Automatisierte Erstellung von Pressedossiers durch Textmining

Projektbericht

Marcel Schöneberg

marcel.schoeneberg@haw-hamburg.de

Hochschule für Angewandte Wissenschaften Hamburg (HAW)
Fakultät für Technik und Informatik
Department Informatik
März 2015

Zusammenfassung

Der folgende Bericht befasst sich mit den Arbeitsergebnissen des zweiten Projekts (Hauptprojekt) im Masterstudium des Autors. Zunächst wird kurz eine Einführung in die Thematik samt der vom Autor gesetzten Ziele gegeben. Daraufhin werden fachliche, sowie technische Grundlagen vorgestellt. Im Anschluss wird auf die praktische Umsetzung, sowie deren Hürden eingegangen. Im Weiteren werden die gewonnenen Ergebnisse vorgestellt und interpretiert. Abschließend resümiert das Fazit die Ergebnisse des Projektes und stellt noch einmal weitere Schritte vor.

1 Einführung in die Thematik

Dieses Dokument stellt die Ergebnisse des vom Autor durchgeführten Moduls 'Projekt 2' (Hauptprojekt) im Masterstudiengang Informatik vor.

Diese Ausarbeitung gliedert sich in fünf Abschnitte. Nach einer Einführung in die Vision und Ziele des durchgeführten Projekts werden im zweiten Abschnitt Grundlagen gelegt. Diese umfassen neben fachlichem Domänenwissen auch eine grundlegende Hypothese, welche in dieser Arbeit untersucht wird. Darüber hinaus wird auch ein Basiswissen über die genutzte Technik vermittelt. Der dritte Abschnitt befasst sich detailliert mit der technischen Umsetzung des Projekts und geht hierbei auf die verschiedenen Arbeitsschritte ein und stellt aufgetretene Probleme, sowie Lösungen vor. Die Ergebnisse welche im Verlauf der Arbeit entstanden sind werden im fünften Abschnitt aufgegriffen und interpretiert. Ein kritisches Fazit welches auch auf weitere mögliche Schritte eingeht bildet den Schluss der Arbeit. Im Anhang A.1 ist eine Danksagung, sowie ein Beispielartikel (Seite 13) zu finden, ebenso sind die Resultate der durchgeführten Experimente im Anhang B zu finden.

1.1 Vision

Die Vision des Projekts ist die (semi)automatisierte Erstellung von Pressedossiers. Hierbei sollen aus einem gegebenen Archiv von Presseartikeln ähnliche Dokumente (ausgehend von einem Leitartikel) gefunden werden, diese Funde sollen Journalisten als Hilfsmittel zur Erstellung von Dossiers dienen (vgl. Grafik 1).

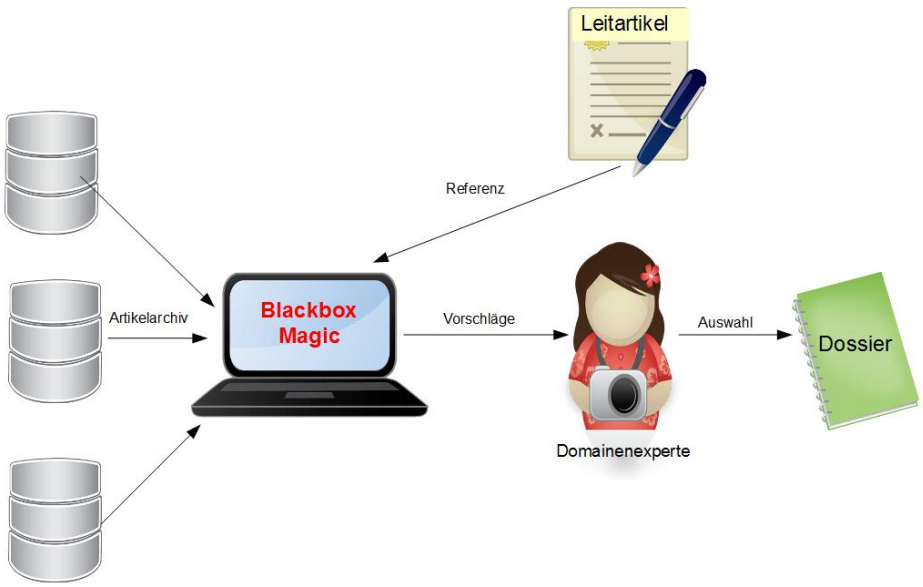


Fig. 1: Realisierbarer Workflow

Bei der Realisierung des geplanten Vorhabens ist sowohl fachliches Wissen (Journalistik) als auch technisches Wissen (Informatik) nötig. Die Schnittstelle dieser beiden Gebiete ist eine Distanzfunktion [FS06, vgl. Seite 29] welche ermittelt wie ähnlich sich gegebene Artikel sind.

Eine detaillierte Erläuterung der Vision sowie ihrer Herkunft kann [Sch15] entnommen werden.

1.2 Ziele

Das in diesem Papier beschriebene Projekt verfolgt mehrere Zielsetzungen. Zum einen soll eine Arbeitsumgebung geschaffen werden, welche es ermöglicht mit den vorhandenen Daten zu experimentieren.

Diese Umgebung soll es ermöglichen auf Basis der gegebenen Daten verschiedene Distanzfunktionen (und Vorverarbeitungen) zu evaluieren um deren Ergebnisse miteinander zu vergleichen.

Darüber hinaus soll einer im weiteren Verlauf vorgestellte Hypothese auf ihre Aussagekraft hin untersucht werden. Anhand der gewonnenen Ergebnisse sollen Rückschlüsse auf etwaige Fehler und Verbesserungen möglich sein.

Grundlegend wurden vom Autor in [Sch15] eine Reihe von Fragen erarbeitet. Diese umfassen neben der Definition von Dossiers auch grundlegende Überlegungen zur Machbarkeit der automatisierten Erstellung von Pressedossiers. Dieser Projektbericht soll daher allgemein auch eine Basis zur Beantwortung dieser Fragen schaffen und mögliche Probleme und interessante Fragestellungen aufdecken.

