Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterthesis

Marcel Schöneberg

Konzepte zur semi-automatisierten Erstellung von Pressedossiers

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Marcel Schöneberg

Konzepte zur semi-automatisierten Erstellung von Pressedossiers

Masterthesis eingereicht im Rahmen der Masterprüfung

im Studiengang Master of Science Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. rer. nat. Kai von Luck
Zweitgutachter: Dr.-Ing. Sabine Schumann

Eingereicht am: 21. September 2015

Marcel Schöneberg

Thema der Arbeit
Konzepte zur semi-automatisierten Erstellung von Pressedossiers

Stichworte
Pressedossier, Text Mining, Distanzfunktion, Empfehlungssystem

Kurzzusammenfassung
Dieses Dokument behandelt die Thematik der semi-automatisierten Erstellung von Pressedossiers. Hierzu werden zunächst Grundlagen vorgestellt und eine Vision samt Fragenkatalog entwickelt. Auf dieses Basis werden verschiedene Konzepte vorgestellt und untersucht, welche es ermöglichen sollen ein Pressearchiv für Journalisten, anhand eines Leitartikels, thematisch einzugrenzen. Sodass diese nur eine vergleichsweise geringe Menge von Dokumenten betrachten müssen, um thematisch abgegrenzte Dossier zu erstellen. Das Ende dieser Arbeit bildet ein Fazit, welches Erfolge, sowie Probleme, resümiert und einen Ausblick auf mögliche weitere Arbeiten bietet.

Marcel Schöneberg

Title of the paper
Concepts for semi-automated creation of press dossiers

Keywords
press dossier, text mining, distance function, recommender system

Abstract
This document deals with the topic of semi-automatized creation of press dossiers. For this the basics are introduced and a vision, including a catalog of questions, is developed. On this basis a number of concepts is presented and analysed, to allow journalists to create a topically bounded dossier, based on an editorial. So that they only need to examine a relativly small amount of documents to create a topically bounded dossier. The end of this theses is a summary, which concludes the archievements, failures, as well as arised problems and gives an outlook of topics of possible future works.

Inhaltsverzeichnis

1. Einleitung

1.1. Motivation

Im heutigen Zeitalter der Digitalisierung sind der Öffentlichkeit mehr Informationen zugänglich als je zu vor. Diese Informationsmenge umfasst unter anderem Nachrichten aus aller Welt. Die hieraus entstehende Informationsflut ist, sowohl für einzelne Menschen, sowie für Presseagenturen und Journalisten, kaum zu verarbeiten. Um diesem „Information Overload“ entgegenzutreten braucht es daher die Hilfe von automatischen Verfahren, welche relevante Daten vorfiltern können, so dass ein Benutzer weniger Informationen zu verarbeiten hat.

Ein weiterer Aspekt der fortschreitenden Digitalisierung und der damit verbunden Informationsflut sind die ungeahnten Möglichkeiten, welche sich daraus ergeben. So eröffnen diverse Open-Data-Initiativen [beispielsweise: Tra; Gov] die Chance eine große Menge von Daten auf ungeahnte Weise zu verknüpfen. Diese Kontextualisierung von Informationen ermöglicht es ggf. auch komplett neue Anwendungsfälle zu erschließen und Zusammenhänge aufzudecken. Ein simples Beispiel für die Bedeutung des Kontextes einer Information ist die Verknüpfung der Daten *Temperatur* mit *Ort* und *Zeit*. Diese Daten ergeben isoliert betrachtet kein vollständiges Bild. So ist beispielsweise eine Temperatur von *20°C* in *Hamburg* durchaus eine Information. Allerdings ergibt sich ein gänzlich anderes Bild wenn eine Temperatur von *20°C in Hamburg im Dezember* gemessen wird. Diese Anomalie ergibt sich allerdings erst im Kontext der Daten. Konkreter können beispielsweise kontextualisierte öffentlich verfügbare Daten genutzt werden um im Fall von Naturkatastrophen die Ersthilfe zu beschleunigen. Weitere Szenarien zeigt unter anderem der Open-Data-Showroom [siehe: Ope] auf.

Neben der erwähnten Kontextualisierung erlaubt die enorme Menge an verfügbaren Daten auch eine Individualisierung. So zeigen Werbetreibende schon seit geraumer Zeit möglichst nur Anzeigen, welche für den Werbeempfänger von Interesse sein könnten. Ein ähnliches

Konzept lässt sich ggf. übertragen, so dass Daten nach ihrem Inhalt für Nutzer gefiltert werden. Denkbar ist eine Filterung von Nachrichten basierend auf den Interessen des Nutzers.

Diese Arbeit soll daher einen weiteres Element, zur Bewältigung der oben geschilderten Ideen (und vieler weiterer Szenarien), hinzufügen. So sollen Chancen genutzt werden, um Vorgänge zu automatisieren, welche vorher nur durch Menschen zu realisieren waren.

Ein Ansatz zur Verfolgung der geschilderten Möglichkeiten bietet das Gebiet des „Text Minings“, welches näher betrachtet werden soll. Dieses umfasst die Analyse von Texten hinsichtlich benötigter Informationen. Hierzu wird zunächst ein konkret vorhandenes Problem vorgestellt.

1.2. Problemstellung und Rahmen der Arbeit

Die obig beschriebene Informationsflut stellt unter anderem für Journalisten massive Probleme dar. Nachrichtenagenturen und Zeitschriften haben große Archive von Artikeln aus diversen Quellen. Allerdings ist es nicht ohne weiteres möglich diese zu überblicken. Dieses führt dazu, dass verwandte Informationen zu einem gegebenen Thema nicht immer problemlos auffindbar sind. Ein Redakteur, welcher an ähnlichen Informationen zu einem gegebenen Thema interessiert ist, um seine Kenntnisse zu erweitern oder Hintergrundinformationen zu finden, muss bei der Erstellung eines solchen Dossiers daher große Mengen von Text lesen und verarbeiten. Diese Arbeit ist von Menschen nicht zu bewältigen, aus diesem Grund stellt sich die Frage, ob das dargestellte Problem mithilfe von Technik angegangen werden kann. Diese Frage soll nachfolgend näher beleuchtet werden. Hierzu wird deshalb zunächst umrissen was diese Arbeit leisten soll und wo ihre Grenzen liegen.

Im Folgenden soll die Problemstellung sowie der Rahmen dieser Arbeit kurz abgesteckt werden. Das Problem der Verarbeitung großer Dokumentmengen soll angegangen werden indem ein *Archiv durchsucht* wird, um ausgehend von einem *Leitartikel ähnliche Dokumente ausfindig zu machen.* Hierbei sollen das Problem und mögliche Lösungen evaluiert werden. Im Kern beschäftigt sich der Autor daher mit Distanzen über Dokumenten, welche deren Ähnlichkeit zueinander beschreiben. Hierzu wird auf Anregungen von Domänenexperten und eigene Ideen eingegangen, mit dem Ziel, die Ähnlichkeitserkennung zwischen Artikeln zu verbessern. Der Aspekt der hierbei im Fokus steht, ist die Annäherung an Verfahren, welche die

Dimension des Problems verkleinern. Ein Journalist soll die Möglichkeit haben, ähnlich gute Rechercheergebnisse mit weniger Arbeit zu erreichen.

Als Arbeitsgrundlage existiert bereits eine konkrete Datenbasis, welche auf dem Eurozine-Korpus basiert. Darüber hinaus existieren diverse, Pressedossiers ähnlichen, Datensammlungen, welche zur Überprüfung der Ergebnisse herangezogen werden sollen. Nähere Informationen zu diesen Daten werden in Abschnitt 2.2.1 und 4.3.1 gegeben. Auf Basis dieser Informationen sollen verschiedene konkrete Verfahren untersucht und die Ergebnisse, bezüglich ihrer Nützlichkeit, evaluiert werden.

Was diese Arbeit nicht als Zielsetzung formuliert, ist die perfekte Lösung des dargestellten Problems. Darüber hinaus erhebt diese Arbeit auch keinen Anspruch darauf, alle denkbaren fachlichen Aspekte zu berücksichtigen. Diese sind zu vielfältig und zu komplex, als dass sie im Rahmen der Arbeit betrachtet werden könnten. Darüber hinaus sind die Facetten des Themas nicht nur von rein technischen Gegebenheiten abhängig, sondern von fachliche Elementen geprägt. Daher ist beispielsweise die Frage nach der „richtigen“ Lösung des Problems im Endeffekt eine rein journalistische Angelegenheit. Der Autor, als Informatiker, kann nur sehr begrenzt beurteilen, welche Informationen für einen Journalisten von Bedeutung sind.

Der fachliche Grundstein dieser Arbeit wurde von einer Domänenexpertin gelegt ([vgl. Hä14; Hä15]) und befindet sich zurzeit noch in der Entwicklung. Daher stellt auch diese Arbeit nur einen Ausschnitt aller möglichen Anforderungen dar.

1.3. Struktur dieser Arbeit

Diese Arbeit gliedert sich grob in sechs Kapitel, welche nachfolgend kurz umrissen werden.

Das vorliegende erste Kapitel dient der Einführung in die Thematik sowie der Beschreibung der Problematik. Darauf basierend wurde kurz der Rahmen der Arbeit abgesteckt.

Kapitel 2 erläutert die Entstehung der Vision dieser Arbeit und stellt daraufhin die benötigten fachlichen und technischen Grundlagen vor.

Kapitel 3 dient dazu das Ziel des Autors herauszuarbeiten und zu präzisieren, hierzu werden zunächst die Kernfragen dieser Arbeit vorgestellt, im Anschluss wird der geplante Anwendungsfall vorgestellt.

Kapitel 4 befasst sich mit der praktischen Umsetzung, welche die Erfüllung der Vision der Arbeit sicherstellen soll. Hierzu werden sowohl die Hauptbestandteile der Implementierung als auch die Methodik zur Bewertung von Ergebnissen vorgestellt.

Kapitel 5 geht auf bereits vorhandene Lösungsansätze (und Grundlagen) sowie in dieser Arbeit zu analysierende Ideen ein. Hierbei ist es das Ziel zu evaluieren, inwieweit diese zu einer Verbesserung der Resultate führen.

Kapitel 6 bietet ein Fazit der vorliegenden Arbeit. Hierzu wird der Kern der einzelnen Kapitel und die Essenz der Lösungsvorschläge resümiert. Daraufhin wird ein möglicher Ausblick auf zukünftige Entwicklungen dieser Arbeit geboten.

2. Grundlagen

Die folgenden Abschnitte beschreiben die Grundlagen dieser Arbeit. Hierbei bilden sowohl fachliche als auch technische Aspekte die Basis für die vorliegende Thesis.

2.1. Entwicklung einer Vision

Die angestrebte Arbeit, mit dem Ziel der semi-automatisierten Dossiererstellung, ist ein domänenübergreifendes Projekt. Dieses umfasst sowohl Informatik auf technischer Seite, als auch (im weitesten Sinne) die Journalistik, auf fachlicher Seite. Aus diesem Grund wurde die ursprüngliche (mittlerweile abgewandelte) Vision von fachlicher Seite inspiriert. Diese Idee umfasste die Zusammenstellung und Analyse von Presseerzeugnissen hinsichtlich verschiedener dargestellter Sichtweisen auf ein Thema. Ziel hierbei sollte es sein zu prüfen, ob Textmining basierte Dossiers dem Wachsen einer gemeinsamen europäischen Erzählung dienen können [vgl. auch: Hä14].

Die ursprünglich fachliche Vision wurde im Dialog mit dem Autor zunächst herunter gebrochen. Das hieraus entstandene primäre Ziel ist nun eine von Mitteln der Informatik gestützte Erstellung von Pressedossiers. Die generierten Dossiers sollen der Kontextualisierung eines Ausgangsartikels (Leitartikel) dienen. Dieser soll dabei für Domänenexperten aufbereitet werden, so dass er in einem Kontext steht und damit weitere Arbeiten erleichtert werden. Hierbei sind ausschließlich textbasierte Artikel und Dossiers im Fokus, daher bleiben multimediale Inhalte außen vor.

2.2. Fachliche Grundlagen

Diese geschilderte Zielsetzung erfordert zunächst die Erarbeitung einiger Grundlagen. Zu diesen gehört vor allem eine Datenbasis, auf welcher gearbeitet werden kann. Diese wird im Folgenden vorgestellt.

2.2.1. Artikelarchiv

Das fachliche Grundgerüst bildet neben dem Wissen eines Domänenexperten das Artikelarchiv des Kulturnetzwerkes Eurozine [Eurb]. Die vom Autor genutzten ca. 2700 Artikel wurden von professionellen Journalisten verfasst. Die Dokumente sind (teils als Übersetzung) in englischen Sprache verfügbar und es besteht darüber hinaus die Möglichkeit Metainformationen zu den Artikeln zu erhalten. Diese umfassen z.B. Verlinkungen auf die Inhaltsverzeichnisse der Ursprungszeitschrift und redaktionell erstellte Archive. Die Artikel selber liegen Form von XML-Dateien vor und weisen eine Semistrukturiertheit auf, so können unter anderem Informationen wie Autor, Kurzzusammenfassung sowie Überschriften etc. direkt dem Dokument entnommen werden. Ein Beispielartikel ist ab Seite 71 zu finden.

Trotz der genannten Vorteile ist das Archiv nicht makellos, weshalb eine Vorverarbeitung von Nöten war. Zu erwähnen ist, dass nicht alle ursprünglich vorhandenen ca. 7600 Artikel in englischer Sprache vorhanden waren. Darüber hinaus enthielt das ursprüngliche Archiv auch eine Reihe von Zusammenfassungen, Inhaltsangaben und Rezensionen bereits erschienener Artikel, welche sich daher nicht für die Analyse anbieten. Bei der Auswahl eines kleineren Testkorpus fielen darüber hinaus einige Gedichte auf, welche im Vergleich zum Rest eine extrem verkürzte Länge sowie eine inhärent andere Art der Sprache verwenden. Aus technischer Sicht ist zu bedenken, dass das XML-Markup nicht bei allen Artikel valide ist, dieses musste in einigen Fällen korrigiert werden.

Aus fachlicher Sicht ist anzumerken, dass die Artikel des Archivs einem Netzwerk aus Kulturzeitschriften zuzuordnen sind. Aus diesem Grund unterscheiden sie sich inhärent von anderen Presseerzeugnissen von Agenturen wie z.B. der Deutschen Presseagentur. Die Artikel des Eurozinearchivs weisen eine große thematische Breite auf und stehen auch zeitlich nicht zwangsläufig in Bezug zu aktuellen Nachrichten. Dieses zeigt sich unter anderem auch in den verwendeten Focalpoints (redaktionell zusammengestellte Dossiers [Eura]. Diese behandeln

sowohl Themen wie den Mauerfall, die Bologna-Reform, europäische Geschichte und auch den Klimawandel. Daher ist eine thematische und zeitliche Einordnung schwer und führt ebenfalls zu überlappenden Themengebieten [vgl. Hä15]. Dieses wird ebenso Auswirkungen darauf haben, mit welchen Mitteln und welchem Erfolg die Dokumente des Archivs analysiert werden können.

Das Archiv selber unterliegt einer Vertraulichkeitserklärung, deshalb kann der Autor nur begrenzt konkrete Artikelbeispiele (vgl. Abschnitt A) benennen. Allerdings lässt die Vertraulichkeitserklärung, bei Interesse, weitere wissenschaftliche Arbeiten zu, sofern Eurozine zustimmt.

2.2.2. Alternative Korpi

Der Eurozinekorpus ist nur einer von einer ganzen Reihe von möglichen Pressearchiven. Eine der bekanntesten Alternativen sind die Korpi der Nachrichtenagentur Reuters. Diese wurden für die Nutzung in der Forschung und Entwicklung von Natural Language Processing, Information Retrieval und Verfahren des maschinellen Lernens freigegeben und werden häufig verwendet [siehe: Reu]. Es existieren verschiedene Versionen der Reuters Korpi, welche im referenzierten Link samt ihrer zugehörigen Informationen aufgeführt sind.

Auswahl des Eurozine Korpus

Vergleicht man die vorgestellten Korpi, so zeigt sich, dass der Reuters Korpus deutlich größer ist. Darüber hinaus sind die Artikel einer Nachrichtenagentur naturgemäß sachlicher gehalten als die eines Kulturmagazins, auch die zeitliche und thematische Begrenzung ist im Reuters Korpus eher gegeben.

Dennoch hat sich der Autor dieser Arbeit dazu entschieden, das Eurozinearchiv als Basis der vorliegenden Thesis zu nutzen. Dieses geschah auch, da die fachliche Visionen und Ziele zusammen mit einer Fachexpertin entwickelt worden sind, welche direkte Einblicke in den Eurozinekorpus hat. Daher ermöglicht diese Zusammenarbeit es vom Wissen der Domänenexpertin zu profitieren. Dieses wäre bei Nutzung des Reuterskorpus nur eingeschränkt möglich gewesen.

2.3. Technische Grundlagen

Nach dieser Grobeinführung in die fachlichen Grundlagen widmen sich die folgenden Abschnitte der technischen Basis der vorliegenden Arbeit. Hierbei geht es vor allem um den Begriff des Text Mining und die technischen Realisierung des Vorhabens.

2.3.1. Text Mining

Da diese Arbeit dem Bereich des Text Mining zuzuordnen ist, soll dieser zunächst grob umrissen werden, um ein Verständnis zu vermitteln. Der Begriff Text Mining beschreibt eine Reihe von Verfahren, welche sich anbieten um dem Problem der immer größer werdenden Informationsmenge und deren Verarbeitung entgegenzutreten. Ziel ist es, durch Nutzung von Methoden wie z.B. der Verarbeitung von natürlicher Sprache (Natural Language Processing; NLP), Information Retrieval usw. relevante Informationen aus Eingangsdaten zu extrahieren. Im Gegensatz zu allgemeinem Data Mining ist die Datenbasis beim Text Mining ein Dokumentarchiv (Korpus) von rein textbasierten Informationen [FS06].

Während des Text Mining Prozesses können Verfahren wie das Extrahieren von „interessanten“ Termen, Aufteilung in Cluster sowie Nutzung von linguistischen Informationen (Ontologien, etc.) und weitere Ideen genutzt werden. Eine Übersicht über die grundlegenden Verfahrensweise ist der Grafik 2.1 zu entnehmen.

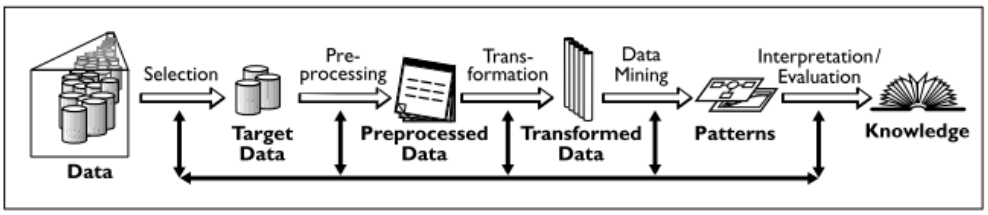


Abbildung 2.1.: Schritte des KDD Prozess [Fay+96]

Die obige Grafik 2.1 beschreibt den KDD (Knowledge Discovery in Databases) Prozess. Dieser illustriert in fünf Schritten den groben Ablauf des Data Mining bzw. Text Mining, so wie er auch in dieser Arbeit genutzt wird.

1. Kennenlernen der Domäne.

2. Auswahl von geeigneten Arbeitsdaten (siehe: Abschnitt 3.2) aus dem Artikelarchiv. Dieses geschieht auf Grundlage von Wissen, welches der Domänenexperte (Journalist) bereitstellt (siehe 1.).

3. Datenbereinigung (wie das Herausfiltern von irrelevanten Informationen (detaillierter in Abschnitt 4.1))

4. Datenreduktion / Transformationen, unter anderem Auswahl geeigneter Features, (Analyseobjekte) zur Untersuchung (konkreter ab Abschnitt 4.1).

5. Auswahl von Analysemethoden und Interpretation der gewonnenen Erkenntnisse (siehe auch: Kapitel 5).

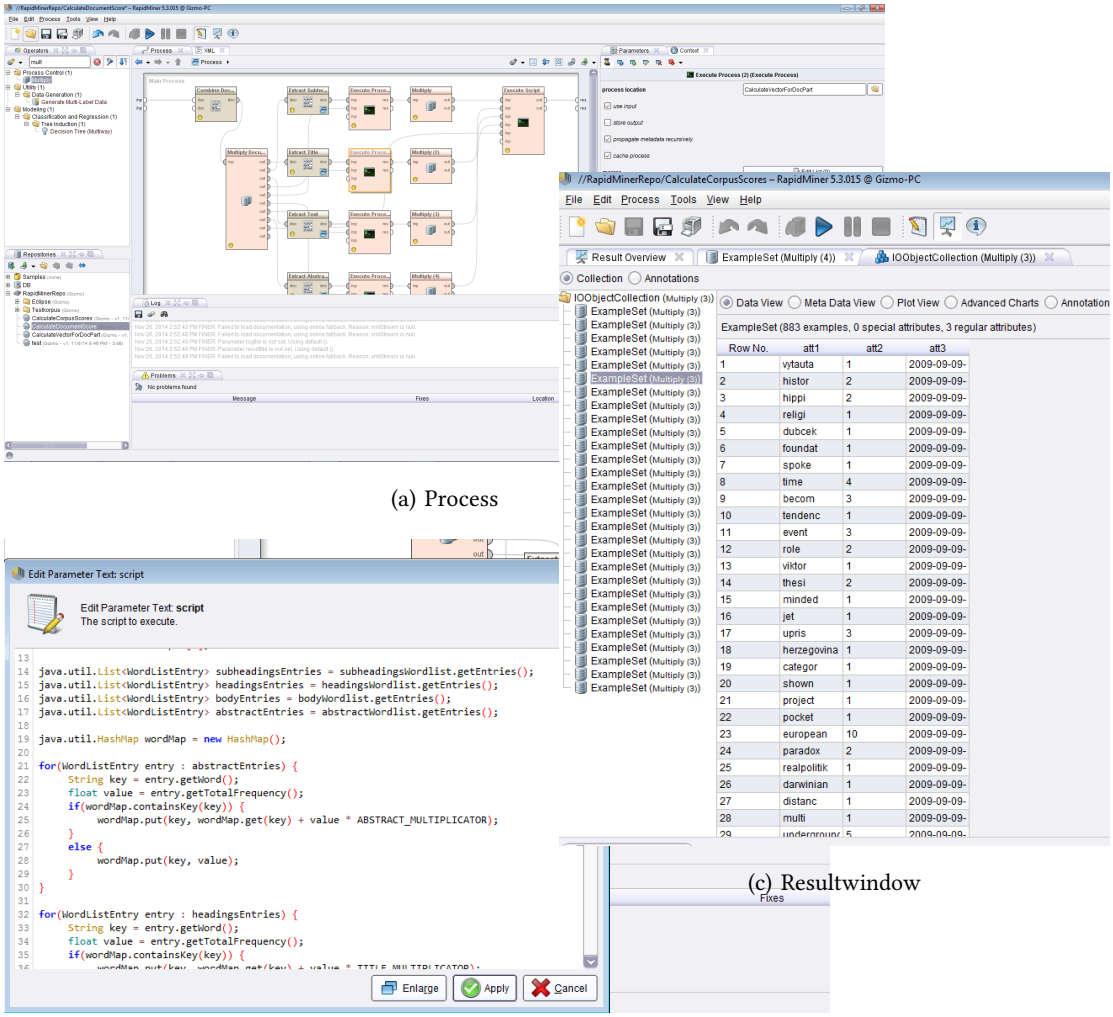
Der vorgestellte Prozess ist allerdings nur als Referenzmodell zu verstehen, eine konkrete Anwendung ist dementsprechend komplexer. Allerdings schafft das Modell ein grundlegendes Verständnis für den Ablauf, welcher vom Autor genutzt wird.

2.3.2. RapidMiner

Da das verfolgte Ziel auch praktisch umgesetzt werden soll, ist eine Plattform für die Realisierung notwendig. Diese bildet die Data Mining-Umgebung RapidMiner [siehe Rap], verwendet wurde Version 5.3. Das Framework wurde ursprünglich von der TU Dortmund unter dem Namen „Yet Another Learning Environment“ (YALE) entwickelt. Die Zielsetzung hierbei war eine Software, welche es erlaubt, eine Vielzahl Methoden des Data Mining auf einfache Art und Weise bereitzustellen, so dass ein Nutzer sich direkt dem Problem widmen kann, welches es zu lösen gilt. Die diversen Algorithmen, welche zur Verfügung stehen, sind daher in Form von Operatoren umgesetzt. Diese sind Bausteine, die direkt zu Prozessen verbunden werden können. Diese Umsetzung ermöglicht es auf (meist) einfache Weise, komplexe Analyseprozesse zusammenzusetzen.

RapidMiner selbst lässt sich über eine Reihe von Plugins erweitern, um Zusatzfunktionalität (z.B. die Text Mining Extensions) hinzuzufügen. Darüber hinaus lässt sich selbst entwickelter

Code einpflegen, um eigene Ideen umzusetzen. Der gesamte Ablauf lässt sich innerhalb einer graphischen Entwicklungsumgebung sowie in Java realisieren.



(a) Process

(c) Resultwindow

(b) Script

Abbildung 2.2.: Einblicke in RapidMiner

Ein Beispiel für die Anwendung von RapidMiner ist in Abbildung 2.2 zu sehen. Diese zeigt einen RapidMiner-Prozess, mit diversen verbundenen Operatoren. Darüber hinaus ist ein Skript zu sehen, welches die Funktionalität durch entwickelte Logik erweitert. Der letzte Teil der Abbildung zeigt eine RapidMiner-Tabelle mit den Ergebnissen des Prozesses.

2.3.3. Alternative Frameworks

RapidMiner ist natürlich nicht die einzige Toolsuite, welche es erlaubt, komplexe Analyseprozesse aufzubauen. Es gibt eine ganze Reihe von alternativen Frameworks, welche ähnliche Möglichkeiten haben. Hierzu gehören beispielsweise:

- Apache Mahout [Mah]
- Apache Tika [Tik]
- KNIME [Kni]
- R [Rpr]
- Weka [Wek]

Diese alternativen Werkzeuge (und diverse weitere) bieten vielfältige Möglichkeiten zur Realisierung. Allerdings kann nur ein Werkzeug genutzt werden und die jeweiligen Vor- und Nachteile eines Werkzeugs zeigen sich häufig erst während dessen Benutzung. Da ein möglichst bekanntes Framework die Wahrscheinlichkeit erhöht, aufkommende Probleme zu lösen, blieben für den Autor „R“ und RapidMiner in der engeren Auswahl. Da „R“, allerdings aufgrund von früheren Erfahrungen ([vgl. Sch14]), ungeeignet erschien fiel die Wahl daher schlussendlich auf RapidMiner.

3. Zielsetzung und Anwendungsfall

Nachdem die grundlegende Vision des Autors sowie fachliche und technische Grundlagen erläutert wurden, ist es wichtig, eine konkrete Zielstellung zu erarbeiten. Da das dargestellte Ziel vor allem auf fachlichem Wissen fußt, ist es nötig, einen Fragenkatalog zu erarbeiten. Dieser soll es ermöglichen zu klären, was Kernaspekte des Ziels sind und wie ein erfolgreicher Abschluss dieser Arbeit aussehen kann. Daher dienen die folgenden Abschnitte dazu, das Ziel des Autors herauszuarbeiten. Hierzu werden zunächst die Kernfragen dieser Arbeit vorgestellt, im Anschluss wird der geplante Anwendungsfall erörtert.

3.1. Fragenkatalog

Der folgende Abschnitt dient dazu eine Reihe an Fragen herauszuarbeiten, welche als Leitfaden der Masterarbeit genutzt werden sollen. Diese beinhalten punktuell zum einen eher fachliche, zum anderen eher technische Aspekte. Die aufgestellten Fragen sind soweit möglich entsprechend ihrer logischen Reihenfolge aufgeführt.

Den Kern bildet zunächst die Frage nach der **Machbarkeit** der (teil-)automatisierten Erstellung von Pressedossiers. Hierbei umfasst das Kriterium *Machtbarkeit* mehr als die reine technische Umsetzung. Ein entscheidender Punkt ist die fachliche Sinnhaftigkeit. Hierzu muss beantwortet werden, ob eine (teil-)automatisierte Erstellung von Pressedossiers als Hilfsmittel zur Einschränkung der Auswahl von Artikeln (siehe: Abschnitt 3.2) **hilfreich** für Journalisten ist. Die Antwort auf diese Fragen hängt von einer Reihe feingranularerer Aspekten ab, welche zum Teil allerdings erst im Laufe der Arbeit zu beantworten sind. Diese Aspekte umfassen unter anderem journalistische Wünsche, welche derzeit noch erarbeitet werden [vgl. Hä15]. Andere Fragen wie z.B. die der technischen Realisierung, bilden einen Schwerpunkt dieser

Arbeit. Daher ist es möglich, dass das gesetzte Ziel der (teil-)automatisierten Erstellung von Pressedossiers, schlussendlich nicht oder nur eingeschränkt zu erreichen ist.

Die nun logisch folgende nächste Frage befasst sich mit der **Definition von Pressedossiers** bzw. dem Nutzungsszenario auf fachlicher Seite. Was ein Dossier ist und wie dieses genutzt wird, ist wichtig für die weitere Arbeit des Autors. Allerdings ist dieser Punkt nicht allein von technischer Seite zu beantworten. Hierzu bedarf es wiederum der Hilfe durch einen Domänenexperten. Die dargelegten Fragen bilden daher die Basis für eine Schnittstelle, welche die fachliche mit der technischen Sicht verbindet. Der so entstehende Verbindungspunkt ist die Distanzfunktion, welche verschiedene Artikel in Bezug zueinander setzt und ihre Ähnlichkeit in Form eines Zahlenwertes repräsentiert. Für diese Funktion müssen fachliche Wünsche mithilfe von Mitteln der Informatik in Algorithmen überführt werden.

Nun, da eine grobe Schnittstelle definiert ist, existiert eine Überleitung zur nächsten Frage. Diese untersucht, ob **fachlich definierte Anforderungen** an ein Dossier sich grundsätzlich in einen **Algorithmus überführen** lassen. Sind die fachlichen Wünsche im Allgemeinen eher auf eine Art „Bauchgefühl“ zurückzuführen bzw. schwammig formuliert, kann man dieses nicht umsetzen, hierzu bedarf es konkreterer Informationen.

Dieses führt zu der eng verbundenen Fragestellung nach der **Machbarkeit der Umsetzung**. Ein mögliches Szenario, welches sich aus der Zusammenarbeit von Informatik und Journalistik entwickelt, könnte eine recht präzise Definition eines Dossiers sein, welche grundlegend genug Informationen enthält, um einen Algorithmus zu entwerfen. Allerdings ist es möglich, dass ein Fazit dieses Szenarios eine nicht realisierbare Lösung ist. Da aufgrund von mangelnder Datenbasis, beschränkter Ressourcen oder technischer Mittel etc. keine praktische Umsetzung innerhalb der geplanten Zeit bzw. den vorhandenen Ressourcen möglich ist.

Grundsätzlich ist darüber hinaus zu beantworten, wie man ein generiertes Ergebnis bewertet, um seine **Qualität** zu ermitteln. Darüber hinaus muss das Bewertungsmaß die Möglichkeit bieten, eine Verbesserung oder Verschlechterung der Ergebnisse bei Änderungen zu erkennen.

Neben der Beantwortung der obigen Fragen wäre es wünschenswert, einen zu **entwickelten Algorithmus möglichst übersichtlich** zu halten, so dass dieser durchaus (für Fachexperten) nachvollziehbar ist. Sofern dieses nicht gelingt, entsteht eine (für Nicht-Informatiker) „Blackbox-Lösung“, welche umso schwerer auf ggf. auftretende weitere Wünsche anpassbar ist.

3.2. Definition von Dossiers und geplanter Anwendungsfall

Nach der Vorstellung der Leitfragen dieser Arbeit ist es wichtig, zunächst die Frage nach der Definition eines Dossiers zu klären. Die Beantwortung dieses Kernpunktes ist allerdings nicht trivial, so gibt es verschiedene (Experten-)Meinungen, darüber was ein Dossier ausmacht ([vgl. Hä14]). Beispiele für Definitionen umfassen „eine Sammlung von Dokumenten in einer Pappmappe“ sowie stark fachliche und von einander abweichende Erklärungen und Wünsche, welche ein Dossier erfüllen muss.

Die vom Autor verfolgte Arbeitsthese ist, dass ein Dossier zunächst eine Zusammenstellung von ähnlichen Artikeln im weitesten Sinne ist. Diese haben eine inhaltliche/semantische Nähe, welche mithilfe von Distanzfunktionen ermittelt werden kann. Dieser Ansatz soll als Grundlage für weitere Arbeiten dienen, da er die bestmögliche Abstraktion darstellt. Er erfüllt allerdings nicht alle möglichen Anforderungen, welche sich ergeben können. Aus diesem Grund ist die Zielsetzung eingeschränkt zu betrachten, dieses illustriert Abbildung 3.1.

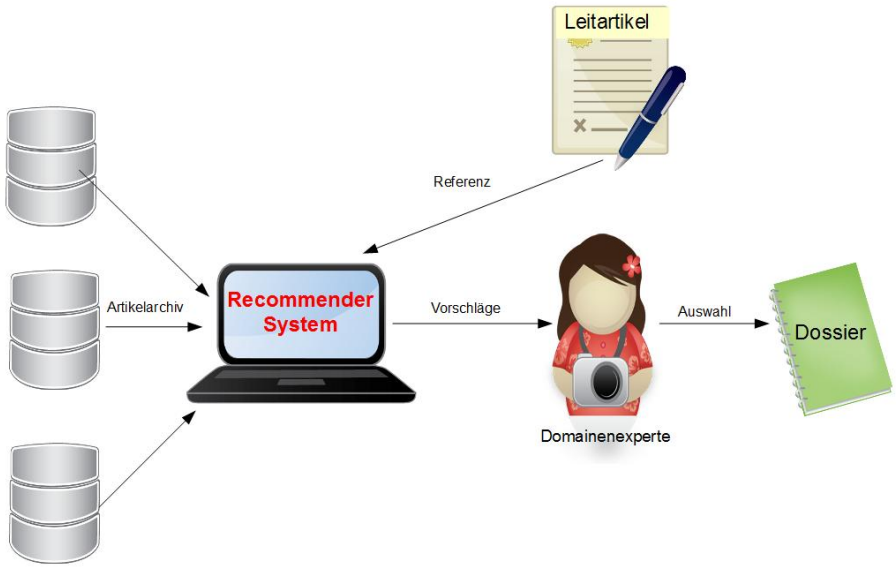


Abbildung 3.1.: Realisierbarer Workflow

Die bereits geschilderte Vision wird durch die obige Grafik noch einmal visualisiert und dient dazu, ein besseres Verständnis für einen möglichen praktischen Anwendungsfall zu erlangen. Dieser beschreibt sich wie folgt:

Ausgehend von einem **bestehenden Artikelarchiv** und einem **gegebenen Leitartikel** (welcher als Referenz für die gesuchte Ähnlichkeit dient), sollen **Vorschläge** (basierend auf der in einer im System berechneten Distanz) generiert werden. Diese werden einem **Domänenexperten** unterbreitet, welcher mithilfe der Vorschläge ein Dossier erstellen kann. Das geschilderte Vorgehen ermöglicht eine semi-automatisierte Dossiererstellung, ohne sich komplett auf die hinter liegende „Intelligenz“ der entworfenen Methode oder den Domänenexperten verlassen zu müssen. Das Verfahren soll hierbei durch weitere Arbeiten fortlaufend verbessert werden, so dass der Domänenexperte im Optimalfall nur noch eine geringe Anzahl von Artikeln sichten muss, um ein brauchbares Dossier zu einem Thema zu erstellen.

4. Architektur und Bewertung

Basierend auf den erarbeiteten Grundlagen und den vorliegenden Leitfragen ist es nun wichtig, sich einer praktischen Umsetzung zu nähern. Hierzu wird zunächst die grundlegende Architektur der Realisierung vorgestellt. Darüber hinaus wird das entworfene Bewertungsschema vorgestellt, welches es ermöglichen soll, zu prüfen wie erfolgreich die verschiedenen durchzuführenden Versuche verlaufen sind.

4.1. Hauptbestandteile des Systems

Die folgenden Abschnitte widmen sich den Grundlagen der Umsetzung der Vision in RapidMiner. Hierzu werden die Hauptbestandteile der Prozesskette umrissen, welche genutzt wird, um das gesetzte Ziel zu erreichen.

Grundsätzlich umfasst das entstandene RapidMiner-Projekt verschiedene Anwendungsfälle, welche sich voneinander erheblich unterscheiden können. So ist beispielsweise die Prozesskette für die Gewichtung einzelner Textbestandteile und die Auswertung der daraus entstehenden Ergebnisse nur eingeschränkt mit der Prozesskette für Clustering-Versuche vergleichbar. Daher wird im Folgenden ein konkreter Versuch detaillierter betrachtet. Dieser orientiert sich grundsätzlich am bereits vorgestellten KDD-Prozess 2.1.

Viele der durchgeführten Versuche basieren auf einer ähnlichen Prozesskette, welche mehrere Bestandteile besitzt. Die wichtigsten sind in Abbildung 4.1 zu sehen. Diese Grafik stellt daher auch einen groben Überblick über das gesamte RapidMiner-Projekt dar. Die einzelnen Teilbereiche des Prozesses werden nachfolgend näher erläutert.

