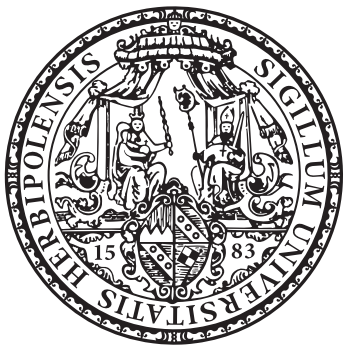
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Anwendung von Verfahren zur Visualisierung von Textcorpora mit Hilfe von Topic Models

Seminararbeit

Eingereicht von:	Wolf, Peter
Studiengang:	Master Wirtschaftsinformatik
Matrikelnummer:	2128372
Betreuer:	Prof. Dr. Frédéric Thiesse
Bearbeitungszeitraum:	24.02.2016 – 21.04.2016



Julius-Maximilians-Universität Würzburg

Lehrstuhl für Wirtschaftsinformatik und Systementwicklung
Josef-Stangl-Platz 2, 97070 Würzburg

Zusammenfassung

Die Zusammenfassung dient dem Leser einen groben Überblick über die Inhalte zu gewinnen (kurze Problemstellung, Herangehensweise, Lösungsansätze und evtl. der Schlüsselerkenntnisse). Der Umfang sollte ca. eine halbe Seite betragen.

Auf der nächsten Seite soll eine Übersetzung der Zusammenfassung als Abstract in englischer Sprache erfolgen.

Abstract

Zusammenfassung der Ausarbeitung in englischer Sprache.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CDMA Code Division Multiple Access

Bei der Erstellung des Abbildungsverzeichnisses muss darauf geachtet werden, dass die Abkürzungen noch in alphabetische Reihenfolge gebracht werden müssen. Weiterhin ist zu beachten, dass nur solche Abkürzungen im Verzeichnis aufgeführt werden, die auch im Text verwendet wurden.

1 Einleitung

Durch das Internet sind immer mehr Informationen vor allem in Textform verfügbar. Um diese Informationen zu verarbeiten, verwendet man Verfahren des Topic Modeling. Das Ergebnis dieser Verfahren sind Topic Models. Bei ihnen handelt es sich um eine statistische Auswertung von Textkörpern. In der Literatur werden unterschiedliche Verfahren dazu beschrieben. Topic Models lassen sich in der Form, wie sie von den verschiedenen Verfahren nur schwer interpretieren und besitzen unter Umständen Topics ohne Aussagegehalt. Ziel dieser Arbeit ist es, Methoden darzustellen, die Topic Models visualisieren können und mit deren Hilfe man bestimmte Probleme mit Topic Models beseitigen kann. Dazu werden im ersten Teil die gängigen Verfahren dargestellt, mit deren Hilfe die Topic Models erstellt werden. Im zweiten Teil wird anschließend auf die unterschiedlichen Verfahren zur Visualisierung von Topic Models eingegangen. Das Augenmerk soll dabei auf den unterschiedlichen Perspektiven liegen, die mit dem jeweiligen Verfahren dargestellt wird. Der vierte Teil umfasst einen Vergleich der Verfahren und eine Diskussion ihrer Vor- und Nachteile. Abschließend erfolgt noch eine beispielhafte Anwendung der Verfahren, die sich als die Verfahren mit den meisten Vorzügen herausgestellt haben. Dazu wird ein Test Corpus erstellt und auf diesen werden die jeweiligen Verfahren angewendet.

Dieser Teil der Arbeit sollte folgenden Inhalte besitzen:

- Einführung in die Problemstellung
- Motivation und Herleitung des Themas
- Aufbau der Arbeit

Hinweis: Es hat sich als hilfreich erwiesen, die Einleitung mit der Zusammenfassung bzw. dem Abstract und der Schlussfolgerung zu vergleichen. Damit stellt man sicher, dass diese inhaltlich im Bezug auf Zielsetzung und Motivation übereinstimmen. Der Umfang sollte ca. 5% der gesamten Arbeit betragen.

2 Hauptteil

Hier soll jeweils eine kurze Einführung erfolgen, die den Zusammenhang des Kapitels zur Arbeit herstellt.

Generelle Hinweise: Verwenden Sie stets eindeutige Begrifflichkeiten achten Sie auf eine logische Herleitung ihrer Argumentationen.

2.1 Grundlagen

Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl.

2.1.1 Topic Models

2.1.2 Darstellungsverfahren

2.2 Zweiter Unterpunkt

Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl.

2.2.1 Dritte Ebene

Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl. congue. Curabitur et sapien.

2.2.1.1 Vierte Ebene

Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl.

Fünfte Ebene Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl.

2.3 Beispiele

Im Folgenden sind verschiedene Beispiele bezüglich der Formatierung aufgeführt. Diese Vorlage enthält zahlreiche weitere Möglichkeiten der Formatierung, welche exemplarisch in der Datei *Demo.pdf* vorgestellt werden.

Nach Möglichkeiten sollten jedoch nur die nachfolgenden Möglichekiten der Formatierung genutzt werden.