2 Grundlagen

Die folgenden Abschnitte stellen einige Grundlagen vor auf welche im Rahmen dieser Arbeit zurückgegriffen wird. Auf fachlicher Seite ist dieses vor allem das genutzte Artikelarchiv. Darüber hinaus wird die verfolgte fachliche Hypothese erläutert, welche näher untersucht wird. Die technische Basis des Projekts bildet die Datamining-Umgebung 'RapidMiner' (https://rapidminer.com/), diese wird daher ebenfalls in ihren Grundzügen vorgestellt.

2.1 Fachliche Basis: Artikelarchiv

Das fachliche Grundgerüst bildet, neben dem später wichtigen Wissen eines Domänenexperten, das Artikelarchiv des Kulturnetzwerkes Eurozine www.eurozine.com. Die vom Autor genutzten ca. 3700 Artikel wurden von professionellen Journalisten verfasst. Die Dokumente sind (teils als Übersetzung) in englischen Sprache verfasst, und es besteht darüber hinaus weitergehend die Möglichkeit Metainformationen zu den Artikeln zu erhalten (z.B. Verlinkungen auf die Inhaltsverzeichnisse der Ursprungszeitschrift, redaktionell erstellte Archive etc.). Die Artikel selber liegen Form von XML-Dateien vor und weisen eine Semistrukturiertheit auf, so können u.A. Informationen wie Autor, Kurzzusammenfassung, sowie Überschriften etc. direkt dem Dokument entnommen werden. Ein Beispielartikel ist afu Seite 13 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 7500 Artikel in englischer Sprache verfasst sind, darüber hinaus enthält das Archiv auch eine Reihe von Zusammenfassungen, Inhaltsangaben, sowie Rezensionen bereits erschienener Artikel. 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.

Das Archiv selber unterliegt einer Vertraulichkeitserklärung, daher kann der Autor nur begrenzt konkreten Artikelbeispiele (A.2) benennen. Allerdings lässt sich die Vertraulichkeitserklärung bei Interesse weitere wissenschaftliche Arbeiten zu sofern Eurozine zustimmt.

2.2 Fachliche Hypothese

Die durchgeführten Analysen dienen zum einen dazu die Datenqualität abzuschätzen, zum anderen sollen erste Experimente den Weg zu passablen Distanzfunktionen ebnen. Hierzu wurde eine fachliche Hypothese erarbeitet, welche auf dem semantischen Markup der vorhandenen Artikel basiert. Dieses erlaubt die gezielte Extraktion und Nutzung verschiedener Artikelinformationen, diese umfassen u.A. den Abstract, den Titel, sowie diverse Paragraphenüberschriften des Artikels. Die genannten Informationen sind aus Sicht des Autors signifikante Abschnitte eines Artikels und bergen einen großen Teil der Gesamtinformationen des Dokumentes in sich bzw. fassen diesen hinreichend gut zusammen. Aus diesem Grund ist die Annahme, dass eine verstärkte Berücksichtigung dieser Aspekte ein guter Anhaltspunkt für die Ähnlichkeit von Artikeln ist. Diese These soll im Projekt untersucht werden und als erster Baustein einer auszubauenden Distanzfunktion genutzt werden.

Weitere Schritte wie Kategorien welche Artikeln zugeordnet werden, sowie die Nutzung des Wissens eines Domänenexperten befinden sich in Planung.

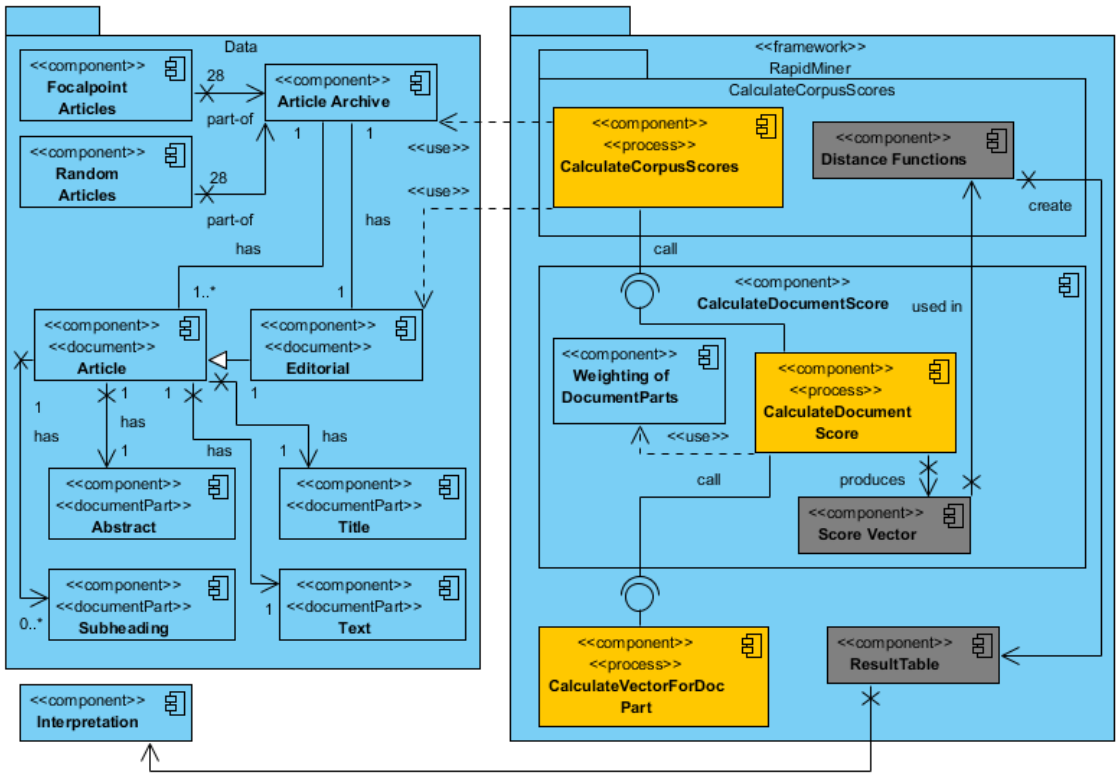


Fig. 2: Komponenten und Workflow

2.3 Technische Basis

Die technische Basis für das durchgeführte Projekt bildet die Datamining-Umgebung RapidMiner (verwendet wurde Version 5.3). Diese (ursprünglich von der TU Dortmund als 'Yet Another Learning Environment' (YALE) entwickelte) Software erlaubt die Nutzung einer Vielzahl von Operatoren. Diese ermöglichen es auf (meist) einfache Weise komplexe Analyseprozesse zusammenzusetzen. RapidMiner selbst lässt sich über eine Reihe von Plugins erweitern um Zusatzfunktionalität (z.B. durch die verwendeten Textmining 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.

