc-diff=0.0033
        exclusivity=0.1368
  servic        word-length=6.0000     coherence=-0.6897
        uniform_dist=0.0819    corpus_dist=0.0256     token-doc-diff=0.0001
        exclusivity=0.1417
```

```
Topic 32        tokens=12991.0000      document_entropy=0.6634        word-
length=7.2000  coherence=-3.8342      uniform_dist=4.1726
        corpus_dist=3.2448     eff_num_words=437.3889 token-doc-diff=0.0268
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.5745
  vertr word-length=5.0000     coherence=0.0000       uniform_dist=0.1445
        corpus_dist=0.0502     token-doc-diff=0.0033  exclusivity=0.3867
  versicherung  word-length=12.0000    coherence=0.0000
        uniform_dist=0.1249    corpus_dist=0.0519     token-doc-diff=0.0011
        exclusivity=0.2600
  zivil word-length=5.0000     coherence=-0.6862      uniform_dist=0.0692
        corpus_dist=0.0478     token-doc-diff=0.0022  exclusivity=0.9449
  optio word-length=5.0000     coherence=0.0000       uniform_dist=0.0599
        corpus_dist=0.0207     token-doc-diff=0.0200  exclusivity=0.2851
  triebwerk     word-length=9.0000     coherence=-1.0894
        uniform_dist=0.0518    corpus_dist=0.0382     token-doc-diff=0.0002
        exclusivity=0.9956
```

```
Topic 33        tokens=57185.0000      document_entropy=3.5201        word-
length=8.0000  coherence=0.0000       uniform_dist=4.6290
        corpus_dist=2.2861     eff_num_words=290.7183 token-doc-diff=0.0176
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.4395
  lieferung     word-length=9.0000     coherence=0.0000
        uniform_dist=0.1918    corpus_dist=0.0779     token-doc-diff=0.0115
        exclusivity=0.7408
  erwerb        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0982    corpus_dist=0.0355     token-doc-diff=0.0000
        exclusivity=0.4373
  leistung      word-length=8.0000     coherence=0.0000
        uniform_dist=0.0830    corpus_dist=0.0203     token-doc-diff=0.0007
        exclusivity=0.5013
  schuld        word-length=6.0000     coherence=0.0000
        uniform_dist=0.0708    corpus_dist=0.0248     token-doc-diff=0.0021
        exclusivity=0.2464
  betrieblich   word-length=11.0000    coherence=0.0000
        uniform_dist=0.0633    corpus_dist=0.0222     token-doc-diff=0.0034
        exclusivity=0.2719
```

```
Topic 34        tokens=7600.0000       document_entropy=1.1812        word-
length=12.2000 coherence=-3.2870      uniform_dist=4.4242
        corpus_dist=3.6641     eff_num_words=430.1780 token-doc-diff=0.0049
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.3386
  covid word-length=5.0000     coherence=0.0000       uniform_dist=0.0897
        corpus_dist=0.0305     token-doc-diff=0.0000  exclusivity=0.1084
  indirek       word-length=7.0000     coherence=-0.3357
        uniform_dist=0.0774    corpus_dist=0.0403     token-doc-diff=0.0000
        exclusivity=0.6722
```

```
  zahlungsmittelgenerier       word-length=22.0000    coherence=-0.4043
        uniform_dist=0.0662    corpus_dist=0.0349     token-doc-diff=0.0002
        exclusivity=0.3812
  anteilsbasier word-length=13.0000    coherence=-0.5581
        uniform_dist=0.0632    corpus_dist=0.0358     token-doc-diff=0.0014
        exclusivity=0.3586
  zahlungsmittel       word-length=14.0000    coherence=-0.3357
        uniform_dist=0.0475    corpus_dist=0.0192     token-doc-diff=0.0033
        exclusivity=0.1727
```

```
Topic 35       tokens=406080.0000     document_entropy=3.6041        word-
length=7.4000  coherence=0.0000       uniform_dist=4.3319
        corpus_dist=0.9026     eff_num_words=560.0486 token-doc-diff=0.0009
        rank_1_docs=0.4737     allocation_ratio=0.0000
        allocation_count=0.3421        exclusivity=0.3280
  erwar word-length=5.0000     coherence=0.0000       uniform_dist=0.0519
        corpus_dist=0.0096     token-doc-diff=0.0003  exclusivity=0.3235
  prufung      word-length=7.0000     coherence=0.0000
        uniform_dist=0.0491    corpus_dist=0.0095     token-doc-diff=0.0001
        exclusivity=0.4767
  enthal       word-length=6.0000     coherence=0.0000
        uniform_dist=0.0454    corpus_dist=0.0078     token-doc-diff=0.0000
        exclusivity=0.3629
  veranderung  word-length=11.0000    coherence=0.0000
        uniform_dist=0.0454    corpus_dist=0.0079     token-doc-diff=0.0000
        exclusivity=0.2137
  leistung     word-length=8.0000     coherence=0.0000
        uniform_dist=0.0388    corpus_dist=0.0063     token-doc-diff=0.0005
        exclusivity=0.2632
```