2.3.1 Aufzählungen

- Erster Punkt
  - Erster Unterpunkt
  - Zweiter Unterpunkt
  - Dritter Unterpunkt
- Zweiter Punkt
- Dritter Punkt

1. Erster Punkt
   a) Erster Unterpunkt
   b) Zweiter Unterpunkt
   c) Dritter Unterpunkt
2. Zweiter Punkt
3. Dritter Punkt

2.3.2 Abbildungen

Wichtig ist, Abbildungen immer im Text zu erläutern. Dies gilt auch für Tabellen. Bei fremden Abbildungen und Tabellen ist zudem die ursprüngliche Quelle anzugeben. Dies kann beispielsweise direkt innerhalb der Bildunterschrift erfolgen.

Abbildung 2.1: Ausführliche Bezeichnung der Abbildung direkt unterhalb der Abbildung.

(a) Erste Abbildung **(b)** Zweite Abbildung

Abbildung 2.2: Ausführliche Bezeichnung der Abbildung direkt unterhalb der Abbildungen.

Tabelle 2.1: Beispiel für eine Tabelle.

Tabellenkopf	Tabellenkopf	Tabellenkopf	Tabellenkopf	Tabellenkopf
Beschreibung	Inhalt	Inhalt	Inhalt	Inhalt
Beschreibung	Inhalt	Inhalt	Inhalt	Inhalt

2.3.3 Tabellen

2.3.4 URLs

Sofern Links zu Webseiten angegeben werden, sollten diese als solche kenntlich gemacht werden, was auch den Vorteil mit sich bringt, dass diese direkt im PDF anklickbar sind. Am elegantesten geschieht dies mithilfe einer Fußnote zur jeweiligen Webseite[1].

2.3.5 Abkürzungen

Abkürzungen müssen zunächst in das Abkürzungsverzeichnis eingepflegt werden. Werden diese anschließend verwendet, wird bei der ersten Verwendung die Langform als Fußnote ausgegeben. Anschließend jeweils nur noch die Kurzform.

CDMA[2] ist ein Codemultiplexverfahren, das die gleichzeitige Übertragung verschiedener Datenströme auf einem gemeinsamen Frequenzbereich ermöglicht. Hirbei unterscheidet man zwischen synchronome und synchronem CDMA.

1 `http://www.bwl.uni-wuerzburg.de/lehrstuehle/wise/`
2 Code Division Multiple Access

2.3.6 Zitate

Ein zentrales Kriterium wissenschaftlichen Arbeitens ist die Unterscheidung zwischen eigenen und aus Literaturquellen entnommenen Beiträgen.

Zur Kennzeichnung einer Literaturquelle ist die „Harvard Citation“ oder „amerikanische Zitierweise“ weit verbreitet und sollte daher verwendet werden. Bei der amerikanischen Zitierweise werden Quellen durch die Nennung des Nachnamens des Autors/der Autorin, das Erscheinungsjahr des Textes sowie die jeweilige(n) Seitenzahl(en), auf die man sich bezieht, direkt im Fließtext angegeben. Zitiert man mehrere Autoren, so werden diese mit einem Semikolon voneinander getrennt.

Die vollständigen bibliographischen Informationen werden dann im Literaturverzeichnis übersichtlich dargestellt.

- [**?** , 24]
  Das Zitat bzw. der Verweis bezieht sich auf eine Textstelle auf der Seite 24.
- [**?** , 24 f.]
  Das Zitat bzw. der Verweis bezieht sich auf eine Textstelle, die sich von Seite 24 auf Seite 25 erstreckt.
- [**?** , 24 ff.]
  Das Zitat bzw. der Verweis bezieht sich auf eine Textstelle, die sich von Seite 24 bis auf Seite 26 erstreckt.
- [**?** , 24-29]
  Dieser Verweis bezieht sich auf die Seiten 24 bis 29. Diese Form der Seitenangabe verwendet man, wenn man sich auf eine Textstelle bezieht, die sich über mehr als drei Seiten erstreckt.
- [**?** , 24 und 27]
  Der Verweis bezieht sich auf Textstellen auf den Seiten 24 und 27.

2.3.6.1 Direkte, wörtliche Zitate

Bei einem direkten Zitat muss der zitierte Text originalgetreu wiedergegeben werden, d.h. Rechtschreibfehler oder eine veraltete Orthographie werden unverändert wiedergegeben. Der zitierte Text steht immer in Anführungszeichen. Wird innerhalb eines Zitates ebenfalls zitiert (Zitat im Zitat), so steht das innen stehende Zitat in einfachen Anführungszeichen.