3 Umsetzung

Die folgenden Abschnitte widmen sich der Umsetzung der erläuterten Vision. Hierbei werden sowohl theoretische Überlegungen, sowie praktische Umsetzungen in RapidMiner näher betrachtet.

3.1 Aufbau

Grundsätzlich besteht das erstellte Framework aus mehreren Bestandteilen, die wichtigsten sind in Abbildung 2 zu sehen. Diese Grafik stellt daher auch einen Überblick über das RapidMiner-Projekt dar.

Datenbasis Als Datenbasis dient das Artikelarchiv, sowie ein Leitartikel. Dieser dient als Ausgangspunkt für die Distanzfunktion die einen Wert berechnet, welche die Nähe zu ihm repräsentiert. Sowohl das Archiv, als auch der Leitartikel liegen in Form von XML-Files in einem separaten Ordner des Projekts. Das Archiv selber ist aufgeteilt in zufällige Artikel (markiert durch einen Namenspräfix), sowie von Menschen zusammengestellte Dokumente eines Focalpoints zu einem konkreten Thema (Demokratie). Diese Zusammenstellungen werden später zur Überprüfung der Ergebnisse benutzt.

RapidMiner Die folgenden Abschnitte beschreiben die RapidMiner-Prozesse, welche die Logik zur Distanzberechnung enthalten. Diese Prozesse beschreiben einen Ablauf von Operatoren (Algorithmen) und sind teils geschachtelt.

CalculateCorpusScores Der Gesamtablauf 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. Die Ergebnisse dieser Berechnung werden zur Distanzfunktionsberechnung weitergereicht. Welche die Resultate (die Distanzen aller Dokumente zum Leitartikel) in eine Ergebnistabelle einträgt.

Die Ergebnistabelle enthält aktuell alle analysierten Dokumente inklusive der verschiedenen berechneten Distanzen. Diese können im Weiteren interpretiert (3.4) werden.

Zu beachten ist, dass im beschriebenen Prozess die Pfade zum Dokumentarchiv, sowie zum Leitartikel im Operator 'Loop testcorpus' bzw. 'Read editorial' gesetzt werden müssen.

CalculateDocumentScores Dieser Prozess berechnet für jedes Inputdokument einen (gewichteten) 'Bag-Of-Words' [FS06, vgl. Seite 68] - einen (Feature)Vektor mit allen enthaltenen Wörtern, sowie ihren Häufigkeiten. Hierbei 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.

Diese Gewichtung ist in Formel 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}:

$$tf_{ges}(w) = (x_1 * tf_{sec_1}(w)) + (x_2 * tf_{sec_2}(w)) + \cdots + (x_n * tf_{sec_n}(w)) \tag{1}$$

Die Parameter für die Gewichtung sind hierbei momentan direkt im Skript 'Weighted BoW' einzutragen.

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

- 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

Im Anschluss wird als Resultat ein Wortvektor mit den vorhanden Worten und deren Häufigkeit im Dokument erstellt.

Interpretation Dieser Verarbeitungsschritt ist kein RapidMiner-Prozess, sondern eine Auswertung durch einen Menschen. Konkret werden im momentanen Projektstand die Einträge der Ergebnistabelle genutzt um anhand der Bewertungskriterien (3.4) Schlüsse auf die Qualität der Ergebnisse zu ziehen.

3.2 Distanzen

Eines der Kernelemente dieses Projekts bildet der Test von verschiedenen Distanzfunktionen. Das Ziel des Projektes ist es erste Analysen durchzuführen und basierend auf den gewonnenen Erkenntnissen und dem Wissen eines Domänenexperten diese zu verbessern. Hierbei soll die Funktion im Rahmen des Ziels so einfach wie möglich gehalten werden. Aus diesem Grund wurde zunächst eine einfache euklidische Distanz gewählt [FS06, Seite 85]. Diese berechnet sich als:

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

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.

Die obig 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 berücksichtigen. 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 mit Hilfe einer Normierung der Termfrequency bzw. dem TF-IDF Maß (Termfrequency-Inverse Document Frequency) angehen [MRS08, siehe, Seite 117 ff.]. Diese sind im folgenden kurz dargestellt und bilden jeweils eine 'neue' Distanzfunktion durch Ersetzung der x- bzw. y-Werte (in der Gleichung 2) durch die in den Formeln 3 bzw. 4 berechneten Werte.

Eine normierte Termfrequency berechnet sich als:

$$ntf_{t,d} = \alpha + (1 - \alpha) * \frac{tf_{t,d}}{tf_{\max}(d)} \tag{3}$$

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:

$$\text{tf-idf}_{\text{t,d}} = tf_{\text{t,d}} * idf_t \tag{4}$$

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

$$idf_t = log\frac{N}{df_t} \tag{5}$$

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

3.2.1 Beispiel

1. Das folgende Beispiel soll kurz erläutern wie eine Distanz zwischen zwei Dokumenten zustanden kommt. Hierbei soll als Basis die reine Termfrequenz benutzt werden, weiterhin sollen die Vergleichsdokumente folgende Wortvorkommen haben:
$$\begin{pmatrix} Dokumentnummer & 'apple' & 'banana' & 'cat'3 & 'window' \\ 1 & 2 & 2 & 0 & 4 \\ 2 & 1 & 3 & 2 & 0 \end{pmatrix}$$
2. Diese werden nun gewichtet, hierbei sollen die vorkommenden Worte in ihrer Reihenfolge genau den Textabschnitten 'Abstract', 'Title', 'Subheadings' und 'Body' entsprechen. Zu bedenken ist, dass jedes Wort im Beispiel 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:

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

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

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

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

3.3 Erweiterungen

Die geschilderten Distanzfunktionen basieren auf der euklidischen Distanz und verwenden als Featurevektor nur den (gewichteten) 'Bag-of-Words' der gegebenen Dokumente. Über diese recht simple Methode hinaus ist eine Vielzahl von Erweiterungen möglich, die umfassen diverse Kriterien wie Neuheit von Artikeln oder auch thematische Breite [BOHG13, vgl. Seite 3] eines zu erstellenden Dossiers. Auch eine Erweiterung durch ein System von Kategorien, welche einem Artikel zugeordnet werden (z.B. mithilfe von kategoriespezischen Schlagwörtern) ist denkbar. Ebenso soll im späteren Verlauf des Projektes das konkrete Wissen von Domänenexperten genutzt werden, sofern dieses relevant und umsetzbar ist [Hä15, vgl.]. All diese Überlegungen sind allerdings im momentanen Projektstand noch nicht umgesetzt worden, befinden sich dennoch in Planung.