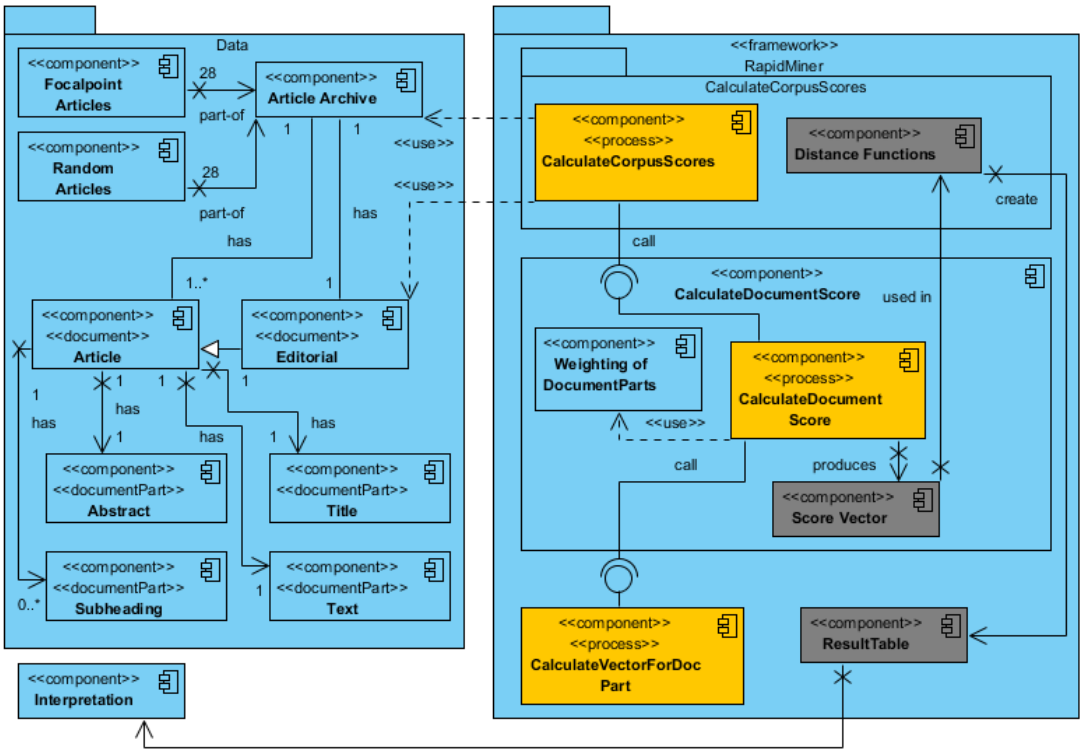


Abbildung 4.1.: Komponenten und Workflow

Datenbasis Als Datenbasis dient das Artikelarchiv (vgl. Abschnitt 2.2.1) sowie ein Leitartikel. Dieser dient als Ausgangspunkt für eine Distanzfunktion, die einen Wert berechnet, welcher die Nähe eines Artikels zum Leitartikel repräsentiert. Sowohl das Archiv als auch der Ausgangsartikel liegen in Form von XML-Files in einem separaten Ordner des Projekts. Das Archiv selber ist aufgeteilt in zufällige Artikel (markiert durch einen Namenspräfix) und von Menschen zusammengestellte Dokumente eines Focalpoints zu einem konkreten Thema. Diese Zusammenstellungen werden später zur Überprüfung der Ergebnisse benutzt. Hierbei ist anzumerken, dass es verschiedene Focalpoints gibt, so dass mehrere Testreihen möglich sind. Dieses verringert die Abhängigkeit zu einem bestimmten Datensatz und verhindert darauf zugeschnittenen Untersuchungen.

RapidMiner Nachdem fachliche Grundaspekte der Architektur geklärt sind, widmen sich die folgenden Abschnitte den RapidMiner-Prozessen, welche die Logik zur Distanzberechnung enthalten. Diese Prozesse beschreiben einen Ablauf von Operatoren (Algorithmen) und sind teils ineinander geschachtelt. Die folgenden Erklärungen umreißen die Hauptbestandteile des Gesamtprozesses, um einen Eindruck zu vermitteln.

CalculateCorpusScores Der Gesamtablauf des Experiments startet in diesem Prozess. Hier werden sowohl der Leitartikel als auch (nacheinander) die Einzeldokumente des Archivs eingelesen. Die Dokumente werden an den Prozess „CalculateDocumentScores“ weitergegeben, welcher einen „Bag-Of-Words“ berechnet. Die Ergebnisse dieser Berechnung werden zur Distanzfunktionsberechnung weitergereicht, welche die Resultate (die Distanzen aller Dokumente zum Leitartikel) in eine Ergebnistabelle einträgt.

Die Ergebnistabelle enthält alle analysierten Dokumente inklusive der verschiedenen berechneten Distanzen. Diese können im Weiteren interpretiert werden (vgl. Abschnitt 4.3.2.3).

CalculateDocumentScores Nach dem Einlesen der Daten berechnet der Prozess „CalculateDocumentScores“ für jedes Input-Dokument einen (gewichteten) „Bag-Of-Words“ [vgl. FS06, Seite 68] einen (Feature-)Vektor mit allen im Dokument enthaltenen Wörtern und ihren Häufigkeiten. Das geschilderte grundlegende Modell ist auch unter dem Namen „Vector Space Model“ (VSM) bekannt [vgl. FS06; LGS, S. 85].

Bei der Berechnung des Feature-Vektors werden vier verschiedene Dokumentteile getrennt behandelt (Abstract, Titel, Unterüberschriften sowie der eigentliche Text). Diese werden mit XPath-Ausdrücken aus dem XML-Artikel entnommen und an den Prozess „CalculateVectorForDocPart“ weitergereicht. Die Ergebnisse dieser Berechnung werden daraufhin an ein Script weitergegeben, welches eine Gewichtung der Einzelbestandteile durchführt. Dieses kann dazu benutzt werden Bestandteile (wie z.B. im Abstract) im „Bag-Of-Words“ stärker hervorzuheben.

CalculateVectorForDocPart Dieser Prozess führt ein Preprocessing [vgl. FS06; FPS96] des Eingabedokumentes durch. Dieses umfasst zunächst:

- Tokenizing: Zerlegung des Inputs in Token (konkret: Wörter)
- Stopword removal: Entfernung von Worten, welche häufig vorkommen, allerdings wenig Bedeutung für den Dokumentinhalt haben (z.B. Artikel).
- Stemming: Reduzierung aller vorhandenen Wörter auf den Wortstamm (z.B. comput: compute, computes, computed, computing, computable …)
- Transform cases: Entfernung von Groß und Kleinschreibung

Diese Vorverarbeitung reduziert den entstehenden Feature-Vektor auf die relevanten Daten hierbei ist die Verarbeitung jederzeit erweiterbar, sofern sich neue Anforderungen ergeben. Ein Beispiel für eine Erweiterung wäre, dass nur die n häufigsten Wörter zur Generierung des „Bag-Of-Words“ genutzt werden.

Im Anschluss an die Verarbeitung wird als Resultat ein Wortvektor mit den vorhandenen Worten und deren Häufigkeit im Eingangsdokument erstellt und zurückgegeben.

4.2. Distanzen

4.2.1. Grundlage: Euklidische Distanz

Ein Kernelemente des Projekts bilden Distanzfunktionen. Das Ziel der folgenden Abschnitte ist daher, kurz die Idee einer Distanzfunktion zu umreißen und Alternativen vorzustellen. Dieses ist die Basis, welche benötigt wird, um die später vorgestellte Methodik der Bewertung von Ergebnissen zu verstehen.

Da es zunächst gewünscht ist, die Funktion im Rahmen des Ziels so einfach wie möglich zu halten, wird zunächst eine einfache *euklidische Distanz* gewählt [FS06, Seite 85]. Diese berechnet sich als:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{4.1}$$

Hierbei sind x bzw. y die jeweiligen (gewichteten) Vorkommen eines Wortes (einerseits im Leitartikel, andererseits im Vergleichsartikel). Die Einzelsummanden der Gleichung (x_1 bis x_n) repräsentieren hierbei alle Worte (bzw. ihr Vorkommen) in den Dokumenten. Anzumerken ist, dass die Gesamtmenge aller Wörter beider Dokumente als Basis für den Vergleich dient, taucht also ein Wort in einem der Artikel auf, im anderen allerdings nicht, so wird seine Häufigkeit im entsprechenden Artikel als 0 angesehen.

Die oben beschriebene Variante der Distanzfunktion geht vom reinen Vorkommen eines Wortes aus (der Termfrequency). Diese berücksichtigt allerdings keine Normierungen, um beispielsweise Häufungen des gleiches Wortes in einem längeren Artikel zu kompensieren. Ebenso kann eine Normierung darauf eingehen, dass ein Wort, in einem Set von Dokumenten häufig vorkommt und damit weniger wertvoll ist, als ein Wort das in dem Set nur wenige Male vorkommt. In letzterem Fall hat das Wort bezogen auf das Dokumentset eine höhere Bedeutung.

Die obigen beiden Problematiken lassen sich mithilfe einer Normierung der Termfrequency bzw. dem TF-IDF Maß (Termfrequency - Inverse Document Frequency) angehen [siehe MRS08, Seite 117 ff.]. Diese sind im Folgenden kurz dargestellt und bilden jeweils veränderte „neue“ Distanzfunktion durch Ersetzung der x- bzw. y-Werte (in der Gleichung 4.1), durch die in den Formeln 4.2 bzw. 4.3 berechneten Werte.

Hierbei berechnet sich eine normierte Termfrequency als:

$$ntf_{t,d} = \alpha + (1 - \alpha) * \frac{tf_{t,d}}{tf_{\text{max}}(d)} \tag{4.2}$$

Wobei α ein Glättungsfaktor im Interval [0,1] ist und tf_{max} das Vorkommen des am häufigsten verwendeten Terms über alle Dokumente.

Das TF-IDF Maß basiert auf dem IDF-Maß [FS06, Seite 68], welches sich wie folgt berechnen lässt:

$$\text{tf-idf}_{\text{t,d}} = tf_{\text{t,d}} * idf_t \tag{4.3}$$

Hierbei ist $tf_{\text{t,d}}$ die Termfrequency eines Wortes und idf_t die Inverse-Document-Frequency, berechnet durch:

$$idf_t = log\frac{N}{df_t} \tag{4.4}$$

Wobei N = Anzahl von Dokumenten im Set und df = Anzahl von Dokumenten mit Term t ist.

Beispiel

1. Das folgende Beispiel soll kurz erläutern wie eine (euklidische) Distanz zwischen zwei Dokumenten zustande kommt, so dass der Leser ein Verständnis für Distanzmaße entwickeln kann. Hierbei soll als Basis die reine Termfrequenz benutzt werden, weiterhin sollen die Vergleichsdokumente folgende Wortvorkommen haben:

$$\begin{pmatrix} Dokumentnummer & `apple` & `banana` & `cat` & `window` \\ 1 & 2 & 2 & 0 & 4 \\ 2 & 1 & 3 & 2 & 0 \end{pmatrix}$$

2. Diese Häufigkeiten werden nun gewichtet. Hierbei sollen die vorkommenden Worte in ihrer Reihenfolge genau den Textabschnitten „Abstract“, „Title“, „Subheadings“ und „Body“ entsprechen.

Dieses bedeutet, dass im ersten Dokument das Wort „apple“ genau 2 mal im Abstract vorkommt, „banana“ 2 mal im Titel, „cat“ 0 mal in den Unterüberschriften benutzt wird und das Wort „window“ 4 mal im restlichen Text auftritt.

Zu bedenken ist, dass jedes Wort im Beispiel (zur Vereinfachung) nur in genau einem Abschnitt vorkommt. Die gewählten Parameter zur Gewichtung seien 4,3,2,1. Hieraus ergibt sich die folgende Matrix mit den jeweils gewichteten Vorkommen:

$$\begin{pmatrix} Dokumentnummer & `apple_1 & `banana_1 & `cat_1 & `window_1 \\ 1 & 2*4 & 2*3 & 0*2 & 4*1 \\ 2 & 1*4 & 3*3 & 2*2 & 0*1 \end{pmatrix}$$

3. Berechnet man nun die euklidische Distanz des zweiten Artikels, zum ersten (dieser dient als Leitartikel), so ergibt sich folgende Gleichung:

$$\begin{aligned} d(x,y) &= \sqrt{(8-4)^2 + (6-9)^2 + (0-4)^2 + (4-0)^2} \\ d(x,y) &= \sqrt{4^2 + -3^2 + -4^2 + 4^2} \\ d(x,y) &= \sqrt{16 + 9 + 16 + 16} \\ d(x,y) &= \sqrt{57} \\ d(x,y) &= 7,55 \end{aligned}$$

Die Distanz der beiden Dokumente, unter Berücksichtigung der Gewichtung, ist daher 7,55.

Dieses Beispiel ist, wie erwähnt, stark vereinfacht und dient nur dazu eine Vorstellung der Funktionsweise der Berechnung zu gewinnen. Die Information, dass die beiden Artikel eine Distanz von 7,55 haben, ist ohne weitere Informationen, zunächst nicht sehr hilfreich. Die Nützlichkeit von Distanzen zeigt sich zumeist erst im Vergleich von mehreren Werten. Dieses würde allerdings den Rahmen eines Beispiels sprengen.

4.2.2. Erweiterung um die Kosinus Distanz

Ausgehend von dem obigen Vorwissen soll nun eine weitere Distanz (bzw. Ähnlichkeit) eingeführt werden, die so genannte Kosinus-Distanz (bzw. Ähnlichkeit) [MRS08, Seite 121]. Diese eignet sich besonders gut, wenn die zu vergleichenden Dokumente eine unterschiedliche Länge aufweisen. Während sich diese Länge bei Verfahren wie der euklidischen Distanz aufsummiert, produziert das Kosinus-Maß keine negativen Effekte aufgrund der Längendifferenz. Die Grundidee dieser Distanz basiert auf der Interpretation des Feature-Vektors eines Dokumentes als Richtungsvektor (in ein einem n-dimensionale Raum). Zwei Feature-Vektoren können dementsprechend verglichen werden, indem man den Kosinus des Winkels der beiden Vektoren berechnet. Je kleiner dieser Winkel ist, umso größer ist die Ähnlichkeit der Dokumente.

Die Ähnlichkeit berechnet sich wie folgt:

$$sim(doc1, doc2) = \frac{\vec{v}(doc1) \cdot \vec{v}(doc2)}{|\vec{v}(doc1)||\vec{v}(doc2)|} \tag{4.5}$$

Hierbei ist v der jeweilige Feature-Vektor der zu vergleichenden Dokumente und $\cdot$ das Skalarprodukt. Um die so berechnete Ähnlichkeit in eine Distanz zu überführen, kann die folgende Funktion genutzt werden:

$$dist(doc1, doc2) = 1 - sim(doc1, doc2) \tag{4.6}$$

Sowohl das Ähnlichkeits- als auch das Distanzmaß der Kosinusfunktion bewegen sich in einem Intervall zwischen 0 und 1 (da keine negativen Werte in den Feature-Vektoren möglich sind).

4.3. Bewertung

Nachdem erläutert wurde, was erreicht werden soll und wie der geplante Workflow (vgl. Graphik 3.1) Vorschläge für einen Fachexperten produziert, welche dieser verwerten kann, ist es essentiell zu wissen, ob etwaige Veränderungen des Algorithmus zu Verbesserungen führen. Hierzu soll nachfolgend auf die Methodik der Bewertung von Ergebnissen eingegangen werden.

4.3.1. Focalpoints und Korpi

Um produzierte Resultate auf ihre Sinnhaftigkeit zu prüfen, benötigt man Hintergrundwissen. Dieses hat nur ein Domänenexperte. Da es allerdings nicht möglich ist jedes Resultat Journalisten vorzulegen, muss auf andere Mittel zurückgegriffen werden. Konkret werden hierzu redaktionell erstellte Focalpoints genutzt, welche von Journalisten erstellt wurden und so das benötigte Wissen implizit enthalten. Die Focalpoints enthalten ca. 30 Artikel zu einem festgelegten Thema, diese Sammlungen stellen zwar kein konkretes Dossier dar, bieten allerdings eine gute Basis für die Überprüfung der gewonnenen Ergebnisse.

Für den Aufbau eines Testkorpus wurde daher jeweils ein gewählter Focalpoint mit diversen anderen Artikeln aus dem Archiv verschmolzen. Die vom Verfahren vorgeschlagenen Artikel (ausgehend von einem Leitartikel, der ebenfalls dem Focalpoint entstammt) sollen nun im besten Fall genau die Artikel des Focalpoints sein. Diese These ergibt sich, da der Leitartikel sowie die anderen Focalpoint-Artikel, ein gemeinsames Thema verfolgen und aus diesem Grund bereits als Dossier zusammengestellt wurden. Deshalb ergibt sich, dass die Focalpoint-Artikel als Positivbeispiele und die zufällig zusammengestellten Artikel als Negativbeispiel dienen.

Grundsätzlich gibt es verschiedene Focalpoints, welche genutzt werden können. Dieses erlaubt eine gewisse Vielfalt der Experimente. In früheren Projekten ([Sch15a]) wurde bereits der Focalpoint „Democracy“ genutzt, daher ist dieser der Ausgangspunkt für Untersuchungen, da er dem Autor bereits bekannt ist.

Im Verlauf dieser Arbeit wurden verschiedene Datensätze von Dokumenten (Korpi) genutzt. Grundsätzlich basiert jeder Korpus auf einem Focalpoint sowie einer Auswahl an zufälligen Artikeln. In den aufgeführten Versuchen wurden die folgenden Datensätze benutzt:

- **Korpus A:** 29 Artikel des Focalpoint „Democracy“; 29 zufällige Artikel
- **Korpus B:** 19 Artikel des Focalpoint „Bologna“; 29 zufällige Artikel
- **Korpus C:** 19 Artikel des Focalpoint „Bologna“; 19 zufällige Artikel
- **Demokratie:** 29 Artikel des Focalpoint „Democracy“
- **Bologna (reduziert):** 19 Artikel des Focalpoint „Bologna“

- **Zufällig:** 29 zufällige Artikel (nicht in den Focalpoints enthalten)

Zu beachten ist, dass die diversen Korpi verschiedene Verhältnisse von Focalpoint-Artikeln zu zufälligen Artikeln haben. Dieses kann ein Ergebnis stark beeinflussen, da die Wahrscheinlichkeit, einen „falschen“ Artikel (einen zufälligen) aus dem Gesamtkorpus zu wählen, steigt.

4.3.2. Methodik

Nachdem die Grundlagen der Bewertung von Ergebnissen erläutert sind, wird im Folgenden auf die gewählte Methodik der Bewertung eingegangen. Hierzu werden verschiedene Ideen aus dem Gebiet des „Information Retrieval“ genutzt, da sie sich anbieten, um eine sinnvolle Bewertung zu entwerfen. Diese Methoden werden im Weiteren erläutert.

4.3.2.1. Precision-Recall

Um eine Bewertung zu ermöglichen, benötigt man konkrete Zahlenwerte, welche die Qualität eines Resultats ausdrücken. Eine Möglichkeit dieses Ziel in Zahlenwerte zu übertragen, sind die Werte Recall (Trefferquote, Gleichung 4.8) und Precision (Genauigkeit, Gleichung 4.7) aus dem Umfeld des Information Retrieval [vgl. MRS08, Seite 155]. Hierbei benötigt man die Menge der „gefundenen“ Dokumente (welche als Ergebnis zurückgegeben werden) sowie die Menge der tatsächlich relevanten Treffer (Artikel des Focalpoints). Der Recall drückt hierbei (das Verhältnis) aus, wie viele relevante Ergebnisse, aus der Gesamtmenge der relevanten Dokumente, ausgewählt wurden. Die Precision hingegen ist ein Verhältnismaß, welches aussagt, wie viele der ausgewählten Ergebnisse relevant sind.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{4.7}$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \tag{4.8}$$

Der Zusammenhang der Werte kann durch die Nutzung von „gefunden“ Dokumenten sowie der Klassifizierung in „true/false positives“ sehr anschaulich dargestellt werden. Anhand von

Tabelle 4.3.2.1 und den Formeln lassen sich Precision und Recall leicht berechnen und mit den Begrifflichkeiten „true bzw. false negatives“ verbinden.

	Relevant	Nicht relevant
Retrieved	true positives	false positives
Not Retrieved	false negatives	true negatives

Darüber hinaus kann im Zweifelsfall zur Bewertung von Ergebnissen auch auf das Wissen eines Domänenexperten zurückgegriffen werden. Dieses ermöglicht es zu prüfen bzw. zu verstehen, warum ein Resultat anders ausfällt als erhofft.

4.3.2.2. Precision-Recall-Kombination

Die Werte Precision und Recall alleine reichen dem Autor nicht zur Bewertung, da diese zunächst nur eine begrenzte Aussagekraft haben. Dieses lässt sich allerdings recht leicht ändern, indem man die Werte in einem **Precision-Recall-Diagramm** [siehe MRS08, Seite 158] kombiniert. Ein solches zeigt die Abbildung 4.2, hierbei zeigt die x-Achse den Recall, während auf der y-Achse die Precision aufgetragen ist. Im Verlauf wird der Recall schrittweise gesteigert, indem mehr Dokumente in die Berechnung mit einbezogen werden. Dieses Verfahren ermöglicht es, beide Werte auf einen Blick zu erfassen und daraus Schlussfolgerungen zu ziehen.

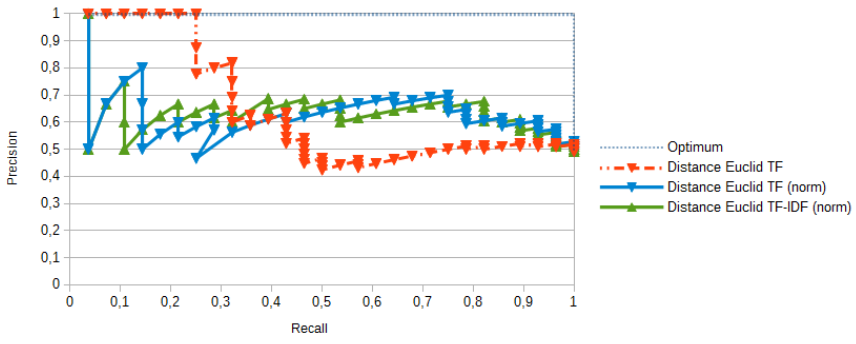


Abbildung 4.2.: Precision-Recall Diagramm

Zu beachten ist, dass der aufgetragene Graph nicht funktional ist, da einem x-Wert (Recall) durchaus mehrere y-Werte (Precision) zugeordnet sein können. Dieses ergibt sich aus der Definition des Recalls, welcher nur steigt, sofern ein weiteres relevantes Dokument in seine Berechnung einfließt.

Grundsätzliche Eigenschaften von Precision und Recall lassen sich besonders gut aus den Diagrammen ablesen. So wird recht schnell deutlich, dass der Recall sich monoton steigend verhält, während die Precision im Optimalfall monoton fallend ist. Vor allem bei der Precision ist diese Eigenschaft enorm hilfreich, da so „frühe Treffer“ in der Distanzfunktion automatisch besser bewertet werden als spätere. Dieses lässt sich besonders gut verstehen, wenn man die Zielsetzung bedenkt: Im Optimalfall berechnet die Distanzfunktion ein Ergebnis bei dem die Focalpoint-Artikel die beste Entfernung zum Leitartikel aufweisen. Diese stehen daher in einem sortierten Ergebnis ganz oben. Die zufällig gewählten Artikel (welche keine Ähnlichkeit zum Leitartikel aufweisen und deshalb keine „Treffer“ sind) bilden die untere Hälfte der Ergebnisse. Bedenkt man nun das Verhalten der Precision, verschlechtert ein „frühes und falsches“ Ergebnis (ein zufälliger Artikel weit oben in der Ergebnisliste) das Gesamtergebnis (z.B. den Durchschnitt, bezogen auf die Precision). Ein richtiges Ergebnis hingegen erhält die Precision und erhöht den Recall.

Aufgrund der oben beschriebenen Eigenschaft gehen alle Distanzen in die Bewertung mit ein (dieses schließt daher auch die zufälligen Artikel ein). Deshalb kommt es selbst im Optimalfall (alle „richtigen“ Ergebnisse bilden die obere Hälfte der sortierten Distanzen) zu einem Abfall der Precision. Ein beispielhaftes Ergebnis der Kombination von Precision und Recall ist in Grafik 4.2 zu sehen. Hierbei wurden drei verschiedene Varianten der euklidischen Distanzfunktion (siehe: Abschnitt 4.2) benutzt, zusätzlich wurde das optimale Ergebnis mit aufgetragen.

Zusätzlich zur visuellen Verdeutlichung der Kombination von Precision und Recall durch Diagramme, nutzt der beschriebene Aufbau auch die **durchschnittliche** (arithmetisches Mittel) **Präzision** als Indikator für die Qualität eines Ergebnisses. Je höher die durchschnittliche Precision der Dokumente bei steigendem Recall ist, umso besser ist das Gesamtergebnis zu bewerten.

4.3.2.3. Erläuterungen zur Auswertung von Ergebnissen

Um detaillierte Analysen und Interpretationen von Ergebnissen anzugehen, sind weitere Informationen von Nöten. Diese sollen nachfolgend dargestellt werden und beinhalten sowohl die Berechnung von Precision-Recall-Diagrammen aus Distanzen als auch generelle Informationen zu den Resultaten welche in Abschnitt B aufgeführt sind.

Herleitung der Precision-Recall-Diagramme

Wie bereits bekannt, wird zur Auswertung von Ergebnissen die Kombination der Werte Precision und Recall genutzt. Diese beruhen auf den Distanzen, welche zunächst in Tabellen festgehalten werden. Diese Werte können in Precision-Recall Diagramme (vgl. Abschnitt 4.3.2.3) übertragen werden. Diese Übertragung ist recht einfach und läuft folgendermaßen ab:

1. Zunächst werden die einzelnen Werte pro Distanzfunktion aufsteigend sortiert.

2. Mithilfe der Formeln 4.7 und 4.8 lässt sich nun für jeden Distanzwert die Precision und der Recall berechnen. Hierbei werden für jeden Wert seine Vorgänger mit in die Berechnung einbezogen.

    - Hierzu lässt sich der Wert für die relevanten und erhaltenen Werte abzählen. Dokumente, die im Rahmen des Ziels als korrekt angesehen werden lassen, sich anhand des Dateinamens identifizieren: Dieser enthält jeweils das Erscheinungsdatum (Format: YYYY-MM-DD), gefolgt vom Autor und der Sprache. Die zufälligen (nicht im Focalpoint enthaltenen) Artikel enthalten zusätzlich vor dem Datum noch eine zufällige Zahl.

3. Die berechneten Werte werden im letzten Schritt in ein Precision-Recall-Diagramm eingetragen. Hierbei ist der Wertebereich für beide Zahlen auf das Intervall $[0,1]$ begrenzt.

Generelles

Die Auswertungen der Precision-Recall Diagramme erfordert einige zusätzliche nützliche Informationen, welche dieser Abschnitt liefert. Diese umfassen Standardwerte der Precision, die einen Vergleich der Ergebnisse mit zufälligen sowie optimalen Ergebnissen erlaubt. Diese Daten ermöglichen einen besseren Überblick über die Qualität der Ergebnisse.

So ist zu beachten, dass die angegebene durchschnittliche Precision als prozentualer Anteil des Optimums zu verstehen ist. Hierbei erreicht die optimale Precision einen Wert von 0,84. Dieses ist dadurch zu erklären, dass die Negativbeispiele, welche nicht aus der Sammlung des Focalpoints stammen, auf jeden Fall in das Ergebnis mit einfließen und damit die Precision senken.

Von Interesse ist auch die durchschnittliche Precision im Fall einer zufälligen Anordnung der Dokumente im Ergebnisranking. Unter der vereinfachten Annahme, dass die „richtigen“ und „falschen“ Artikel abwechselnd vorkommen, ergibt sich eine durchschnittliche Präzision von ~50%. Unter dieser Annahme ist jede signifikante Verbesserung dieser Erfolgsrate ein Fortschritt.

Eine weitere Besonderheit der Ergebnisse ist die jeweils erste Distanz, welche in jeder Testreihe und Funktion 0 beträgt. Dieses ist dadurch zu erklären, dass der entsprechende Artikel der Leitartikel ist. Dieser dient wie erläutert als Vergleichsbasis und weist daher zu sich selber eine Entfernung von 0 auf.

4.3.3. Risikobetrachtung

Grundsätzlich müssen auch potenzielle Schwachstellen der gewählten Bewertungsmethodik untersucht werden. Trotz der positiven Bilanz der Versuche des Autors in [Sch15a] sollten die Ergebnisse hinsichtlich ihrer Stabilität auf anderen Dokumentensets untersucht werden. Darüber hinaus ist es ebenso denkbar, dass die gewählte Auswertungsmethodik (welche frühe Fehler stark in die Ergebnisse mit einfließen lässt) und Zusammensetzungen des Testkorpus (50% „richtige“ sowie 50% „falsche“ Dokumente) die Ergebnisse maßgeblich beeinflussen.

Ebenfalls ist zu bedenken, dass eine Erweiterung der Distanzfunktion das Ergebnis stark beeinflussen kann. So ist es denkbar, dass eine Erweiterung, z.B. um Kategorien, das Ergebnis verschlechtert, da diese inkompatibel zu anderen Ideen (wie der verstärkten Gewichtung von Textanteilen) ist.

5. Evaluierung von Lösungsansätzen

Auf Basis des zuvor erarbeiteten Wissens sollen nun konkrete Lösungsansätze zur Erfüllung der geschilderten Vision vorgestellt und untersucht werden. Hierbei wird zunächst auf bereits bekannte Ideen eingegangen. Daraufhin werden eigene Vorschläge analysiert. Das übergeordnete Ziel ist zu evaluieren in wieweit diverse Ansätze zu einer Verbesserung der Resultate führen bzw. herauszufinden, wo Stärken und Schwächen von Verfahren liegen.

5.1. Recommender Systems

Wie bereits erwähnt ist die Aufgabe des zu entwickelnden Systems, Vorschläge für seinen Benutzer zu generieren. Solche „Recommender Systems“ sind keineswegs eine Neuheit, sondern bereits länger Gegenstand der Forschung [Bob+13].

Hierbei wird im Groben zwischen den Ansätzen der kollaborativen sowie der inhaltsbasierten Filterung unterschieden. Während der kollaborative Ansatz darauf basiert, dass das Verhalten einer Menge von Nutzern analysiert wird, um Vorschläge für neue Nutzer zu generieren, basiert der Inhalts-getriebene Ansatz darauf, potenziell interessante Objekte möglichst gut zu beschreiben, um sie so, auf Basis eines Benutzerprofils, zu identifizieren. Darüber hinaus existieren hybride Ansätze, welche die Ideen beider Methoden vereinen.

Grundsätzlich lassen sich viele der Ideen von Systemen wie [CR12; Ban+12] usw. auf die dargestellte Problemstellung übertragen. Allerdings formulieren viele bekannte Ansätze oft sehr spezifische Probleme oder aber arbeiten auf völlig anderen Grundstrukturen als dieses Projekt. Daher sollen zunächst grundlegend bekannte Verfahren genutzt werden. Auf einige dieser Basisverfahren wird deshalb im Folgenden eingegangen. Diese (und im Weiteren vorgestellte Ideen) können daraufhin in zukünftigen Arbeiten verbessert werden.

5.2. Basisansätze des Mining: Klassifikation und Clustering

Zunächst sollen bereits bekannte Verfahren vorgestellt werden, welche einen Mehrwert für die Erfüllung des geschilderten Ziels haben können. Daher dient das folgende Kapitel als Einführung in die Thematik der Klassifizierung bzw. des Clustering, einem bekannten Teilgebiet des Datamining. Für diese Probleme existieren bereits diverse Algorithmen sowie Toolboxen, welche gute Ergebnisse liefern können. Diese Verfahren sind wohldefiniert und haben häufig eine Anzahl von Parametern, mit denen das jeweilige Ergebnis (maßgeblich) beeinflusst werden kann. Da diese Verfahren eine gute Grundlage für weitere Arbeiten bilden, wird es als sinnvoll erachtet in den nachfolgenden Abschnitten näher auf konkrete Verfahren und Frameworks einzugehen.

Ziel ist es, dem Leser einen Eindruck über die vorhandenen Verfahren zu vermitteln. Die geschilderten Methoden sind allerdings keine konkreten Lösungsvorschläge für das in der Vision geschilderte Problem. Sie sind vielmehr Basiswissen, welches das Verständnis des Lesers erweitern soll und besonders für weitere Arbeiten mit ähnlicher Zielsetzung von Nutzen ist.

Zunächst ist es wichtig die grundlegenden Gemeinsamkeiten und Unterschieden zwischen Klassifikation und Clustering zu verstehen, bevor eine detailliertere Einführung in die jeweilige Thematik folgt. Beide Herangehensweisen gruppieren Objekte aufgrund von einer oder mehreren Eigenschaften.

Die Grundannahme der Klassifikation ist hierbei, dass bereits „Label“ für einige Objekte bestehen. Konkret heißt dies, dass man einen Satz von Daten hat, sodass einzelne Artikel bereits (verschiedenen) Klassen (beispielsweise Themen) zugewiesen sind. Gesucht wird daher eine „Regel“, welche neue Objekte **den bestehenden** Klassen zuweist. Dieses erfordert zwingend die vorklassifizierten Daten (das Trainingsset) und ist daher eine Technik des überwachten Lernens. Die Trainingsdaten repräsentieren sozusagen einen Lehrer, welcher die korrekte Antwort bereits kennt.

Im Gegenteil hierzu braucht das Clustering keine Trainingsdaten. Es existieren deshalb keine bereits vorher festgelegten Gruppen. Diese innerhalb der Daten zu finden ist die Hauptaufgabe des Clusterings. Die Cluster werden hierbei durch die Ähnlichkeit von Objekten zueinander gebildet, so dass Objekte, welche näher beieinander liegen in einen gemeinsamen Cluster fallen.

Diese Technik stammt aus dem Gebiet des unüberwachten Lernens [vgl. CL14, Kapitel 5,6], da kein weiteres Vorwissen vorhanden ist.

5.2.1. Klassifikation

Eine der häufigsten Aufgaben des Datamining ist die Klassifikation (auch Kategorisierung) von Elementen. Hierbei geht es darum, Datensätze in vordefinierte Kategorien einzusortieren. Dieses kann im konkreten Fall das Einordnen von Zeitungsartikeln in Kategorien wie „Sport“, „Politik“ und „Wirtschaft“ sein.

Grundsätzlich existieren verschiedene Hauptansätze zur Bewältigung einer Klassifikationsaufgabe. Ein Ansatz basiert auf Expertenwissen, welches in den Kategorisierungsprozess aufgenommen wird (z.B. durch Systeme von Regeln). Ein anderer Ansatz beruht auf maschinellem Lernen. Bei diesem Verfahren soll das klassifizierende System die Informationen aus gegebenen Beispielen erlernen (ein stark vereinfachtes Beispiel hierfür könnte sein, dass Begriffe, welche mit Sport assoziiert werden, häufiger in Beispielartikeln der Kategorie Sport zu finden sind als in der Gruppe Wirtschaft).

Weiterhin kann man Klassifikationsverfahren daran unterscheiden, ob einem Dokument jeweils eine Kategorie oder eine Mischung aus mehreren Themengebieten zugewiesen wird [FS06, Seite 7ff.].

Im Folgenden werden einige Grundideen zu Klassifikationsverfahren aufgeführt. Diese sollen nicht mehr als einen groben Eindruck der vorhandenen Verfahren vermitteln.

5.2.1.1. K-Nearest Neighbor-Algorithmus

Ein sehr grundlegendes Klassifikationsverfahren ist der K-Nearest-Neighbor-Algorithmus. Für die Klassifizierung eines Objektes x werden hierbei die k nächsten „Nachbarn“ des Objektes betrachtet, diese stammen aus den vorgegebenen Trainingsdaten. Eine Mehrheitsentscheidung dieser Nachbarn legt die Kategorie des Objektes x fest. Im konkreten Fall von Dokumenten würde so der Feature-Vektor eines Dokuments mit seinen n Nachbarn verglichen. Sind dabei beispielsweise 3 von 5 Nachbarn der Kategorie „Sport“ zugewiesen, so behandelt das Dokument

selber höchstwahrscheinlich auch ein Thema der Kategorie „Sport“ und wird daher dieser zugewiesen.

Die Kernaufgabe bei diesem Verfahren besteht hierbei in der Bestimmung der nächsten Nachbarn. Dieses ist sehr einfach, da nur mithilfe einer Distanzfunktion (vgl. Abschnitt 4.2) Dokumente bzw. deren Feature-Vektoren verglichen werden müssen. Die Objekte mit der geringsten Distanz zum einzuordnenden Feature-Vektor sind die nächsten Nachbarn.

5.2.1.2. Decision Trees

Eine weitere häufig genutzte Methode zum Klassifizieren von Objekten sind sogenannte symbolische Klassifizierer. Diese haben unter anderem das Ziel „verständlicher“ für Menschen zu sein als alternative Ansätze. Ein Beispiel für solche Methoden sind Decision Trees (Entscheidungsbäume), hierbei werden die inneren Knoten des Baums mit Labeln versehen, welche aus dem Feature-Raum der zu klassifizierenden Objekte stammt. Die Blattknoten des so entstehenden Baumes bilden die Kategorien in welche Objekte eingeordnet werden.

Ein Beispiel für einen solchen Baum ist in Graphik 6.2 zu sehen. Dieser beschreibt die Klassifikation von Früchten.

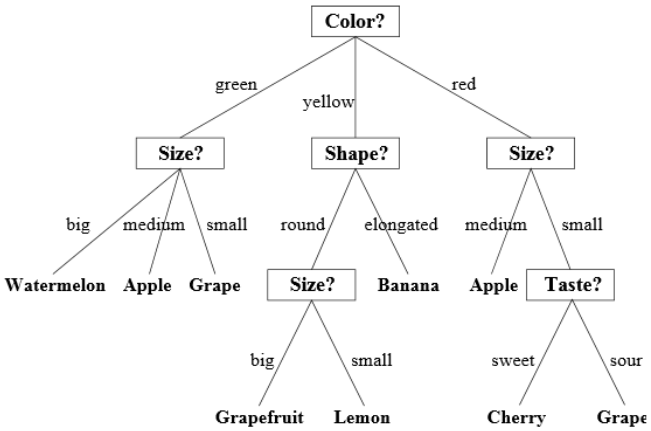


Abbildung 5.1.: Beispiel eines Entscheidungsbaums [Dec]

An diesem Beispiel lässt sich ebenfalls gut die Entstehung eines Decision Trees nachvollziehen. Dieser basiert auf dem Hintergrundwissen , welches zur zu treffenden Entscheidung verfügbar ist, sowie wichtigen Kernfragen. Entsprechend der ausgewählten Eigenschaften, anhand welcher unterschieden werden soll werden Zweige des Baums gebildet, deren Knoten jeweils die Werte der Eigenschaft darstellen. Konkret ist die erste Kernfrage der Klassifizierung die Farbe eines Objektes, welche laut dem vorhandenen Wissen drei Werte annehmen kann. Daher ist die Farbe der Wurzelknoten des Unterbaums und die Kindknoten bilden die Werte ab, welche eingenommen werden können. Die anhand der Entscheidung entstehenden Unterbäume ermöglichen nach dem abgebildeten Wissensstand weitere Entscheidungen entsprechend der Form bzw. Größe der einzuordnenden Elemente. So teilen sich die verschiedenen Teilbäume immer weiter anhand von definierten Kernfragen bis eine Wurzel erreicht wird. Im Kern wird bei diesem Vorgehen daher auf Basis von Klassifikationseigenschaften (den Fragen) und den entsprechend möglichen Werten (Antworten) der Baum aufgebaut.