„Wirtschaftsinformatiker können vielseitig in allen Unternehmensbereichen und Branchen eingesetzt werden, in denen ein hoher IT-Bezug gegeben ist.“ [**?** , 76]

Vor allem bei längeren Zitaten empfiehlt sich folgende Zitierweise:

> „Wirtschaftsinformatiker können vielseitig in allen Unternehmensbereichen und Branchen eingesetzt werden, in denen ein hoher IT-Bezug gegeben ist. Allgemein können zwei Einsatzgebiete unterschieden werden. Zum einen ist

> der Einsatz innerhalb der IT-Abteilung eines Unternehmens möglich. [...] Zum anderen ist ein Einsatz außerhalb der IT-Abteilung an verschiedenen Schnittstellen möglich.“ [**?** , 76]

Auslassungen bzw. Ellipsen in einem direkten Zitat sind erlaubt, wenn der Sinn der ursprünglichen Belegstelle nicht verstellt wird. Auslassungen werden durch eine eckige Klammer mit drei Punkten [...] gekennzeichnet.

2.3.6.2 Indirekte, sinngemäße Zitate

Das indirekte Zitat, d.h. die Wiedergabe eines fremden Gedankens mit eigenen Worten, gibt die Meinung eines Autors sinngemäß wieder.

Nach Ansicht von [**?**] stellen die neuesten Erkentnisse der Experten am Forschungszentrum CERN ein große Entdeckung dar.

Aufgrund der vorgestellten Ergebnisse ist mit großer Wahrscheinlichkeit davon auszugehen, dass das lange gesuchte Higgs-Teilchen endlich gefunden worden ist [**?**].

2.3.7 Literaturverzeichnis

Zu jeder wissenschaftlichen Arbeit gehört ein Literaturverzeichnis, in dem alle zitierten Quellen in alphabetischer Reihenfolge enthalten sind. Die Literaturangaben werden alphabetisch nach Zuname des Autors, dann chronologisch geordnet. Dabei wird bei der Art der Literaturquelle unterschieden.

Die Aufgabe der Formatierung wird weitestgehend durch BibTeX weitestgehend automatisiert. Wichtig ist jedoch, alle verwedeten Quellen in die Literatur-Datenbank einzupflegen. Für die Erstellung und Pflege dieser Datenbank haben sich vor allem Citavi [3] und JabRef [4] als hilfreich verwiesen.

3 `http://www.citavi.com`
4 `http://jabref.sourceforge.net/`

3 Schlussfolgerung

In der Schlussfolgerung sollten folgende Punkte deutlich werden:

- die Themenstellung
- der gewählte Ansatz
- die Eregbnisse der Arbeit
- eine kritische Stellungnahme/Einschätzung
- nächste Schritte

Hinweis: Die Schlussfolgerung sollte mit der Zusammenfassung bzw. dem Abstract und der Einleitung abgeglichen werden. Es sollte immer eine Zusammenfassung der wesentlichen Erkenntnisse der eigenen Arbeit sein, die den Forschungsbeitrag darstellt. Der Umfang der Schlussfolge-rung sollte ähnlich wie die Einleitung ca. 5% der gesamten Arbeit betragen.

Literaturverzeichnis

Anhang

Ein Anhang zur wissenschaftlichen Arbeit ist notwendig, wenn Materialien, die die Arbeit als Ganzes oder auch größere Teile derselben betreffen, jedoch nur schwer im Ausführungsteil unterzubringen sind. Das ist insbesondere dann der Fall, wenn sie aufgrund ihres Umfangs den Gesamtzusammenhang der Ausführung stören würden. Inhaltlich darf im Anhang nichts stehen, was zum Verständnis des Textes notwendig ist, der Text der Arbeit darf an dieser Stelle nicht „unter anderen Vorzeichen“ fortgesetzt werden. Er sollte nicht dazu verwendet werden, der Arbeit einen größeren Umfang zu geben und diese „dicker“ erscheinen zu lassen!

Der Anhang eignet sich für ergänzende Dokumente und Materialien, vor allem, falls diese für den Leser nur schwer oder gar nicht zugänglich sind, wie bspw. unveröffentlichte Betriebsunterlagen.

Vor allem in den empirischen Arbeiten kann der Anhang dazu dienen, verwendete Datensätze, eingesetzte mathematisch-statistische Verfahren oder Programme näher zu kennzeichnen. Werden im Rahmen der Untersuchungen Befragungen durchgeführt, sind die Fragestellungen und Ergebnisse im Anhang zu dokumentieren. Auf Gespräche darf im Rahmen der Ausführungen nur dann Bezug genommen werden, wenn ein vom Gesprächspartner unterzeichnetes Ergebnis-Protokoll im Anhang der Arbeit beigefügt ist.

Besteht der Anhang aus mehreren Elementen, so sind die einzelnen Elemente durch Nummerierung voneinander zu trennen.

A.1 Materialien

Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl.

A.2 Gesprächsprotokolle

Fusce vitae quam eu lacus pulvinar vulputate. Suspendisse potenti. Aliquam imperdiet ornare nibh. Cras molestie tortor non erat. Donec dapibus diam sed mauris laoreet volutpat. Sed at ante id nibh consectetuer convallis. Suspendisse diam tortor, lobortis eget, porttitor sed, molestie sed, nisl.

Eidesstaatliche Erklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate deutlich kenntlich gemacht zu haben.
Ich erkläre weiterhin, dass die vorliegende Arbeit in gleicher oder ähnlicher Form noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Würzburg, den 21. März 2016 Wolf, Peter