3.4 Probleme

Während der Implementierung des obigen Konzeptes traten zahlreiche Hürden auf, darüber hinaus wurden Designentscheidungen gefällt die nicht immer vorteilhaft waren. Diese Probleme sollen im folgenden kurz erläutert werden um eine Weiterbenutzung und Entwicklung des Werkzeugs durch andere Personen zu erleichtern.

Zunächst muss erwähnt werden, dass die Erweiterung von RapidMiner um eigene Bestandteile (beispielsweise Bag-Of-Words Gewichtung,Distanzfunktionen) nicht immer problemlos möglich war. Diese Erweiterungen sind in Form von Java bzw. Groovy Skripten umgesetzt und arbeiten auf den internen Datenstrukturen von RapidMiner um Berechnungen durchzuführen und an das Programm zurückzugeben. Die Verwendung der internen Datentypen ist allerdings nur bedingt dokumentiert [RM-08a] und veraltet. Darüber hinaus existieren generelle Dokumentationen [RM-08b] welche u.A. die vorhandenen Operatoren der RapidMiner-Basis (ohne Erweiterungen der Community etc.) erläutern. Allerdings ist auch deren Aussagekraft nur teilweise nützlich wenn es um eigens entwickelte Erweiterungen geht. Zudem ist zu erwähnen, dass die API-Dokumentation zwischenzeitlich nicht online zu erreichen war, so dass auf eine Communityversion [RM-] bzw. den Quelltext zurückgegriffen werden musste. Zu bedenken ist auch, dass es dem Autor bei Nutzung von RapidMiner-Skripten nicht möglich war deren Output an einen weiteren RapidMiner-Operator weiterzureichen, allerdings war es möglich den Output eines Operators an ein weiteres Skript zu leiten. Dieses liegt an den Eigenarten des RapidMiner-Skript-Operators (welcher die Groovy basierten Skripte enthält), da dessen Output nicht so typisiert werden kann, dass ein weiterer Operator diesen interpretieren kann.

Anzumerken ist auch, dass der Autor die Möglichkeit der Erweiterung durch kurze Skripte der Alternative des Schreibens von eigenen Operatoren vorgezogen hat. Es ist möglich eigene RM-Operatoren zu entwickeln, welche daraufhin ebenso in der RapidMiner-Umgebung genutzt werden können. Diese

scheinen allerdings im Gegensatz zu den Skripten typisierbaren Output zu unterstützen. Darüber hinaus sollten sie die Möglichkeit bieten eigene Parameter zu definieren, welches an einigen Stellen von Vorteil (aus Sicht der einfachen Benutzung) wäre. Die ursprünglich gewählte Designalternative des Autors war daher suboptimal und kann bei späterem Bedarf geändert werden. Ebenfalls besteht die Möglichkeit, dass das Problem der schlechten Dokumentation in der kommerziellen Version von RapidMiner weniger ausgeprägt ist, dieses müsste allerdings geklärt werden. Grundsätzlich waren (im momentanen Projektstand) alle auftretenden Hürden zu lösen und stellten daher zur eine Verzögerung dar.

3.5 Bewertung

Der folgende Abschnitt geht auf die Bewertung von gewonnenen Ergebnissen ein. Der geplante Workflow (vgl. Graphik 1) produziert Vorschläge für einen Fachexperten, welche dieser verwerten kann. Allerdings ist es für den Autor dieser Arbeit wichtig zu wissen ob etwaige Veränderungen an der 'Blackbox' des Algorithmus zu Verbesserungen führen. Zu diesem Zweck wurde ein redaktionell erstellter Focalpoint (Themengebiet: Demokratie) genutzt, dieser enthält ca. 30 Artikel zum Thema. Diese Sammlung stellt zwar kein konkretes Dossier dar, bietet allerdings eine gute Basis für die Validierung und Verifikation der Ergebnisse. Für den Aufbau eines Testkorpus wurde daher der Focalpoint mit der gleichen Anzahl von zufälligen Artikeln aus dem Archiv verschmolzen. Die von der Blackbox vorgeschlagenen Artikel (ausgehend von einem Leitartikel der ebenfalls dem Focalpoint entstammt) sollen nun im besten Fall genau die Artikel des Focalpoints sein, da diese ein gemeinsames Thema verfolgen.

3.5.1 Precision-Recall

Eine Möglichkeit dieses Ziel in Zahlenwerte zu übertragen sind die Werte Recall (Trefferquote, Gleichung 7) und Precision (Genauigkeit, Gleichung 6) aus dem Umfeld des Information Retrieval [MRS08, vgl. Seite 155]. Hierbei benötigt man (im konkreten Fall) die Menge der 'gefundenen' Dokumente, sowie die Menge der tatsächlich relevanten Treffer (Artikel des Focalpoints). Der Recall drückt hierbei (im Verhältnis) aus wie viele relevante Ergebnisse aus der Gesamtmenge der relevanten Dokumente ausgewählt wurden, während die Precision ein Verhältnismaß ist, welches aussagt wie viele der ausgewählten Ergebnisse relevant sind.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{6}$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \tag{7}$$

Der Zusammenhang der Werte kann durch die Nutzung von 'gefunden' Dokumenten, sowie der Klassifizierung in 'true/false positives' sehr anschaulich dargestellt werden. Anhand der Tabelle 3.5.1 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 zur Bewertung von Ergebnissen auch auf das Wissen eines Domänenexperten zugegriffen werden um zu prüfen bzw. um zu verstehen warum ein Resultat anders ausfällt als erhofft.

3.5.2 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 in dem man die Werte in einem **Precision-Recall Diagramm** [MRS08, siehe Seite 158] kombiniert. Ein solches zeigt die Abbildung 3, hierbei zeigt die x-Achse den Recall, während auf der y-Achse die Precision aufgetragen ist. Hierbei wird der Recall schrittweise gesteigert, indem mehr Dokumente in die Berechnung mit einbezogen werden.