5.2.1.3. Neuronale Netzwerke

Eine weitere Methode Texte zu klassifizieren, sind neuronale Netzwerke. Diese bestehen aus mehreren Schichten von künstlichen Neuronen und Verbindungen, welche mit Gewichten belegt sind. Die Grundfunktionalität besteht nun darin, dass am „Eingangslayer“ Daten angelegt werden deren Kategorisierung am Ausgang des Netzes bekannt ist. Das Netz muss nun „lernen“, die Gewichtungen der Verbindungen der Zwischenschichten anzupassen um den gewünschten Output zu erzielen. Ebenso kann der Schwellwert, welchen es zu erreichen gilt entsprechend angepasst werden.

Das auf diese Weise angelernte Netz kann daraufhin auf unbekannte Dokumente angewandt werden, um eine Klassifikation pro Dokument zu berechnen. Ein großer Nachteil dieser Methode ist allerdings die Komplexität, welche unter anderem beim Modellieren eines Netzes auftritt. Darüber hinaus ist ein Neuronales Netz nicht so verständlich wie beispielsweise ein Entscheidungsbaum.

5.2.1.4. Support Vector Maschines

Eine Support Vector Maschine (SVM) ist ein Klassifikationsverfahren, welches die Aufgabe hat seine Eingangsdaten möglichst klar voneinander abzugrenzen. Hierzu wird die Menge von Trainingsdaten, deren Klassifikation bekannt ist, als Basis verwendet. Die zu klassifizierenden Elemente werden jeweils durch einen Vektor beschrieben, der ihre Eigenschaften wiedergibt. Die gegebenen Vektoren bilden einen Raum, welcher mithilfe der SVM durch eine imaginäre Ebene in zwei Klassen geteilt werden soll. Ziel ist es nun die Ebene so zu legen, dass eine möglichst gute Trennung erzielt wird, sodass der Raum zwischen den Elementen möglichst „breit“ ist.

Grundsätzlich sind SVMs ein sehr generelles Mittel, welches auf viele Probleme angewendet werden kann. Allerdings ist bei ihrer Nutzung zu beachten, dass sie im Regelfall dafür gedacht sind, eine Menge von Objekten in genau zwei Klassen zu teilen. Soll eine Einteilung in mehr Kategorien passieren, muss dieses in der Modellierung des Problems berücksichtigt werden, so dass beispielsweise die Menge der Objekte erst in Kategorie eins und zwei geteilt wird und diese wiederum weiter geordnet werden.

5.2.1.5. Evaluierung in Bezug auf die Vision

Die kurz vorgestellten Verfahren aus dem Bereich der Klassifikation, zeigen ein großes allgemeines Potenzial. Allerdings ist die bisher verfolgte Vision des Autors (vgl. Abschnitt 3.2) darauf ausgelegt, nur auf Basis eines Leitartikels hilfreiche Ergebnisse für Fachexperten zu generieren. Dieses beinhaltet keine Trainingsdaten, welche für eine Klassifikationsaufgabe zwingend benötigt werden. Daher sind die dargelegten Methoden im konkreten Fall nicht anwendbar. Allerdings bilden die vorgestellten Verfahren wichtige Grundlagen, welche in dieser Arbeit vorgestellt werden müssen.

Allerdings ist anzumerken, dass es für weiterführende Ziele durchaus hilfreich sein kann, auf Klassifikationsverfahren zurückzugreifen. Im konkreten Anwendungsfall können hierbei die Focalpoints des Eurozine-Archivs als Ausgangspunkt für weitere Untersuchungen dienen. Diese haben allerdings nur einen recht kleinen Umfang, der ggf. nicht als Trainingsbasis ausreicht, dennoch stellen sie einen guten Einstiegspunkt dar.

5.2.2. Clustering

Als Clustering bezeichnet man einen unüberwachten Prozess, der Objekte in eine Anzahl von Gruppen (Cluster) einordnet. Hierbei liegen nicht wie im Fall von Klassifikation Trainingsdaten vor, so dass die Gruppierung in Cluster ohne weitere Informationen (außer den zu clusternden Daten selber) geschehen muss. Die grundlegende Hypothese die allen Clusterverfahren zu Grunde liegt, nennt sich Clusterhypothese. Diese besagt, dass relevante Dokumente, untereinander dazu tendieren ähnlicher zu sein als nicht relevante [FS06, Seite 82].

Ein abstraktes Beispiel für die Relevanz von Clustering-Verfahren ist das übergeordnete Thema dieser Arbeit. Große Dokumentarchive, die Redakteuren zugänglich gemacht werden müssen, aber nicht mehr händisch zu verarbeiten sind. Im konkreten Fall ist es denkbar, die Menge der Dokumente in verschiedene Gruppen einzuteilen, um so jeweils kleinere Mengen von Dokumenten zu erhalten, welche daraufhin weiter unterteilt und durchsucht werden können. Allerdings liegt ein Problem bei einer solchen Einteilung darin, dass ein Nutzer im Optimalfall bestimmen muss wie die Daten (entsprechend seiner Interessen) zu unterteilen sind. Dieses Wissen lässt sich allerdings nur schwer automatisieren.

Grundsätzlich ist das Ziel eines guten Clusterings, Elemente so zu gruppieren, dass ähnliche Objekte in einen Cluster fallen. Dieses erfordert im Allgemeinen eine Ähnlichkeits- bzw. Distanzfunktion (siehe: Abschnitt 4.2). Darüber hinaus ist das Clustering grundsätzlich ein Optimierungsproblem, welches versucht, die entstehenden Cluster möglichst gut voneinander abzugrenzen.

Betrachtet man verschiedene Clusteringverfahren, so kann man zwischen rein partitionierenden Verfahren, welche eine Menge von Elementen in n Gruppen teilen und Verfahren, welche diese Teilung hierarchisch durchführen, so das jeder Cluster selber in weitere Untercluster zerlegt wird (z.B. würde hierbei ein Cluster Politik weiter zerlegt in Innenpolitik und Außenpolitik) unterscheiden. Darüber hinaus kann ein Objekt entweder zu genau einem Cluster gehören, oder aber seine Zugehörigkeit verteilt sich prozentual auf mehrere Cluster.

Die nächsten Abschnitte stellen grundlegende Verfahren und Frameworks aus dem Bereich des Clusterings kurz vor.

5.2.2.1. K-Means

Der K-Means-Algorithmus teilt einen Korpus von Vektoren (im Beispiel von Dokumenten: Den Bag-Of-Words als Feature-Vektor) in eine Menge von Clustern. Hierbei benötigt der Algorithmus die Anzahl k, welche festlegt wie viele Cluster gebildet werden sollen, als Input. Die Vorgehensweise des Algorithmus ist hierbei wie folgt:

Initialisierung:

1. Wähle k zufällige Seeds (Means), diese bilden die initialen Clusterschwerpunkte.
2. Jeder andere Vektor wird dem Cluster mit dem naheliegensten Mean zugewiesen.

Iteration:

1. Neuberechnung der Clusterschwerpunkte unter Nutzung der Distanzfunktion.
2. Neuzuordnung aller Vektoren zu ihren (neu berechneten) naheliegensten Schwerpunkten.

Abbruchbedingung:
(Nahezu) keine weiteren Änderungen der Schwerpunkte.

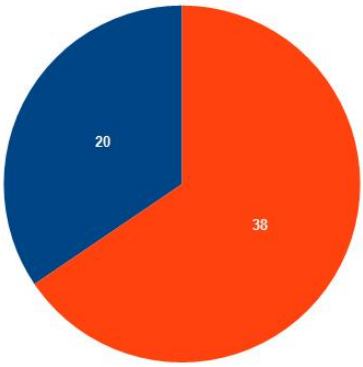
Algorithmus 1: K-Means-Algorithmus

Evaluierung
Der K-Means-Algorithmus ist aufgrund seiner Einfachheit und Effizienz sehr beliebt, allerdings ist er sehr anfällig für Änderungen der initial gewählten Seeds. Sobald diese suboptimal gewählt werden, verschlechtern sich die Ergebnisse ebenso. Grundsätzlich existieren eine Reihe von Erweiterungen und Anpassungen [FS06, Seite 86 ff.] zur Verbesserung des grundlegenden Algorithmus.

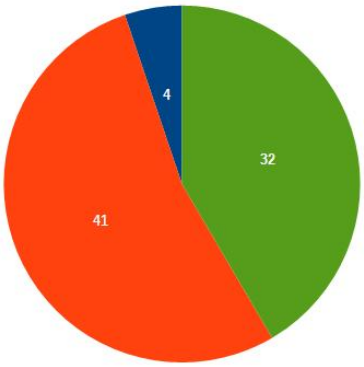
Zur Untersuchung der Nützlichkeit des vorgestellten Algorithmus im Kontext der geschilderten Vision wurden einige Experimente durchgeführt. Konkret wurden verschiedene Zusammenstellungen der in Abschnitt 4.3.1 geschilderten Korpi mithilfe des K-Means-Algorithmus geclustered. Hierbei wurden verschiedene Parameter verändert. Diese umfassen den jeweils genutzten Korpus, die Anzahl der zu generierenden Cluster sowie die genutzte Distanzfunktion.

Die folgenden Graphiken visualisieren die Ergebnisse. Hierbei wird jeweils dargestellt wie viele Dokumente einem Cluster zugewiesen worden.

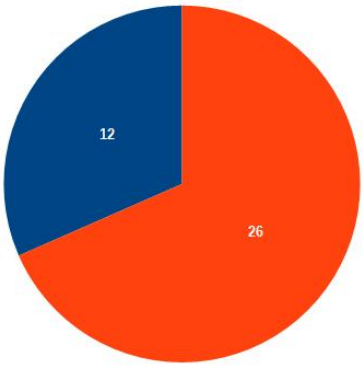
Abbildung 5.2.: Ergebnisübersicht des K-Means-Clustering



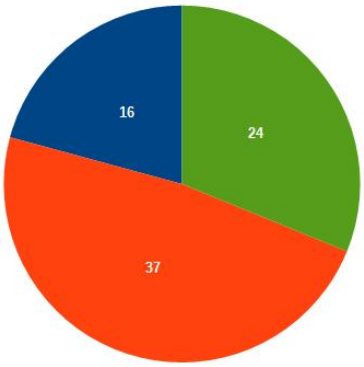
(a) Demokratie ∪ Zufällig
Gesamtanzahl der Dokumente: 58
Distanz: Euklid und Kosinus



(c) Bologna (reduziert) ∪ Demokratie ∪ Zufällig
Gesamtanzahl der Dokumente: 77
Distanz: Euklid



(b) Bologna (reduziert) ∪ Zufällig
Gesamtanzahl der Dokumente: 38
Distanz: Euklid und Kosinus



(d) Bologna (reduziert) ∪ Demokratie ∪ Zufällig
Gesamtanzahl der Dokumente:77
Distanz: Kosinus

Das (vereinfachte) Ergebnis des Versuchs ist in Graphik 5.2 zu sehen. Dieses ist als wenig erfolgreich zu bewerten. Ein optimales Clustering hätte jeweils die Einzelbestandteile des jeweiligen Korpus voneinander separiert, so dass a) und b) jeweils 2 Cluster mit jeweils 29 bzw. 19 Artikeln aufgewiesen hätten. Die Ergebnisse zeigen allerdings ein Verhältnis von 20:38 bzw. 12:26 Artikeln. Daher wurde offensichtlich keine gute Trennung erreicht, da mindestens 9 bzw. 7 Artikel falsch zugeordnet wurden. Diese erste Analyse ist allerdings sehr oberflächlich, da nur die Gesamtanzahl der zugeordneten Dokumente pro Cluster bewertet wird. Eine detaillierte Analyse kann allerdings keinesfalls bessere Ergebnisse liefern. Es könnte zwar analysiert

werden, welcher Cluster, welche Dokumente beherbergt, allerdings ändert dieses nichts an der grundlegenden Tatsache, dass die Mehrzahl der Dokumente falsch zugeordnet ist. Ein weiterer bemerkenswerter Punkt ist, dass das Clustering für die Unterversuche a) und b) mit der Euklidischen Distanz einerseits und mit der Kosinus-Distanz andererseits, das gleiche Ergebnis liefern.

Betrachtet man die Experimente c) und d), so wäre eine Optimalverteilung auf die Cluster ein Verhältnis von 19:29:29. Auch dieses Ziel wird verfehlt. Dieses wird besonders im Teilversuch c) deutlich, da nur 4 Dokumente dem Cluster 0 zugeordnet wurden. Derselbe Versuch unter Nutzung der Kosinus Distanz zeigt hier wesentlich bessere Ergebnisse, auch wenn dennoch eine recht große Diskrepanz zum Optimalergebnis besteht.

Zusammenfassend liefert das K-Means-Clustering im konkreten Fall mit wenig Aufwand schnelle Ergebnisse. Diese sind allerdings schon unter recht einfachen Betrachtungen unbrauchbar, so dass weitere Alternativen untersucht werden müssen.

5.2.2.2. EM-basiert

Die Grundfunktion des EM-Verfahrens (Expectation Maximization) ist ein iteratives Vorgehen, welches das Ziel hat, eine in den Daten vorhandene Wahrscheinlichkeitsverteilung zu bestimmen. Diese definiert die Zugehörigkeit der einzelnen Datenelemente (wie z.B. Dokumente eines Korpus) zu verschiedenen Modellen (Im konkreten Fall: Cluster).

Die Grundannahme hierbei ist, dass die Dokumente sowohl Eigenschaften aufweisen, die gemessen werden können (wie Häufigkeit von Wörtern), als auch latente Eigenschaften, welche nicht direkt ersichtlich sind (wie beispielsweise Korrelationen zwischen der Häufigkeitsverteilung von Wörtern über den Korpus).

Der Algorithmus führt in jeder Iteration einen Expectation- sowie einen Maximization-Schritt durch. Hierbei ordnet der E-Schritt die Daten neuen Clustern zu, während der M-Schritt die Parameter des Modells verbessert. Diese werden daraufhin in der nächsten Iteration

verwendet. Sobald keine wesentliche Verbesserung mehr stattfindet, hat der Algorithmus sein Ziel erreicht.

Initialisierung:
Wähle zufällige Daten für alle Modelle.

Iteration:

E-Schritt:
Weise die Daten dem Model zu, zu dem sie am besten passen.

M-Schritt:
Aktualisiere die Parameter des Models, nur mit den E-Schritt zugewiesenen Daten.

Abbruchbedingung:
Es finden (nahezu) keine weiteren Änderungen der Parameter mehr statt.

Algorithmus 2: EM-Algorithmus

Fazit
Der beschriebene Algorithmus ist ein häufig genutztes Standardverfahren, auf welchem diverse weitere Verfahren aufsetzen. Allerdings ist eine der größten Nachteile der oft hohe Ressourcenbedarf, welcher durch die langsame Konvergenz des Verfahrens entsteht. Dieses äußert sich darin, dass die Abbruchbedingung des Algorithmus erst nach geraumen Durchläufen erreicht wird. Praktisch war es dem Autor daher nicht möglich ein Clustering mithilfe des EM-Verfahrens innerhalb von RapidMiner durchzuführen. Dieses gelang auch nicht nach signifikanter Einschränkung der zu clusternden Daten, der maximalen Anzahl von Durchläufen und der Verschlechterung der zu erreichenden Schwelle der Abbruchbedingung.

5.2.2.3. Latente Dirichlet Allocation

Die latente Dirichlet Allocation (LDA) ist ein generatives Wahrscheinlichkeitsmodell für einen Korpus (z.B. eine Sammlung von Textdokumenten). Die grundlegende Idee ist, dass jedes Dokument aus einer Anzahl von Topics (auch Themen genannt) besteht (welche nach außen nicht direkt sichtbar (d.h. latent) sind). Weiterhin ist jede Topic eine Mischung von Wörtern anzusehen, diese bilden das Thema [vgl. BNJ03].

In diesem Modell sollen daher die verschiedenen Wörter (und letztendlich auch die entsprechenden Dokumente) eines Korpus mit möglichst hoher Wahrscheinlichkeit einem Thema zugeordnet werden. Die nun zugeordneten Themen bilden die späteren Cluster. Basierend auf der Zuordnung von Wörtern und Dokumenten zu Themen (und damit Clustern) lässt sich die Topiczusammensetzung eines Dokumentes bestimmen (beispielsweise: 20% Topic A, 70% Topic B sowie 10% Topic C). Darüber hinaus kann man anhand der Keywords pro Cluster (z.B. die am häufigsten benutzten Worte) Schlagwörter ermitteln, welche approximiert den Inhalt eines Topics (bzw. Clusters) wiedergegeben.

Grundlegend basiert die Themenzuordnung von LDA auf einem Lernverfahren, welches auf bayesscher Statistik fußt, und den Methoden des unüberwachten Lernens zuzuordnen ist. Weiterhin ist die Grundidee ein Bag-Of-Words Ansatz, welcher ein Dokument nur als Ansammlung von Wörtern, allerdings ohne Semantik, ansieht.

Fazit

Da das LDA-Verfahren mit dem EM-Verfahren verwandt ist, leidet es unter denselben Nachteilen, welche unter anderem einen extremen Rechenaufwand umfassen. Ein für den Autor hinderlicher Fakt ist darüber hinaus, dass keine Implementierung des Algorithmus für das RapidMiner Framework vorhanden ist. Allerdings ist zu erwarten, dass diese ähnliche Probleme verursacht wie die vorhandene Umsetzung des EM-Algorithmus. Aus diesem Grund kann der Author dieser Arbeit nur auf die allgemeine Nützlichkeit des LDA-Verfahrens hinweisen, allerdings keine konkreten Untersuchungen anbieten.

Tiefergehende Informationen und Experimente (auf einem Datensatz von Kurznachrichten) mit dem umrissenen Verfahren wurden vom Autor dieser Arbeit in [Sch14] beschrieben.

5.2.2.4. Carrot^2

Carrot^2 ([Car]) ist ein Open-Source-Framework zur Clusterung von Suchresultaten. Hierbei teilt es Textkorpi in thematische Kategorien. Darüber hinaus bietet das Framework die Möglichkeit, direkt auf APIs von Google, Bing sowie Lucene und diverse anderen zuzugreifen.

Das Carrot^2-Framework unterstützt verschiedene Clusteralgorithmen. Diese sind im Nachfolgenden aufgeführt.

Algorithmus	Author	Hierarchisches Clustering	Paper
Lingo	Stanislaw Osinski	×	[OW05; Osi04]
STC	Oren Zamir	×	[SW03]
Lingo3G	Stanisław Osinski	✓	× (kommerziell)

Tabelle 5.1.: Clusteralgorithmen im Carrot^2-Framework

Nutzen

Das vorgestellte Framework erfüllt zwar nicht direkt die Anforderungen der dargelegten Vision, allerdings ist es ebenfalls eine interessante Basis für weitere Arbeiten, welche der Autor nicht verschweigen möchte.

5.2.2.5. Overview-Project

Ein weiteres interessantes Projekt, welches an dieser Stelle erwähnt werden soll, ist das Overview-Project [siehe Oveb; Bre+14]). Dieses von der „The Associated Press“ ins Leben gerufene Vorhaben, ist ein Open-Source-Framework, welches speziell für Journalisten entworfen wurde, um Inhalte in großen Dokumentenmengen aufzudecken. Hierbei werden die Daten in Themenbereiche geordnet und graphisch aufbereitet.
Das Kernvorgehen des von Overview genutzten Vorgehens ist ein hierarchisches Clustering. Eine Menge von Dokumenten wird in Untermengen aufgeteilt, welche wiederum geteilt werden usw. Diese Mengen werden jeweils mit möglichst beschreibenden Labeln versehen, um dem User einen leichteren Überblick zu ermöglichen.

Nutzen

Ebenso wie das Carrot^2-Framework ist das Overview-Project nicht direkt in Bezug auf die Vision von Bedeutung, stellt aber dennoch einen interessanten Anhaltspunkt für weitere Forschungen dar. Dieses zeigt sich auch dadurch, dass das Projekt speziell für die Vorsortierung von Textdokumenten entworfen wurde und mithilfe von journalistischen Fachwissen entstand. Darüber hinaus wurde das Verfahren in diversen Kontexten genutzt, Beispielprojekte hierfür sind auf der Projektseite aufgelistet ([Ovea]).

5.2.2.6. Zusammenfassung

Das Vorstellen der Verfahren diente zum einen dazu, einen Überblick in vorhandenen Methoden zu geben, zum anderen zeigt sich allerdings auch, dass die Verfahren nicht der gestellten Aufgabe direkt entsprechen. Für die Nutzung von Klassifikationsverfahren fehlt zunächst das

nötige Trainingsset. Clustering-Algorithmen sind zwar relativ nah an der Vision, liefern allerdings letztendlich nur Cluster ohne direkten Bezug (Distanz) zum in der Vision vorgegebenen Leitartikel.
Darüber hinaus bilden die diversen Algorithmen und Frameworks allerdings eine wichtige Grundlage für ähnliche Aufgaben. Mit diesem vorhanden Basiswissen sollen im Folgenden weitere Ansätze untersucht werden, welche zur Erfüllung der Vision beitragen können und die bereits vorhandenen Verfahren ergänzen können. Hierbei soll zunächst auf fachliches Wissen von Domänenexperten zurückgegriffen werden.

5.3. Gewichtung von diversen Artikelabschnitten

5.3.1. Ursprünglicher Ansatz

Der nachfolgend erläuterte Versuch stellt eine Erweiterung der in Vorversuchen [siehe: Sch15a] durchgeführten Experimente dar. Diese basieren auf der fachlich getriebenen Hypothese, dass verschiedene Textanteile (wie Abstract, Überschriften und Text) eine unterschiedlich wichtige Bedeutung für die Distanz zweier Artikel haben. Diese These begründet sich in der Idee, dass diese Bestandteile einen großen Teil der Gesamtinformationen des Dokumentes in sich tragen bzw. diesen hinreichend gut zusammenfassen. Aus diesem Grund ist die Annahme, dass eine verstärkte Berücksichtigung dieser Aspekte ein guter Anhaltspunkt für die Ähnlichkeit von Artikeln ist.
Diese auf Basis der Textanteile durchgeführte Gewichtung ist in Formel 5.1 formal beschrieben. Hierzu werden Parameter eingeführt, welche jeweils angeben wie „wichtig“ der Abstract, Titel, die Unterüberschriften sowie der Rest des Textes sind. In der Berechnung wird die **Worthäufigkeit** tf für **Wort** w, welches im Abschnitt mit dem **Parameter** x_n vorkommt, mit x_n multipliziert (aufgrund der These, dass das Wort für den Artikel ausschlaggebender ist als andere Wörter). Die **Gesamthäufigkeit** $tf_{ges}(w)$ eines Wortes w im Artikel ergibt sich daher als Summe über die gewichteten Vorkommen pro **Abschnitt** tf_{sec_n}:

$$tf_{ges}(w) = (x_1 * tf_{sec_1}(w)) + (x_2 * tf_{sec_2}(w)) + \cdots + (x_n * tf_{sec_n}(w)) \tag{5.1}$$

5.3.2. Frühere Resultate

Die Ergebnisse des Projektes waren grundsätzlich positiv zu bewerten. Zusammengefasst deuteten sie darauf hin, dass eine verstärkte Beachtung der Überschriften das Ergebnis signifikant verbessern kann. Allerdings zeigte sich auch dass eine stärkere Gewichtung des Abstracts

sowie (abgeschwächt) der Überschriften wenig hilfreich ist. Dieses Vorgehen zeigte Tendenzen, die Ergebnisse bei überhöhten Parametern zu verschlechtern.
Näheres zur Durchführung und den Ergebnissen dieser Versuche ist dem entsprechenden Bericht [Sch15a] zu entnehmen.

5.3.3. Überprüfung der Ergebnisse durch Hinzunahme weiterer Focalpoints

Die im Folgenden dargestellten Ergebnisse stellen die Fortsetzung der in Vorstudien durchgeführten Versuche dar. Die Versuchsreihen sollen vor allem dazu dienen, die Ergebnisse des Projekts auf anderen Daten zu überprüfen, um auszuschließen, dass die Resultate durch zufällig gute Testparameter verfälscht wurden. Hierzu wurden weitere Ausschnitte des Eurozinearchivs als Testkorpi benutzt. Diese unterscheiden sie sich durch die Wahl der genutzten Focalpoints und durch die Gesamtgröße.
Die Versuchsdurchführung entspricht der in Vorversuchen geschilderten Weise ([vgl. Sch15a]) und wird ebenfalls wie in Abschnitt 4.3.2.3 beschrieben bewertet. Die Auswahl der Testreihen entspricht hierbei den jeweils erfolgversprechendsten Parametern des Projektes 2. Um im Weiteren eine einheitliche Begriffsgebung zu nutzen, wird der Begriff „zusammenfassende Anteile“ für die Bestandteile „Abstract“, „Titel/Überschrift“ sowie „Subheadings“ benutzt, da diese in gewissem Maße Abschnitte des Artikels in Kurzform wiedergeben. Das Gegenteil hierzu ist der restliche Text des Artikels, welcher den Großteil der Wörter enthält.

5.3.4. Ergebnisse

Die Ergebnisse der beschriebenen Testreihen sind ab Seite 82 zu finden. Dort sind für verschiedene Parameter jeweils die Ergebnistabellen für den Korpus A, B und C aufgelistet.
Grundsätzlich ist festzustellen, dass die positiven Ergebnisse der Vorstudien sich nicht reproduzieren lassen.
Dieses ist unter anderem daran zu erkennen, dass die Ergebnisparameter, welche zuvor die durchschnittliche Precision optimiert haben (Hervorhebung von Überschriften ; vgl. Seite 106), nicht den selben Erfolg auf einem anderen Korpus (B) liefern (vgl. Seite 109). Die erfolgreichsten Parameter für diesen neuen Korpus sind hingegen jene, welche den Textanteil der Artikel stärker gewichten als die zusammenfassenden Anteile (vgl. Seite 101). Dieses deutet daraufhin, dass die alleinige Gewichtung von Textanteilen keine grundsätzliche Verbesserung der Distanzfunktion bewirkt, da das Verfahren nicht auf jeden Korpus verallgemeinerbar ist.
Anzumerken ist, dass der Korpus B sich nicht nur durch den Focalpoint von der Dokumentzusammenstellung der Vorstudien unterscheidet. Ein weiterer wichtiger Faktor, welcher Einfluss

auf das Ergebnis haben kann, ist das Verhältnis von „richtigen“ zu „falschen“ (dem Focalpoint angehörigen bzw. nicht angehörigen) Artikeln innerhalb des Testkorpus. Dieses Verhältnis entspricht in den Testdaten der Vorstudien (A) ca. 50:50, während es im Datenset B ca. 20 zu 30 Artikel sind. Da dieses, wie erwähnt, großen Einfluss haben kann, wurde eine weitere Testreihe mit einem weiteren Datenset durchgeführt. Dieses basiert auf demselben Focalpoint wie der normale Korpus „Masterthesis“, reduziert allerdings die Anzahl der zufälligen Artikel, so das ein Verhältnis von 1:1 von „richtigen“ zu „falschen“ Artikeln entsteht.
Betrachtet man die Ergebnisse auf dem beschriebenen neuen Korpus, so zeigt sich, dass sie durchgehend besser sind, als jene auf dem nicht reduzierten Korpus (vgl. Seite 112 mit 109). Trotzdem sind auch im reduzierten Testset jene die besten Resultate, deren Parameter den Textanteil besonders stark hervorheben (siehe Seite 104). Festzuhalten ist daher, dass sowohl der genutzte Korpus, als auch dessen Verhältnis, von als richtig zu bewertenden zu falschen Artikeln, maßgeblichen Einfluss auf das Ergebnis hat. Aufgrund dieser Tatsache, ist das Verfahren momentan nicht auf beliebige Korpi verallgemeinerbar und deshalb nicht zwingend zielführend.
Ein positiver Aspekt der Vergleichsuntersuchung ist, dass die TF-IDF basierte Distanz grundsätzlich die stabilsten Resultate liefert. Dieses zeigt sich in den diversen Diagrammen (bspw. Seite 101 und 112), in denen die anderen Maße recht schlecht abschneiden und sich ggf. stark unterscheiden. Diese Stabilität und auch Qualität des Distanzmaßes, scheint Korpus übergreifend zu gelten.

5.4. Reduktion des Feature-Vektors

Nachdem die beschriebene Testreihe, mit dem Thema Gewichtung, zum einen fachlich getrieben ist und zum anderen einen groß ausgelegten Feature-Vektor („Bag-Of-Words“) hat, soll eine weitere Versuchsreihe die gewonnen Erkenntnisse nutzen. In der Retroperspektive ergibt sich aus den vorherigen Suchergebnissen, dass der genutzte Feature-Raum deutlich zu groß ist, da er alle Terme in allen Dokumenten mit einbezieht. Eine logische Weiterentwicklung ist deshalb die Reduktion des genutzten Features. Dieses Problem soll im Weiteren angegangen werden. Hierbei wird vor allem auf potenzielle Probleme eingegangen sowie Lösungsvorschläge erarbeitet.

5.4.1. Problemstellung und Lösungsansätze

Die in Abschnitt 5.3 geschilderten Versuche zeigen zwar Lösungsideen auf, liefern allerdings kein zufriedenstellendes Ergebnis. Dieses Phänomen zeigt sich ebenfalls bei ersten Versuchen

eines Clusterings der Artikel in eine Anzahl von Kategorien, welches einen weiteren Ansatz zur automatisierten Dossiererstellung bilden sollte (siehe: Abschnitt 5.2).
Die Ursachen für das Scheitern der Versuche können nicht genau bestimmt werden, da die Erstellung von Pressedossiers kein gut durchschaubarer Prozess ist. Allerdings basiert die gesamte Arbeit auf dem „Bag-Of-Words"-Ansatz, welcher alle in den Artikeln verwendeten Wörter sowie deren Häufigkeit enthält. Dieses Konstrukt bildet den Feature-Vektor auf welchem Distanzen gebildet werden. Dieser Vektor ist entsprechend groß, so hat der Feature-Vektor des Testdokumentes in Abschnitt A selbst nach der Vorverarbeitung (siehe: Abschnitt 4.1) noch eine Dimension von ca. 1100. Hierbei summieren sich die Häufigkeiten der Wörter auf ca. 7300.
Betrachtet man die bisher genutzte Distanzfunktion (Euklid, siehe: Abschnitt 4.2) so ist erkennbar, dass diese auf dem quadrierten Abstand der Komponenten des Featurvektors arbeitet. Diese Distanz wächst daher besonders stark mit jedem Wort das Dokument A enthält, Dokument B allerdings nicht. Dieses Verhalten ist durchaus wünschenswert, führt aber dazu, dass die Distanz beim Vergleich längerer Dokumente (und damit i.A. größerer Feature-Vektoren) sehr schnell ansteigt. Dieser Einfluss wird bisher nicht berücksichtigt. Zwar tragen Maße wie TF-IDF dem Fakt Rechnung, dass gewisse Wörter im Gesamtkorpus häufig vorkommen (und deshalb weniger relevant in einem Artikel sind), allerdings löst dieses nicht das Problem von Wörtern, welche zwar nur in wenigen Dokumenten vorkommen, ggf. aber keine tragende Semantik vermitteln.
Grundsätzlich zeigt sich daher, dass die Größe des Feature-Vektors einen negativen Einfluss auf die Distanz haben kann. Es wäre deshalb potenziell sinnvoll, die Dimension des Feature-Vektors zu verringern, ohne allerdings dabei zu starken Einfluss auf die Semantik der untersuchten Artikel zu nehmen. Um das Problem eines zu großen Feature-Vektors anzugehen, muss im speziellen Fall eine Möglichkeit gefunden werden, die Wortanzahl (Dimension) im Bag-Of-Words zu verringern.
Hierfür sind verschiedene Ansätze denkbar, diese umfassen:

- Reduktion durch Auswahl der Features: Auf die nach TF(-IDF) n (z.B. 10) „besten" Terme beschränken.
- Mapping von ähnlichen Wörtern auf semantischer Ebene.

Der erste Ansatz ist sehr effektiv, da durch die Auswahl von Kernfeatures eine immense Verkleinerung des Feature-Vektors stattfindet. Allerdings ist er auch recht simpel und berücksichtigt daher keine semantischen Beziehungen zwischen verschiedenen Wörtern. So ist es denkbar, dass ein Dokument A das Thema „Demokratie" behandelt. Neben offensichtlich hierfür wich-

tigen Worten wie „Demokratie“ werden allerdings auch Begriffe wie „Politik“, „Regierung“ oder „Republik“ genutzt. Diese Worte werden einzeln im Feature-Vektor vermerkt und in die Berechnung einer Distanz zu einem anderen Dokument mit einbezogen. Bei diesem Verfahren geht allerdings die semantische Nähe der genannten Begriffe verloren, so dass ein anderes Dokument B die selben Terme enthalten muss, um ein hohes Ähnlichkeitsmaß zu Dokument A zu erreichen.
Dieser Verlust an Informationen soll im Weiteren mithilfe des zweiten Ansatzes (semantisches Mapping) näher untersucht werden, um eine Lösung anzubieten. Grundsätzlich gilt es hierbei eine Möglichkeit zu finden, semantische Ähnlichkeiten zu erkennen, um so Terme auf semantischer Ebene auf *einen* Begriff abzubilden.

5.4.2. Domänenwissen und dessen Probleme

Um das beschriebene Ziel, Terme auf semantische Ebene zu vereinheitlichen, zu erreichen, stellt sich zunächst die Frage, wie man das hierzu notwendige Wissen erlangen kann. Diese Informationen basieren darauf, dass man die semantische Ebene eines Textes versteht - eine Aufgabe, welche Domänenwissen erfordert. Dieses ist generell nur mit einigem Aufwand zu erlangen, kann aber grundsätzlich in diversen Arten auftreten [vgl. FS06, Seite 41 ff.]. So kann es konkret bei Experten erfragt werden, oder es liegt ggf. schon in verallgemeinerter Form vor. Dieses Themengebiet ist entsprechend groß und kann in dieser Arbeit daher nur angerissen werden, um ein Verständnis zu vermitteln. Hierzu sollen einige Begriffe, welche im Folgenden eine Rolle spielen, kurz eingeführt und erläutert werden.

Taxonomie

Eine Taxonomie ist ein Wörterbuch, welches eine hierarchische Struktur aufweist. Die Terme einer Taxonomie sind untereinander, wenn möglich, durch Relationen wie „Ober- zu Unterbegriff“ (bzw. andersherum / beidseitig) verbunden. Aus diesem Grund können Taxonomien gut als Bäume dargestellt werden. Hierbei ist allerdings zu beachten, dass ein Term an mehreren Stellen des Baumes auftauchen kann. Darüber hinaus kann eine Taxonomie auch Hinweise auf Synonyme für Terme enthalten. Dieses bildet eine weitere interessante Quelle für Informationen zu einem Wort.

Thesaurus

Ein Thesaurus ist ein Wörterbuch, in dem alle Terme in Beziehung zueinander stehen. Diese Beziehungsrelationen haben typischerweise drei Ausprägungen: hierarchisch (Ober- /Unterbegriff), assoziativ („siehe ebenfalls“) und Äquivalenz (Synonyme). Zusätzlich kann ein Thesaurus

Informationen darüber enthalten, was ein Begriff semantisch darstellt, wie er verwendet bzw. unter welchen Randbedingungen er genutzt wird. Ebenso können historische Wortbedeutungen aufgelistet sein.

Ontologie
Eine Ontologie ist, ähnlich einem Thesaurus, eine Taxonomie, die eine Struktur und Relationen zwischen den verschiedenen Begriffen beinhaltet. Die Beziehungen in einer Ontologie sind vielfältiger und spezifischer, diese können zum Beispiel darstellen, ob ein Begriff einer Organisation oder einem Ort zugeordnet werden kann. Ontologische Beziehungen werden häufig in komplexen Systemen genutzt und werden oft als Bestandteil des „semantic web“ angesehen.

5.4.3. Lösungsversuch Wordnet – Reduktion durch Ausnutzung von Semantik

Da die Erlangung von Domänenwissen, wie bereits erwähnt, eine komplexe Aufgabenstellung ist, kann diese nicht ohne weiteres vom Autor gelöst werden. Eine detaillierte Ausarbeitung zu diesem Thema würde viel Zeit kosten und bedürfte einer stark ausgeprägten Zusammenarbeit mit einem Domänenexperten. Grundsätzlich existieren allerdings bereits (nicht fachspezifische) Ansätze zur Dimensionreduktion eines Feature-Vektors. Diese arbeiten häufig auf Ontologien, um ein Mapping zwischen Begriffen zu erreichen (z.B. [San+06; ZAMA08]). Ein solches Wissen kann, wie erwähnt, entweder mit einem Fachexperten zusammen erarbeitet werden (um in Ontologien festgehalten zu werden), oder es wird ein allgemeinerer Ansatz genutzt. Einen solchen bietet „WordNet“, eine von der Universität Princeton entwickelte, lexikalische Datenbank der englischen Sprache, welche auf psycho-linguistischen Erkenntnissen [vgl. auch Ros+76] basiert.

5.4.3.1. Einführung WordNet

Aufgrund der erwähnten Komplexität ist es dem Autor nicht ohne weiteres möglich, eine maßgeschneiderte Lösung für journalistisches Domänenwissen zu entwerfen. Daher soll im Folgenden der allgemeinere WordNet-Ansatz auf seine Nützlichkeit untersucht werde. Hierzu muss zunächst dargestellt werden, was genau WordNet ist und welche Möglichkeiten es bietet. Hierzu stellen die folgenden Absätze eine kurze Einführung in WordNet sowie dessen Aufbau dar. Diese Daten sind größtenteils den Quellen [Uni00; Mil95] entnommen.
WordNet beschreibt sich selber als lexikalische Datenbank englischer Nomen, Verben, Adjektive und Adverben. Es gruppiert diese Terme in Mengen von „kognitiven Synonymen“ (Synsets). Diese stellen jeweils ein einzigartiges semantisches Konzept dar. Synsets sind untereinander

durch semantische und lexikalische Verknüpfungen verbunden, so dass sie ein Netzwerk von aussagekräftiges Begriffen und Konzepten darstellen (ein Wortnetz). Grundsätzlich ähnelt WordNet deshalb einem Thesaurus, der Begriffe nach Bedeutung gruppiert. Hierbei ist zu beachten, dass nicht nur Wörter, sondern ganze Konzepte (semantische Einheiten) aufeinander verlinkt werden, welche sich semantisch unterscheiden. Die so entstehenden Verbindungen tragen daher ein beschreibendes semantische Label.

