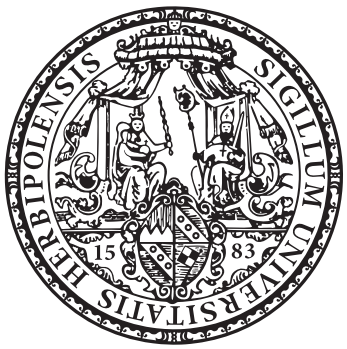
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

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

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

Einführung in die Problemstellung 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.

Motivation und Herleitung des Themas Topic Models lassen sich in der Form, wie sie von den verschiedenen Verfahren nur schwer interpretieren und besitzen unter Umständen Topics ohne Aussagegehalt.

Aufbau der Arbeit 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.

Sammlung der Ideen

- [Cha12, 1] Das Verstehen und organisieren von großen Sammlungen von Dokumenten ist zu einer wichtigen Tätigkeit in vielen Bereichen geworden. Viele Sammlungen sind aber nicht sinnvoll organisiert und das Organisieren von Hand ist zu umständlich.
- [Sie14, 1] In letzter Zeit wurde der Darstellung von Ergebnissen der Latent Dirichlet Allokation (LDA) den Topic Models viel Aufmerksamkeit zu Teil. (Gardner, Chaney, Chuang, Gretarsson) Eine solche Darstellung ist aber nicht einfach umzusetzen, aufgrund der hohen Abstraktion der Ergebnisse. Diese Ergebnisse lassen sich nur vollständig und kompakt darstellen, wenn eine interaktive Darstellungsform gewählt wird.
- [Sny13, 1] Wenn Nutzer Texte analysieren, benötigen sie eine intuitive Methode um Texte zu verstehen und zusammenzufassen. Werkzeuge, die dies ermöglichen lassen

sich in zwei Kategorien unterteilen. Die eine basiert auf strukturierten Metadaten und die Andere auf Informationen, die aus den Texten extrahiert wurden.

- [Ble09, 1] Topic Models sind hierarchische Bayes Modelle die einen betrachteten Textkorpus als eine kleine Verteilung von Wörtern darstellen. Die Idee hinter Topic Models ist, sich einen zufälligen Prozess vorzustellen aus dem die versteckte Struktur der Themen und die beobachtete Sammlung an Dokumenten zu Stande kommt. Dieser Prozess wird anschließend umgekehrt um die posteriore Verteilung der versteckten Themen Struktur zu schließen.
- [New10, 1] Während viele Nutzer konkrete Informationen zu einem Thema finden möchten, gibt es eine große Anzahl an Nutzern die alle Informationen zu einem bestimmten Thema finden und verstehen möchten und die Reichweite ihrer Suche (in Tiefe und Breite) kennen möchten. Eine genaue und intuitive Visualisierung der Suchergebnisse kann zu diesem Verständnis beitragen.

Gedanke In den letzten Jahren wurden Topic Models immer beliebter. Sie lassen sich jedoch nur schwer interpretieren. Aus diesem Grund wurden unterschiedliche Verfahren zur visuellen Aufbereitung der Topic Models entwickelt. Diese Visualisierungen unterscheiden sich in bestimmten Punkten.

Gedanke Das Ergebnis der Verfahren zur Erstellung von Topic Models ist eine Matrix mit den Wahrscheinlichkeiten der einzelnen Wörter. Als solche ist es nur schwer möglich diese zu interpretieren. Beleg?

2 Grundlagen

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 Topic Models

2.1.1 Probabilistic Latent Semantic Analysis

2.1.2 Latent Dirichlet Allocation

2.2 Darstellungsverfahren

2.2.1 Topic Browser

2.2.2 LDAvis

2.2.3 Termite

3 Visualisierung der Ergebnisse

3.0.0.1 Vierte Ebene

Fünfte Ebene

4 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

[Ble09] Blei, David M. und Lafferty, John D.: Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013* (2009)

[Cha12] Chaney, Allison June-Barlow und Blei, David M.: Visualizing Topic Models, in: *ICWSM*

[New10] Newman, David; Baldwin, Timothy; Cavedon, Lawrence; Huang, Eric; Karimi, Sarvnaz; Martinez, David; Scholer, Falk und Zobel, Justin: Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web* (2010), Bd. 8(2): S. 169–175

[Sie14] Sievert, Carson und Shirley, Kenneth E.: LDAvis: A method for visualizing and interpreting topics, in: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, S. 63–70

[Sny13] Snyder, Justin; Knowles, Rebecca; Dredze, Mark; Gormley, Matthew R. und Wolfe, Travis: Topic Models and Metadata for Visualizing Text Corpora, in: *HLT-NAACL*, S. 5–9

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 22. März 2016 Wolf, Peter