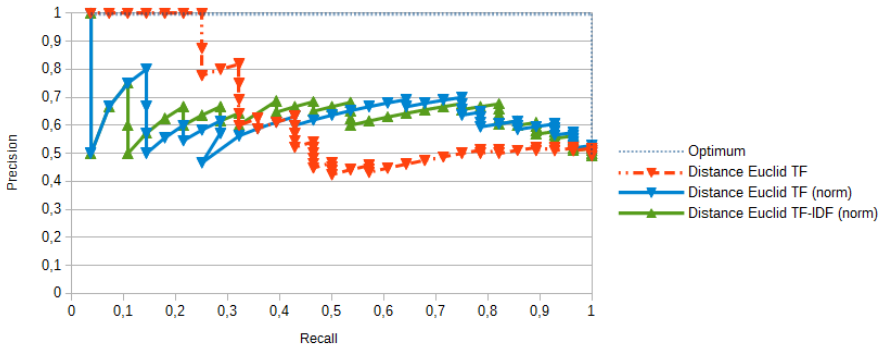


Fig. 3: Precision-Recall Diagramm

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, da dieser 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 daher 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.

Grundsätzlich gehen aufgrund der oben beschriebenen Eigenschaft im momentanen Projektaufbau alle Distanzen in die Bewertung mit ein (dieses schließt daher auch die zufälligen Artikel mit ein). Daher 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 3 zu sehen. Hierbei wurden drei verschiedene Varianten der euklidischen Distanzfunktion (siehe: 3.2) benutzt, zusätzlich wurde das optimale Ergebnis mit aufgetragen.

Zusätzlich zur visuellen Verdeutlichung der Kombination von Precision und Recall durch Diagramme nutzt der momentane 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 Ergebnisse

Im Folgenden sollen die Ergebnisse welche mit dem Versuchsaufbau erzielt wurden erläutert werden. Die entsprechenden Ergebnistabellen sowie Diagramme sind unter im Anhang unter B zu finden.

4.1 Herleitung der Diagramme

Die im Anhang B aufgelisteten Parameter und Distanzen lassen sich leicht in Precision-Recall Diagramme übertragen. Diese Übertragung ist recht einfach:

1. Zunächst werden die einzelnen Werte pro Distanzfunktion aufsteigend sortiert.
2. Mithilfe der Formeln 6 und 7 lässt sich nun für jeden Distanzwert die Precision, sowie der Recall berechnen. Hierbei werden für jeden Wert seine Vorgänger in die Berechnung mit 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 eingetragen. Hierbei ist der Intervall für beide Werte auf 0 - 1 begrenzt.

4.2 Generelles

Bei den Testreihen im Anhang B 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 'richtige' und 'falsche' Artikel abwechselt 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 weißt daher zu sich selber eine Entfernung von 0 auf.

Im Folgenden 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.

4.3 Testreihen mit diversen Gewichtungsparametern

Die im Anhang B aufgeführten Tabellen gliedern sich in mehrere Abschnitte, zunächst werden ab B.1 verschiedene Basis- und Sonderfälle für die Gewichtung von Abstract, Title, Subheading und Text aufgeführt. Im Anschluss wird ab B.2 mit diversen Werten experimentiert, welche zusammenfassende Abschnitte (Abstract und (Unter)-überschrift) stärker gewichten. Basierend auf den Ergebnissen der verstärkten Beachtung des Abstracts wird ebenso eine Untersuchung der (Unter-)überschriften durchgeführt, diese findet sich ab Seite 41. Eine weitere Versuchsreihe wird ab B.3 gezeigt, hier werden die vorherigen Ergebnisse auf ihre Plausibilität geprüft indem der reine Text höher gewichtet wird.

4.4 Interpretation

Die folgenden Abschnitte widmen sich der Interpretation der im Anhang aufgelisteten Ergebnisse.

Grundlegend ist festzustellen, dass einige Distanzen offensichtlich nicht berechnet werden konnten, so dass diese mit 'NaN' (Not A Number) markiert wurden. Der Autor ist zum momentanen Zeitpunkt nicht in der Lage dieses Phänomen zu erklären. Dieses erschwert die Versuchsinterpretation, macht sie allerdings nicht unmöglich, da dieser Fehler nur in den Sonderfällen auftritt (welche dem Autor u.A. zum Auffinden von groben Fehlern dienen).

Darüber hinaus fällt zunächst auf, dass die normierte euklidische Distanz (basierend auf der Termfrequency (TF)) sich ähnlich verhält wie die normierte TF-IDF Variante der selben Funktion. dieses ist über alle im Anhang gezeigten Diagramme mit verschiedenen Gewichtungen der Fall. Hingegen neigt die rein TF basierte Variante der euklidischen Distanz zu stärkeren Ausbrüchen und verhält sich daher deutlich anders als die beiden anderen Versionen (z.B. Seite 33. Besonders ausgeprägt ist dieses in den Sonderfällen der Gewichtung zu sehen (z.B. nur Gewichtung des Abstracs; siehe Seiten 23, 27).

Eine weitere Auffälligkeit im Verhalten der TF basierten Distanz ist der Verlauf der Precision bei niedrigem Recall. Dieser scheint bei wenig extremen Gewichtungen der Zusammenfassungen (also einem

guten Verhältnis von normalem Text zum Rest) zunächst die Precision hoch zu halten und daraufhin stark abzufallen (siehe Seiten: B.1, B.2, B.3).

Des Weiteren lässt die Testreihe den Schluss zu, dass eine gut gewählte stärkere Gewichtung des Abstracts (und in geringerem Maße der (Unter-)Überschriften) eines Artikels die durchschnittliche Precision verbessern kann (siehe Seite 33). Hierbei wurde die durchschnittliche Precision der normierten TF Distanz um 4% verbessert, die TF-IDF basierte Distanz verbesserte sich um 0,7% gegenüber der Basissituation. Allerdings fällt innerhalb dieses Szenarios der Gewichtung die normale Termfrequency basierte durchschnittliche Precision stark ab (zwischen 6% und 22%; siehe Seiten: 33, 37). Bei dieser generellen verstärkten Beachtung der zusammenfassenden Anteile des Artikels tritt daher keine signifikante Verbesserung ein.