Struktur

Die Hauptbeziehung zwischen Wörtern in WordNet stellt „Synonymität“ dar. Synonyme, welche das selbe Konzept darstellen und in vielen Fällen direkt austauschbar sind, werden in den Synsets abgebildet. WordNet enthält ca. 117.000 Synsets, welche jeweils zu anderen Sets verlinken. Diese Synsets enthalten hierbei eine Kurzbeschreibung des semantischen Konzeptes, welches sie beschreiben sowie ein Beispiel, das die Verwendung in einem Satz beschreibt. Da Wörter je nach Kontext verschiedene Bedeutungen haben, treten diese u.U. in mehreren Synsets auf.

Beziehungen

Die am häufigsten verwendete Beziehung, zwischen Synsets, ist die Ober-Unterbegriffs-Verbindung (diese ist auch als „Ist-Ein“ oder **Hyperonym**-Beziehung bekannt). Diese verbindet beispielsweise Oberbegriffe wie „political orientation“ mit konkreteren Ausprägungen wie „democracy“. Hierbei führen alle Hierarchien von Nomen letztendlich auf den Root-Knoten „entity“. Zu erwähnen ist, dass die Ober-Unterbegriffs-Relation transitiv ist.
Neben der Hyperonym Beziehung existiert auch die **Meronym** Relation, welche eine „Ist ein Teil von“ Verbindungen umsetzt. So sind z.B. „Reifen“ Teile eines Autos. Diese Beziehungsart vererbt sich, so dass ein „Kleinwagen“ Reifen hat, da sein Oberbegriff „Auto“ diese ebenfalls besitzt.
Verben werden von WordNet ebenfalls in Synset-Hierarchien eingeteilt. Die Verben werden hierbei immer spezifischer, je tiefer der Baum wird (Beispiel: „kommunizieren“-„reden“-„flüstern“). Hierbei ist die Spezialisierung jeweils von der semantischen Bedeutung abhängig (im obigen Beispiel: Lautstärke).
WordNet gruppiert Adjektive anhand von **Antonymen**. So drücken Paare wie „feucht“-„trocken“ den semantischen Gegensatz des Paares aus. Jedes dieser so vernetzten, gegensätzlichen Adjektive verlinkt wiederum auf eine Reihe von semantisch ähnlichen Worten. Darüber hinaus verweisen sogenannte „pertainyms“ (englisches Originalwort, ohne deutsche Entsprechung) auf

die zugehörigen Nomen, so dass dem Adjektiv „kriminell“ das Nomen „Verbrechen“ zugeordnet ist.

Part-of-speech übergreifende Beziehungen
Der Großteil der WordNet Wordbeziehung verbindet Wörter vom selben POS-Typ (Part-Of-Speech), daher werden Nomen mit anderen Nomen verbunden usw. WordNet besteht konkret aus vier Unternetzen, jeweils für Nomen, Verben, Adjektive und Adverben. Diese enthalten einige POS-übergreifende Verbindungen, welche letztendlich auf den selben Wortstamm zurückzuführen sind, ein Beispiel hierfür ist „überwachen“ (Verb), überwacht (Adjektiv), Überwachung (Nomen).
Diese Erklärung ist recht umfangreich, allerdings werden für das beschriebene Problem des semantischen Mappings nicht alle Aspekte der Beschreibung benötigt. Die Erläuterung zeigt allerdings zugleich die Komplexität des Themengebiets und die Schwierigkeit dieses auf bestimmte Aspekte einzuschränken. Daher wurde die Erklärung auch gegeben um einen tieferen Einblick zu gewähren.

5.4.3.2. Idee, Probleme und Durchführung

Auf Basis des nun erlangten Vorwissens soll eine konkrete Idee dargestellt werden, welche daraufhin umgesetzt sowie analysiert werden soll. Grundsätzlich ist die Idee der semantischen Reduktion nicht neu, es gibt bereits diverse Ansätze den Feature-Vektor (Bag-Of-Words) mithilfe von WordNet zu verkleinern. Hierbei verfolgen Paper wie [San+06; ZAMA08] den Ansatz, eine Basis von vordefinierten Klassen (concept-classes) zu nutzen, um auf diese andere Wörter abzubilden. Diese und ähnliche Ansätze verlassen sich in ihrem Vorgehen auf Domänenwissen (konkret: existierende Ontologien). Aus diesem Wissen werden die definierten concept-classes gewonnen, welche als Basis für das semantische Mapping dienen.
Eine solche Basis ist im vorliegenden Fall zurzeit nicht vorhanden, daher scheidet ein solcher Ansatz aufgrund des mangelnden Vorwissens aus. Zwar gibt es diverse bereits zusammengestellte Ontologien für verschiedene Domänen [beispielsweise: Bbc]), diese sind allerdings auf einen konkreten Anwendungszweck zugeschnitten und deshalb nicht ohne Weiteres auf das vorliegende Archiv anwendbar.
Ein anderes Problem, welches an dieser Stelle erwähnt werden muss, ist die so genannte **Word Sense Disambiguation** [vgl. Nav09], welche das Problem der Mehrdeutigkeit natürlicher Sprache beschreibt. Da ein Wort in verschiedenen Kontexten verschiedene Bedeutungen haben kann, ist es für eine Maschine schwer, den semantischen Unterschied festzustellen und das Wort dementsprechend einzuordnen. Ein Beispiel hierfür wären die beiden Sätze „Der Mann saß

auf der *Bank*.“ und „Der Mann beraubte die *Bank*“, hierbei wird das Wort *Bank*, zum einen im Sinne von Sitzgelegenheit, im anderen im Sinne von Aufbewahrungsort für Geld verwendet.
Die obigen beiden Probleme erschweren das Ziel der Verkleinerung des Feature-Vektors enorm. Aus diesem Grund nutzt der Autor im Folgenden eine recht simple Methode, welche allerdings dennoch vielversprechend scheint.
Die nachfolgend genutzte Grundidee ergibt sich aus der These, dass ein Großteil der sinntragenden Terme eines Textes, sich durch die verwendeten Nomen ergibt. Diese beschreiben allerdings häufig diverse Facetten eines Themas. Im Beispielartikel in Abschnitt A sind unter den am häufigsten verwendeten Wörtern unter anderem (jeweils in ihrer gestemmten Form): „govern“, „polit“, „peopl“ sowie „societi“. Diese und weitere konkrete Ausprägungen werden jeweils einzeln gezählt und in den Feature-Vektor aufgenommen. Die grundsätzlichen „Themen“ dieser Begriffe lassen sich allerdings zusammenfassen in zwei Gebiete: *Politik* und *Gesellschaft*. Eine solche Zusammenfassung von Begriffen würde den Feature-Vektor erheblich verkleinern, ohne dabei auf Domänenwissen zurückzugreifen.
Eine Möglichkeit zur Umsetzung dieser Zusammenfassung von Begriffen bietet WordNet. Dieses erlaubt die Extraktion von Oberbegriffen (Hypernymen) zu einem gegebenen Wort. Allerdings ist anzumerken, dass ein Term wie *Demokratie* durchaus mehrere Hypernyme haben kann. Im konkreten Fall wären Oberbegriffe laut WordNet beispielsweise: *politische Orientierung*, *Regierungsform* sowie *Mehrheitsprinzip*. Dieses Beispiel zeigt, dass das gewählte Verfahren weiterhin Mehrdeutigkeiten erlaubt, da das Problem der Word Sense Disambiguation nicht aufgelöst wird.
Die konkret genutzte Umsetzung nimmt daher alle Nomen eines Inputdokumentes, greift auf entsprechende Oberbegriffe zurück und nutzt die Resultate der Anfrage als Eintrag im Feature-Vektor. So werden im Optimalfall Terme (bzw. deren Häufigkeiten) verschmolzen, sofern sie einen gemeinsamen Oberbegriff aufweisen. Dieses soll die Menge der vorhandenen Features drastisch reduzieren.

5.4.3.3. Resultate

Nutzt man das geschilderte Verfahren zur Reduktion des Feature-Vektors durch Verschmelzung von Begriffen zu einem Oberbegriff, so sind die Resultate nicht so gut wie erhofft. Dieses liegt vor allem daran, dass WordNet für diverse Wörter keinen Stem (gemeinsamen Wortstamm) findet bzw. keinen (gemeinsamen) Oberbegriff. Des Weiteren existieren für einige Terme mehrere gültige Stems. Dieses Verhalten führte in der Testreihen des Autors zu diversen Ergebnissen, welche das Resultat stark beeinflussen.

Sofern Terme berücksichtigt werden, für die kein Stem bzw. kein Oberbegriff gefunden werden kann und mehrdeutige Stems erlaubt sind, so beträgt die Dimension des Feature-Vektors 1178. Filtert man nun mehrdeutige Stemformen heraus (und nimmt nur die erste Form zur Berechnung), so reduziert sich der Vektor auf 945 Einträge. Entfernt man nun noch Terme, welche nicht auf einen Stem reduziert werden können bzw. Terme für die kein Oberbegriff gefunden werden kann, so beträgt die Dimension 591. Diese Resultate zeigen ein breites Spektrum an Ergebnissen, welche nicht ohne Weiteres erklärt werden können. Da es sich beim Ausgangsartikel um ein journalistisches Erzeugnis guter Qualität handelt, kann davon ausgegangen werden, dass die Ergebnisspanne nicht auf Rechtschreibfehler etc. zurückzuführen ist.
Darüber hinaus offenbarten die Testreihen eine weitere Problematik, welche vom Autor im Vorhinein nicht als derartig dramatisch eingeschätzt wurde. Erste Versuche führten zu starken Häufungen der Oberbegriffe *metallisches Element* bzw. *lineare Einheit*. Diese Begriffe haben allerdings keinen Bezug zur Thematik „Demokratie“, welche im Ausgangsartikel behandelt wird. Untersuchungen dieses Phänomens zeigten, dass das Wort *be* als Abkürzung für *Beryllium* interpretiert wurde, wodurch der Oberbegriff *metallisches Element* gewählt wurde. Ähnlich verhielt es sich beim Term *in*, welcher als *Inch* interpretiert wurde.
Diese und ähnliche Phänomene zeigen, dass eine solche Verarbeitung von natürliche Sprache für das gesetzte Ziel nicht hilfreich ist. Es konnte der ursprüngliche Feature-Vektor von einer Dimension der Größe 1091 stark reduziert werden, auch wurden näherungsweise die vom Autor erwarteten Oberbegriffe für die zwei Kernthemen *Politik* und *Gesellschaft* generiert (vgl. Tabelle 5.2 und 5.3). Allerdings treten hierbei starke negative Effekte ein, welche das Ergebnis weiterer Versuche durchaus stark negativ beeinflussen können. Zur Reduktion des Feature-Vektors sind daher bessere Verfahren und detailliertes Domänenwissen unerlässlich.

Stem	Anzahl
transpar	55
govern	50
polit	46
trust	37
citizen	33
democrat	27
democraci	25
peopl	25
societi	25
inform	23
power	23

Tabelle 5.2.: Häufigste Wörter (Stems) eines Dokuments

Hypernym	Anzahl
hyper:polity	55
hyper:physical_phenomenon	46
hyper:property	42
hyper:national	33
hyper:social_relation	32
hyper:social_group	30
hyper:person	29
hyper:activity	28
hyper:quality	27
hyper:group	26
hyper:message	26

Tabelle 5.3.: Häufigste WordNet-Oberbegriffe eines Dokuments

5.4.4. Reduktion durch Extraktion von Kernfeatures

Die folgenden Abschnitte beschreiben ein weiteres Verfahren zur Reduktion des Feature-Vektors eines Dokumentes. Diese Idee wurde bereits in Abschnitt 5.4.1 erwähnt und stellt im Vergleich zur semantischen Reduktion (z.B. mit WordNet) den simpleren Ansatz dar. Bei diesem Verfahren wird der Feature-Raum, welcher zur Berechnung der Ähnlichkeit genutzt wird, auf eine überschaubare Anzahl von Kerntermen begrenzt. Das beschriebene Verfahren arbeitet nur auf den Häufigkeiten von Termen eines Dokumentes, daher wird kein Domänenwissen gefordert.

5.4.4.1. Idee und Durchführung

Die bisherigen Lösungsansätze stellten jeweils komplexe Ideen in den Vordergrund, so wurde versucht semantische Eigenschaften wie Wichtigkeit von Teilbereichen eines Dokumentes oder inhaltliche Ähnlichkeit von Begriffen zu nutzen. Hierbei wurde ein recht naheliegender Ansatz nicht berücksichtigt, welcher die Distanzberechnung auf die Kernbegriffe eines Artikels beschränkt.
Dieser Ansatz definiert einen *Kernbegriff* eines Artikels anhand seiner Vorkommenshäufigkeit. So werden im durchgeführten Experiment die jeweils 10 bzw. 30 häufigsten Begriffe extrahiert. Diese Auswahl wird nach der, bereits beschriebenen, Vorverarbeitung (siehe: Ab-

schnitt 4.1) durchgeführt, so dass die ausgewählten Begriffe möglichst vereinheitlicht und von den Stopwords bereinigt sind.
Der so entstehende Raum von zu vergleichenden Elementen ist drastisch verkleinert. Während bisher die Distanz von je zwei Artikeln in ihrer Gesamtgröße berechnet wurde, beschränkt sich die Distanz nun auf die Vereinigungsmenge der einzelnen Feature-Vektoren. Die Vereinigung ist nötig, da zwei Artikel im Allgemeinen nicht die gleichen Kernbegriffe enthalten. So hat Dokument A beispielsweise die Feature-Menge {democracy, transparency, politics}, während Artikel B die Features {bologna, politics, education} besitzt. Um diese Feature-Sets vergleichbar zu machen, wird daher die Vereinigung der beiden Mengen gebildet, wobei jeweils die Häufigkeit für die im Artikel X nicht vorkommende Terme auf 0 gesetzt wird. Die auf diese Weise entstehende Termmenge ist im Beispiel nach oben auf maximal 6 Terme beschränkt (sofern beide Feature-Sets disjunkt) sind, im Falle von gleichen bzw. sich überschneidenden Feature-Sets ist die Größe dementsprechend kleiner.
Darüber hinaus wurde das Experiment, basierend auf den vorherigen Versuchen, eingeschränkt, so dass nun nur noch das TF-IDF Maß verwendet wird (welches die besten Resultate produziert). Weiterhin wurde, neben der bereits mehrfach genutzten euklidischen Distanz, eine weiteres Ähnlichkeitsmaß untersucht, die sogenannte Kosinus-Distanz (siehe: Abschnitt 4.2).
Im Laufe des Experiments wurden, als Datenbasis, die bereits bekannten Korpi A und C (siehe: Abschnitt 4.3.1) wiederverwendet. Darüber hinaus wurde ein weiterer Testkorpus geschaffen, welcher die beiden Testssets vereint um eine größere Menge von Artikeln für weiterführende Tests zu erhalten (Demokratie ∪ Bologna-reduziert ∪ Zufällig).

5.4.4.2. Resultate

Die Ergebnisse des durchgeführten Experiments sind im Anhang unter Abschnitt B.2 zu finden. Zu sehen sind, wie bereits in vorherigen Versuchen, die entsprechenden Precision-Recall Ergebnisdiagramme samt der durchschnittlichen Precision pro Parameter und Distanzfunktion.
Betrachtet man die Ergebnisse des reduzierten Feature-Sets zunächst im Hinblick auf die euklidische Distanzfunktion (mit TF-IDF Maß) und im Vergleich zu den Resultaten des gesamten Feature-Raums, so zeigt sich, dass sich die Resultate nicht signifikant unterscheiden (siehe Seiten: 82, 114, 119). Die durchschnittliche Precision variiert im Vergleich zwischen den Experimenten trotz des wesentlich reduzierten Feature-Raums nur um wenige Prozent. So schwankt die Präzision für den Testkorpus A zwischen ~69% und ~73% bzw. auf dem Testkorpus C zwischen ~67% und ~73%.
Vergleicht man die Resultate der euklidischen Distanzfunktion mit denen der Kosinus-Funktion, so stellt man eine massive Verbesserung der Ergebnisse fest. Die Resultate auf dem Testkorpus

A verbessern sich von ~69% auf ~94%, wenn man die 10 häufigsten Terme als Berechnungsgrundlage nimmt. Für den Testkorpus Bologna verbessern sich die Resultate beim selben Grundsetting von ~72% auf ~100% (was einem perfekten Ergebnis entspricht, so dass die Artikel des Focalpoints die besten Distanzen zum Leitartikel aufweisen).
Eine ähnlich gute Erfolgsbilanz zeigt sich auch, wenn man im Grundsetting die 30 häufigsten Terme (statt 10) als Berechnungsbasis nimmt (vgl. Seiten 119, 128). Für die Korpi A und C ergeben sich auch in diesem Fall Verbesserungen von ~22% bzw. ~33%.
Hieraus lässt sich ableiten, dass die signifikante Verbesserung nicht an einen bestimmten Korpus gebunden, sondern auf die Distanzfunktion zurückzuführen ist.
Untersucht man dieses Phänomen genauer, so zeigt sich, dass die Kosinus-Distanzfunktion unempfindlicher gegenüber unterschiedlichen Dokumentlängen (und damit auch der Größe des Feature-Vektors) ist. Daher erklärt die Heterogenität dieser Größen die bisher schlechten Ergebnisse unter Nutzung der euklidischen Distanz.
Auf Basis der guten Ergebnisse wurde vom Autor, wie erwähnt, ein neuer Korpus eingeführt. Diese Vereinigung der Korpi Demokratie, Bologna und „Zufällig“ dient dazu, die Anzahl der negativ-Ergebnisse (nicht zum Focalpoint gehörend) zu vergrößern. Hiermit wird untersucht, wie stark die Ergebnisse vom Verhältnis „richtig“ (zugehörig zum Focalpoint) zu „falsch“ abhängen. Das Verhältnis in den ursprünglichen Korpi ist ~50:50, während im neuen Korpus ein Verhältnis von 38:62 herrscht. Die Resultate sind ab Seite 132 zu finden.
Zu sehen ist, dass die durchschnittliche Präzision zwar sinkt (und somit nicht mehr im Bereich von 95% - 100% liegt - bezogen auf die Kosinus Distanz). Dennoch sind die Resultate bei einem Ergebnis von ~82% des Optimums durchaus weiterhin zufriedenstellend .
Ebenfalls festzuhalten ist, dass mit den durchgeführten Versuchen kein eindeutiger Einfluss auf die Festlegung des Eingangsparameters zu treffen ist. Sowohl die 10 sowie die 30 häufigsten Terme liefern auf dem Bologna Korpus, unter Nutzung von Kosinus-Distanz, perfekte Ergebnisse. Auf den beiden anderen Korpi erzielt die Kosinus-Distanz minimal bessere Ergebnisse, sofern mehr Terme in die Berechnung mit einfließen (vgl. Seiten 128, 138). Allerdings sind diese Unterschiede zu vernachlässigen, wenn man bedenkt, dass die Anzahl der eingehenden Terme verdreifacht wurde.

5.5. Weitere Lösungsansätze

Die folgenden Abschnitte sollen darstellen was, weitere Arbeiten auf dem gewählten Themengebiet, untersuchen könnten. Einige der Ansätze würden hierbei diese Arbeit direkt fortführen, während andere grundsätzlich neue Zweige untersuchen würden.

Wie dargestellt ist das gesamte Themengebiet der vorhandenen Klassifizierungs- und Clusteringverfahren ein Bereich, welcher weiter untersucht werden sollte. Einige der vorgestellten Frameworks wurden für einen sehr ähnlichen Anwendungsfall, wie den in dieser Arbeit genutzten, entwickelt. Diese sollten daher Priorität genießen. So ist vor allem das Overview-Project interessant. Durch seine Eigenschaft als Open-Source-Projekt könnten mögliche Erkenntnisse vergleichsweise direkt produktiv eingesetzt werden.
Ein weiterer während der Arbeit aufgekommener Ansatz ist eine andere Art der Ermittlung von Kernfeatures. Grundlegend sind Kernfeatures (konkret: Wörter) diejenigen, welche die Semantik eines Dokumentes tragen. Daher ergibt sich die Frage, wie diese zu ermitteln sind. Eine konkrete Idee war es deshalb, den gesamten Korpus als Basis zu nehmen und zu untersuchen, welche Terme am stärksten vom mittleren Vorkommen aller Wörter abweichen. Die resultierenden Terme wären nun die, welche besonders häufig bzw. besonders selten genutzt werden. Dieses könnte ein Indiz für ihre Wichtigkeit sein.
Ein gänzlich anderer Ansatz, welcher zur Debatte stand, ist die Erstellung von neuen Focalpoints auf Basis von bereits vorhandenen. Die Kernidee ist, ein vorhandenes Dossier zu nehmen und durch Hinzufügen bzw. Entfernen von „Kernfeatures“ neue, aber ähnlich gelagerte Dossiers zu generieren. Ein Beispiel wäre, die Basis des Focalpoints „End of Democracy“ zu erweitern. Denkbar wäre eine Einengung der Thematik um örtliche Bezugspunkte wie „Deutschland“, so dass das Ergebnis eine Untermenge des Ursprungs-Focalpoints ist, welcher sich allerdings regional auf Deutschland beschränkt. Ebenso denkbar ist es den Focalpoint „End of Democracy“ auszuweiten, sodass ein Dossier des Themas „Democracy“ entsteht. Dieses Beispiel lässt sich beliebig erweitern und soll daher nur einen Einstiegspunkt bilden. Weiterhin ist zu beachten, dass eine gegebene Menge von Kernfeatures natürlich nie genau einen gegebenen Focalpoint exakt konstruiert, sondern ihn nur annähert. Aus diesem Grund stellen Ausweitungen bzw. Einengungen des Themas eine Herausforderung dar.
Weitere Ideen, welche während der Bearbeitung aufkamen, bezogen sich beispielsweise auf die Nutzung von neuronalen Netzen, um optimale Gewichtungsfaktoren für Textanteile zu berechnen bzw. aus Beispielen zu erlernen. Darüber hinaus ist es ebenfalls denkbar, dass „Named Entities“ (beispielsweise Personen oder Firmennamen) eine thematische Eingrenzung eines recht großen Focalpoints erreichen können, da viele Thematiken mit Personen oder Institutionen verknüpft sind. Weitere Möglichkeiten bestehen darin, die dargestellten Ansätze auf andere Korpi (z.B. Reuters) zu übertragen. Dieses würde das Problem des sehr breiten Themenfeldes und zeitlichen Bezuges des Eurozine-Korpus lösen.
Darüber hinaus ist es möglich, das Wissen von Domänenexperten stärker einzubinden. Denkbar wäre beispielsweise die Zusammenstellung von fachlichen Ontologien, welche als Basis für

thematische Cluster genutzt werden können. Eine Menge von Schlagwörtern kann so von Experten jeweils zu Gruppen zusammengefasst werden, sodass auf dieser Grundlage fachlich bestimmte Cluster definiert werden. Die Dokumente können daraufhin auf die entsprechenden Wörter analysiert werden, um Cluster zu finden.
Ebenso wurde in den vorgestellten Ansätzen noch nicht auf die vorhandenen Metainformationen von Artikeln zurückgegriffen. Diese zusätzlichen Informationen erlauben es ggf. auf Basis der Artikelautoren, Herkunftsländer, Zeitschriften etc. bessere Ergebnisse zu produzieren.
Weiterhin ist es denkbar, dass einige in dieser Arbeit vorgestellte Verfahren auf Basis der gewonnenen Erkenntnisse verbessert werden können. Dieses würde ggf. eine produktive Nutzung der Methoden ermöglichen.

6. Fazit und Ausblick

6.1. Zusammenfassung

Ziel dieser Arbeit war es, sich in das Thema der semi-automatisierten Erstellung von Pressedossiers einzuarbeiten, hierbei verschiedene Fragestellungen zu konkretisieren und vor allem diverse Ideen zu untersuchen sowie sie durch Experimente zu analysieren.
Dieses Ziel wurde nach Ansicht des Autors durchaus erfüllt. Während Kapitel eins und zwei die Problemstellung und eine Vision und Grundlagen absteckten, wurde in Kapitel drei ein konkreter Fragenkatalog entworfen. Diese Erkenntnisse flossen in Kapitel vier und fünf in die konkrete Realisierung einer Architektur und Bewertung, sowie in die eigentlichen Lösungsansätze des Problems mit ein.
Betrachtet man die einzelnen Abschnitte dieser Arbeit detaillierter, so zeigt sich zunächst, dass die Erfüllung der Vision dieser Arbeit inhärent schwer zu messen ist. Dieser Punkt wird auch durch die thematische und zeitliche Breite des benutzten Korpus verstärkt, da dieses die Analyse erschwert. Dasselbe Phänomen zeigt sich bei der Ausformulierung konkreterer Fragen sowie der Definition von Dossiers. So kann man sagen, dass eine semi-automatisierte Erstellung von Dossiers durchaus machbar ist und für Journalisten hilfreich sein kann, allerdings lassen sich andere Fragen weniger gut beantworten. So ist die verwendete Definition des Begriffs Dossier recht offen und lässt sich auf vielfache Weise interpretieren. Dieses ist allerdings wie gezeigt eine inhärente Eigenschaft. Daher sind auch die fachlichen Anforderungen an ein Dossier schwer zu fassen. Dieses macht den Entwurf eines übersichtlichen Verfahrens, welches qualitativ hochwertige Ergebnisse liefert, schwer. Aus diesen Gründen müssen in weiteren Arbeiten zum Thema Mittel und Wege gefunden werden, dieses Problem zu lösen. Aufbauende Arbeiten sollten sich daher u.A. damit beschäftigen, eine detailliertere Definition von Dossiers zu erarbeiten, welche es erlaubt, den Erfolg direkter zu messen. Ob ein solches Unterfangen, ohne die Vision zu sehr einzuengen, machbar ist, bleibt allerdings fraglich. Dieser Punkt muss sehr genau mit Domänenexperten abgestimmt werden, um deren Wünschen optimal entgegenzukommen.

Eine weitere Erkenntnis der vorliegenden Arbeit betrifft den Teilbereich Architektur und technische Realisierung. Einen allgemeinen Aufbau für eine Analyse mit einem breiten Spektrum an Experimenten zu entwerfen, stellt sich als schwerer heraus, als ursprünglich vermutet. In der Retroperspektive ist das Grobgerüst des dargestellten KDD-Prozesses, die treffenste Architektur. Diese ist nur schwerlich verfeinert darzustellen, da Details von Experiment zu Experiment stark variieren. Im Nachhinein war ebenfalls die technische Realisierung mithilfe von „RapidMiner“ oft problematischer als erwartet. Grundlegende Schritte sind zwar schnell realisiert, sobald allerdings eigene Logik hinzugefügt werden muss, ergeben sich diverse Probleme. Zu diesen gehören beispielsweise die Handhabung von RapidMiner internen Datenstrukturen, welche unnötig kompliziert erscheinen. Darüber hinaus ist die Dokumentation des Tools oft wenig hilfreich. Für weitere Untersuchungen würde der Autor daher überlegen, eine Alternative zur benutzen Toolsuite zu untersuchen.
Ein weiterer Bereich über den man im Nachgang dieser Arbeit reflektieren muss, ist der gesamte Bereich der verfolgten Lösungsansätze. Zunächst ist der Kernpunkt der Experimente die Methode mit welcher der Erfolg der Versuche gemessen wird. Diese erscheint dem Autor nach wie vor sinnvoll, allerdings sollte, in potenziell weiteren Arbeiten, überlegt werden, ob das verwendete Verfahren nicht verbessert werden kann.
Untersucht man die konkret gewählten Lösungsansätze, so zeigt sich, dass eine Gewichtung einzelner Textabschnitte durchaus die Resultate verbessern kann. Allerdings kann eine falsche Gewichtung auch zu negativen Ergebnissen führen. Dieses stellt ein Problem dar, welches näher untersucht werden sollte.
Ein weiterer Aspekt, welcher durch die Analysen offengelegt wurde, ist die Problematik eines auf Domänenwissen fußenden Ansatzes. Der mit WordNet verfolgte Ansatz hat nicht funktioniert. Allerdings zeigte die Beschäftigung mit der Gesamtthematik, dass der zu Grunde liegende Ansatz ein großes Potenzial hat. Für eine erfolgreiche Nutzung des Domänenwissens wäre, allerdings laut Ansicht des Autors wie bereits erwähnt eine komplexe fachliche Ontologie nötig, welche die spezifischen Begriffe von Kulturzeitschriften bündelt.
Eine weitere Kernerkenntnis dieser Arbeit ist, dass es essentiell notwendig ist, den Feature-Raum eines Dokumentes möglichst weit zu verkleinern, ohne dabei wichtige Kerninformationen zu verlieren. Diese an sich einfache Erkenntnis stellt einen der Kernpunkte dar, welcher in weiteren Arbeiten unter keinen Umständen unterschätzt werden sollte. Ähnlich wichtig ist die Wahl der Distanzfunktion, welche die Ergebnisse maßgeblich beeinflusst.
Eine weiterer bemerkenswerter Punkt ergibt sich aus der Vorstellung von diversen Klassifikations- und Clusteringverfahren. Diese sind zwar nach wie vor nicht die erste Wahl des Autors, da die Toolboxen jeweils nur schlecht an fachliche Bedürfnisse angepasst werden

können. Allerdings zeigen sich interessante Möglichkeiten, vor allem in Bezug auf bereits existierende Projekte, welche konkret für Journalisten entwickelt worden sind. Diese sollten definitiv in weiteren Arbeiten detaillierter auf ihre Nützlichkeit untersucht werden. Die entsprechenden Analysen sollte in enger Zusammenarbeit mit Fachexperten geschehen, um auf etwaige Wünsche und Ideen konkret eingehen zu können.
Das Fazit der gesamten Arbeit sieht der Autor durchaus positiv und als Erfolg. Grundsätzlich wurde gezeigt, dass die dargestellten Verfahren einen positiven Nutzen für die Arbeit von Journalisten haben können. So ist es möglich große Artikelarchive durchaus effektiv auf konkrete Fragestellungen hin zu untersuchen, ohne jedes Dokument sichten zu müssen. Ähnliche Arbeiten sollten einen Fokus auf ein gut abgegrenztes Ziel legen, welches auf Basis der Erkenntnisse dieser Arbeit möglich ist. Darüber hinaus sollte die Evaluierung von vorhandenen Algorithmen und Frameworks einen weiteren Kernpunkt darstellen. Auch wenn diese Methoden häufig nicht trivial sind, ist es sinnvoll, sich der Verbesserung der vorhandenen Algorithmen zu widmen und daraufhin zu überlegen, wie man fachlichen Wünschen nachkommen kann. Ein solches Vorgehen kann bereits bekanntes Wissen nutzen und schnell gute Ergebnisse liefern, welche daraufhin fachlich aufgewertet werden können.
Abschließend sieht der Autor im gewählten Anwendungsfeld weiterhin großes Potenzial für weitere Forschungsarbeiten. Allerdings ist anzumerken, dass eine recht hohe Einarbeitungszeit von Nöten ist, um die vorhandenen Verfahren kennenzulernen. andererseits liefern auch recht simple Ansätze schon sehr gute Ergebnisse, auf welchem aufgesetzt werden kann.

6.2. Ausblick

Grundsätzlich zeigt sich, dass die vorliegende Arbeit nur einen kleinen Teil der vorhandenen Möglichkeiten untersucht hat. Konkret wurde die Kontextualisierung von Presseartikeln eines bestehenden Archivs behandelt und selbst dieser Teilbereich erlaubt eine Vielzahl von Erweiterungen (wie bereits in Abschnitt 5.5) dargestellt wurde. Über dieses Feld hinaus bieten sich diverse weitere Forschungsgebiete an.
Zum einen sind Ausdehnungen der Untersuchungen auf das Themenfeld der Individualisierung von Vorschlagsystemen [vgl. z.B. KFD12] möglich. Die dargestellten Verfahren berücksichtigen im besten Fall zwar die fachlichen Bedürfnisse einer gesamten Fachgruppe, allerdings werden hierbei Wünsche von einzelnen Redakteuren außen vor gelassen. Dieses kann in bestimmten Szenarien allerdings ein essentielles Bedürfnis sein, welches momentan nicht bedient wird.
Darüber hinaus ist die verwendete Datenbasis ein gut gepflegtes und umfängliches Archiv, allerdings könnten Forschungen auch andere Datenbestände nutzen. Mögliche weitere Quellen

umfassen die diversen Open-Data-Initiativen, welche gänzlich andere Daten bieten. Diese beinhalten Informationen zu verschiedenen Sachgebieten, die dementsprechend ggf. eine völlig andere Struktur als ein journalistisches Pressearchiv aufweisen. Darüber hinaus wären auch soziale Medien eine denkbare Informationsquelle [beispielsweise: Sch14], welche zudem die Herausforderung beinhaltet, dass innerhalb von sehr kurzer Zeit viele neue Daten generiert werden, die zudem nicht redaktionell gepflegt sind.
Eine aus anderen Datenquellen folgende Konsequenz wären zudem völlig neue Anwendungsfälle, welche nichts mehr mit journalistischen Erzeugnissen zu tun haben. So ist es denkbar, dass die vorgestellten Verfahren auch in komplett andere Domänen portiert werden können.
Darüber hinaus lässt sich auch ein weiter gefasster Ausblick geben, welcher weitere interessante Fragen aufwirft. So kann man sich beispielsweise mit der Thematik beschäftigen welche Chancen und Risiken die vom Autor verfasste Vision bietet. In diesem Kontext stellt sich die Frage wie effektiv ein Verfahren sein müsste, um die redaktionelle Erstellung von Dossiers durch rein automatisierte Verfahren abzulösen. Ebenfalls interessant ist die verwandte Fragestellung nach der Monetarisierbarkeit solcher Verfahren. Dieses beinhaltet die Gefahr, dass entsprechend billige Verfahren, qualitativ hochwertiger menschlicher Arbeit vorgezogen werden um die Kosten zu senken, solange dieses die Gesamtbilanz verbessert. Weiterhin muss auch die Frage nach der Verwendung solcher Verfahren gestellt werden, so könnten unter Umständen entsprechend effektive und effiziente Methoden für ethisch fragwürdige Untersuchungen missbraucht werden, obwohl eine andere Idee im Vordergrund der Entwicklung stand.
Die obigen Ideen dienen nur als Beispiel für die vielfältigen denkbaren Verbesserungen, Erweiterungen, Portierungen in andere Themengebiete sowie Fragestellungen welche sich ergeben. Daher hegt der Autor die Hoffnung, dass diese Arbeit einen Beitrag leistet, weitere Ideen zu entwickeln und umzusetzen.

Danksagung

Hiermit möchte ich mich noch einmal bei ganz herzlich bei „Eurozine – Gesellschaft zur Vernetzung von Kulturmedien mbH“ und den Verantwortlichen Carl Henrik Fredriksson (Chefredakteur) und Veronika Leiner (Geschäftsführung) bedanken. Erst durch die Freigabe des Eurozine-Archivs zur Verwendung im Umfeld der Masterarbeit des Autors wurden die dargestellten Untersuchungen möglich.

Literatur

[BNJ03] David M. Blei, Andrew Y. Ng und Michael I. Jordan. „Latent Dirichlet Allocation“. In: **J. Mach. Learn. Res.** 3 (03/2003), S. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[Ban+12] Cristian Bancu u. a. „ARSYS – Article Recommender System“. In: **Proceedings of the 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing**. SYNASC '12. Washington, DC, USA: IEEE Computer Society, 2012, S. 349–355. ISBN: 978-0-7695-4934-7. DOI: 10.1109/SYNASC.2012.38. URL: http://dx.doi.org/10.1109/SYNASC.2012.38.

[Bbc] **BBC Ontologies**. URL: http://www.bbc.co.uk/ontologies (besucht am 10. Sep. 2015).

[Bob+13] J. Bobadilla u. a. „Recommender Systems Survey“. In: **Know.-Based Syst.** 46 (07/2013), S. 109–132. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2013.03.012. URL: http://dx.doi.org/10.1016/j.knosys.2013.03.012.

[Bre+14] M. Brehmer u. a. „Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists“. In: **Visualization and Computer Graphics, IEEE Transactions on** 20.12 (12/2014), S. 2271–2280. ISSN: 1077-2626. DOI: 10.1109/TVCG.2014.2346431.

[CL14] J. Cleve und U. Lämmel. **Data Mining**. Gruyter, de Oldenbourg, 2014. ISBN: 9783486713916. URL: https://books.google.de/books?id=4i2nngEACAAJ.

[CR12] Sidharth Chhabra und Paul Resnick. „CubeThat: News Article Recommender“. In: **Proceedings of the Sixth ACM Conference on Recommender Systems**. RecSys '12. New York, NY, USA: ACM, 2012, S. 295–296. ISBN: 978-1-4503-1270-7. DOI: 10.1145/2365952.2366020. URL: http://doi.acm.org/10.1145/2365952.2366020.

[Car] **Carrot Square**. URL: http://project.carrot2.org/index.html (besucht am 10. Sep. 2015).

[Dec] **Decision Tree**. URL: http://cvpr.uni-muenster.de/teaching/ws08/mustererkennungWS08/script/ME08.pdf (besucht am 2. Aug. 2015).

[Dil] **Dilbert Questions**. URL: http://dilbert.com/strip/2010-11-22 (besucht am 25. Aug. 2015).

[Eura] **Eurozine Focalpoints**. URL: http://www.eurozine.com/comp/FocalPoints.html (besucht am 10. Sep. 2015).

[Eurb] **Eurozine**. URL: www.eurozine.com (besucht am 10. Sep. 2015).

[FPS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro und Padhraic Smyth. „From Data Mining to Knowledge Discovery: An Overview“. In: **Advances in Knowledge Discovery and Data Mining**. 1996, S. 1–34.

[FS06] Ronen Feldman und James Sanger. **The Text Mining Handbook**. Cambridge Books Online. Cambridge University Press, 2006. ISBN: 9780511546914. URL: http://dx.doi.org/10.1017/CBO9780511546914.

[Fay+96] Usama Fayyad u. a. „The KDD Process for Extracting Useful Knowledge from Volumes of Data“. In: **Communications of the ACM** 39 (1996), S. 27–34.

[Gov] **GOVDATA**. URL: https://www.govdata.de/ (besucht am 10. Sep. 2015).

[Har+13] Sébastien Harispe u. a. „Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis“. In: **CoRR** abs/1310.1285 (2013). URL: http://arxiv.org/abs/1310.1285.

[Hä14] Nina Hälker. **Dienen textminingbasierte Dossiers dem Wachsen einer gemeinsamen europäischen Erzählung?** Ausarbeitung. 2014.

[Hä15] Nina Hälker. **Halbautomatisierte Erstellung von Dossiers auf der Basis von Textmining-Verfahren**. Masterarbeit Arbeitspapier. 2015.

[KFD12] Evan Kirshenbaum, George Forman und Michael Dugan. „A Live Comparison of Methods for Personalized Article Recommendation at Forbes.Com“. In: **Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II**. ECML PKDD'12. Berlin, Heidelberg: Springer-Verlag, 2012, S. 51–66. ISBN: 978-3-642-33485-6. DOI: 10.1007/978-3-642-33486-3_4. URL: http://dx.doi.org/10.1007/978-3-642-33486-3_4.