```
Topic 36       tokens=30596.0000      document_entropy=1.9573        word-
length=6.2000  coherence=-2.0334      uniform_dist=4.5302
        corpus_dist=2.4123     eff_num_words=315.0899 token-doc-diff=0.0057
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.4265
  risiko       word-length=6.0000     coherence=0.0000
        uniform_dist=0.1661    corpus_dist=0.0530     token-doc-diff=0.0037
        exclusivity=0.3187
  verlust      word-length=7.0000     coherence=-0.2875
        uniform_dist=0.0827    corpus_dist=0.0243     token-doc-diff=0.0000
        exclusivity=0.5026
  aktiva       word-length=6.0000     coherence=-0.2075
        uniform_dist=0.0755    corpus_dist=0.0367     token-doc-diff=0.0006
        exclusivity=0.4993
  priva word-length=5.0000     coherence=-0.2335      uniform_dist=0.0753
        corpus_dist=0.0330     token-doc-diff=0.0009  exclusivity=0.5020
  investm      word-length=7.0000     coherence=-0.2999
        uniform_dist=0.0695    corpus_dist=0.0286     token-doc-diff=0.0005
        exclusivity=0.3099
```

```
Topic 37       tokens=21395.0000      document_entropy=0.2410        word-
length=8.4000  coherence=-5.0650      uniform_dist=4.5711
        corpus_dist=2.6566     eff_num_words=253.3388 token-doc-diff=0.0114
        rank_1_docs=0.1000     allocation_ratio=0.0000
        allocation_count=0.1000        exclusivity=0.4487
  medical      word-length=7.0000     coherence=0.0000
        uniform_dist=0.1518    corpus_dist=0.0671     token-doc-diff=0.0002
        exclusivity=0.3568
```

```
  nichtfinanziell      word-length=15.0000    coherence=0.0000
        uniform_dist=0.1518    corpus_dist=0.0549     token-doc-diff=0.0013
        exclusivity=0.2720
  angab word-length=5.0000     coherence=-0.1333      uniform_dist=0.1398
        corpus_dist=0.0404     token-doc-diff=0.0006  exclusivity=0.3842
  vamed word-length=5.0000     coherence=-1.3811      uniform_dist=0.0987
        corpus_dist=0.0599     token-doc-diff=0.0088  exclusivity=0.9975
  beschaftig    word-length=10.0000    coherence=-0.4690
        uniform_dist=0.0932    corpus_dist=0.0354     token-doc-diff=0.0005
        exclusivity=0.2330
```

```
Topic 38       tokens=5550.0000       document_entropy=0.4145        word-
length=7.8000  coherence=-8.0408      uniform_dist=4.4745
        corpus_dist=3.7065     eff_num_words=374.5167 token-doc-diff=0.0487
        rank_1_docs=0.0000     allocation_ratio=0.0000
        allocation_count=0.0000        exclusivity=0.4124
  beschaftig    word-length=10.0000    coherence=0.0000
        uniform_dist=0.1418    corpus_dist=0.0584     token-doc-diff=0.0003
        exclusivity=0.2970
  chanc word-length=5.0000     coherence=-0.5090      uniform_dist=0.1065
        corpus_dist=0.0343     token-doc-diff=0.0032  exclusivity=0.2923
  textziff      word-length=8.0000     coherence=-1.5983
        uniform_dist=0.0724    corpus_dist=0.0422     token-doc-diff=0.0152
        exclusivity=0.3514
  erklarung     word-length=9.0000     coherence=-0.2224
        uniform_dist=0.0601    corpus_dist=0.0180     token-doc-diff=0.0291
        exclusivity=0.1341
  express       word-length=7.0000     coherence=-1.9340
        uniform_dist=0.0587    corpus_dist=0.0501     token-doc-diff=0.0009
        exclusivity=0.9870
```

```
Topic 39       tokens=18334.0000      document_entropy=0.2981        word-
length=8.2000  coherence=-2.4151      uniform_dist=4.3922
        corpus_dist=2.5808     eff_num_words=307.0062 token-doc-diff=0.0130
        rank_1_docs=0.0833     allocation_ratio=0.0000
        allocation_count=0.0833        exclusivity=0.3889
  integrier     word-length=9.0000     coherence=0.0000
        uniform_dist=0.1790    corpus_dist=0.0815     token-doc-diff=0.0022
        exclusivity=0.6059
  mitarbei      word-length=8.0000     coherence=-0.1176
        uniform_dist=0.1570    corpus_dist=0.0621     token-doc-diff=0.0020
        exclusivity=0.1885
  nachhaltigkei word-length=13.0000    coherence=-0.1333
        uniform_dist=0.1451    corpus_dist=0.0573     token-doc-diff=0.0002
        exclusivity=0.4375
  cloud word-length=5.0000     coherence=-0.4047      uniform_dist=0.0740
        corpus_dist=0.0418     token-doc-diff=0.0030  exclusivity=0.6046
  servic        word-length=6.0000     coherence=-0.4690
        uniform_dist=0.0607    corpus_dist=0.0174     token-doc-diff=0.0057
        exclusivity=0.1082
```