Betrachtet man die Versuchsreihe, welche verstärkt auf die (Unter-)überschriften eingeht (ab Seite 41) fällt auf, dass diese das Potenzial besitzt die Resultate für zwei von drei Distanzfunktionsvarianten stark zu verbessern. Wie bereits in den anderen Versuchsreihen verschlechtert sich das Resultat der rein Termfrequency basierten Funktion, im Gegenzug verbessern sich die normierten Varianten allerdings um ca. 8% bzw. 5% im direkten Vergleich zur normalen Gewichtung (Seite 20).

Grundsätzlich zeigt sich, dass die Versuchsreihen abhängig von den Parametern die Ergebnisse durchaus verbessern können. Allerdings ist das Testen verschiedener Kombinationen aufwändig und fehlerträchtig, da eine Übergewichtung der Parameter die Ergebnisse wie beschrieben auch verschlechtern kann.

Darüber hinaus ist auch zu bedenken, dass die Versuchsreihen auf anderen Korpi wiederholt werden sollten um u.A. die Stabilität der Ergebnisse gegen zu prüfen.

5 Fazit

Die vorliegende Arbeit stellte zunächst das Thema, sowie fachliche und technische Grundlagen vor. Daraufhin wurde eine Umsetzung des verfolgten Ansatzes, samt seiner theoretischen Basis vorgestellt.

Zunächst ist es gelungen eine praktische Umsetzung der dargestellten Theorie mit dem Tool RapidMiner zu erreichen. Diese hat allerdings, wie geschildert, durchaus offenes Verbesserungspotenzial, auch wenn dieses keinen direkten Einfluss auf die Effektivität des erstellten Frameworks hat.

Darüber hinaus wurde im Rahmen der Arbeit am Artikelarchiv ein tieferes Verständnis für dieses gewonnen, sowie diverse Probleme identifiziert und soweit nötig behoben. Diese umfassten wie geschildert z.B. ein invalides XML-Markup.

Betrachtet man die Interpretation der Versuchsreihen so belegen diese, dass diverse Gewichtungen von speziellen Textanteilen eindeutig dazu beitragen können die Ergebnisse (im Vergleich zum Ausgangsfall der Normalgewichtung) zu verbessern. Dieses zeigt sich besonders wenn man eine abwechselnde Anordnung von 'richtigen' und 'falschen' Ergebnissen (Erfolgsquote 50%) als Vergleichsbasis nutzt. Anzumerken ist, dass das Ermitteln der richtigen Parameter zur Gewichtung zeitaufwändig ist, so dass sich potenziell gute Versuchsreihen bei überhöhten Gewichtungen auch negativ weiterentwickeln können (siehe: Wichtung des Abstracs, Seite 33 ff.). Auch zeigt sich, dass die Sonderfälle der Gewichtung (29 ff.) Indikatoren sein können, die anzeigen welche Parameter potenziell erfolgversprechend sind. Dieses sollte in weiteren Arbeiten berücksichtigt werden.

Trotz der guten Ergebnisse gibt es einige Punkte zu bedenken. Zum einen sind die vorliegenden Testreihen ein gutes Indiz für die Nützlichkeit von diversen Gewichtungsparametern zur Verbesserung von Distanzfunktionen unter Dokumenten, allerdings sollten die Ergebnisse hinsichtlich ihrer Stabilität auf anderen Dokumenten untersucht werden. Darüber hinaus wäre es ebenso denkbar, dass die gewählte Auswertungsmethodik (welche frühe Fehler stark in die Ergebnisse mit einfließen lässt), sowie die Zusammensetzung des Testkorpus (50% 'richtige', sowie 50% 'falsche' Dokumente) die Ergebnisse maßgeblich beeinflussen.

Zusammenfassend zeigt sich, dass das durchgeführte Projekt durchaus erfreuliche Erfolge aber auch Raum für Verbesserungen mit sich bringt. Sodass das Forschungsfeld weiterhin interessant ist und weitere Arbeiten durchaus denkbar und sinnvoll sind.

5.1 Weitere Schritte

Die vorliegende Arbeit lässt sich in vielfältiger Weise erweitern. Diverse neue Features sind in Planung, diese umfassen u.A. ein Kategoriensystem. Dieses soll Schlagwörter aus mehreren semantischen Kategorien (wie z.B. Sport, Politik etc.) erkennen und mit einem beliebigen Faktor hervorheben und diese daraufhin in die Distanzfunktion mit einfließen lassen.

Weiterhin sollen die Erkenntnisse einer Domänenexpertin genutzt werden um die Distanzfunktion um journalistische Wünsche und Einflussfaktoren für Pressedossiers zu erweitern. Die Arbeiten hierzu werden in [Hä15] durchgeführt. Inwieweit diese Erkenntnisse umsetzbar und hilfreich sind soll innerhalb der Masterarbeit des Autors geklärt werden.

Über die genannten weiteren Schritte hinaus besteht eine Vielzahl von Möglichkeiten welche es wert wären untersucht zu werden. Diese umfassen beispielsweise eine erweiterte Vorverarbeitung mit Hilfe von Natural Language Processing', sowie die Nutzung von Ontologien.

Literatur

[BOHG13] Bobadilla, J. ; Ortega, F. ; Hernando, A. ; GutiéRrez, A.: Recommender Systems Survey. In: **Know.-Based Syst.** 46 (2013), Juli, 109–132. http://dx.doi.org/10.1016/j.knosys.2013.03.012. – DOI 10.1016/j.knosys.2013.03.012. – ISSN 0950–7051

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

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

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

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

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

[RM-] : RapidMiner API-Dokumentation (inoffiziell). http://fossies.org/dox/rapidminer-5.3.013/index.html. Community. – API Dokumentation aus Quellcode

[RM-08a] : RapidMiner API Dokumentation (offiziell). Version: 2008. http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/WS0809/rm-api/overview-summary.html. TU Dortmund, 2008. – API Dokumentation

[RM-08b] : RapidMiner Dokumentation. Version: 2008. http://docs.rapidminer.com/. RapidMiner Inc., 2008. – Dokumentation

[Sch15] Schöneberg, Marcel: Automatisierte Erstellung von Pressedossiers durch Textmining. 2015. – Ausarbeitung

A Dokumente

A.1 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 Archivs zur Verwendung im Umfeld der Masterarbeit des Autors wurden die dargestellten Untersuchungen möglich.