[Kni] **KNIME**. URL: https://www.knime.org/ (besucht am 10. Sep. 2015).

[Kra] Ivan Krastev. **The transparency delusion**. Zeitungsartikel. URL: http://www.eurozine.com/articles/2013-02-01-krastev-en.html.

[LGS] Pasquale Lops, Marco de Gemmis und Giovanni Semeraro. In: **Recommender Systems Handbook**.

[Li+10] Lihong Li u. a. „A Contextual-bandit Approach to Personalized News Article Recommendation“. In: **Proceedings of the 19th International Conference on World Wide Web**. WWW '10. New York, NY, USA: ACM, 2010, S. 661–670. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772758. URL: http://doi.acm.org/10.1145/1772690.1772758.

[MR02] Christopher Manning und Prabhakar Raghavan. **Text Retrieval and Mining (CS27A) - Lecture 8**. Vorlesung. Stanford, 09/2002. URL: https://web.stanford.edu/class/cs276a/handouts/lecture8.pdf.

[MRS08] Christopher D. Manning, Prabhakar Raghavan und Hinrich Schütze. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715. URL: http://www-nlp.stanford.edu/IR-book/.

[Mah] **Apache Mahout**. URL: http://mahout.apache.org/ (besucht am 10. Sep. 2015).

[McD83] John P. McDermott. „Extracting Knowledge From Expert Systems.“ In: **IJCAI**. Hrsg. von Alan Bundy. William Kaufmann, 1983, S. 100–107. URL: http://dblp.uni-trier.de/db/conf/ijcai/ijcai83.html#McDermott83.

[Mil95] George A. Miller. „WordNet: A Lexical Database for English“. In: **Commun. ACM** 38.11 (11/1995), S. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: http://doi.acm.org/10.1145/219717.219748.

[Nav09] Roberto Navigli. „Word Sense Disambiguation: A Survey“. In: **ACM Comput. Surv.** 41.2 (02/2009), 10:1–10:69. ISSN: 0360-0300. DOI: 10.1145/1459352.1459355. URL: http://doi.acm.org/10.1145/1459352.1459355.

[Nik12] Stanislav Nikolov. „Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series“. In: (11/2012). URL: http://dspace.mit.edu/bitstream/handle/1721.1/85399/870304955.pdf.

[Nor12] Matthew North. **Data mining for the Masses**. 2012. ISBN: 978-0615684376.

[OW05] Stanislaw Osinski und Dawid Weiss. „A Concept-Driven Algorithm for Clustering Search Results“. In: **IEEE Intelligent Systems** 20.3 (2005), S. 48–54. DOI: 10.1109/MIS.2005.38. URL: http://doi.ieeecomputersociety.org/10.1109/MIS.2005.38.

[Ope] **Open Data Showroom**. URL: http://opendata-showroom.org/de/ (besucht am 10. Sep. 2015).

[Osi04] Stanislaw Osinski. „Dimensionality Reduction Techniques for Search Results Clustering“. Magisterarb. Department of Computer Science, University of Sheffield, UK, 2004. URL: http://www.cs.put.poznan.pl/dweiss/carrot/xml/publications.xml?lang=en.

[Ovea] **Overview Project: Completed Stories**. URL: https://blog.overviewdocs.com/completed-stories/ (besucht am 24. Juli 2015).

[Oveb] **Overview Project**. URL: https://blog.overviewdocs.com/ (besucht am 24. Juli 2015).

[PL10] Chi-Chieh Peng und Duen-Ren Liu. „Combining Reputation and Content-based Filtering for Blog Article Recommendation in Social Bookmarking Websites“. In: **Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business**. ICEC '10. New York, NY, USA: ACM, 2010, S. 8–14. ISBN: 978-1-4503-1427-5. DOI: 10.1145/2389376.2389379. URL: http://doi.acm.org/10.1145/2389376.2389379.

[Pr] **Precision-Recall diagram**. URL: http://wikis.gm.fh-koeln.de/wiki_ir/uploads/InformationRetrieval/Recall/ir_bild.jpg (besucht am 26. Nov. 2014).

[Rap] **RapidMiner**. URL: https://rapidminer.com/ (besucht am 10. Sep. 2015).

[Reu] **Reuters Korpi**. URL: http://trec.nist.gov/data/reuters/reuters.html (besucht am 10. Sep. 2015).

[Rm a] **RapidMiner API-Dokumentation (inoffiziell)**. API Dokumentation aus Quellcode. URL: http://fossies.org/dox/rapidminer-5.3.013/index.html.

[Rm b] **RapidMiner API Dokumentation (offiziell)**. API Dokumentation. 2008. URL: http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/WS0809/rm-api/overview-summary.html.

[Rm c] **RapidMiner Dokumentation**. Dokumentation. 2008. URL: http://docs.rapidminer.com/.

[Ros+76] Eleanor Rosch u. a. „Basic objects in natural categories“. In: **Cognitive Psychology** (1976).

[Rpr] **R-Project**. URL: https://www.r-project.org/ (besucht am 10. Sep. 2015).

[SW03] Jerzy Stefanowski und Dawid Weiss. „Carrot and Language Properties in Web Search Results Clustering“. In: **Web Intelligence, First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003, Proceedings**. 2003, S. 240–249. DOI: 10.1007/3-540-44831-4_25. URL: http://dx.doi.org/10.1007/3-540-44831-4_25.

[San+06] Tratz Sanfilippo u. a. „Ontological Annotation with WordNet“. In: SemAnnot 2005, 5th International Workshop on Knowledge Markup and Semantic Annotation. Galway, Ireland, 2006. URL: http://www.osti.gov/bridge/servlets/purl/908503-NCEc8D/908503.pdf.

[Sch05] Katharina Schwarz. **Domain model enhanced search - A comparison of taxonomy, thesaurus and ontology**. Masterthesis. 07/2005.

[Sch14] Marcel Schöneberg. **Erkennung von Trends in sozialen Netzwerken**. Projektbericht. 10/2014. URL: http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-proj/schoeneberg.pdf.

[Sch15a] Marcel Schöneberg. **Automatisierte Erstellung von Pressedossiers durch Textmining**. Projektbericht. 02/2015. URL: http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2015-proj/schoeneberg.pdf.

[Sch15b] Marcel Schöneberg. **Automatisierte Erstellung von Pressedossiers durch Textmining**. Ausarbeitung. 02/2015. URL: http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-sem/schoeneberg/bericht.pdf.

[Sch15c] Marcel Schöneberg. **Konzepte zur semi-automatisierten Erstellung von Pressedossiers**. Masterthesis. 09/2015.

[Sip15] Sigurd Sippel. „Recommendations for cocktail recipes“. In: (02/2015). URL: http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-sem/sippel/bericht.pdf.

[TCY09] Joe Tekli, Richard Chbeir und Kokou Yetongnon. „Survey: An Overview on XML Similarity: Background, Current Trends and Future Directions“. In: **Comput. Sci. Rev.** 3.3 (08/2009), S. 151–173. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2009.03.001. URL: http://dx.doi.org/10.1016/j.cosrev.2009.03.001.

[Tik] **Apache Tika**. URL: https://tika.apache.org/ (besucht am 10. Sep. 2015).

[Tra] **Transparenzportal Hamburg**. URL: http://transparenz.hamburg.de/open-data/ (besucht am 10. Sep. 2015).

[Uni00] Princeton University. „About WordNet“. In: (2000). URL: http://wordnet.princeton.edu/wordnet/.

[Wek] **Weka**. URL: http://www.cs.waikato.ac.nz/ml/weka/ (besucht am 10. Sep. 2015).

[YLZ14] Ming Yang, Ying-ming Li und Zhongfei(Mark) Zhang. „Scientific articles recommendation with topic regression and relational matrix factorization“. English. In: **Journal of Zhejiang University SCIENCE C** 15.11 (2014), S. 984–998. ISSN: 1869-1951. DOI: 10.1631/jzus.C1300374. URL: http://dx.doi.org/10.1631/jzus.C1300374.

[ZAMA08] Elberrichi Zakaria, Rahmoun Abdelattif und Bentaalah Mohamed Amine. „Using WordNet for Text Categorization“. In: **The International Arab Journal of Information Technology** 5.1 (08/2008). URL: http://iajit.org/PDF/vol.5,no.1/3-37.pdf.

Tabellenverzeichnis

Abbildungsverzeichnis

Anhang

A. Beispielartikel

Der folgende Artikel stellt einen realen Artikel des Archivs dar und dient der Erläuterung des Aufbaus. Dieser ist auch online abrufbar ([Kra]).

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE article SYSTEM "article.dtd">
<article lang="en">
<imprint>
<author>Ivan Krastev</author>
<copyright>Ivan Krastev</copyright>
<copyright>Eurozine</copyright>
<firstin><i>In Mistrust We Trust: Can Democracy Survive When We Don't Trust Our
    Leaders?</i> TED Books 2013</firstin>
<pubdate>2013-02-02</pubdate>
</imprint>
<title>The transparency delusion</title>
<blurb>Disillusionment with democracy founded on mistrust of business and
    political elites has prompted a popular obsession with transparency. But the
    management of mistrust cannot remedy voters' loss of power and may spell the
    end for democratic reform.</blurb>
<body>

<motto>There is strong shadow where there is much light.<br/>Goethe</motto>

A well-known French engraving of 1848, the year French citizens received the
    universal right to vote, epitomizes the dilemmas of European democracies at
    their birth. The engraving pictures a worker with a rifle in one hand and a
    ballot in the other. The message is clear: bullets for the nation's enemies
    and ballots for the class enemies. Elections were meant to be the instrument
    for inclusion and nation building. They integrated workers into the nation by
     sharing power with them. The man with a rifle in one hand and the ballot in
    the other symbolized the arrival of democracy in France because he was, at
    once, both a Frenchman and a worker, a representative of a nation and a
    social position absorbed in class struggle. He understood that the person who
     would stand beside him on the barricades would also be a worker and a
    Frenchman with a clear idea who the enemy was. His rifle was not only a
    symbol of his constitutional rights, it was evidence that the new democratic
    citizen was prepared to defend both his fatherland and his class interest. He
     knew that the power of his vote was dependent on the firepower of his gun.
```

```
    The ballot was an additional weapon because elections were a civilized form
    of civil war. They were not simply mechanisms for changing governments. They
    were tools for remaking the world.<p/>
<includebox>/XML/infobox/democracybox.htm</includebox>
The ubiquitous smartphone of today may not be a rifle, but it has the capacity to
    perform its own kind of shooting. It can document abuses of power and make
    them public. It can connect and empower people. And it can spread truth. It
    is hardly accidental that the recent wave of popular protests around the
    world coincided with the spread of smartphones. Innocent photos posted on
    social networks triggered many of our current political scandals. In China,
    Brother Wristwatch and Uncle House are some of the latest victims of the
    citizen with the smartphone. Both of them are low-ranking officials who were
    exposed for suspected corruption this year by Internet mobbing. Brother Watch
     was captured in several photos wearing very expensive watches, some of which
     cost more than his annual salary. Uncle House, who was in charge of a
    district urban management bureau in the southern city of Guangzhou, was
    exposed for collecting real estate — 22 properties in all. The smartphone-
    equipped citizens ousted both of them. In Russia, the legitimacy of the
    Russian Orthodox Church was undermined when a blogger posted a photo on
    Facebook showing the patriarch donning an expensive watch, and it declined
    further when Russians learned that the patriarch's public relations team
    doctored videos to conceal this fact from the public. In Syria, citizens
    armed with smartphones documented the massively heinous crimes of the regime.
     And in the United States, a smartphone recorded Governor Mitt Romney's
    infamous "47 per cent comment" that outraged the other half of America (and,
    one would hope, some of that original 47 per cent, too).<p/>

<imgfloat name="krastev_ted1_468w.jpg"/><p/>

The smartphone can also function as a citizen's personal lie detector. A voter, in
     real time, can fact-check the various claims and assertions politicians make
    , from the most vital political issues to the more mundane personal anecdotes
    . When Republican vice presidential candidate Paul Ryan "misremembered" his
    first marathon time — he claimed he ran it in under three hours when it
    really took him more than four hours — his "mistake" inspired immediate
    questions about the candidate's credibility. It is not that politicians can't
     fool people anymore, but they do it at the risk of looking like fools
    themselves. The outsized influence of fact-checking websites during the last
    US presidential campaign is a classic illustration of the power of the
    smartphone to unearth the truth — or at least to pretend to present factual
    truth to the public.<p/>
<infobox>
<h1>Further information</h1>
<br/>
<a href="http://www.ted.com/pages/tedbooks_library"><img src="../../UserFiles/
    Image/krastev_ted_100.jpg" style="float: centre;margin: 5 px 5 px 5 px 5 px;"
     /></a><br/>

Ivan Krastev's book <i>In Mistrust We Trust</i> is based on his June 2012 TED talk
     <a href="http://www.ted.com/talks/
```

```
    ivan_krastev_can_democracy_exist_without_trust.html"><b>"Can democracy exist
    without trust?"</b></a>
For further information on the book, please visit the <a href="http://www.ted.com/
    pages/tedbooks_library">TED Books Library</a>.
</infobox>
The smartphone also empowers citizens to speak and express their views and
    opinions. They can call, email, and tweet their judgments and thus contribute
     to a broader political conversation in real time. Each of the three debates
    between the two candidates in the recent American presidential election
    generated, just for the duration of the debate, more than seven million
    tweets. Life may not be more enlightened, but it is far more entertaining in
    the age of Twitter.<p/>

But perhaps most critically, the new citizens can use their smartphones to
    mobilize public action, to ask other citizens to come to the streets and to
    collectively defend their interests. The Arab Spring was the ultimate
    manifestation of the power of citizens armed with smartphone power to
    overthrow tyrants and to make history. Smartphones can't maim or kill, but
    they do make it more costly for the governments to do so themselves. At the
    same time, the Arab Spring represented significant limits to the power of the
     smartphone. The person with the smartphone never knows who might respond to
    his appeal for political action. He may have his Facebook friends, but he
    lacks a genuine political community and political leaders. You can tweet a
    revolution, but you can't tweet a transition. It turned out, of course, that
    Islamist political parties that relied on traditional party structures and
    clear ideologies were the winners of the post-revolutionary elections in the
    Middle East.<p/>

Today, it is the person with the smartphone in one hand and the blank ballot in
    the other that symbolizes our democratic condition. Yet he or she is not a
    recognizable member of any particular class or ethnic group, and the ballot
    is no longer a weapon at his or her disposal. We don't think in terms of
    barricades, and we have vague ideas of who are "comrades" and who are enemies
    . Both the ballot and the smartphone are instruments of control, not
    instruments of choice. The actual fear of the smartphone voter is that the
    people he or she votes for will serve only their selfish interests. The
    citizen with the smartphone doesn't confront the tough ideological choices
    his predecessors faced. While the expansion of choices has radically
    increased in recent decades, in politics it has been the reverse. For the
    politically committed citizen of yesterday, changing one's party or political
     camp was as unthinkable as swapping one's religion. To move from the Left to
     the Right today, or the other way around, is as simple as traversing the
    border between France and Germany — it's a high-speed highway with no
    passport control.<p/>

So does the citizen with the smartphone represent the power we have accrued or the
     power we have lost? Should we be nostalgic for the decline of ideological
    politics or liberated by its burden? And can we trust the smartphone to be an
     effective new instrument to defend our rights?
```

```
<subheading>Transparency is the new religion</subheading>
Is the citizen with the smartphone the one who can restore our trust in democracy
    and democratic institutions? I am sceptical. Smartphones may make it easier
    for us to control our politicians, but trust refers to the confidence in the
    operation of institutions that people cannot directly monitor and control. We
     don't trust our families and friends because we are able to control them.
    The increased capacity of people to control their representatives doesn't
    translate easily into trust in democracy. Lenin used to believe that "trust
    is good, control is better," but the Bolshevik titan is not widely known for
    his model of democratic governance. And while it is likely that today's
    crisis of trust is probably less dramatic than the surveys tell us (and the
    current public debate suggests), sociologist Niklas Luhmann has argued that
    trust is "a basic fact of social life," without which one could not get out
    of the bed in the morning. It is also clear that the increased ability of
    citizens to control their governments has not led to more trust in democracy.
     Unfortunately, most of the initiatives that claim to rebuild civic trust are
     in reality helping arouse a democracy of mistrust. This trend is nowhere
    more evident than in today's popular obsession with transparency.<p/>

Transparency is the new political religion shared by a majority of civic activists
     and an increasing number of democratic governments. The transparency
    movement embodies the hope that a combination of new technologies, publicly
    accessible data, and fresh civic activism can more effectively assist people
    control their representatives. What makes transparency so attractive for
    different civic groups is the exciting premise that when people "know," they
    will take action and demand their rights. And it is fair to admit that the
    advancement of the transparency movement in many areas has demonstrated
    impressive results. Governmental legislation that demanded companies to
    disclose the risks related to their products empowered customers and made
    life safer (we have today's often reviled Ralph Nader as one early person to
    thank here). Demand for disclosure has also transformed the relations between
     doctors and patients, teachers and students. Now patients have a greater
    capacity to keep doctors accountable, and parents can more effectively decide
     which school to select for their children. The new transparency movement has
     empowered the customers.<p/>

Thus it is logical to assume that, stripped of the privilege of secrecy,
    governments will be irreversibly changed. They will become more honest. Where
     the government maintains too many secrets, democracy becomes brittle, even
    when competitive elections produce, ex ante, uncertain outcomes. Only
    informed citizens can keep governments accountable. In short, it is
    unsurprising that democracy activists have invested so much hope that
    transparency itself can restore trust in democratic institutions. As American
     legal scholar and activist Lawrence Lessig stated in his essay "Against
    Transparency": "How could anyone be against transparency? Its virtues and its
     utilities seem so crushingly obvious." But while the virtues of transparency
     are obvious, the risks should not be ignored, as Lessig powerfully argues.<p
    />
```

```
The notion that transparency will restore public trust in democracy rests on
    several problematic assumptions, primarily the presupposition that "if only
    people knew" everything would be different. It is not so simple. The end of
    government secrecy does not mean the birth of the informed citizen, nor does
    more control necessarily suggest more trust in public institutions. For
    instance, when American voters learned that the US had started a war with
    Iraq without proof of weapons of mass destruction, they still re-elected the
    president who led the way. And when Italians kept Silvio Berlusconi in power
    for more than a decade, they had long been saturated with news of all the
    wrongdoings that anti-Berlusconi activists hoped would be enough to get rid
    of the guy. But in politics, "knowing everything" still means knowing
    different things. And the very fact that governments are compelled to
    disclose information does not necessarily translate to people knowing more or
     understanding better. Inundating people with information is a time-tested
    way to keep people uninformed. If you don't trust me, ask your accountant. He
     will tell you that the best way to discourage any tax inspector to look into
     the workings of your company is to give him all available information
    instead the needed and the useful items. When it comes to the relations
    between trust and control, the issue is even more complex. Does control
    create trust, or is it simply a substitute for it? Do authoritarian
    governments increase their capacity to control society in order to trust them
     more?<p/>

Contrary to the claim of transparency advocates who insist that it is possible to
    reconcile the demand for the opening of government with the protection of
    citizens' privacy, I contend that wholly transparent government denotes a
    wholly transparent citizen. We can't make the government fully transparent
    without sacrificing our privacy. In contrast to those advocates who believe
    that a politics of full disclosure improves the quality of public debate, I
    think that injections of huge flows of information make public conversation
    more complicated, shifting the focus away from the moral competence of the
    citizen to his expertise in one or another area. Contrary to the expectations
     of the transparency movement that full disclosure of government information
    will make public discourse more rational and less paranoid, my argument is
    that a focus on transparency will only fuel conspiracy theories. There is
    nothing more suspicious than the claim of absolute transparency. And nobody
    can honestly say that when our governments have become more transparent our
    debates have become less paranoid. The rise of the transparency movement has
    the potential to remake democratic politics, but we should be sure we are in
    agreement as to the direction of the change. Is the transparency movement
    capable of restoring trust in democratic institutions, or is it,
    alternatively, going to make "mistrust" the official idiom of democracy?

<subheading>A society of spies</subheading>
Crucially, our extreme focus on transparency influences the very way democracy
    works. It may even contribute to a process of replacing representative
    democracy with political regimes that limit themselves only to citizen
    control of the executive. Contrary to its stated ambition to restore trust in
     democratic institutions, the transparency movement may accelerate the
    process of transforming democratic politics into the management of mistrust.
```

```
    The politics of  transparency is not an alternative to a democracy without
    choices; it is its justification and blurs the distinction between democracy
    and the new generation of market-friendly authoritarian regimes. It is not
    surprising that Chinese leaders enthusiastically endorse the idea of
    transparency. What they oppose is the competition of parties and ideas and
    the search for political alternatives to the Communist rule.<p/>

<imgfloat name="krastev_ted2_468w.jpg"/><p/>
In the late eighteenth century, British philosopher and social theorist Jeremy
    Bentham designed an institutional form he dubbed the panopticon. The concept
    of the design was to allow a watchman to observe all inmates in an
    institution — whether a prison, school, or hospital — without them being
    able to recognize whether or not they were being watched. The panopticon soon
     became the symbol of our modern understanding of power as the control over
    dangerous individuals or groups. The twentieth century's famous anti-utopias
    — portrayed in Aldous Huxley's <i>Brave New World</i>, Yevgeny Zamyatin's <i
    >We</i>, George Orwell's <i>1984</i> — are, by and large, stories of
    transparent societies in which the government has the capacity of total
    control. Knowing everything is the government's utopia of absolute power.<p/>

If the idea of the "naked" society is the dream of governments, the idea of a
    naked government and denuded corporations represent the wish fulfilment of
    many democracy activists. Initiatives such as Publish What You Pay, Open
    Government Initiative, or radical political efforts such as WikiLeaks are the
     best studies making the case that when armed with the "right" information,
    people can keep governments accountable. Louis Brandeis' oft-quoted line that
     "sunlight is said to be the best of disinfectants" succinctly summarizes the
     philosophy of the transparency movement. The movement aims to build a
    reverse panopticon whereby it is not government that will monitor society but
     society that will monitor those in power. The totalitarian utopia of people
    spying for the government is now replaced by the progressive utopia of people
     spying on the government.<p/>

The problem, however, is that spying is spying, regardless of who is spying on
    whom (just as the winner of a rat race is, alas, still a rat). Should we
    concede our right to privacy in order to get better public services? Is it
    fundamentally different from the demand of totalitarian regimes to proscribe
    individual choice in order to achieve national greatness and a more equal
    society? The debate over WikiLeaks' published cables brought into full view
    the moral dimension of the war against secrecy. As a rule, governments
    monitor people. When you make such efforts transparent, you also open up to
    the world those citizens who spoke with or were monitored by the government.
    It is impossible to publish authentic documents without putting at risk
    government sources. And it is impossible to open state files without reading
    the information they have collected about its citizens. The opening of secret
     police files in post-communist societies is the classical example of the
    dilemmas behind any politics of disclosure. Should everyone know what others
    have been doing during the communist period? Should only the files of public
    figures be opened? How reliable is the information collected by the secret
    police? Will the knowledge about others produce moral catharsis in society,
```

```
    or will it be used simply as "kompromat" (compromise) in sordid power games?
    These are not easy questions.<p/>

Modern society was built on the hope that one day we will trust strangers and
    institutions as if they were members of our families. Recent experience shows
    , however, that the reverse is true. We have begun treating our families with
     the mistrust earlier reserved for criminals. What we are witnessing is how
    the combination of mistrust and new technologies is remaking our private
    lives. Mistrust is now the default option even in family relations. Indeed,
    lawyers now say that technology is turning divorce into an arms race.
    Kitchens and bedrooms are now bugged like the American embassy in Moscow was
    in the days of the Cold War. Thus, while the promise of transparency was to
    restore trust in public institutions, in reality it spread mistrust into the
    sphere of private life.

<subheading>The age of spin</subheading>
The late US Senator and public intellectual Daniel Patrick Moynihan was one of the
     first to analyze the impact of government secrecy on the way society trusts
    its institutions. He argued convincingly that secrecy should be understood as
     any other form of regulation. In his view, the performance of the US
    government was negatively affected during the Cold War by those in power
    deploying considerable forms of secrecy. Secrecy was responsible, he
    suggested, for the paranoid turn in American politics during the McCarthy era
     and badly hurt the readiness of citizens to trust their government. Moynihan
    's contention that in order to trust the government, citizens should see its
    full profile is therefore hard to dispute. But while the argument for
    transparency is a powerful one, the notion of full disclosure is not
    unproblematic. Is every unveiling not, at the same time, a veiling of another
     sort? Is the information that governments collect with the understanding
    that it will become immediately public as reliable as the information
    collected when they knew it would be kept secret? Would, say, the Pentagon
    Papers have been the blockbuster that it was if the government released it on
     its own?<p/>

Further, the availability of information is no guarantee that people will have
    more trust in the decision-making process, because information never comes
    without interpretation. Reading the same raw data, Republicans and Democrats
    in the US or secularists and the Muslim Brotherhood in Egypt will spin it
    differently, because policy making cannot be divorced from the interests and
    values of the decision makers. "Ours, it appears, is an Age of Obsessions,"
    write the anthropologists Jean and John Comaroff in the Afterword to the
    collection <i>Transparency and Conspiracy</i>. "It is an age in which people
    almost everywhere seem preoccupied, simultaneously, with transparency and
    conspiracy."<p/>

The ambiguity of the politics of trust is best observed in the case of Russia's
    recent presidential elections. In December 2011, the country's parliamentary
    elections ended in civic explosion. Hundreds of thousands of people went onto
     the streets of Moscow and other big cities asking for fair elections and
    real choices. The escalating crisis of legitimacy of the regime forced the
```

```
        government to invent imaginative ideas to justify its power. The central
        proposal was ingenious. In order to guarantee the fairness of the vote, the
        Kremlin proposed that webcams be installed at all polling stations; every
        citizen could personally monitor the fairness of the process. As China's
        Xinhua wire service enthusiastically reported, "From Kamchatka to Kaliningrad
         and from Chechnya to Chukotka, more than 2.5 million net surfers registered
        to view live streaming from at least 188,000 webcams installed in more than
        94,000 polling stations on Russian territory." In the words of one Finnish
        observer, what happened was a lesson in transparency: "a landmark in the
        history of democracy and democratic elections."<p/>

    It is hardly difficult to argue that in the context of Vladimir Putin's regime,
        where the government decides who will run and who will not, the installation
        of webcams was little more than a farce. Far more important is the ambiguity
        of the presence of the webcams. Viewed from Moscow and the West, the webcams
        are perceived as an instrument to keep the government under control — to
        allow people knowledge about what the government is doing. But from the point
         of view of a post—communist Russian voter living in the deep countryside,
        the webcam sent a different message: government knows how you vote. In a way,
         then, Putin succeeded twice. He succeeded to look transparent in the eyes of
         the West and threatening to most of his own citizens. In short, the webcams
        during Russia's elections were simultaneous acts of transparency and
        conspiracy.<p/>

    In Bulgaria in the summer of 2009, a new government came into office. The promise
        of openness was high on its agenda. In his first days, the new prime minister
         decreed that all the discussions at the Council of Ministers would be made
        available on the government's website within 48 hours. Civic organizations
        were euphoric. But the consequence was wholly unexpected.<p/>

    Armed now with the understanding that government information will be almost
        immediately put online, ministers were unduly careful what they said and how
        their words could be construed. Soon, the government began to use the
        openness policy as a kind of public relations instrument. The prime minister
        spent government meetings attacking his opponents or making speeches. Further
        , most decisions were taken with hardly any discussion. This perverse
        consequence of transparency was that the "real" decisions were taken outside
        of the Council of Ministers and that openness worked to strengthen the
        personal power of the prime minister.<p/>

    The transparency—conspiracy axis is perhaps best revealed in the character and
        mindset of today's great soldiers in the war against government secrecy.
        Julian Assange, the founder of WikiLeaks, described his organization as an "
        open source democratic intelligence agency." In many ways Assange resembles
        someone straight out of a Joseph Conrad conspiracy novel. Of the dozens of
        recently published books about Assange, not to mention his own autobiography,
         the radical transparency activist comes off as a secretive, paranoid,
        authoritarian figure. He is someone you might admire but not someone you can
        trust. Assange has made deception his passion and his profession. His
        preferred strategy is to avoid distinguishing between democratic and
```

```
    authoritarian governments; in his conception, all governments are
    authoritarian. Is it possible that Assange's worldview could be a starting
    point for restoring trust in democracy?<p/>

At the moment when government information is designed to be immediately open to
    everybody, its value as information stands in decline and its value as an
    instrument of manipulating the public increases. Just remember how gangsters
    in crime movies talk when they know that their rooms are bugged. They speak
    clearly and offer banalities while at the same time exchange secret notes
    under the table. This is how governments work in the age of transparency. The
     obvious question begged here is why the influx of information fails to
    change the quality of democracy. In his study of truth telling in ancient
    Greece, Michel Foucault points out that the act of truth telling can't be
    reduced to citizens learning something they didn't know before. Paradoxically
    , truth in politics is something that everybody knows but nobody dares to
    express or pay attention to. People hardly need additional data to realize
    that inequality is rising or that immigrants are mistreated. The WikiLeaks
    cables didn't help us learn something about America's policies we hadn't
    known. Rather, it is the decision of someone to take personal risks and
    confront the authorities or his or her community and not some "unknown" truth
     that makes a speech politically powerful. Living in truth can't be reduced
    to having access to full information. It is the person daring to say the
    truth and not the truth itself that will ultimately bring change.

<subheading>Transparency and anti-politics</subheading>
"You can be sure that in the nearest future, someone will create software that
    will make it almost impossible for politicians to lie," my old friend Scott
    Carpenter, a deputy director of Google Ideas, told me only half-jokingly.
    Recently, Google established Google Ideas, a think tank that works to put
    technology into the service of citizens. For years, politics was the art of
    telling people what they want to hear. Carpenter's suggestion was that in the
     age of transparency, this should no longer be possible. What my friend had
    in mind was that the new software would track all the statements and
    positions taken by a politician on a certain issue so that when he changes
    his position and starts to flip-flop, the voter can punish him for his
    opportunism. Not only that, we would know whom the politician meets, who
    contributes to his campaign, and whether his spouse or kids serve on the
    boards of the government's favoured companies.<p/>

Transparency then stands less in opposition to secrecy but to deception and lies.
    The promise of the transparent society is no different from the promise of
    the science-fictional Truth Machine. It is the promise of a society without
    lies. You can never eliminate the liars, but you can eliminate the lie and
    its attendant power to subvert society. What is disturbing in the growing
    hope that transparency will improve our societies is something T.S. Eliot
    observed almost a century ago: how the advocates of transparency are "
    dreaming of systems so perfect so no one will need to be good." In this
    imagining, trust comes not from shared goals or experience or from certain
    ethics but from the mastery of the institutional design. Rather than believe
```

```
        in the self-correcting nature of democratic society, they hold out faith for
        the establishment of societies that make no mistakes.<p/>

    If the Enlightenment philosophers once tried to understand man — his heart, his
        mind, his fears — the new generation of democratic reformers have lost
        interest in people. In their world of institutions and incentives, changing
        your mind is only a sign of political opportunism. But isn't changing one's
        mind the very essence of democratic politics? Is consistency more important
        for democratic politics than the readiness to change your point of view when
        presented with new information or new circumstances? Imagine how the world
        would look if Woodrow Wilson or FDR hadn't revisited their early pledges that
         America would remain on the sidelines. The original sin of the transparency
        movement is just this neglect for the psychological complexity of democratic
        politics.<p/>

    The trap of the current transparency-centred reform movement is the assumption
        that it is enough to know who is giving money to politicians or whom they
        meet for dinner to arrive at a clear picture of the nature of the decision-
        making process. The fact that a congressman has received, say, $50,000 from a
         defence contractor simply can't guarantee that it was this donation that
        determined the legislator's support for the increase of the military budget.
        But in our Age of Transparency people are tempted to take shortcuts. "Tell me
         his donors, and I will tell you his politics" is the regrettable shorthand
        for today's political environment. But politics cannot be reduced in this way
        . All this new information and state-of-the-art digital technologies don't
        help fashion a better understanding of democratic politics. Rather, this
        approach risks that the public will start treating its own representatives as
         dangerous criminals who should be monitored round the clock. The problem
        with the assumption that trust depends mostly on our ability to control our
        politicians has the disastrous consequence that most of our gifted and civic-
        minded citizens are appalled at the very thought of ever running for office.
        Is it possible to restore trust in democracy by treating politicians not as
        national leaders but as persons to be distrusted by definition?<p/>

    "When we really wish to know how the world is going," once wrote the philosopher
        and mystery writer G.K. Chesterton, "it is not a bad test to take some tag or
         current phrase of the press and reverse it, substituting the precise
        contrary, and see whether it makes more sense that way." In our case, does it
         make more sense that transparency will restore trust in democratic
        institutions or that it will reduce politics to simply the management of
        mistrust? The transparency-centred reform of democracy is not ultimately an
        alternative to the democracy of mistrust — a way out, so to speak — but is
        instead its major justification. It is the outcome of the incapacity of the
        average voter to bring change and to have a meaningful choice in democratic
        politics in the age of "no alternatives." It tacitly accepts that democratic
        politics is no longer about clashing visions of the "good society" or
        conflicting interests and values. It is simply the process of controlling
        those in power. But transparent decision making is not the same as good
        policy. Transparency is not a simulacrum for the public interest.
        Transparency can be one of the instruments of social reform, but it cannot be
```

```
      the goal and content of democratic reform. How we take decisions won't
     replace the fundamental question of what is best for society.