A.2 Beispielartikel

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

Listing 1: Beispielartikel

```
<?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
```

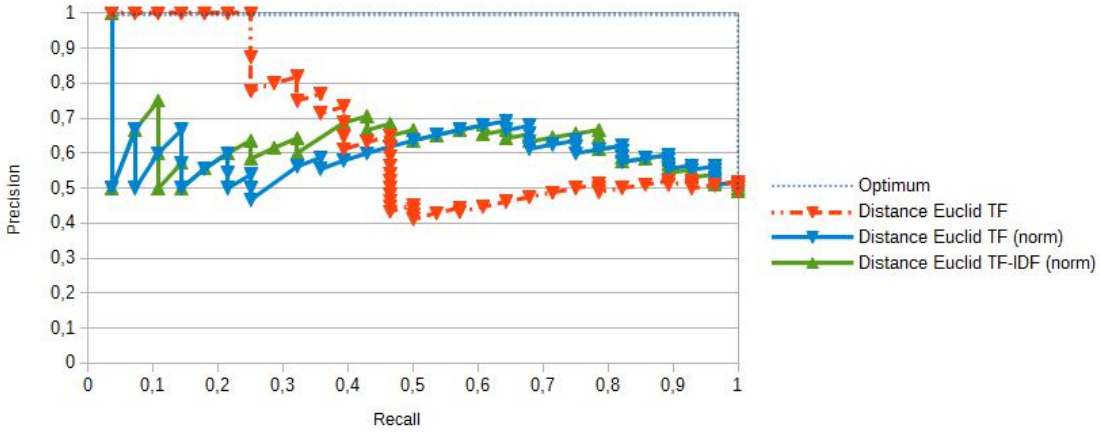
```
change.

<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>
```

B Testergebnisse

B.1 Basis- und Sonderfälle

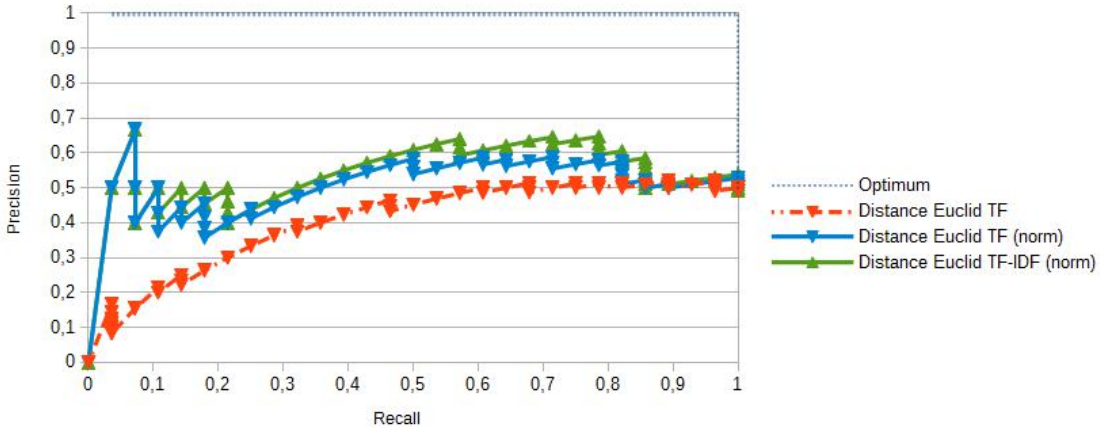
Parameter	Wert	Durchschnitt Precision	
Abstract	1	Optimal	
Title	1	Euclid TF	72,68%
Subheadings	1	Euclid TF (norm)	70,31%
Body	1	Euclid TF-IDF	72,79%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid 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

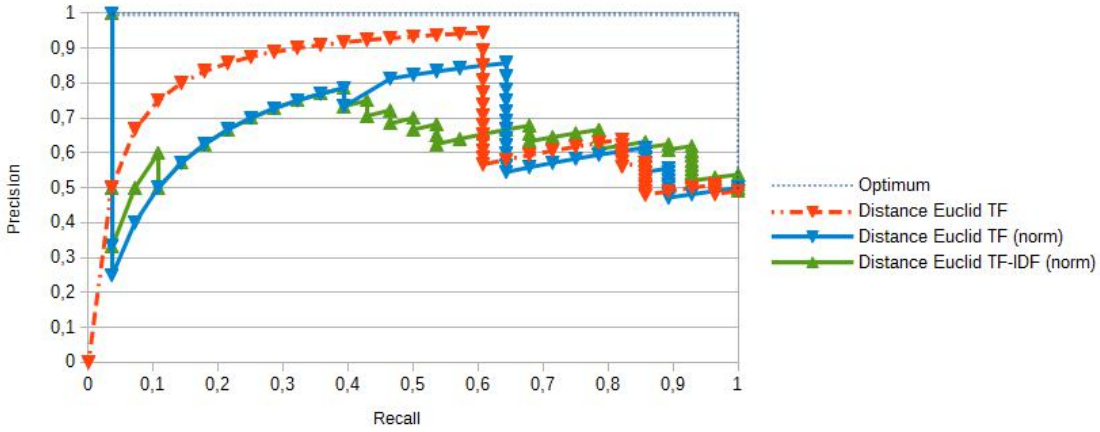
Parameter	Wert	Durchschnitt Precision	
Abstract	1	Optimal	
Title	0	Euclid TF	42,44%
Subheadings	0	Euclid TF (norm)	60,03%
Body	0	Euclid TF-IDF	63,80%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	6,782	3,391	6,681
3	2008-05-02-wennerhag-en.xml	8,307	2,077	3,501
4	2008-11-21-leggewiewelzer-en.xml	7,416	3,708	7,376
5	2009-04-21-fraser-en.xml	9,381	4,690	8,524
6	2009-07-14-biscione-en.xml	8,485	2,828	5,155
7	2009-09-09-kavaliauskas-en.xml	6,403	6,403	14,007
8	2010-09-14-ditchev-en.xml	6,856	3,428	5,944
9	2011-07-11-bluhdorn-en.xml	6,557	3,279	6,564
10	2011-11-02-G1000-en.xml	8,062	2,687	4,907
11	2011-11-10-sierakowski-en.xml	6,708	3,354	7,175
12	2011-12-19-amirpur-en.xml	7,071	7,071	14,714
13	2012-01-25-halmai-en.xml	8,062	4,031	6,620
14	2012-02-08-elsenhans-en.xml	7,000	3,500	6,386
15	2012-09-05-jahanbegloo-en.xml	6,856	6,856	14,511
16	2012-11-21-holmes-en.xml	6,403	3,202	6,506
17	2013-02-08-wallerstein-en.xml	6,633	3,317	6,327
18	2013-02-19-leggewie-en.xml	6,928	3,464	6,613
19	2013-02-26-james-en.xml	6,557	3,279	6,400
20	2013-05-03-muller-en.xml	6,782	3,391	7,335
21	2013-06-14-pomerantsev-en.xml	7,483	1,871	4,422
22	2013-07-29-gole-en.xml	6,633	3,317	7,233
23	2013-08-13-krastev-en.xml	6,481	3,240	6,819
24	2013-08-20-leggewienanz-en.xml	7,416	3,708	6,226
25	2013-09-11-deniztekin-en.xml	6,633	6,633	14,699
26	2013-11-08-vidanava-en.xml	6,928	3,464	7,251
27	2013-11-22-offe-en.xml	5,916	2,958	6,059
28	2013-12-12-margetts-en.xml	6,782	6,782	12,488