<subheading>Exit and voice</subheading>
"It is happier to be cheated sometimes," observed the proverb—happy Samuel Johnson
     , "than not to trust." And he was right, because a society of mistrust is a
     society of powerless citizens. In his classic study "Exit, Voice and Loyalty,
     " the great economist and social thinker Albert Hirschman argues that there
     are two kinds of responses to the deterioration of services or the
     performance of institutions: exit and voice. To paraphrase Hirschman, "exit"
     is the act of leaving because a better good or service is provided by another
      firm or organization. Indirectly and unintentionally "exit" can cause a
     deteriorating organization to improve its performance. "Voice" is the act of
     complaining, petitioning or protesting, with the intention of achieving a
     restoration of the quality that has been impaired. Easy availability to exit
     is inimical to voice, for by comparison with exit voice is costly in terms of
      effort and time. Moreover, to be effective, voice often requires group
     action and is thus subject to all the well—known difficulties of organization
      — namely, representation and free riding.<p/>

Voice and exit thus distinguish the world of politics from the world of the market
     . The politics of voice is what we call political reform. But in order for
     political reform to succeed, there are several important preconditions.
     People must feel committed to invest themselves in changing their societies
     by feeling a part of that society. And for the voice option to function
     properly, people should strategically interact with others and work to make
     change together. Commitment to one's group is critically important for the
     messy and methodical politics of change to work properly. What worries me
     most at present is that citizens react to the failures of democracy in a way
     similar to how they react when disappointed with the market. They simply exit
     . They exit by leaving the country or stopping voting or, indeed, voting with
      blank ballots. The citizen with the smartphone acts in the world of politics
      the same way he acts in  the sphere of the market. He tries to change
     society simply by monitoring and leaving. But it is the readiness to stay and
      change reality that is at the heart of democratic politics. It is this basic
      trust that allows society to advance. This is why democracy cannot exist
     without trust and why politics as the management of mistrust will stand as
     the bitter end of democratic reform.<p/>

<i>This is a slightly edited excerpt from Ivan Krastev's book </i>In Mistrust We
     Trust: Can Democracy Survive When We Don't Trust Our Leaders?
</body></article>
```

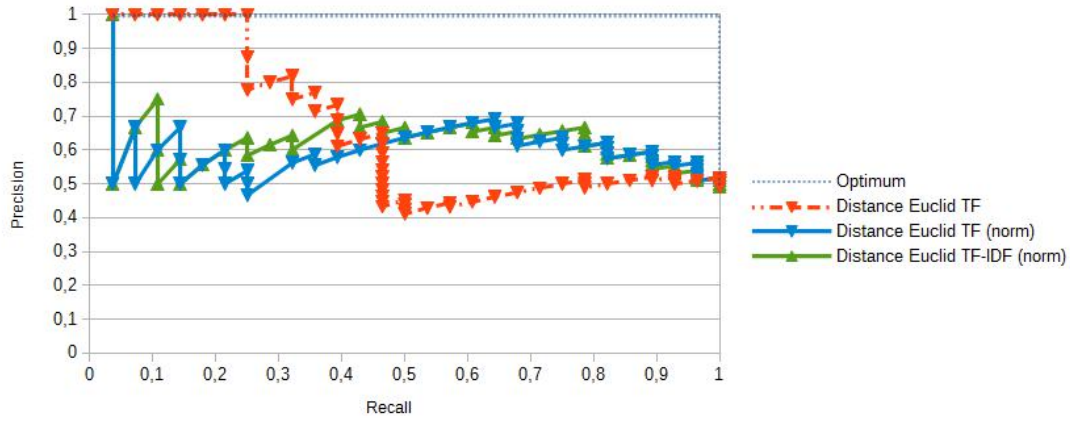
Listing 7.1: Beispielartikel

B. Untersuchungsergebnisse

B.1. Gewichtung von Textanteilen

B.1.1. Basisfall

Parameter	Wert	Durchschnitt Precision	
Korpus	A		
Abstract	1	Optimal	
Title	1	Euklid TF	72,68%
Subheadings	1	Euklid TF (norm)	70,31%
Body	1	Euklid TF-IDF	72,79%

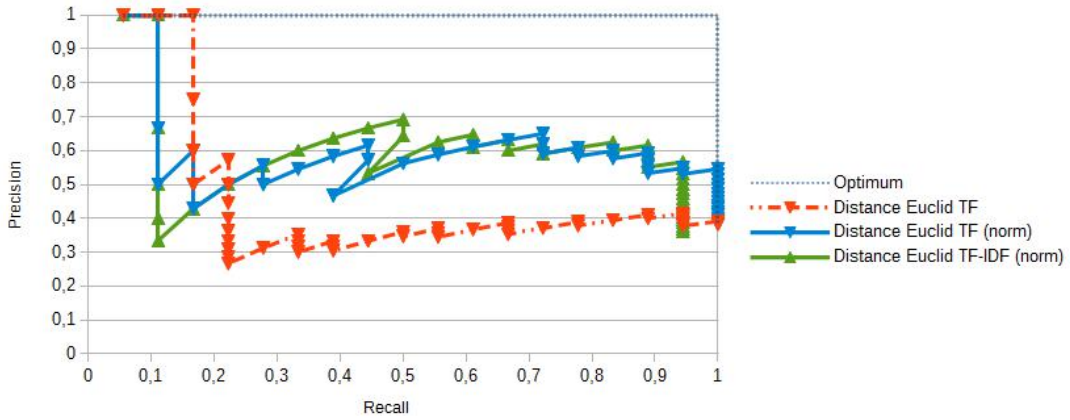


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	175,111	3,648	5,154
3	2008-05-02-wennerhag-en.xml	183,014	2,104	2,649
4	2008-11-21-leggewiewelzer-en.xml	137,310	4,577	7,167
5	2009-04-21-fraser-en.xml	204,203	1,791	2,298
6	2009-07-14-biscione-en.xml	159,984	3,721	6,517
7	2009-09-09-kavaliauskas-en.xml	156,490	3,557	5,790
8	2010-09-14-ditchev-en.xml	138,268	7,682	12,250
9	2011-07-11-bluhdorn-en.xml	151,139	2,606	4,144

10	2011-11-02-G1000-en.xml	136,741	2,791	4,596
11	2011-11-10-sierakowski-en.xml	143,482	2,609	3,960
12	2011-12-19-amirpur-en.xml	163,954	2,876	5,340
13	2012-01-25-halmai-en.xml	215,986	1,565	2,005
14	2012-02-08-elsenhans-en.xml	165,015	5,323	7,687
15	2012-09-05-jahanbegloo-en.xml	151,526	4,329	7,137
16	2012-11-21-holmes-en.xml	144,665	2,494	3,989
17	2013-02-08-wallerstein-en.xml	158,745	5,121	7,053
18	2013-02-19-leggewie-en.xml	190,628	3,026	4,644
19	2013-02-26-james-en.xml	149,817	6,514	10,153
20	2013-05-03-muller-en.xml	158,190	3,228	4,742
21	2013-06-14-pomerantsev-en.xml	161,861	3,174	5,431
22	2013-07-29-gole-en.xml	181,940	2,067	3,752
23	2013-08-13-krastev-en.xml	142,464	6,194	9,487
24	2013-08-20-leggewienanz-en.xml	165,076	2,663	3,828
25	2013-09-11-deniztekin-en.xml	148,020	14,802	21,032
26	2013-11-08-vidanava-en.xml	148,125	8,713	13,270
27	2013-11-22-offe-en.xml	141,142	2,433	3,816
28	2013-12-12-margetts-en.xml	163,576	2,371	3,559
29	2013-12-12-pogonyi-en.xml	149,117	4,660	6,759
30	1117-2007-07-06-lapin1-en.xml	156,394	17,377	23,683
31	1193-2007-11-02-boulbina-en.xml	169,797	4,354	7,218
32	1270-2008-04-09-miklosi-en.xml	159,962	4,443	6,781
33	1344-2008-08-07-seymour-en.xml	229,460	2,550	4,676
34	2100-2011-09-27-scruton-en.xml	151,934	9,496	13,407
35	211-2002-12-20-verene-en.xml	174,645	2,460	3,929
36	2163-2012-01-11-ohlheiser-en.xml	154,612	4,685	8,001
37	2200-2012-03-20-mondediploo-en.xml	152,089	19,011	26,390
38	223-2003-01-31-des-en.xml	157,038	5,609	7,579
39	2447-2013-04-12-sanchez-en.xml	149,496	8,794	12,924
40	2495-2013-06-25-zhurzhenko-en.xml	196,235	2,066	5,325
41	2517-2013-08-13-osteuropa-en.xml	150,499	7,921	11,476
42	256-2003-02-11-kaplinski-en.xml	155,917	5,997	8,625

43	266-2003-02-16-mangasassen-en.xml	153,496	2,476	3,560
44	2666-2014-04-03-knausgard-en.xml	331,056	1,733	2,024
45	294-2003-03-04-ursic-en.xml	156,426	31,285	41,628
46	335-2003-05-15-henard-en.xml	145,499	9,094	14,346
47	414-2003-10-20-bogdanovic-en.xml	188,101	2,351	3,214
48	441-2003-11-28-abraham-en.xml	155,461	7,403	10,183
49	479-2004-03-03-senyener-en.xml	157,550	6,850	10,791
50	480-2004-03-04-cakmak-en.xml	155,775	5,372	8,092
51	505-2004-04-05-uys-en.xml	146,263	6,648	10,753
52	540-2004-06-21-peters-en.xml	198,230	2,227	2,740
53	62-2001-04-01-mistry-en.xml	227,886	2,713	4,864
54	661-2005-07-14-revista-en.xml	154,777	10,318	14,097
55	772-2006-02-01-boutang-en.xml	148,876	8,757	13,206
56	785-2006-02-16-sambrook-en.xml	150,047	4,689	7,244
57	80-2001-11-14-blecher-en.xml	150,612	4,564	6,705
58	904-2006-08-17-eder-en.xml	156,908	4,241	5,955

Parameter	Wert	Durchschnitt Precision	
Korpus	B		
Abstract	1	Optimal	
Title	1	Euklid TF	57,04%
Subheadings	1	Euklid TF (norm)	73,82%
Body	1	Euklid TF-IDF	73,61%

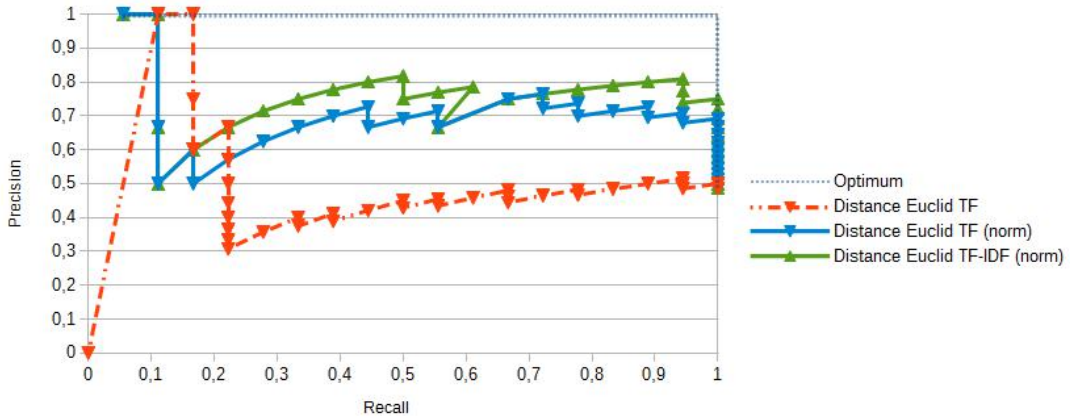


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	151,119	1,737	2,268
2	bologna-2010-03-11-manchev-en.xml	91,940	2,485	3,572
3	bologna-2010-04-13-dorling-en.xml	110,869	2,520	3,784
4	bologna-2010-07-01-balan-en.xml	49,163	2,235	3,222
5	bologna-2010-07-01-benhamou-en.xml	76,603	1,781	2,334
6	bologna-2010-07-01-bikbov-en.xml	120,250	1,240	1,526
7	bologna-2010-07-01-bot-en.xml	50,249	2,185	2,917
8	bologna-2010-07-01-calderwilliams-en.xml	82,043	2,735	4,326
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	87,721	1,790	2,190
11	bologna-2010-07-01-herwigEtAl-en.xml	86,093	2,208	3,532
12	bologna-2010-07-01-lichtenbergerh-en.xml	59,405	1,747	2,407
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	104,010	2,261	2,691

14	bologna-2010-07-01-munch-en.xml	89,185	1,652	2,240
15	bologna-2010-07-01-power-en.xml	125,547	2,369	4,051
16	bologna-2010-07-01-samalavicius-en.xml	88,011	1,796	2,327
17	bologna-2010-07-01-santos-en.xml	242,946	1,233	1,102
18	bologna-2010-07-01-schwan-en.xml	72,911	2,804	3,312
19	bologna-2010-07-01-vonosten-en.xml	51,971	2,260	3,053
20	1117-2007-07-06-lapin1-en.xml	59,296	6,588	8,375
21	1193-2007-11-02-boulbina-en.xml	103,073	2,643	5,317
22	1270-2008-04-09-miklosi-en.xml	87,721	2,437	4,334
23	1344-2008-08-07-seymour-en.xml	207,714	2,308	4,005
24	2100-2011-09-27-scruton-en.xml	65,841	4,115	5,391
25	211-2002-12-20-verene-en.xml	115,352	1,625	2,838
26	2163-2012-01-11-ohlheiser-en.xml	98,306	2,979	6,179
27	2200-2012-03-20-mondediploo-en.xml	58,566	7,321	9,326
28	223-2003-01-31-des-en.xml	66,106	2,361	2,725
29	2447-2013-04-12-sanchez-en.xml	69,412	4,083	5,996
30	2495-2013-06-25-zhurzhenko-en.xml	178,227	1,876	4,879
31	2517-2013-08-13-osteuropa-en.xml	68,037	3,581	5,477
32	256-2003-02-11-kaplinski-en.xml	88,966	3,422	4,277
33	266-2003-02-16-mangasassen-en.xml	101,637	1,639	1,787
34	2666-2014-04-03-knausgard-en.xml	329,437	1,725	1,846
35	294-2003-03-04-ursic-en.xml	58,138	11,628	12,178
36	335-2003-05-15-henard-en.xml	80,716	5,045	8,419
37	414-2003-10-20-bogdanovic-en.xml	135,967	1,700	2,112
38	441-2003-11-28-abraham-en.xml	65,276	3,108	4,082
39	479-2004-03-03-senyener-en.xml	72,256	3,142	6,614
40	480-2004-03-04-cakmak-en.xml	78,275	2,699	4,643
41	505-2004-04-05-uys-en.xml	83,379	3,790	6,053
42	540-2004-06-21-peters-en.xml	156,965	1,764	1,741
43	62-2001-04-01-mistry-en.xml	211,974	2,524	4,412
44	661-2005-07-14-revista-en.xml	61,327	4,088	4,914
45	772-2006-02-01-boutang-en.xml	70,704	4,159	6,608
46	785-2006-02-16-sambrook-en.xml	88,843	2,776	4,271

47	80-2001-11-14-blecher-en.xml	82,831	2,510	3,211
48	904-2006-08-17-eder-en.xml	76,740	2,074	3,236

Parameter	Wert	Durchschnitt Precision	
Korpus	C		
Abstract	1	Optimal	
Title	1	Euklid TF	57,91%
Subheadings	1	Euklid TF (norm)	79,56%
Body	1	Euklid TF-IDF	84,25%

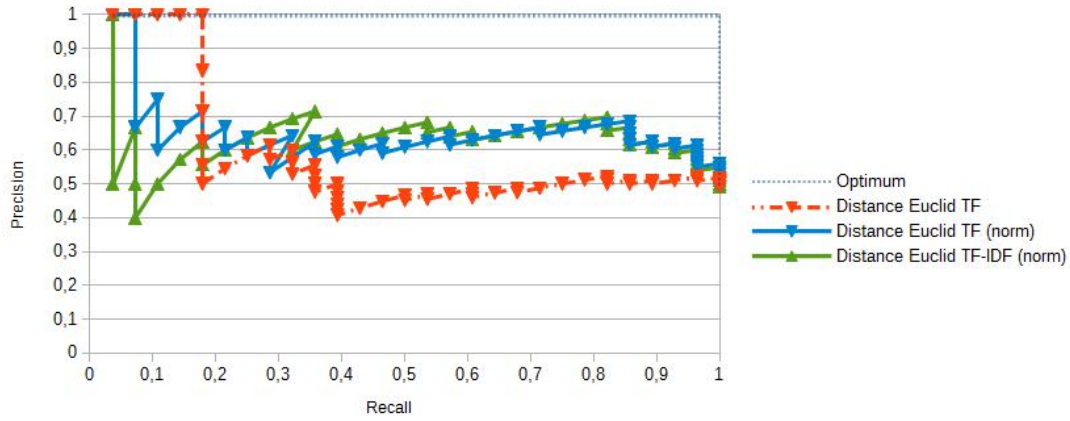


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	151,119	1,737	2,169
2	bologna-2010-03-11-manchev-en.xml	91,940	2,485	3,578
3	bologna-2010-04-13-dorling-en.xml	110,869	2,520	3,547
4	bologna-2010-07-01-balan-en.xml	49,163	2,235	3,047
5	bologna-2010-07-01-benhamou-en.xml	76,603	1,781	2,222
6	bologna-2010-07-01-bikbov-en.xml	120,250	1,240	1,434
7	bologna-2010-07-01-bot-en.xml	50,249	2,185	2,775
8	bologna-2010-07-01-calderwilliams-en.xml	82,043	2,735	4,216
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	87,721	1,790	2,045
11	bologna-2010-07-01-herwigEtAl-en.xml	86,093	2,208	3,328
12	bologna-2010-07-01-lichtenbergerh-en.xml	59,405	1,747	2,313
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	104,010	2,261	2,552

14	bologna-2010-07-01-munch-en.xml	89,185	1,652	2,134
15	bologna-2010-07-01-power-en.xml	125,547	2,369	3,848
16	bologna-2010-07-01-samalavicius-en.xml	88,011	1,796	2,227
17	bologna-2010-07-01-santos-en.xml	242,946	1,233	1,056
18	bologna-2010-07-01-schwan-en.xml	72,911	2,804	3,157
19	bologna-2010-07-01-vonosten-en.xml	51,971	2,260	2,887
20	1117-2007-07-06-lapin1-en.xml	59,296	6,588	7,904
21	1270-2008-04-09-miklosi-en.xml	87,721	2,437	4,230
22	2100-2011-09-27-scruton-en.xml	65,841	4,115	5,117
23	211-2002-12-20-verene-en.xml	115,352	1,625	2,923
24	2163-2012-01-11-ohlheiser-en.xml	98,306	2,979	6,131
25	2200-2012-03-20-mondediploo-en.xml	58,566	7,321	8,853
26	223-2003-01-31-des-en.xml	66,106	2,361	2,616
27	2447-2013-04-12-sanchez-en.xml	69,412	4,083	5,770
28	2495-2013-06-25-zhurzhenko-en.xml	178,227	1,876	4,607
29	2517-2013-08-13-osteuropa-en.xml	68,037	3,581	5,207
30	266-2003-02-16-mangasassen-en.xml	101,637	1,639	1,797
31	2666-2014-04-03-knausgard-en.xml	329,437	1,725	1,944
32	335-2003-05-15-henard-en.xml	80,716	5,045	8,133
33	441-2003-11-28-abraham-en.xml	65,276	3,108	3,865
34	479-2004-03-03-senyener-en.xml	72,256	3,142	6,464
35	505-2004-04-05-uys-en.xml	83,379	3,790	5,936
36	62-2001-04-01-mistry-en.xml	211,974	2,524	4,551
37	661-2005-07-14-revista-en.xml	61,327	4,088	4,647
38	785-2006-02-16-sambrook-en.xml	88,843	2,776	4,120

B.1.2. Hervorhebung von zusammenfassenden Anteilen

Parameter	Wert	Durchschnitt Precision	
Korpus	A		
Abstract	10	Optimal	
Title	8	Euklid TF	66,43%
Subheadings	5	Euklid TF (norm)	74,77%
Body	1	Euklid TF-IDF	73,46%

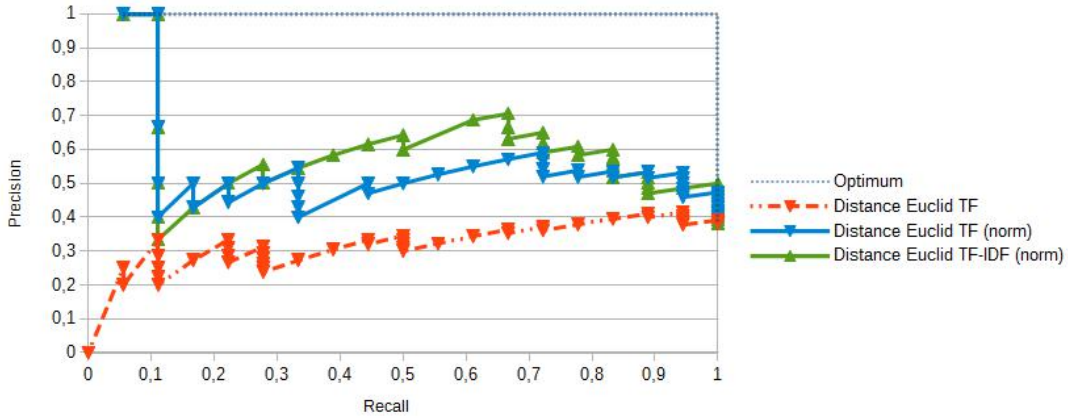


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	211,059	2,741	4,330
3	2008-05-02-wennerhag-en.xml	236,753	1,821	2,454
4	2008-11-21-leggewiewelzer-en.xml	172,020	3,584	6,455
5	2009-04-21-fraser-en.xml	244,049	2,141	3,165
6	2009-07-14-biscione-en.xml	209,136	3,428	6,294
7	2009-09-09-kavaliauskas-en.xml	194,160	3,406	6,223
8	2010-09-14-ditchev-en.xml	170,068	4,724	8,318
9	2011-07-11-bluhdorn-en.xml	188,738	2,169	3,785
10	2011-11-02-G1000-en.xml	183,649	2,449	4,386
11	2011-11-10-sierakowski-en.xml	173,196	3,149	5,560
12	2011-12-19-amirpur-en.xml	208,607	2,575	4,931

13	2012-01-25-halmai-en.xml	265,863	1,583	2,166
14	2012-02-08-elsenhans-en.xml	204,546	2,922	4,643
15	2012-09-05-jahanbegloo-en.xml	193,716	3,798	6,904
16	2012-11-21-holmes-en.xml	176,559	2,207	3,955
17	2013-02-08-wallerstein-en.xml	195,740	3,764	5,745
18	2013-02-19-leggewie-en.xml	241,762	3,060	5,051
19	2013-02-26-james-en.xml	186,917	4,793	8,331
20	2013-05-03-muller-en.xml	196,835	2,982	4,936
21	2013-06-14-pomerantsev-en.xml	196,997	3,863	7,454
22	2013-07-29-gole-en.xml	229,399	2,012	3,832
23	2013-08-13-krastev-en.xml	181,436	3,860	6,933
24	2013-08-20-leggewienanz-en.xml	212,619	3,222	4,882
25	2013-09-11-deniztekin-en.xml	183,437	7,055	11,775
26	2013-11-08-vidanava-en.xml	181,645	5,190	9,194
27	2013-11-22-offe-en.xml	174,270	2,293	3,928
28	2013-12-12-margetts-en.xml	206,698	2,349	3,765
29	2013-12-12-pogonyi-en.xml	191,742	3,364	5,379
30	1117-2007-07-06-lapin1-en.xml	190,764	7,065	11,685
31	1193-2007-11-02-boulbina-en.xml	218,451	4,551	8,331
32	1270-2008-04-09-miklosi-en.xml	194,319	4,318	7,390
33	1344-2008-08-07-seymour-en.xml	262,698	2,550	4,863
34	2100-2011-09-27-scruton-en.xml	187,361	6,044	9,840
35	211-2002-12-20-verene-en.xml	213,434	2,223	3,817
36	2163-2012-01-11-ohlheiser-en.xml	189,549	3,717	6,994
37	2200-2012-03-20-mondediploo-en.xml	180,983	22,623	35,986
38	223-2003-01-31-des-en.xml	201,301	2,917	4,482
39	2447-2013-04-12-sanchez-en.xml	186,757	7,470	12,547
40	2495-2013-06-25-zhurzhenko-en.xml	240,988	2,191	5,615
41	2517-2013-08-13-osteuropa-en.xml	178,748	8,512	13,780
42	256-2003-02-11-kaplinski-en.xml	191,612	5,179	8,199
43	266-2003-02-16-mangasassen-en.xml	189,481	3,056	5,510
44	2666-2014-04-03-knausgard-en.xml	357,249	1,870	2,336
45	294-2003-03-04-ursic-en.xml	184,600	23,075	34,921

46	335-2003-05-15-henard-en.xml	176,929	7,077	12,542
47	414-2003-10-20-bogdanovic-en.xml	232,230	2,037	3,067
48	441-2003-11-28-abraham-en.xml	188,340	6,726	10,668
49	479-2004-03-03-senyener-en.xml	191,726	5,991	11,211
50	480-2004-03-04-cakmak-en.xml	202,598	5,332	9,963
51	505-2004-04-05-uys-en.xml	181,039	4,526	7,924
52	540-2004-06-21-peters-en.xml	276,317	2,303	2,965
53	62-2001-04-01-mistry-en.xml	286,800	2,706	4,605
54	661-2005-07-14-revista-en.xml	183,581	12,239	19,052
55	772-2006-02-01-boutang-en.xml	187,310	4,460	7,532
56	785-2006-02-16-sambrook-en.xml	195,760	4,775	8,570
57	80-2001-11-14-blecher-en.xml	181,135	5,489	9,134
58	904-2006-08-17-eder-en.xml	206,746	3,230	5,100

Parameter	Wert	Durchschnitt Precision	
Korpus	B		
Abstract	10	Optimal	
Title	8	Euklid TF	40,42%
Subheadings	5	Euklid TF (norm)	68,60%
Body	1	Euklid TF-IDF	73,29%

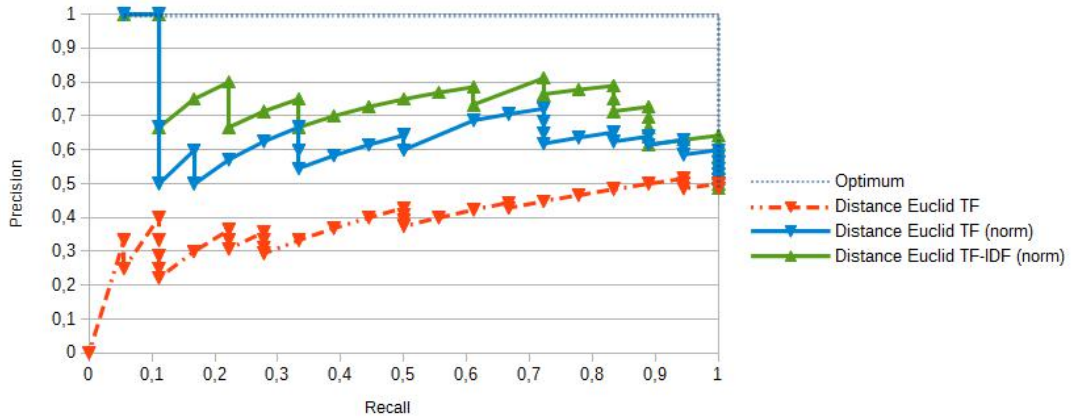


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	187,502	1,720	2,277
2	bologna-2010-03-11-manchev-en.xml	115,862	2,519	4,043
3	bologna-2010-04-13-dorling-en.xml	131,427	2,987	4,700
4	bologna-2010-07-01-balan-en.xml	72,083	2,060	3,573
5	bologna-2010-07-01-benhamou-en.xml	119,105	2,290	3,235
6	bologna-2010-07-01-bikbov-en.xml	162,253	1,330	1,634
7	bologna-2010-07-01-bot-en.xml	75,631	1,939	2,942
8	bologna-2010-07-01-calderwilliams-en.xml	110,127	2,824	5,433
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	122,381	2,353	3,102
11	bologna-2010-07-01-herwigEtAl-en.xml	130,618	2,292	3,976
12	bologna-2010-07-01-lichtenbergerh-en.xml	97,026	2,256	3,317
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	145,812	1,823	2,272

14	bologna-2010-07-01-munch-en.xml	139,377	2,080	2,638
15	bologna-2010-07-01-power-en.xml	154,570	2,916	5,421
16	bologna-2010-07-01-samalavicius-en.xml	131,244	2,263	3,419
17	bologna-2010-07-01-santos-en.xml	292,062	1,248	1,200
18	bologna-2010-07-01-schwan-en.xml	104,221	2,481	3,245
19	bologna-2010-07-01-vonosten-en.xml	95,042	1,980	2,880
20	1117-2007-07-06-lapin1-en.xml	82,310	3,049	5,813
21	1193-2007-11-02-boulbina-en.xml	142,818	2,975	6,309
22	1270-2008-04-09-miklosi-en.xml	107,033	2,379	4,587
23	1344-2008-08-07-seymour-en.xml	226,985	2,204	3,973
24	2100-2011-09-27-scruton-en.xml	92,380	2,980	4,864
25	211-2002-12-20-verene-en.xml	139,011	1,448	2,671
26	2163-2012-01-11-ohlheiser-en.xml	125,646	2,464	5,391
27	2200-2012-03-20-mondediploo-en.xml	64,537	8,067	12,514
28	223-2003-01-31-des-en.xml	108,904	1,578	2,230
29	2447-2013-04-12-sanchez-en.xml	97,509	3,900	6,722
30	2495-2013-06-25-zhurzhenko-en.xml	219,374	1,994	5,057
31	2517-2013-08-13-osteuropa-en.xml	72,877	3,470	5,637
32	256-2003-02-11-kaplinski-en.xml	109,695	2,965	3,904
33	266-2003-02-16-mangasassen-en.xml	125,702	2,027	3,538
34	2666-2014-04-03-knausgard-en.xml	343,029	1,796	1,945
35	294-2003-03-04-ursic-en.xml	62,936	7,867	8,739
36	335-2003-05-15-henard-en.xml	100,439	4,018	7,446
37	414-2003-10-20-bogdanovic-en.xml	170,438	1,495	1,979
38	441-2003-11-28-abraham-en.xml	82,462	2,945	4,887
39	479-2004-03-03-senyener-en.xml	92,644	2,895	7,585
40	480-2004-03-04-cakmak-en.xml	121,417	3,195	7,046
41	505-2004-04-05-uys-en.xml	105,513	2,638	4,374
42	540-2004-06-21-peters-en.xml	230,805	1,923	1,997
43	62-2001-04-01-mistry-en.xml	261,878	2,471	3,997
44	661-2005-07-14-revista-en.xml	66,633	4,442	6,235
45	772-2006-02-01-boutang-en.xml	104,752	2,494	4,299
46	785-2006-02-16-sambrook-en.xml	125,228	3,054	6,147

47	80-2001-11-14-blecher-en.xml	97,417	2,952	4,266
48	904-2006-08-17-eder-en.xml	127,318	1,989	3,535

Parameter	Wert	Durchschnitt Precision	
Korpus	C		
Abstract	10	Optimal	
Title	8	Euklid TF	43,83%
Subheadings	5	Euklid TF (norm)	75,19%
Body	1	Euklid TF-IDF	83,49%

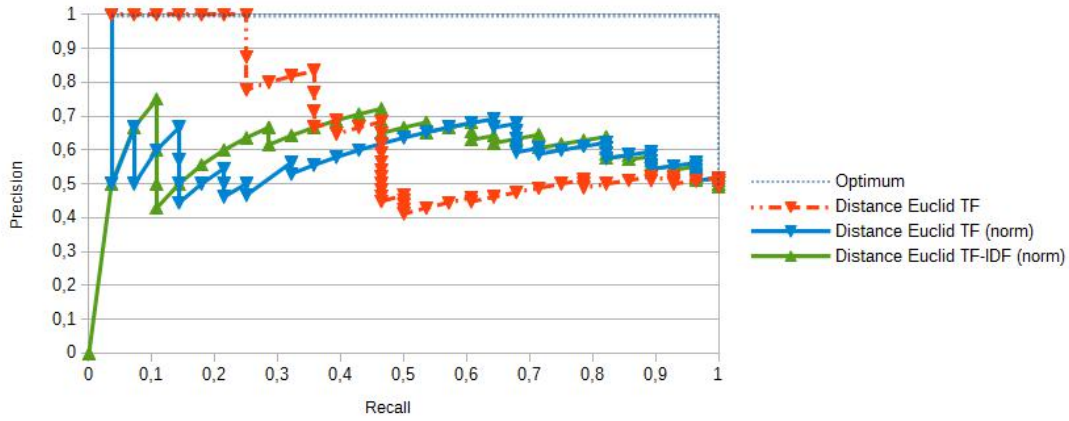


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	187,502	1,720	2,160
2	bologna-2010-03-11-manchev-en.xml	115,862	2,519	4,080
3	bologna-2010-04-13-dorling-en.xml	131,427	2,987	4,406
4	bologna-2010-07-01-balan-en.xml	72,083	2,060	3,322
5	bologna-2010-07-01-benhamou-en.xml	119,105	2,290	3,072
6	bologna-2010-07-01-bikbov-en.xml	162,253	1,330	1,517
7	bologna-2010-07-01-bot-en.xml	75,631	1,939	2,786
8	bologna-2010-07-01-calderwilliams-en.xml	110,127	2,824	5,297
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	122,381	2,353	2,878
11	bologna-2010-07-01-herwigEtAl-en.xml	130,618	2,292	3,741
12	bologna-2010-07-01-lichtenbergerh-en.xml	97,026	2,256	3,125
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	145,812	1,823	2,164

14	bologna-2010-07-01-munch-en.xml	139,377	2,080	2,485
15	bologna-2010-07-01-power-en.xml	154,570	2,916	5,158
16	bologna-2010-07-01-samalavicius-en.xml	131,244	2,263	3,245
17	bologna-2010-07-01-santos-en.xml	292,062	1,248	1,162
18	bologna-2010-07-01-schwan-en.xml	104,221	2,481	3,050
19	bologna-2010-07-01-vonosten-en.xml	95,042	1,980	2,708
20	1117-2007-07-06-lapin1-en.xml	82,310	3,049	5,508
21	1270-2008-04-09-miklosi-en.xml	107,033	2,379	4,434
22	2100-2011-09-27-scruton-en.xml	92,380	2,980	4,629
23	211-2002-12-20-verene-en.xml	139,011	1,448	2,762
24	2163-2012-01-11-ohlheiser-en.xml	125,646	2,464	5,527
25	2200-2012-03-20-mondediploo-en.xml	64,537	8,067	11,885
26	223-2003-01-31-des-en.xml	108,904	1,578	2,176
27	2447-2013-04-12-sanchez-en.xml	97,509	3,900	6,475
28	2495-2013-06-25-zhurzhenko-en.xml	219,374	1,994	4,787
29	2517-2013-08-13-osteuropa-en.xml	72,877	3,470	5,355
30	266-2003-02-16-mangasassen-en.xml	125,702	2,027	3,543
31	2666-2014-04-03-knausgard-en.xml	343,029	1,796	2,029
32	335-2003-05-15-henard-en.xml	100,439	4,018	7,226
33	441-2003-11-28-abraham-en.xml	82,462	2,945	4,676
34	479-2004-03-03-senyener-en.xml	92,644	2,895	7,301
35	505-2004-04-05-uys-en.xml	105,513	2,638	4,322
36	62-2001-04-01-mistry-en.xml	261,878	2,471	4,092
37	661-2005-07-14-revista-en.xml	66,633	4,442	5,868
38	785-2006-02-16-sambrook-en.xml	125,228	3,054	5,893

B.1.3. Hervorhebung des Textanteils

Parameter	Wert	Durchschnitt Precision	
Korpus	A		
Abstract	1	Optimal	
Title	1	Euklid TF	73,07%
Subheadings	1	Euklid TF (norm)	69,35%
Body	10	Euklid TF-IDF	70,57%

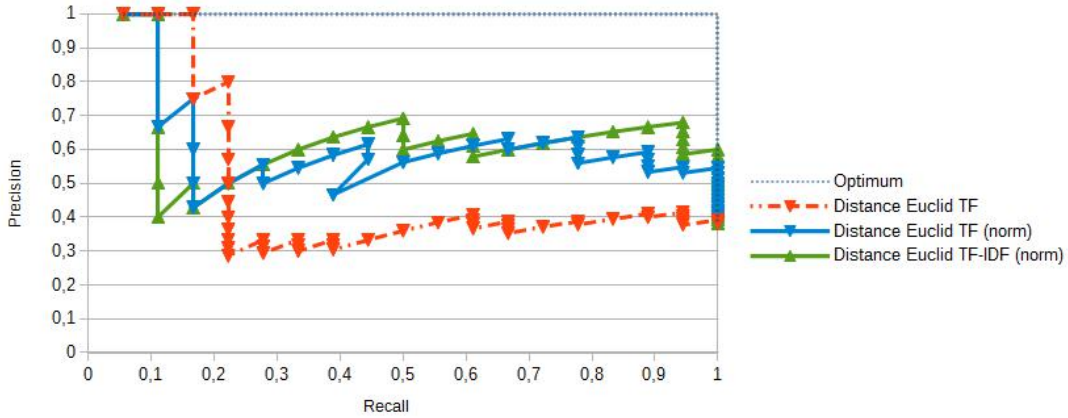


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	1720,259	3,740	5,210
3	2008-05-02-wennerhag-en.xml	1784,126	2,163	2,701
4	2008-11-21-leggewiewelzer-en.xml	1348,188	4,781	7,343
5	2009-04-21-fraser-en.xml	2015,951	1,768	2,232
6	2009-07-14-biscione-en.xml	1559,091	3,784	6,556
7	2009-09-09-kavaliauskas-en.xml	1529,862	3,625	5,814
8	2010-09-14-ditchev-en.xml	1357,022	8,377	13,153
9	2011-07-11-bluhdorn-en.xml	1477,597	2,716	4,263
10	2011-11-02-G1000-en.xml	1330,569	2,766	4,488
11	2011-11-10-sierakowski-en.xml	1412,052	2,567	3,827
12	2011-12-19-amirpur-en.xml	1595,001	2,987	5,535

13	2012-01-25-halmai-en.xml	2105,165	1,577	2,009
14	2012-02-08-elsenhans-en.xml	1619,140	5,223	7,430
15	2012-09-05-jahanbegloo-en.xml	1477,682	4,451	7,245
16	2012-11-21-holmes-en.xml	1418,072	2,564	4,048
17	2013-02-08-wallerstein-en.xml	1554,374	5,164	7,024
18	2013-02-19-leggewie-en.xml	1848,528	2,934	4,453
19	2013-02-26-james-en.xml	1465,557	6,913	10,614
20	2013-05-03-muller-en.xml	1547,213	3,158	4,576
21	2013-06-14-pomerantsev-en.xml	1592,025	3,122	5,280
22	2013-07-29-gole-en.xml	1767,486	2,094	3,775
23	2013-08-13-krastev-en.xml	1394,542	6,641	9,930
24	2013-08-20-leggewienanz-en.xml	1605,002	2,627	3,749
25	2013-09-11-deniztekin-en.xml	1450,600	17,690	24,684
26	2013-11-08-vidanava-en.xml	1453,146	8,548	12,770
27	2013-11-22-offe-en.xml	1382,601	2,460	3,813
28	2013-12-12-margetts-en.xml	1591,943	2,434	3,628
29	2013-12-12-pogonyi-en.xml	1455,593	4,968	7,112
30	1117-2007-07-06-lapin1-en.xml	1532,432	21,284	28,393
31	1193-2007-11-02-boulbina-en.xml	1657,161	4,350	7,131
32	1270-2008-04-09-miklosi-en.xml	1569,737	4,472	6,746
33	1344-2008-08-07-seymour-en.xml	2260,365	2,563	4,674
34	2100-2011-09-27-scruton-en.xml	1488,568	9,858	13,690
35	211-2002-12-20-verene-en.xml	1708,447	2,501	3,962
36	2163-2012-01-11-ohlheiser-en.xml	1516,909	4,862	8,216
37	2200-2012-03-20-mondediploo-en.xml	1491,896	18,649	25,484
38	223-2003-01-31-des-en.xml	1535,294	6,533	8,702
39	2447-2013-04-12-sanchez-en.xml	1462,909	9,086	13,149
40	2495-2013-06-25-zhurzhenko-en.xml	1910,893	2,070	5,321
41	2517-2013-08-13-osteuropa-en.xml	1475,965	7,768	11,104
42	256-2003-02-11-kaplinski-en.xml	1526,812	5,872	8,357
43	266-2003-02-16-mangasassen-en.xml	1507,929	2,432	3,427
44	2666-2014-04-03-knausgard-en.xml	3287,567	1,721	2,000
45	294-2003-03-04-ursic-en.xml	1535,276	30,706	40,182

46	335-2003-05-15-henard-en.xml	1425,960	9,443	14,702
47	414-2003-10-20-bogdanovic-en.xml	1840,508	2,409	3,256
48	441-2003-11-28-abraham-en.xml	1524,095	7,583	10,264
49	479-2004-03-03-senyener-en.xml	1545,178	6,992	10,821
50	480-2004-03-04-cakmak-en.xml	1522,692	5,419	7,986
51	505-2004-04-05-uys-en.xml	1431,912	7,089	11,340
52	540-2004-06-21-peters-en.xml	1909,898	2,236	2,746
53	62-2001-04-01-mistry-en.xml	2223,978	2,648	4,767
54	661-2005-07-14-revista-en.xml	1518,411	10,123	13,617
55	772-2006-02-01-boutang-en.xml	1457,953	10,195	15,171
56	785-2006-02-16-sambrook-en.xml	1463,092	4,704	7,139
57	80-2001-11-14-blecher-en.xml	1478,880	4,481	6,489
58	904-2006-08-17-eder-en.xml	1529,525	4,459	6,162

Parameter	Wert	Durchschnitt Precision	
Korpus	B		
Abstract	1	Optimal	
Title	1	Euklid TF	58,53%
Subheadings	1	Euklid TF (norm)	74,40%
Body	10	Euklid TF-IDF	76,51%

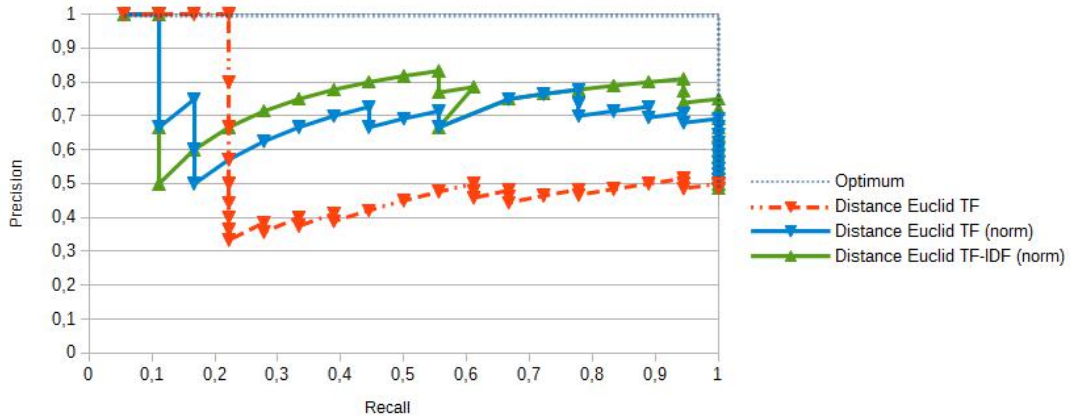


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	1478,910	1,754	2,305
2	bologna-2010-03-11-manchev-en.xml	907,693	2,514	3,589
3	bologna-2010-04-13-dorling-en.xml	1100,378	2,501	3,741
4	bologna-2010-07-01-balan-en.xml	489,768	2,321	3,303
5	bologna-2010-07-01-benhamou-en.xml	744,248	1,768	2,333
6	bologna-2010-07-01-bikbov-en.xml	1170,164	1,241	1,537
7	bologna-2010-07-01-bot-en.xml	493,977	2,330	3,106
8	bologna-2010-07-01-calderwilliams-en.xml	810,077	2,784	4,303
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	860,880	1,757	2,147
11	bologna-2010-07-01-herwigEtAl-en.xml	831,788	2,236	3,537
12	bologna-2010-07-01-lichtenbergerh-en.xml	582,712	1,760	2,422
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	1024,099	2,415	2,874

14	bologna-2010-07-01-munch-en.xml	850,011	1,628	2,253
15	bologna-2010-07-01-power-en.xml	1234,362	2,329	3,914
16	bologna-2010-07-01-samalavicius-en.xml	853,650	1,775	2,263
17	bologna-2010-07-01-santos-en.xml	2372,589	1,238	1,107
18	bologna-2010-07-01-schwan-en.xml	707,566	2,924	3,473
19	bologna-2010-07-01-vonosten-en.xml	495,719	2,338	3,149
20	1117-2007-07-06-lapin1-en.xml	583,187	8,100	9,821
21	1193-2007-11-02-boulbina-en.xml	1006,309	2,641	5,266
22	1270-2008-04-09-miklosi-en.xml	867,840	2,472	4,376
23	1344-2008-08-07-seymour-en.xml	2056,790	2,332	4,028
24	2100-2011-09-27-scruton-en.xml	645,144	4,272	5,520
25	211-2002-12-20-verene-en.xml	1133,827	1,660	2,882
26	2163-2012-01-11-ohlheiser-en.xml	964,204	3,090	6,380
27	2200-2012-03-20-mondediploo-en.xml	583,051	7,288	9,247
28	223-2003-01-31-des-en.xml	638,186	2,716	3,143
29	2447-2013-04-12-sanchez-en.xml	678,851	4,216	6,118
30	2495-2013-06-25-zhurzhenko-en.xml	1728,353	1,873	4,882
31	2517-2013-08-13-osteuropa-en.xml	676,188	3,559	5,418
32	256-2003-02-11-kaplinski-en.xml	876,783	3,372	4,238
33	266-2003-02-16-mangasassen-en.xml	1007,575	1,625	1,726
34	2666-2014-04-03-knausgard-en.xml	3284,102	1,719	1,842
35	294-2003-03-04-ursic-en.xml	578,001	11,560	12,081
36	335-2003-05-15-henard-en.xml	794,280	5,260	8,713
37	414-2003-10-20-bogdanovic-en.xml	1332,436	1,744	2,157
38	441-2003-11-28-abraham-en.xml	644,221	3,205	4,158
39	479-2004-03-03-senyener-en.xml	712,745	3,225	6,643
40	480-2004-03-04-cakmak-en.xml	762,748	2,714	4,542
41	505-2004-04-05-uys-en.xml	818,915	4,054	6,460
42	540-2004-06-21-peters-en.xml	1503,085	1,760	1,751
43	62-2001-04-01-mistry-en.xml	2074,386	2,470	4,346
44	661-2005-07-14-revista-en.xml	610,202	4,068	4,873
45	772-2006-02-01-boutang-en.xml	690,120	4,826	7,667
46	785-2006-02-16-sambrook-en.xml	867,519	2,789	4,152

47	80-2001-11-14-blecher-en.xml	823,446	2,495	3,176
48	904-2006-08-17-eder-en.xml	733,231	2,138	3,246

Parameter	Wert	Durchschnitt Precision	
Korpus	C		
Abstract	1	Optimal	
Title	1	Euklid TF	63,33%
Subheadings	1	Euklid TF (norm)	63,77%
Body	10	Euklid TF-IDF	84,52%

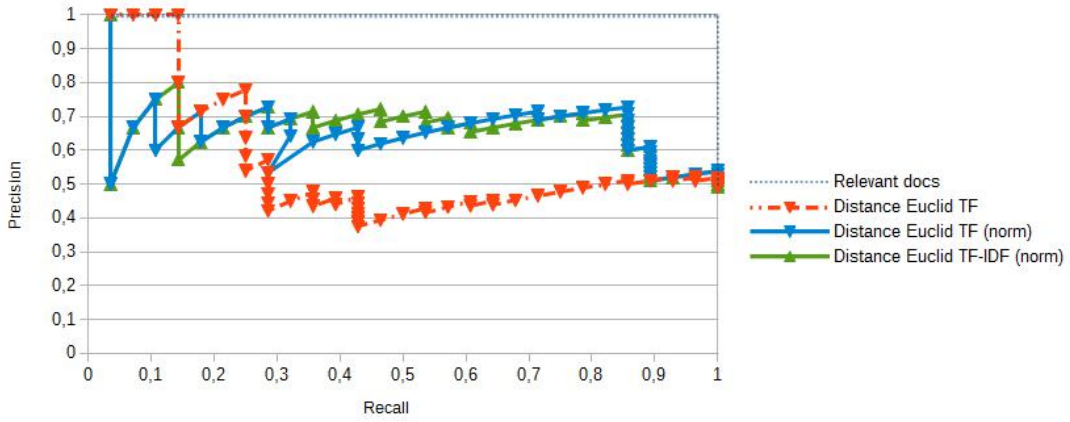


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en,xml	1478,910	1,754	2,206
2	bologna-2010-03-11-manchev-en,xml	907,693	2,514	3,586
3	bologna-2010-04-13-dorling-en,xml	1100,378	2,501	3,507
4	bologna-2010-07-01-balan-en,xml	489,768	2,321	3,125
5	bologna-2010-07-01-benhamou-en,xml	744,248	1,768	2,223
6	bologna-2010-07-01-bikbov-en,xml	1170,164	1,241	1,447
7	bologna-2010-07-01-bot-en,xml	493,977	2,330	2,954
8	bologna-2010-07-01-calderwilliams-en,xml	810,077	2,784	4,188
9	bologna-2010-07-01-editorial-en,xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en,xml	860,880	1,757	2,007
11	bologna-2010-07-01-herwigEtAl-en,xml	831,788	2,236	3,336
12	bologna-2010-07-01-lichtenbergerh-en,xml	582,712	1,760	2,328
13	bologna-2010-07-01-lichtenbergerpadis-en,xml	1024,099	2,415	2,725

14	bologna-2010-07-01-munch-en,xml	850,011	1,628	2,145
15	bologna-2010-07-01-power-en,xml	1234,362	2,329	3,718
16	bologna-2010-07-01-samalavicius-en,xml	853,650	1,775	2,167
17	bologna-2010-07-01-santos-en,xml	2372,589	1,238	1,059
18	bologna-2010-07-01-schwan-en,xml	707,566	2,924	3,314
19	bologna-2010-07-01-vonosten-en,xml	495,719	2,338	2,981
20	1117-2007-07-06-lapin1-en,xml	583,187	8,100	9,267
21	1270-2008-04-09-miklosi-en,xml	867,840	2,472	4,272
22	2100-2011-09-27-scruton-en,xml	645,144	4,272	5,238
23	211-2002-12-20-verene-en,xml	1133,827	1,660	2,964
24	2163-2012-01-11-ohlheiser-en,xml	964,204	3,090	6,299
25	2200-2012-03-20-mondediploo-en,xml	583,051	7,288	8,784
26	223-2003-01-31-des-en,xml	638,186	2,716	3,012
27	2447-2013-04-12-sanchez-en,xml	678,851	4,216	5,883
28	2495-2013-06-25-zhurzhenko-en,xml	1728,353	1,873	4,609
29	2517-2013-08-13-osteuropa-en,xml	676,188	3,559	5,149
30	266-2003-02-16-mangasassen-en,xml	1007,575	1,625	1,731
31	2666-2014-04-03-knausgard-en,xml	3284,102	1,719	1,940
32	335-2003-05-15-henard-en,xml	794,280	5,260	8,410
33	441-2003-11-28-abraham-en,xml	644,221	3,205	3,936
34	479-2004-03-03-senyener-en,xml	712,745	3,225	6,507
35	505-2004-04-05-uys-en,xml	818,915	4,054	6,329
36	62-2001-04-01-mistry-en,xml	2074,386	2,470	4,489
37	661-2005-07-14-revista-en,xml	610,202	4,068	4,611
38	785-2006-02-16-sambrook-en,xml	867,519	2,789	4,007

B.1.4. Hervorhebung von Überschriften

Parameter	Wert	Durchschnitt Precision	
Korpus	A		
Abstract	1	Optimal	
Title	15	Euklid TF	60,41%
Subheadings	15	Euklid TF (norm)	78,00%
Body	1	Euklid TF-IDF	77,76%

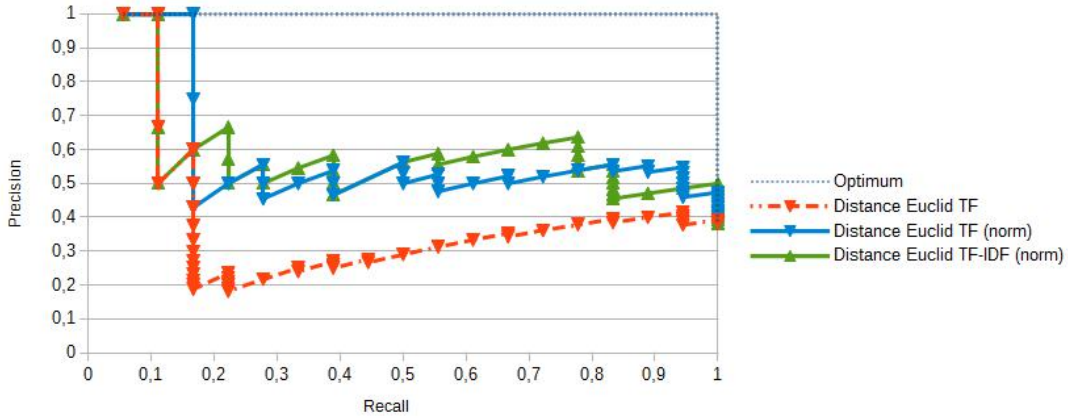


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	214,560	2,823	4,585
3	2008-05-02-wennerhag-en.xml	215,866	2,137	3,144
4	2008-11-21-leggewiewelzer-en.xml	185,078	2,682	4,720
5	2009-04-21-fraser-en.xml	231,635	2,032	3,005
6	2009-07-14-biscione-en.xml	217,989	2,986	5,912
7	2009-09-09-kavaliauskas-en.xml	209,239	2,906	5,998
8	2010-09-14-ditchev-en.xml	173,476	6,939	12,277
9	2011-07-11-bluhdorn-en.xml	200,856	2,336	4,004
10	2011-11-02-G1000-en.xml	181,444	3,128	5,682
11	2011-11-10-sierakowski-en.xml	175,519	3,191	5,604
12	2011-12-19-amirpur-en.xml	222,560	2,248	4,058

13	2012-01-25-halmai-en.xml	273,733	1,521	2,095
14	2012-02-08-elsenhans-en.xml	228,118	2,037	3,198
15	2012-09-05-jahanbegloo-en.xml	192,904	3,937	7,250
16	2012-11-21-holmes-en.xml	196,316	2,727	5,127
17	2013-02-08-wallerstein-en.xml	195,643	4,348	6,817
18	2013-02-19-leggewie-en.xml	273,004	2,167	3,823
19	2013-02-26-james-en.xml	187,353	5,064	8,996
20	2013-05-03-muller-en.xml	200,430	3,084	5,083
21	2013-06-14-pomerantsev-en.xml	193,894	3,802	7,137
22	2013-07-29-gole-en.xml	239,829	2,067	3,969
23	2013-08-13-krastev-en.xml	199,178	3,905	6,950
24	2013-08-20-leggewienanz-en.xml	220,549	2,902	4,513
25	2013-09-11-deniztekin-en.xml	185,014	7,709	12,738
26	2013-11-08-vidanava-en.xml	183,426	10,790	18,493
27	2013-11-22-offe-en.xml	204,834	2,468	3,854
28	2013-12-12-margetts-en.xml	224,537	1,796	2,811
29	2013-12-12-pogonyi-en.xml	185,515	4,033	6,689
30	1117-2007-07-06-lapin1-en.xml	191,120	11,242	18,420
31	1193-2007-11-02-boulbina-en.xml	202,926	5,203	9,153
32	1270-2008-04-09-miklosi-en.xml	194,052	5,390	9,336
33	1344-2008-08-07-seymour-en.xml	270,459	2,601	5,052
34	2100-2011-09-27-scruton-en.xml	187,190	10,399	17,445
35	211-2002-12-20-verene-en.xml	210,886	2,481	4,317
36	2163-2012-01-11-ohlheiser-en.xml	185,712	5,628	10,274
37	2200-2012-03-20-mondediploo-en.xml	188,804	12,587	21,489
38	223-2003-01-31-des-en.xml	198,043	3,536	5,543
39	2447-2013-04-12-sanchez-en.xml	187,768	6,057	10,265
40	2495-2013-06-25-zhurzhenko-en.xml	275,020	2,007	5,240
41	2517-2013-08-13-osteuropa-en.xml	188,494	6,080	10,190
42	256-2003-02-11-kaplinski-en.xml	195,156	3,983	6,374
43	266-2003-02-16-mangasassen-en.xml	189,127	3,050	5,212
44	2666-2014-04-03-knausgard-en.xml	349,208	1,828	2,340
45	294-2003-03-04-ursic-en.xml	190,266	12,684	19,969

46	335-2003-05-15-henard-en.xml	183,385	6,792	11,805
47	414-2003-10-20-bogdanovic-en.xml	223,540	2,378	3,653
48	441-2003-11-28-abraham-en.xml	192,260	5,493	8,826
49	479-2004-03-03-senyener-en.xml	193,375	8,408	14,921
50	480-2004-03-04-cakmak-en.xml	190,174	6,558	11,237
51	505-2004-04-05-uys-en.xml	181,849	8,266	14,650
52	540-2004-06-21-peters-en.xml	264,966	2,572	3,567
53	62-2001-04-01-mistry-en.xml	284,991	2,689	5,050
54	661-2005-07-14-revista-en.xml	190,095	12,673	20,979
55	772-2006-02-01-boutang-en.xml	186,301	6,010	10,206
56	785-2006-02-16-sambrook-en.xml	186,489	5,828	10,056
57	80-2001-11-14-blecher-en.xml	185,742	5,629	9,587
58	904-2006-08-17-eder-en.xml	191,614	5,179	8,361

Parameter	Wert	Durchschnitt Precision	
Korpus	B		
Abstract	1	Optimal	
Title	15	Euklid TF	47,56%
Subheadings	15	Euklid TF (norm)	71,23%
Body	1	Euklid TF-IDF	72,4%

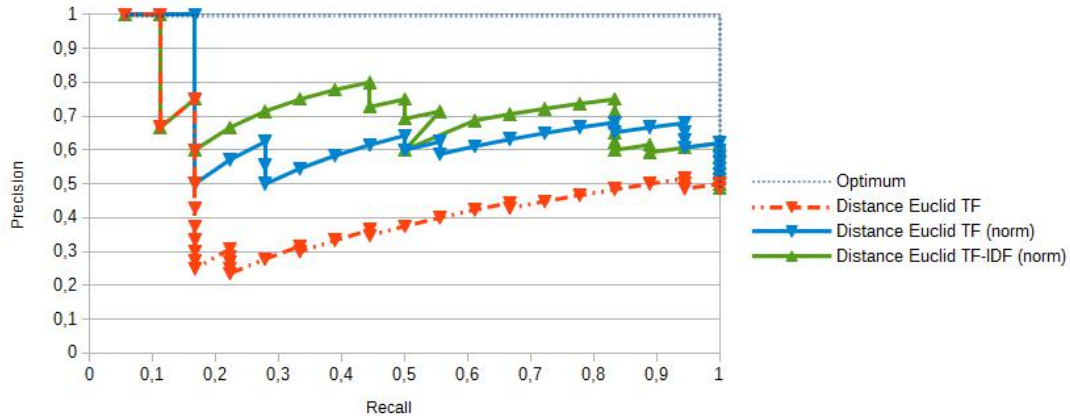


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	184,870	1,830	2,552
2	bologna-2010-03-11-manchev-en.xml	113,230	2,573	4,105
3	bologna-2010-04-13-dorling-en.xml	132,499	2,598	4,059
4	bologna-2010-07-01-balan-en.xml	64,444	2,685	4,544
5	bologna-2010-07-01-benhamou-en.xml	120,033	2,791	4,072
6	bologna-2010-07-01-bikbov-en.xml	147,160	1,326	1,791
7	bologna-2010-07-01-bot-en.xml	69,836	1,887	2,684
8	bologna-2010-07-01-calderwilliams-en.xml	92,774	3,092	5,324
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	99,111	2,023	2,794
11	bologna-2010-07-01-herwigEtAl-en.xml	129,368	2,753	5,279
12	bologna-2010-07-01-lichtenbergerh-en.xml	74,465	2,190	2,927
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	126,925	2,115	2,676

14	bologna-2010-07-01-munch-en.xml	158,398	1,760	2,134
15	bologna-2010-07-01-power-en.xml	143,255	2,703	5,499
16	bologna-2010-07-01-samalavicius-en.xml	121,087	2,242	3,722
17	bologna-2010-07-01-santos-en.xml	303,465	1,199	1,127
18	bologna-2010-07-01-schwan-en.xml	105,546	2,639	3,405
19	bologna-2010-07-01-vonosten-en.xml	101,020	1,329	2,122
20	1117-2007-07-06-lapin1-en.xml	73,621	4,331	7,269
21	1193-2007-11-02-boulbina-en.xml	111,301	2,854	5,640
22	1270-2008-04-09-miklosi-en.xml	97,401	2,706	5,141
23	1344-2008-08-07-seymour-en.xml	233,000	2,240	4,134
24	2100-2011-09-27-scruton-en.xml	78,810	4,378	7,196
25	211-2002-12-20-verene-en.xml	129,422	1,523	2,773
26	2163-2012-01-11-ohlheiser-en.xml	104,518	3,167	6,596
27	2200-2012-03-20-mondediploo-en.xml	76,039	5,069	9,683
28	223-2003-01-31-des-en.xml	94,223	1,683	2,171
29	2447-2013-04-12-sanchez-en.xml	91,356	2,947	4,806
30	2495-2013-06-25-zhurzhenko-en.xml	256,377	1,871	4,746
31	2517-2013-08-13-osteuropa-en.xml	87,161	2,812	4,829
32	256-2003-02-11-kaplinski-en.xml	109,430	2,233	2,692
33	266-2003-02-16-mangasassen-en.xml	114,342	1,844	2,644
34	2666-2014-04-03-knausgard-en.xml	330,710	1,731	1,903
35	294-2003-03-04-ursic-en.xml	70,937	4,729	5,717
36	335-2003-05-15-henard-en.xml	95,200	3,526	5,930
37	414-2003-10-20-bogdanovic-en.xml	152,010	1,617	2,065
38	441-2003-11-28-abraham-en.xml	84,362	2,410	3,525
39	479-2004-03-03-senyener-en.xml	85,563	3,720	8,566
40	480-2004-03-04-cakmak-en.xml	88,516	3,052	5,572
41	505-2004-04-05-uys-en.xml	91,236	4,147	6,902
42	540-2004-06-21-peters-en.xml	215,133	2,089	2,305
43	62-2001-04-01-mistry-en.xml	257,047	2,425	4,405
44	661-2005-07-14-revista-en.xml	76,191	5,079	8,607
45	772-2006-02-01-boutang-en.xml	87,676	2,828	4,566
46	785-2006-02-16-sambrook-en.xml	98,270	3,071	5,098

47	80-2001-11-14-blecher-en.xml	90,890	2,754	4,206
48	904-2006-08-17-eder-en.xml	96,607	2,611	4,199

Parameter	Wert	Durchschnitt Precision	
Korpus	C		
Abstract	1	Optimal	
Title	15	Euklid TF	52,80%
Subheadings	15	Euklid TF (norm)	60,24%
Body	1	Euklid TF-IDF	80,87%

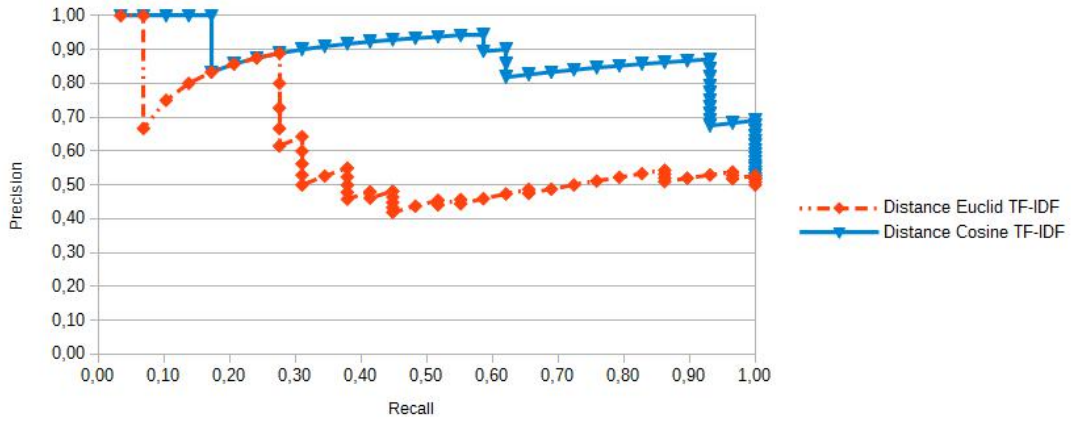


#	Dateiname	Distanz Euklid TF	Distanz Euklid TF (norm)	Distanz Euklid TF-IDF (norm)
1	bologna-2010-02-05-newfield-en.xml	184,870	1,830	2,440
2	bologna-2010-03-11-manchev-en.xml	113,230	2,573	3,972
3	bologna-2010-04-13-dorling-en.xml	132,499	2,598	3,798
4	bologna-2010-07-01-balan-en.xml	64,444	2,685	4,229
5	bologna-2010-07-01-benhamou-en.xml	120,033	2,791	3,881
6	bologna-2010-07-01-bikbov-en.xml	147,160	1,326	1,668
7	bologna-2010-07-01-bot-en.xml	69,836	1,887	2,538
8	bologna-2010-07-01-calderwilliams-en.xml	92,774	3,092	5,264
9	bologna-2010-07-01-editorial-en.xml	0,000	0,000	0,000
10	bologna-2010-07-01-gilbert-en.xml	99,111	2,023	2,584
11	bologna-2010-07-01-herwigEtAl-en.xml	129,368	2,753	4,940
12	bologna-2010-07-01-lichtenbergerh-en.xml	74,465	2,190	2,803
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	126,925	2,115	2,566

14	bologna-2010-07-01-munch-en.xml	158,398	1,760	2,071
15	bologna-2010-07-01-power-en.xml	143,255	2,703	5,206
16	bologna-2010-07-01-samalavicius-en.xml	121,087	2,242	3,524
17	bologna-2010-07-01-santos-en.xml	303,465	1,199	1,083
18	bologna-2010-07-01-schwan-en.xml	105,546	2,639	3,223
19	bologna-2010-07-01-vonosten-en.xml	101,020	1,329	2,014
20	1117-2007-07-06-lapin1-en.xml	73,621	4,331	6,760
21	1270-2008-04-09-miklosi-en.xml	97,401	2,706	4,933
22	2100-2011-09-27-scruton-en.xml	78,810	4,378	6,757
23	211-2002-12-20-verene-en.xml	129,422	1,523	2,849
24	2163-2012-01-11-ohlheiser-en.xml	104,518	3,167	6,519
25	2200-2012-03-20-mondediploo-en.xml	76,039	5,069	9,218
26	223-2003-01-31-des-en.xml	94,223	1,683	2,062
27	2447-2013-04-12-sanchez-en.xml	91,356	2,947	4,662
28	2495-2013-06-25-zhurzhenko-en.xml	256,377	1,871	4,491
29	2517-2013-08-13-osteuropa-en.xml	87,161	2,812	4,603
30	266-2003-02-16-mangasassen-en.xml	114,342	1,844	2,604
31	2666-2014-04-03-knausgard-en.xml	330,710	1,731	1,990
32	335-2003-05-15-henard-en.xml	95,200	3,526	5,752
33	441-2003-11-28-abraham-en.xml	84,362	2,410	3,314
34	479-2004-03-03-senyener-en.xml	85,563	3,720	8,217
35	505-2004-04-05-uys-en.xml	91,236	4,147	6,712
36	62-2001-04-01-mistry-en.xml	257,047	2,425	4,418
37	661-2005-07-14-revista-en.xml	76,191	5,079	8,092
38	785-2006-02-16-sambrook-en.xml	98,270	3,071	5,011

B.2. Reduktion des Feature-Vektors durch Nutzung der häufigsten Worte

Parameter	Wert	Durchschnitt Precision	
Korpus	A	Optimal	
# häufigste Terme	10	Euklid TF-IDF	68,59%
		Kosinus TF-IDF	94,18%



#	Dateiname	Distanz Euklid TF-IDF
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2013-11-22-offe-en.xml	104,547
3	266-2003-02-16-mangasassen-en.xml	106,471
4	demo-2008-11-21-leggewiewelzer-en.xml	108,231
5	demo-2011-11-02-G1000-en.xml	108,365
6	demo-2013-09-11-deniztekin-en.xml	108,471
7	demo-2011-11-10-sierakowski-en.xml	108,972
8	demo-2010-09-14-ditchev-en.xml	109,900
9	demo-2013-12-12-pogonyi-en.xml	111,772
10	335-2003-05-15-henard-en.xml	113,119
11	294-2003-03-04-ursic-en.xml	115,598
12	80-2001-11-14-blecher-en.xml	115,866
13	1117-2007-07-06-lapin1-en.xml	116,151
14	demo-2013-02-26-james-en.xml	116,327
15	2200-2012-03-20-mondediploo-en.xml	116,632

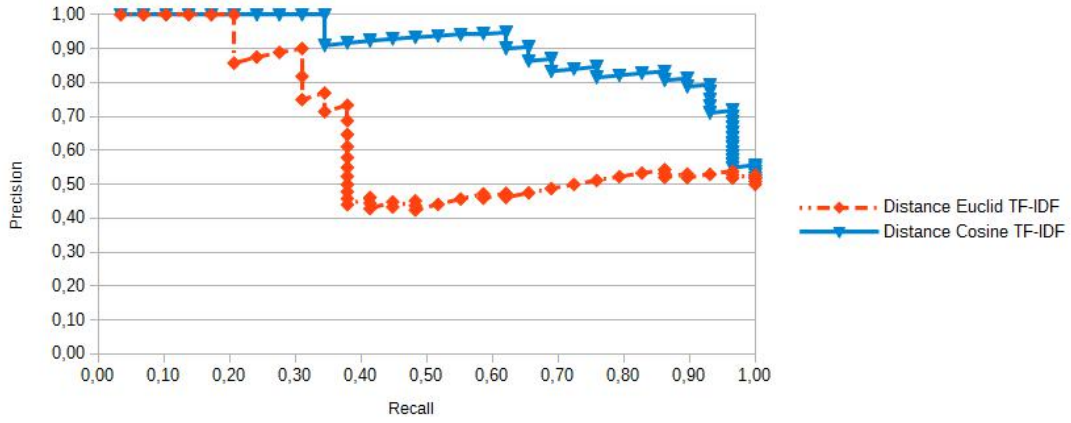
16	2100-2011-09-27-scruton-en.xml	117,090
17	441-2003-11-28-abraham-en.xml	117,188
18	772-2006-02-01-boutang-en.xml	117,716
19	demo-2009-09-09-kavaliauskas-en.xml	118,076
20	demo-2011-07-11-bluhdorn-en.xml	118,157
21	2447-2013-04-12-sanchez-en.xml	118,571
22	505-2004-04-05-uys-en.xml	118,916
23	661-2005-07-14-revista-en.xml	119,306
24	785-2006-02-16-sambrook-en.xml	119,553
25	demo-2013-08-13-krastev-en.xml	119,867
26	223-2003-01-31-des-en.xml	120,058
27	demo-2013-11-08-vidanava-en.xml	120,133
28	479-2004-03-03-senyener-en.xml	120,628
29	2517-2013-08-13-osteuropa-en.xml	121,573
30	904-2006-08-17-eder-en.xml	121,988
31	480-2004-03-04-cakmak-en.xml	122,287
32	demo-2013-06-14-pomerantsev-en.xml	122,385
33	demo-2012-11-21-holmes-en.xml	122,442
34	256-2003-02-11-kaplinski-en.xml	123,826
35	demo-2009-07-14-biscione-en.xml	125,455
36	1270-2008-04-09-miklosi-en.xml	125,535
37	demo-2013-02-08-wallerstein-en.xml	125,936
38	demo-2012-09-05-jahanbegloo-en.xml	126,099
39	demo-2013-05-03-muller-en.xml	126,238
40	2163-2012-01-11-ohlheiser-en.xml	127,656
41	demo-2012-02-08-elsenhans-en.xml	128,020
42	demo-2011-12-19-amirpur-en.xml	129,634
43	demo-2009-04-21-fraser-en.xml	134,220
44	demo-2013-08-20-leggewienanz-en.xml	134,287
45	demo-2001-11-27-rosenberg-en.xml	135,377
46	demo-2013-12-12-margetts-en.xml	138,123
47	211-2002-12-20-verene-en.xml	143,743
48	414-2003-10-20-bogdanovic-en.xml	146,636

49	1193-2007-11-02-boulbina-en.xml	150,486
50	demo-2013-02-19-leggewie-en.xml	155,952
51	demo-2013-07-29-gole-en.xml	160,941
52	demo-2008-05-02-wennerhag-en.xml	165,266
53	540-2004-06-21-peters-en.xml	167,809
54	2495-2013-06-25-zhurzhenko-en.xml	172,511
55	demo-2012-01-25-halmai-en.xml	191,031
56	1344-2008-08-07-seymour-en.xml	193,324
57	62-2001-04-01-mistry-en.xml	206,010
58	2666-2014-04-03-knausgard-en.xml	218,140

#	Dateiname	Distanz Kosinus TF-IDF
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2013-11-22-offe-en.xml	0,518
3	demo-2013-09-11-deniztekin-en.xml	0,562
4	demo-2011-11-02-G1000-en.xml	0,562
5	demo-2011-07-11-bluhdorn-en.xml	0,570
6	266-2003-02-16-mangasassen-en.xml	0,613
7	demo-2011-11-10-sierakowski-en.xml	0,616
8	demo-2008-11-21-leggewiewelzer-en.xml	0,635
9	demo-2012-11-21-holmes-en.xml	0,688
10	demo-2010-09-14-ditchev-en.xml	0,689
11	demo-2009-09-09-kavaliauskas-en.xml	0,692
12	demo-2013-12-12-pogonyi-en.xml	0,697
13	demo-2013-05-03-muller-en.xml	0,702
14	demo-2008-05-02-wennerhag-en.xml	0,717
15	demo-2001-11-27-rosenberg-en.xml	0,724
16	demo-2013-08-20-leggewienanz-en.xml	0,729
17	demo-2011-12-19-amirpur-en.xml	0,731
18	demo-2013-12-12-margetts-en.xml	0,734
19	2495-2013-06-25-zhurzhenko-en.xml	0,766
20	demo-2009-04-21-fraser-en.xml	0,769
21	335-2003-05-15-henard-en.xml	0,773
22	505-2004-04-05-uys-en.xml	0,781
23	demo-2009-07-14-biscione-en.xml	0,789
24	demo-2013-06-14-pomerantsev-en.xml	0,793
25	demo-2013-02-19-leggewie-en.xml	0,835
26	demo-2013-07-29-gole-en.xml	0,845
27	demo-2013-02-26-james-en.xml	0,846
28	demo-2013-08-13-krastev-en.xml	0,864
29	demo-2012-02-08-elsenhans-en.xml	0,873
30	demo-2012-09-05-jahanbegloo-en.xml	0,890

31	demo-2012-01-25-halmai-en.xml	0,905
32	80-2001-11-14-blecher-en.xml	0,911
33	540-2004-06-21-peters-en.xml	0,913
34	772-2006-02-01-boutang-en.xml	0,928
35	2100-2011-09-27-scruton-en.xml	0,931
36	2447-2013-04-12-sanchez-en.xml	0,933
37	785-2006-02-16-sambrook-en.xml	0,934
38	256-2003-02-11-kaplinski-en.xml	0,935
39	904-2006-08-17-eder-en.xml	0,940
40	62-2001-04-01-mistry-en.xml	0,958
41	demo-2013-11-08-vidanava-en.xml	0,958
42	demo-2013-02-08-wallerstein-en.xml	1,000
43	1117-2007-07-06-lapin1-en.xml	1,000
44	1193-2007-11-02-boulbina-en.xml	1,000
45	1270-2008-04-09-miklosi-en.xml	1,000
46	1344-2008-08-07-seymour-en.xml	1,000
47	211-2002-12-20-verene-en.xml	1,000
48	2163-2012-01-11-ohlheiser-en.xml	1,000
49	2200-2012-03-20-mondediploo-en.xml	1,000
50	223-2003-01-31-des-en.xml	1,000
51	2517-2013-08-13-osteuropa-en.xml	1,000
52	2666-2014-04-03-knausgard-en.xml	1,000
53	294-2003-03-04-ursic-en.xml	1,000
54	414-2003-10-20-bogdanovic-en.xml	1,000
55	441-2003-11-28-abraham-en.xml	1,000
56	479-2004-03-03-senyener-en.xml	1,000
57	480-2004-03-04-cakmak-en.xml	1,000
58	661-2005-07-14-revista-en.xml	1,000

Parameter	Wert	Durchschnitt Precision	
Korpus	A	Optimal	
# häufigste Terme	30	Euklid TF-IDF	72,54%
		Kosinus TF-IDF	95,01%



#	Dateiname	Distanz Euklid TF-IDF
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2008-11-21-leggewiewelzer-en.xml	116,288
3	demo-2013-11-22-offe-en.xml	120,499
4	demo-2011-11-02-G1000-en.xml	121,610
5	demo-2010-09-14-ditchev-en.xml	122,274
6	demo-2013-09-11-deniztekin-en.xml	124,435
7	266-2003-02-16-mangasassen-en.xml	124,656
8	demo-2012-11-21-holmes-en.xml	125,722
9	demo-2013-08-13-krastev-en.xml	126,167
10	demo-2011-11-10-sierakowski-en.xml	126,432
11	335-2003-05-15-henard-en.xml	127,475
12	2447-2013-04-12-sanchez-en.xml	127,832
13	demo-2013-12-12-pogonyi-en.xml	128,585
14	80-2001-11-14-blecher-en.xml	129,514
15	demo-2013-02-26-james-en.xml	131,050
16	2200-2012-03-20-mondediploo-en.xml	131,114

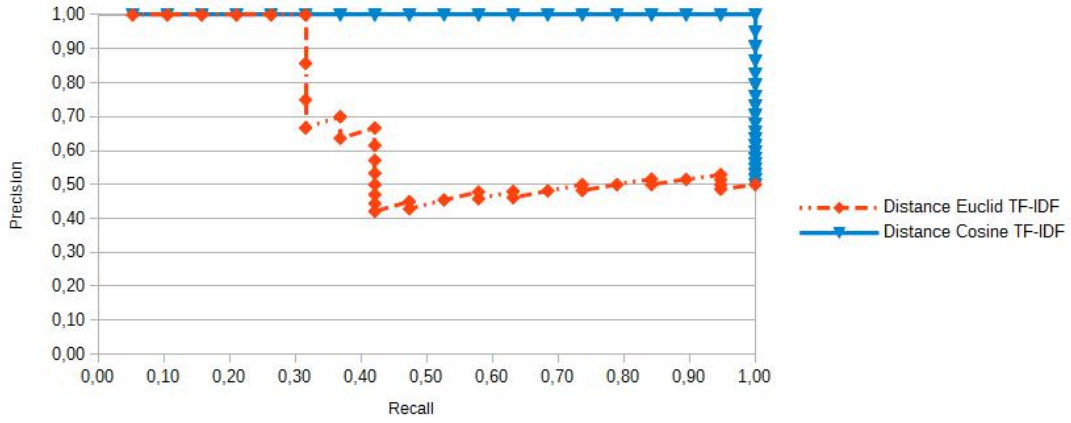
17	294-2003-03-04-ursic-en.xml	131,206
18	441-2003-11-28-abraham-en.xml	131,305
19	2100-2011-09-27-scruton-en.xml	131,370
20	1117-2007-07-06-lapin1-en.xml	131,621
21	505-2004-04-05-uys-en.xml	131,799
22	2517-2013-08-13-osteuropa-en.xml	132,575
23	772-2006-02-01-boutang-en.xml	134,045
24	661-2005-07-14-revista-en.xml	134,510
25	785-2006-02-16-sambrook-en.xml	134,618
26	demo-2013-11-08-vidanava-en.xml	135,044
27	223-2003-01-31-des-en.xml	135,610
28	904-2006-08-17-eder-en.xml	136,407
29	demo-2012-09-05-jahanbegloo-en.xml	136,539
30	479-2004-03-03-senyener-en.xml	136,547
31	demo-2011-07-11-bluhdorn-en.xml	137,015
32	256-2003-02-11-kaplinski-en.xml	137,906
33	480-2004-03-04-cakmak-en.xml	137,960
34	demo-2009-07-14-biscione-en.xml	138,268
35	demo-2013-06-14-pomerantsev-en.xml	140,004
36	demo-2009-09-09-kavaliauskas-en.xml	140,353
37	1270-2008-04-09-miklosi-en.xml	140,424
38	demo-2013-05-03-muller-en.xml	140,588
39	2163-2012-01-11-ohlheiser-en.xml	141,598
40	demo-2011-12-19-amirpur-en.xml	141,771
41	demo-2013-02-08-wallerstein-en.xml	143,265
42	demo-2012-02-08-elsenhans-en.xml	149,850
43	demo-2013-12-12-margetts-en.xml	150,612
44	demo-2013-08-20-leggewienanz-en.xml	151,753
45	demo-2001-11-27-rosenberg-en.xml	152,486
46	demo-2009-04-21-fraser-en.xml	154,055
47	211-2002-12-20-verene-en.xml	159,261
48	414-2003-10-20-bogdanovic-en.xml	160,854
49	demo-2013-07-29-gole-en.xml	163,634

50	1193-2007-11-02-boulbina-en.xml	168,431
51	demo-2013-02-19-leggewie-en.xml	173,868
52	demo-2008-05-02-wennerhag-en.xml	176,731
53	540-2004-06-21-peters-en.xml	188,053
54	2495-2013-06-25-zhurzhenko-en.xml	191,601
55	demo-2012-01-25-halmai-en.xml	201,477
56	1344-2008-08-07-seymour-en.xml	208,259
57	62-2001-04-01-mistry-en.xml	226,113
58	2666-2014-04-03-knausgard-en.xml	245,487

#	**Dateiname**	**Distanz Kosinus TF-IDF**
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2008-11-21-leggewiewelzer-en.xml	0,538
3	demo-2013-11-22-offe-en.xml	0,549
4	demo-2011-11-02-G1000-en.xml	0,549
5	demo-2012-11-21-holmes-en.xml	0,555
6	demo-2011-07-11-bluhdorn-en.xml	0,599
7	demo-2013-09-11-deniztekin-en.xml	0,636
8	demo-2010-09-14-ditchev-en.xml	0,638
9	demo-2011-11-10-sierakowski-en.xml	0,642
10	demo-2011-12-19-amirpur-en.xml	0,677
11	266-2003-02-16-mangasassen-en.xml	0,681
12	demo-2013-05-03-muller-en.xml	0,685
13	demo-2008-05-02-wennerhag-en.xml	0,688
14	demo-2013-08-13-krastev-en.xml	0,688
15	demo-2001-11-27-rosenberg-en.xml	0,696
16	demo-2009-07-14-biscione-en.xml	0,697
17	demo-2013-12-12-margetts-en.xml	0,704
18	demo-2013-07-29-gole-en.xml	0,719
19	demo-2013-12-12-pogonyi-en.xml	0,723
20	505-2004-04-05-uys-en.xml	0,733
21	demo-2013-08-20-leggewienanz-en.xml	0,736
22	335-2003-05-15-henard-en.xml	0,745
23	demo-2009-09-09-kavaliauskas-en.xml	0,751
24	2447-2013-04-12-sanchez-en.xml	0,766
25	demo-2009-04-21-fraser-en.xml	0,782
26	demo-2013-06-14-pomerantsev-en.xml	0,800
27	2495-2013-06-25-zhurzhenko-en.xml	0,800
28	demo-2012-09-05-jahanbegloo-en.xml	0,800
29	demo-2013-02-19-leggewie-en.xml	0,805
30	demo-2013-02-26-james-en.xml	0,818

31	80-2001-11-14-blecher-en.xml	0,830
32	demo-2012-01-25-halmai-en.xml	0,836
33	2517-2013-08-13-osteuropa-en.xml	0,868
34	demo-2012-02-08-elsenhans-en.xml	0,871
35	785-2006-02-16-sambrook-en.xml	0,875
36	256-2003-02-11-kaplinski-en.xml	0,876
37	2100-2011-09-27-scruton-en.xml	0,879
38	62-2001-04-01-mistry-en.xml	0,890
39	demo-2013-11-08-vidanava-en.xml	0,906
40	904-2006-08-17-eder-en.xml	0,916
41	772-2006-02-01-boutang-en.xml	0,916
42	540-2004-06-21-peters-en.xml	0,920
43	2163-2012-01-11-ohlheiser-en.xml	0,920
44	441-2003-11-28-abraham-en.xml	0,921
45	2200-2012-03-20-mondediploo-en.xml	0,923
46	1344-2008-08-07-seymour-en.xml	0,952
47	2666-2014-04-03-knausgard-en.xml	0,957
48	1270-2008-04-09-miklosi-en.xml	0,959
49	414-2003-10-20-bogdanovic-en.xml	0,969
50	661-2005-07-14-revista-en.xml	0,972
51	1117-2007-07-06-lapin1-en.xml	0,973
52	demo-2013-02-08-wallerstein-en.xml	0,975
53	211-2002-12-20-verene-en.xml	0,980
54	480-2004-03-04-cakmak-en.xml	0,981
55	294-2003-03-04-ursic-en.xml	0,990
56	223-2003-01-31-des-en.xml	0,994
57	479-2004-03-03-senyener-en.xml	0,994
58	1193-2007-11-02-boulbina-en.xml	0,994

Parameter	Wert	Durchschnitt Precision	
Korpus	C	Optimal	
# häufigste Terme	10	Euklid TF-IDF	72,27%
		Kosinus TF-IDF	100,00%



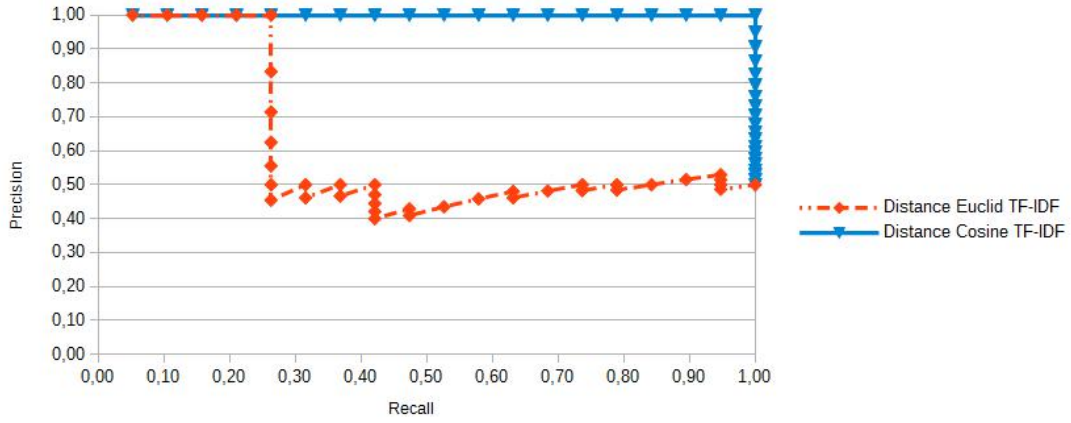
#	Dateiname	Distanz Euklid TF-IDF
1	bologna-2010-07-01-editorial-en.xml	0,000
2	bologna-2010-07-01-balan-en.xml	30,594
3	bologna-2010-07-01-bot-en.xml	32,802
4	bologna-2010-07-01-lichtenbergerh-en.xml	39,421
5	bologna-2010-07-01-vonosten-en.xml	41,713
6	bologna-2010-07-01-calderwilliams-en.xml	42,403
7	1117-2007-07-06-lapin1-en.xml	42,872
8	2200-2012-03-20-mondediploo-en.xml	44,159
9	441-2003-11-28-abraham-en.xml	45,607
10	bologna-2010-07-01-benhamou-en.xml	48,405
11	2100-2011-09-27-scruton-en.xml	50,170
12	bologna-2010-07-01-schwan-en.xml	50,418
13	661-2005-07-14-revista-en.xml	50,804
14	2447-2013-04-12-sanchez-en.xml	51,147
15	223-2003-01-31-des-en.xml	52,545
16	479-2004-03-03-senyener-en.xml	53,833

17	2517-2013-08-13-osteuropa-en.xml	55,920
18	335-2003-05-15-henard-en.xml	56,187
19	785-2006-02-16-sambrook-en.xml	56,921
20	bologna-2010-07-01-lichtenbergerpadis-en.xml	59,346
21	266-2003-02-16-mangasassen-en.xml	60,291
22	bologna-2010-07-01-gilbert-en.xml	60,374
23	bologna-2010-07-01-samalavicius-en.xml	62,642
24	1270-2008-04-09-miklosi-en.xml	64,078
25	bologna-2010-07-01-herwigEtAl-en.xml	67,912
26	2163-2012-01-11-ohlheiser-en.xml	68,140
27	bologna-2010-03-11-manchev-en.xml	69,907
28	bologna-2010-07-01-munch-en.xml	69,993
29	505-2004-04-05-uys-en.xml	75,419
30	bologna-2010-04-13-dorling-en.xml	79,385
31	bologna-2010-07-01-bikbov-en.xml	92,935
32	211-2002-12-20-verene-en.xml	93,899
33	bologna-2010-07-01-power-en.xml	96,690
34	bologna-2010-02-05-newfield-en.xml	112,601
35	2495-2013-06-25-zhurzhenko-en.xml	163,288
36	62-2001-04-01-mistry-en.xml	180,242
37	2666-2014-04-03-knausgard-en.xml	186,107
38	bologna-2010-07-01-santos-en.xml	209,134

#	**Dateiname**	**Distanz Kosinus TF-IDF**
1	bologna-2010-07-01-editorial-en.xml	0,000
2	bologna-2010-07-01-bikbov-en.xml	0,287
3	bologna-2010-07-01-lichtenbergerh-en.xml	0,298
4	bologna-2010-07-01-bot-en.xml	0,300
5	bologna-2010-07-01-balan-en.xml	0,316
6	bologna-2010-07-01-benhamou-en.xml	0,320
7	bologna-2010-07-01-samalavicius-en.xml	0,338
8	bologna-2010-07-01-munch-en.xml	0,370
9	bologna-2010-07-01-santos-en.xml	0,415
10	bologna-2010-07-01-power-en.xml	0,421
11	bologna-2010-07-01-vonosten-en.xml	0,424
12	bologna-2010-07-01-gilbert-en.xml	0,432
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	0,490
14	bologna-2010-07-01-calderwilliams-en.xml	0,504
15	bologna-2010-07-01-schwan-en.xml	0,569
16	bologna-2010-03-11-manchev-en.xml	0,598
17	bologna-2010-02-05-newfield-en.xml	0,678
18	bologna-2010-07-01-herwigEtAl-en.xml	0,822
19	bologna-2010-04-13-dorling-en.xml	0,874
20	2447-2013-04-12-sanchez-en.xml	0,884
21	2666-2014-04-03-knausgard-en.xml	0,914
22	211-2002-12-20-verene-en.xml	0,973
23	1117-2007-07-06-lapin1-en.xml	1,000
24	1270-2008-04-09-miklosi-en.xml	1,000
25	2100-2011-09-27-scruton-en.xml	1,000
26	2163-2012-01-11-ohlheiser-en.xml	1,000
27	2200-2012-03-20-mondediploo-en.xml	1,000
28	223-2003-01-31-des-en.xml	1,000
29	2495-2013-06-25-zhurzhenko-en.xml	1,000
30	2517-2013-08-13-osteuropa-en.xml	1,000

31	266-2003-02-16-mangasassen-en.xml	1,000
32	335-2003-05-15-henard-en.xml	1,000
33	441-2003-11-28-abraham-en.xml	1,000
34	479-2004-03-03-senyener-en.xml	1,000
35	505-2004-04-05-uys-en.xml	1,000
36	62-2001-04-01-mistry-en.xml	1,000
37	661-2005-07-14-revista-en.xml	1,000
38	785-2006-02-16-sambrook-en.xml	1,000

Parameter	Wert	Durchschnitt Precision	
Korpus	C	Optimal	
# häufigste Terme	30	Euklid TF-IDF	67,39%
		Kosinus TF-IDF	100,00%



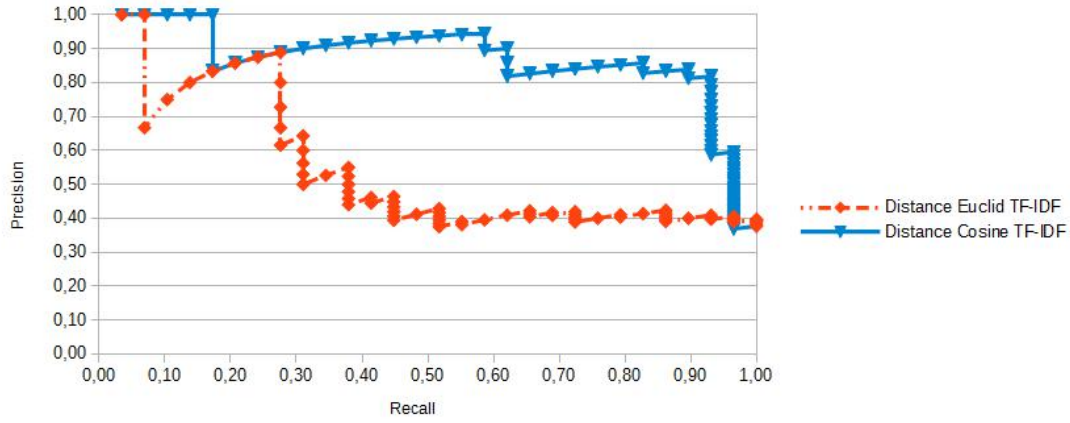
#	Dateiname	Distanz Euklid TF-IDF
1	bologna-2010-07-01-editorial-en.xml	0,000
2	bologna-2010-07-01-bot-en.xml	38,497
3	bologna-2010-07-01-balan-en.xml	38,536
4	bologna-2010-07-01-vonosten-en.xml	44,576
5	bologna-2010-07-01-lichtenbergerh-en.xml	47,749
6	1117-2007-07-06-lapin1-en.xml	48,394
7	2200-2012-03-20-mondediploo-en.xml	49,487
8	441-2003-11-28-abraham-en.xml	51,137
9	2100-2011-09-27-scruton-en.xml	55,390
10	661-2005-07-14-revista-en.xml	56,205
11	223-2003-01-31-des-en.xml	56,267
12	bologna-2010-07-01-calderwilliams-en.xml	56,294
13	2447-2013-04-12-sanchez-en.xml	58,060
14	bologna-2010-07-01-schwan-en.xml	60,341
15	479-2004-03-03-senyener-en.xml	60,589
16	bologna-2010-07-01-benhamou-en.xml	60,597

17	2517-2013-08-13-osteuropa-en.xml	62,161
18	335-2003-05-15-henard-en.xml	64,389
19	266-2003-02-16-mangasassen-en.xml	67,149
20	785-2006-02-16-sambrook-en.xml	68,250
21	bologna-2010-07-01-gilbert-en.xml	70,901
22	1270-2008-04-09-miklosi-en.xml	71,127
23	bologna-2010-07-01-samalavicius-en.xml	74,565
24	bologna-2010-07-01-herwigEtAl-en.xml	76,013
25	bologna-2010-07-01-lichtenbergerpadis-en.xml	78,038
26	2163-2012-01-11-ohlheiser-en.xml	79,825
27	bologna-2010-03-11-manchev-en.xml	80,150
28	bologna-2010-07-01-munch-en.xml	81,320
29	505-2004-04-05-uys-en.xml	85,059
30	bologna-2010-04-13-dorling-en.xml	92,876
31	211-2002-12-20-verene-en.xml	101,173
32	bologna-2010-07-01-bikbov-en.xml	101,459
33	bologna-2010-07-01-power-en.xml	114,991
34	bologna-2010-02-05-newfield-en.xml	125,833
35	2495-2013-06-25-zhurzhenko-en.xml	173,462
36	62-2001-04-01-mistry-en.xml	201,109
37	2666-2014-04-03-knausgard-en.xml	214,802
38	bologna-2010-07-01-santos-en.xml	218,092

#	**Dateiname**	**Distanz Kosinus TF-IDF**
1	bologna-2010-07-01-editorial-en.xml	0,000
2	bologna-2010-07-01-bikbov-en.xml	0,317
3	bologna-2010-07-01-bot-en.xml	0,335
4	bologna-2010-07-01-lichtenbergerh-en.xml	0,342
5	bologna-2010-07-01-vonosten-en.xml	0,373
6	bologna-2010-07-01-benhamou-en.xml	0,384
7	bologna-2010-07-01-balan-en.xml	0,393
8	bologna-2010-07-01-santos-en.xml	0,410
9	bologna-2010-07-01-munch-en.xml	0,417
10	bologna-2010-07-01-samalavicius-en.xml	0,434
11	bologna-2010-07-01-gilbert-en.xml	0,492
12	bologna-2010-07-01-power-en.xml	0,495
13	bologna-2010-07-01-lichtenbergerpadis-en.xml	0,559
14	bologna-2010-03-11-manchev-en.xml	0,588
15	bologna-2010-07-01-calderwilliams-en.xml	0,607
16	bologna-2010-07-01-schwan-en.xml	0,616
17	bologna-2010-02-05-newfield-en.xml	0,651
18	bologna-2010-07-01-herwigEtAl-en.xml	0,748
19	bologna-2010-04-13-dorling-en.xml	0,859
20	2447-2013-04-12-sanchez-en.xml	0,882
21	266-2003-02-16-mangasassen-en.xml	0,924
22	62-2001-04-01-mistry-en.xml	0,924
23	223-2003-01-31-des-en.xml	0,924
24	2666-2014-04-03-knausgard-en.xml	0,925
25	2100-2011-09-27-scruton-en.xml	0,926
26	211-2002-12-20-verene-en.xml	0,934
27	2200-2012-03-20-mondediploo-en.xml	0,936
28	661-2005-07-14-revista-en.xml	0,951
29	335-2003-05-15-henard-en.xml	0,953
30	1270-2008-04-09-miklosi-en.xml	0,955

31	2517-2013-08-13-osteuropa-en.xml	0,955
32	1117-2007-07-06-lapin1-en.xml	0,958
33	441-2003-11-28-abraham-en.xml	0,964
34	505-2004-04-05-uys-en.xml	0,970
35	2495-2013-06-25-zhurzhenko-en.xml	0,972
36	785-2006-02-16-sambrook-en.xml	0,979
37	2163-2012-01-11-ohlheiser-en.xml	1,000
38	479-2004-03-03-senyener-en.xml	1,000

Parameter	Wert	Durchschnitt Precision	
Korpus	Demokratie ∪ Bologna ∪ Zufällig	Optimal	
# häufigste Terme	10	Euklid TF-IDF	58,66%
		Kosinus TF-IDF	81,33%



#	Dateiname	Distanz Euklid TF-IDF
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2013-11-22-offe-en.xml	104,547
3	266-2003-02-16-mangasassen-en.xml	106,471
4	demo-2008-11-21-leggewiewelzer-en.xml	108,231
5	demo-2011-11-02-G1000-en.xml	108,365
6	demo-2013-09-11-deniztekin-en.xml	108,471
7	demo-2011-11-10-sierakowski-en.xml	108,972
8	demo-2010-09-14-ditchev-en.xml	109,900
9	demo-2013-12-12-pogonyi-en.xml	111,772
10	335-2003-05-15-henard-en.xml	113,119
11	294-2003-03-04-ursic-en.xml	115,598
12	80-2001-11-14-blecher-en.xml	115,866
13	1117-2007-07-06-lapin1-en.xml	116,151
14	demo-2013-02-26-james-en.xml	116,327
15	2200-2012-03-20-mondediploo-en.xml	116,632
16	2100-2011-09-27-scruton-en.xml	117,090

17	441-2003-11-28-abraham-en.xml	117,188
18	772-2006-02-01-boutang-en.xml	117,716
19	demo-2009-09-09-kavaliauskas-en.xml	118,076
20	demo-2011-07-11-bluhdorn-en.xml	118,157
21	bologna-2010-07-01-balan-en.xml	118,233
22	2447-2013-04-12-sanchez-en.xml	118,571
23	505-2004-04-05-uys-en.xml	118,916
24	661-2005-07-14-revista-en.xml	119,306
25	785-2006-02-16-sambrook-en.xml	119,553
26	demo-2013-08-13-krastev-en.xml	119,867
27	223-2003-01-31-des-en.xml	120,058
28	demo-2013-11-08-vidanava-en.xml	120,133
29	bologna-2010-07-01-schwan-en.xml	120,279
30	479-2004-03-03-senyener-en.xml	120,628
31	2517-2013-08-13-osteuropa-en.xml	121,573
32	904-2006-08-17-eder-en.xml	121,988
33	480-2004-03-04-cakmak-en.xml	122,287
34	demo-2013-06-14-pomerantsev-en.xml	122,385
35	demo-2012-11-21-holmes-en.xml	122,442
36	bologna-2010-07-01-editorial-en.xml	122,479
37	bologna-2010-07-01-calderwilliams-en.xml	123,333
38	bologna-2010-07-01-bot-en.xml	123,382
39	256-2003-02-11-kaplinski-en.xml	123,826
40	bologna-2010-07-01-vonosten-en.xml	125,208
41	demo-2009-07-14-biscione-en.xml	125,455
42	1270-2008-04-09-miklosi-en.xml	125,535
43	demo-2013-02-08-wallerstein-en.xml	125,936
44	demo-2012-09-05-jahanbegloo-en.xml	126,099
45	demo-2013-05-03-muller-en.xml	126,238
46	2163-2012-01-11-ohlheiser-en.xml	127,656
47	bologna-2010-07-01-lichtenbergerh-en.xml	127,980
48	demo-2012-02-08-elsenhans-en.xml	128,020
49	bologna-2010-07-01-herwigEtAl-en.xml	128,511

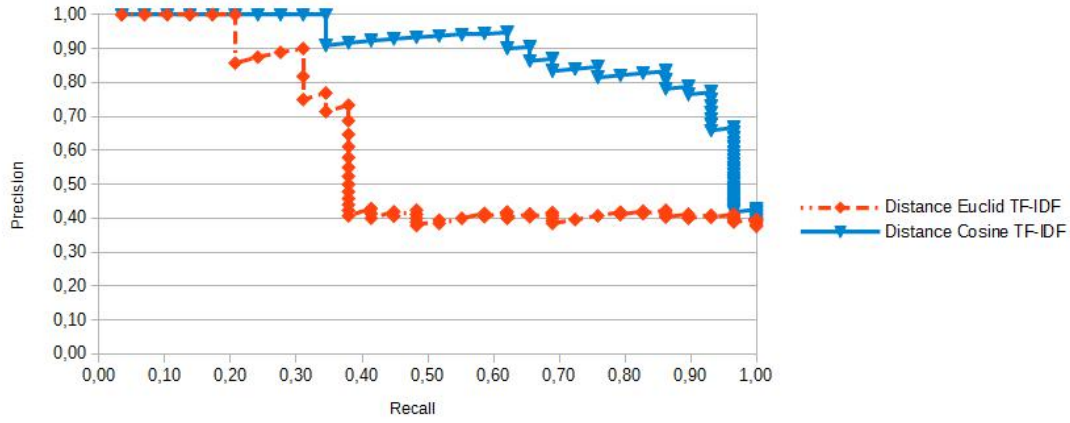
50	demo-2011-12-19-amirpur-en.xml	129,634
51	bologna-2010-07-01-gilbert-en.xml	130,415
52	bologna-2010-03-11-manchev-en.xml	132,136
53	bologna-2010-04-13-dorling-en.xml	132,518
54	bologna-2010-07-01-benhamou-en.xml	132,876
55	demo-2009-04-21-fraser-en.xml	134,220
56	demo-2013-08-20-leggewienanz-en.xml	134,287
57	bologna-2010-07-01-lichtenbergerpadis-en.xml	134,309
58	demo-2001-11-27-rosenberg-en.xml	135,377
59	demo-2013-12-12-margetts-en.xml	138,123
60	bologna-2010-07-01-samalavicius-en.xml	139,280
61	211-2002-12-20-verene-en.xml	143,743
62	bologna-2010-07-01-munch-en.xml	145,258
63	414-2003-10-20-bogdanovic-en.xml	146,636
64	1193-2007-11-02-boulbina-en.xml	150,486
65	demo-2013-02-19-leggewie-en.xml	155,952
66	demo-2013-07-29-gole-en.xml	160,941
67	bologna-2010-07-01-power-en.xml	162,567
68	bologna-2010-07-01-bikbov-en.xml	164,767
69	demo-2008-05-02-wennerhag-en.xml	165,266
70	bologna-2010-02-05-newfield-en.xml	165,729
71	540-2004-06-21-peters-en.xml	167,809
72	2495-2013-06-25-zhurzhenko-en.xml	172,511
73	demo-2012-01-25-halmai-en.xml	191,031
74	1344-2008-08-07-seymour-en.xml	193,324
75	62-2001-04-01-mistry-en.xml	206,010
76	2666-2014-04-03-knausgard-en.xml	218,140
77	bologna-2010-07-01-santos-en.xml	255,664

#	**Dateiname**	**Distanz Kosinus TF-IDF**
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2013-11-22-offe-en.xml	0,518
3	demo-2013-09-11-deniztekin-en.xml	0,562
4	demo-2011-11-02-G1000-en.xml	0,562
5	demo-2011-07-11-bluhdorn-en.xml	0,570
6	266-2003-02-16-mangasassen-en.xml	0,613
7	demo-2011-11-10-sierakowski-en.xml	0,616
8	demo-2008-11-21-leggewiewelzer-en.xml	0,635
9	demo-2012-11-21-holmes-en.xml	0,688
10	demo-2010-09-14-ditchev-en.xml	0,689
11	demo-2009-09-09-kavaliauskas-en.xml	0,692
12	demo-2013-12-12-pogonyi-en.xml	0,697
13	demo-2013-05-03-muller-en.xml	0,702
14	demo-2008-05-02-wennerhag-en.xml	0,717
15	demo-2001-11-27-rosenberg-en.xml	0,724
16	demo-2013-08-20-leggewienanz-en.xml	0,729
17	demo-2011-12-19-amirpur-en.xml	0,731
18	demo-2013-12-12-margetts-en.xml	0,734
19	2495-2013-06-25-zhurzhenko-en.xml	0,766
20	demo-2009-04-21-fraser-en.xml	0,769
21	335-2003-05-15-henard-en.xml	0,773
22	505-2004-04-05-uys-en.xml	0,781
23	demo-2009-07-14-biscione-en.xml	0,789
24	demo-2013-06-14-pomerantsev-en.xml	0,793
25	demo-2013-02-19-leggewie-en.xml	0,835
26	demo-2013-07-29-gole-en.xml	0,845
27	demo-2013-02-26-james-en.xml	0,846
28	demo-2013-08-13-krastev-en.xml	0,864
29	bologna-2010-07-01-schwan-en.xml	0,870
30	demo-2012-02-08-elsenhans-en.xml	0,873

31	demo-2012-09-05-jahanbegloo-en.xml	0,890
32	bologna-2010-07-01-gilbert-en.xml	0,900
33	demo-2012-01-25-halmai-en.xml	0,905
34	bologna-2010-03-11-manchev-en.xml	0,910
35	80-2001-11-14-blecher-en.xml	0,911
36	540-2004-06-21-peters-en.xml	0,913
37	772-2006-02-01-boutang-en.xml	0,928
38	bologna-2010-07-01-balan-en.xml	0,931
39	2100-2011-09-27-scruton-en.xml	0,931
40	bologna-2010-04-13-dorling-en.xml	0,932
41	2447-2013-04-12-sanchez-en.xml	0,933
42	785-2006-02-16-sambrook-en.xml	0,934
43	256-2003-02-11-kaplinski-en.xml	0,935
44	904-2006-08-17-eder-en.xml	0,940
45	bologna-2010-07-01-herwigEtAl-en.xml	0,955
46	62-2001-04-01-mistry-en.xml	0,958
47	demo-2013-11-08-vidanava-en.xml	0,958
48	bologna-2010-07-01-samalavicius-en.xml	0,968
49	bologna-2010-07-01-santos-en.xml	0,980
50	1117-2007-07-06-lapin1-en.xml	1,000
51	1193-2007-11-02-boulbina-en.xml	1,000
52	1270-2008-04-09-miklosi-en.xml	1,000
53	1344-2008-08-07-seymour-en.xml	1,000
54	211-2002-12-20-verene-en.xml	1,000
55	2163-2012-01-11-ohlheiser-en.xml	1,000
56	2200-2012-03-20-mondediploo-en.xml	1,000
57	223-2003-01-31-des-en.xml	1,000
58	2517-2013-08-13-osteuropa-en.xml	1,000
59	2666-2014-04-03-knausgard-en.xml	1,000
60	294-2003-03-04-ursic-en.xml	1,000
61	414-2003-10-20-bogdanovic-en.xml	1,000
62	441-2003-11-28-abraham-en.xml	1,000
63	479-2004-03-03-senyener-en.xml	1,000

64	480-2004-03-04-cakmak-en.xml	1,000
65	661-2005-07-14-revista-en.xml	1,000
66	bologna-2010-02-05-newfield-en.xml	1,000
67	bologna-2010-07-01-benhamou-en.xml	1,000
68	bologna-2010-07-01-bikbov-en.xml	1,000
69	bologna-2010-07-01-bot-en.xml	1,000
70	bologna-2010-07-01-calderwilliams-en.xml	1,000
71	bologna-2010-07-01-editorial-en.xml	1,000
72	bologna-2010-07-01-lichtenbergerh-en.xml	1,000
73	bologna-2010-07-01-lichtenbergerpadis-en.xml	1,000
74	bologna-2010-07-01-munch-en.xml	1,000
75	bologna-2010-07-01-power-en.xml	1,000
76	bologna-2010-07-01-vonosten-en.xml	1,000
77	demo-2013-02-08-wallerstein-en.xml	1,000

Parameter	Wert	Durchschnitt Precision	
Korpus	Demokratie ∪ Bologna ∪ Zufällig	Optimal	
# häufigste Terme	30	Euklid TF-IDF	61,55%
		Kosinus TF-IDF	82,69%



#	Dateiname	Distanz Euklid TF-IDF
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2008-11-21-leggewiewelzer-en.xml	116,288
3	demo-2013-11-22-offe-en.xml	120,499
4	demo-2011-11-02-G1000-en.xml	121,610
5	demo-2010-09-14-ditchev-en.xml	122,274
6	demo-2013-09-11-deniztekin-en.xml	124,435
7	266-2003-02-16-mangasassen-en.xml	124,656
8	demo-2012-11-21-holmes-en.xml	125,722
9	demo-2013-08-13-krastev-en.xml	126,167
10	demo-2011-11-10-sierakowski-en.xml	126,432
11	335-2003-05-15-henard-en.xml	127,475
12	2447-2013-04-12-sanchez-en.xml	127,832
13	demo-2013-12-12-pogonyi-en.xml	128,585
14	80-2001-11-14-blecher-en.xml	129,514
15	demo-2013-02-26-james-en.xml	131,050
16	2200-2012-03-20-mondediploo-en.xml	131,114

17	294-2003-03-04-ursic-en.xml	131,206
18	441-2003-11-28-abraham-en.xml	131,305
19	2100-2011-09-27-scruton-en.xml	131,370
20	1117-2007-07-06-lapin1-en.xml	131,621
21	505-2004-04-05-uys-en.xml	131,799
22	2517-2013-08-13-osteuropa-en.xml	132,575
23	bologna-2010-07-01-balan-en.xml	133,757
24	772-2006-02-01-boutang-en.xml	134,045
25	bologna-2010-07-01-schwan-en.xml	134,376
26	661-2005-07-14-revista-en.xml	134,510
27	785-2006-02-16-sambrook-en.xml	134,618
28	demo-2013-11-08-vidanava-en.xml	135,044
29	223-2003-01-31-des-en.xml	135,610
30	904-2006-08-17-eder-en.xml	136,407
31	demo-2012-09-05-jahanbegloo-en.xml	136,539
32	479-2004-03-03-senyener-en.xml	136,547
33	demo-2011-07-11-bluhdorn-en.xml	137,015
34	bologna-2010-07-01-bot-en.xml	137,703
35	256-2003-02-11-kaplinski-en.xml	137,906
36	bologna-2010-07-01-editorial-en.xml	137,928
37	480-2004-03-04-cakmak-en.xml	137,960
38	demo-2009-07-14-biscione-en.xml	138,268
39	bologna-2010-07-01-vonosten-en.xml	139,316
40	demo-2013-06-14-pomerantsev-en.xml	140,004
41	demo-2009-09-09-kavaliauskas-en.xml	140,353
42	1270-2008-04-09-miklosi-en.xml	140,424
43	demo-2013-05-03-muller-en.xml	140,588
44	2163-2012-01-11-ohlheiser-en.xml	141,598
45	bologna-2010-07-01-calderwilliams-en.xml	141,644
46	demo-2011-12-19-amirpur-en.xml	141,771
47	bologna-2010-07-01-lichtenbergerh-en.xml	141,937
48	demo-2013-02-08-wallerstein-en.xml	143,265
49	bologna-2010-07-01-herwigEtAl-en.xml	144,021

50	bologna-2010-07-01-gilbert-en.xml	147,658
51	bologna-2010-07-01-benhamou-en.xml	148,614
52	bologna-2010-03-11-manchev-en.xml	149,258
53	demo-2012-02-08-elsenhans-en.xml	149,850
54	demo-2013-12-12-margetts-en.xml	150,612
55	demo-2013-08-20-leggewienanz-en.xml	151,753
56	bologna-2010-07-01-lichtenbergerpadis-en.xml	152,053
57	demo-2001-11-27-rosenberg-en.xml	152,486
58	bologna-2010-04-13-dorling-en.xml	152,656
59	demo-2009-04-21-fraser-en.xml	154,055
60	bologna-2010-07-01-samalavicius-en.xml	156,691
61	211-2002-12-20-verene-en.xml	159,261
62	414-2003-10-20-bogdanovic-en.xml	160,854
63	demo-2013-07-29-gole-en.xml	163,634
64	bologna-2010-07-01-munch-en.xml	163,704
65	1193-2007-11-02-boulbina-en.xml	168,431
66	demo-2013-02-19-leggewie-en.xml	173,868
67	bologna-2010-07-01-bikbov-en.xml	176,590
68	demo-2008-05-02-wennerhag-en.xml	176,731
69	bologna-2010-07-01-power-en.xml	182,951
70	bologna-2010-02-05-newfield-en.xml	185,208
71	540-2004-06-21-peters-en.xml	188,053
72	2495-2013-06-25-zhurzhenko-en.xml	191,601
73	demo-2012-01-25-halmai-en.xml	201,477
74	1344-2008-08-07-seymour-en.xml	208,259
75	62-2001-04-01-mistry-en.xml	226,113
76	2666-2014-04-03-knausgard-en.xml	245,487
77	bologna-2010-07-01-santos-en.xml	268,336

#	**Dateiname**	**Distanz Kosinus TF-IDF**
1	demo-2013-02-01-krastev-en.xml	0,000
2	demo-2008-11-21-leggewiewelzer-en.xml	0,538
3	demo-2013-11-22-offe-en.xml	0,549
4	demo-2011-11-02-G1000-en.xml	0,549
5	demo-2012-11-21-holmes-en.xml	0,555
6	demo-2011-07-11-bluhdorn-en.xml	0,599
7	demo-2013-09-11-deniztekin-en.xml	0,636
8	demo-2010-09-14-ditchev-en.xml	0,638
9	demo-2011-11-10-sierakowski-en.xml	0,642
10	demo-2011-12-19-amirpur-en.xml	0,677
11	266-2003-02-16-mangasassen-en.xml	0,681
12	demo-2013-05-03-muller-en.xml	0,685
13	demo-2008-05-02-wennerhag-en.xml	0,688
14	demo-2013-08-13-krastev-en.xml	0,688
15	demo-2001-11-27-rosenberg-en.xml	0,696
16	demo-2009-07-14-biscione-en.xml	0,697
17	demo-2013-12-12-margetts-en.xml	0,704
18	demo-2013-07-29-gole-en.xml	0,719
19	demo-2013-12-12-pogonyi-en.xml	0,723
20	505-2004-04-05-uys-en.xml	0,733
21	demo-2013-08-20-leggewienanz-en.xml	0,736
22	335-2003-05-15-henard-en.xml	0,745
23	demo-2009-09-09-kavaliauskas-en.xml	0,751
24	2447-2013-04-12-sanchez-en.xml	0,766
25	demo-2009-04-21-fraser-en.xml	0,782
26	demo-2013-06-14-pomerantsev-en.xml	0,800
27	2495-2013-06-25-zhurzhenko-en.xml	0,800
28	demo-2012-09-05-jahanbegloo-en.xml	0,800
29	demo-2013-02-19-leggewie-en.xml	0,805
30	demo-2013-02-26-james-en.xml	0,818

31	bologna-2010-07-01-schwan-en.xml	0,825
32	80-2001-11-14-blecher-en.xml	0,830
33	demo-2012-01-25-halmai-en.xml	0,836
34	2517-2013-08-13-osteuropa-en.xml	0,868
35	demo-2012-02-08-elsenhans-en.xml	0,871
36	785-2006-02-16-sambrook-en.xml	0,875
37	256-2003-02-11-kaplinski-en.xml	0,876
38	2100-2011-09-27-scruton-en.xml	0,879
39	62-2001-04-01-mistry-en.xml	0,890
40	bologna-2010-03-11-manchev-en.xml	0,890
41	bologna-2010-07-01-gilbert-en.xml	0,902
42	demo-2013-11-08-vidanava-en.xml	0,906
43	bologna-2010-07-01-herwigEtAl-en.xml	0,907
44	904-2006-08-17-eder-en.xml	0,916
45	772-2006-02-01-boutang-en.xml	0,916
46	bologna-2010-07-01-balan-en.xml	0,918
47	540-2004-06-21-peters-en.xml	0,920
48	2163-2012-01-11-ohlheiser-en.xml	0,920
49	441-2003-11-28-abraham-en.xml	0,921
50	2200-2012-03-20-mondediploo-en.xml	0,923
51	bologna-2010-07-01-lichtenbergerpadis-en.xml	0,932
52	bologna-2010-07-01-bikbov-en.xml	0,934
53	bologna-2010-04-13-dorling-en.xml	0,935
54	bologna-2010-07-01-lichtenbergerh-en.xml	0,939
55	bologna-2010-07-01-santos-en.xml	0,940
56	bologna-2010-07-01-vonosten-en.xml	0,943
57	bologna-2010-07-01-benhamou-en.xml	0,950
58	1344-2008-08-07-seymour-en.xml	0,952
59	2666-2014-04-03-knausgard-en.xml	0,957
60	1270-2008-04-09-miklosi-en.xml	0,959
61	bologna-2010-07-01-bot-en.xml	0,961
62	bologna-2010-07-01-samalavicius-en.xml	0,967
63	414-2003-10-20-bogdanovic-en.xml	0,969

64	661-2005-07-14-revista-en.xml	0,972
65	1117-2007-07-06-lapin1-en.xml	0,973
66	bologna-2010-07-01-power-en.xml	0,973
67	bologna-2010-02-05-newfield-en.xml	0,974
68	demo-2013-02-08-wallerstein-en.xml	0,975
69	bologna-2010-07-01-editorial-en.xml	0,975
70	211-2002-12-20-verene-en.xml	0,980
71	480-2004-03-04-cakmak-en.xml	0,981
72	294-2003-03-04-ursic-en.xml	0,990
73	bologna-2010-07-01-calderwilliams-en.xml	0,990
74	bologna-2010-07-01-munch-en.xml	0,992
75	223-2003-01-31-des-en.xml	0,994
76	479-2004-03-03-senyener-en.xml	0,994
77	1193-2007-11-02-boulbina-en.xml	0,994

Versicherung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 21. September 2015 Marcel Schöneberg