29	2013-12-12-pogonyi-en.xml	7,550	3,775	6,961
30	1117-2007-07-06-lapin1-en.xml	6,325	3,162	7,588
31	1193-2007-11-02-boulbina-en.xml	8,544	2,848	6,122
32	1270-2008-04-09-miklosi-en.xml	6,633	6,633	13,836
33	1344-2008-08-07-seymour-en.xml	6,481	6,481	13,593
34	2100-2011-09-27-scruton-en.xml	6,782	3,391	6,956
35	211-2002-12-20-verene-en.xml	5,916	2,958	6,161
36	2163-2012-01-11-ohlheiser-en.xml	6,928	3,464	7,666
37	2200-2012-03-20-mondediploo-en.xml	4,796	NaN	NaN
38	223-2003-01-31-des-en.xml	7,874	2,625	5,157
39	2447-2013-04-12-sanchez-en.xml	7,141	3,571	7,297
40	2495-2013-06-25-zhurzhenko-en.xml	6,403	6,403	13,685
41	2517-2013-08-13-osteuropa-en.xml	4,796	NaN	NaN
42	256-2003-02-11-kaplinski-en.xml	6,325	6,325	12,869
43	266-2003-02-16-mangasassen-en.xml	7,483	3,742	9,592
44	2666-2014-04-03-knausgard-en.xml	6,782	3,391	6,938
45	294-2003-03-04-ursic-en.xml	4,796	NaN	NaN
46	335-2003-05-15-henard-en.xml	5,831	5,831	13,063
47	414-2003-10-20-bogdanovic-en.xml	7,071	2,357	4,566
48	441-2003-11-28-abraham-en.xml	6,000	6,000	13,113
49	479-2004-03-03-senyener-en.xml	6,325	6,325	15,993
50	480-2004-03-04-cakmak-en.xml	9,055	4,528	11,124
51	505-2004-04-05-uys-en.xml	6,481	3,240	6,089
52	540-2004-06-21-peters-en.xml	10,863	1,810	2,722
53	62-2001-04-01-mistry-en.xml	8,944	2,981	4,647
54	661-2005-07-14-revista-en.xml	4,796	NaN	NaN
55	772-2006-02-01-boutang-en.xml	7,416	3,708	7,392
56	785-2006-02-16-sambrook-en.xml	8,124	2,708	5,991
57	80-2001-11-14-blecher-en.xml	6,164	6,164	12,545
58	904-2006-08-17-eder-en.xml	8,246	2,749	5,145

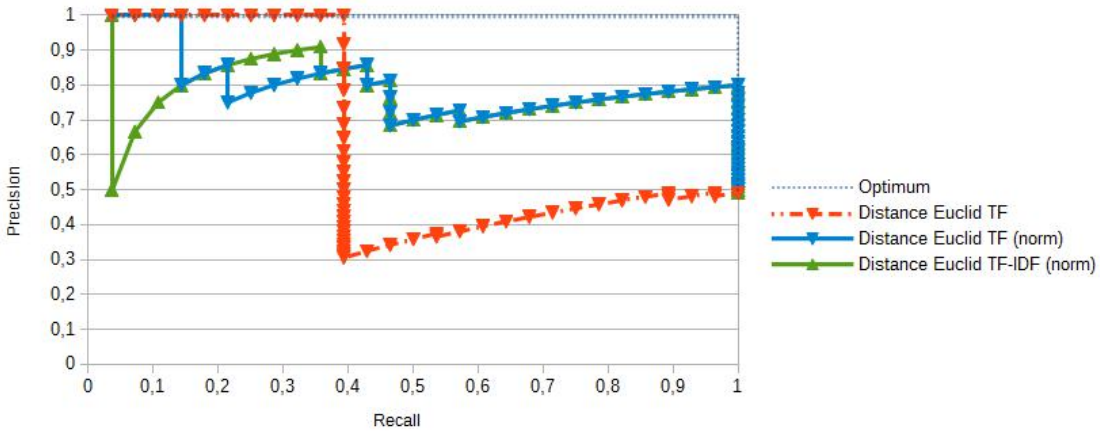
Parameter	Wert	Durchschnitt Precision	
Abstract	0	Optimal	
Title	1	Euclid TF	79,32%
Subheadings	0	Euclid TF (norm)	74,02%
Body	0	Euclid TF-IDF	74,61%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	2,000	2,000	4,903
3	2008-05-02-wennerhag-en.xml	2,236	2,236	4,186
4	2008-11-21-leggewiewelzer-en.xml	2,449	2,449	4,491
5	2009-04-21-fraser-en.xml	2,646	2,646	4,729
6	2009-07-14-biscione-en.xml	2,236	2,236	5,070
7	2009-09-09-kavaliauskas-en.xml	2,000	2,000	5,850
8	2010-09-14-ditchev-en.xml	2,000	2,000	4,106
9	2011-07-11-bluhdorn-en.xml	2,000	2,000	4,283
10	2011-11-02-G1000-en.xml	2,000	2,000	5,283
11	2011-11-10-sierakowski-en.xml	2,236	2,236	4,844
12	2011-12-19-amirpur-en.xml	2,000	2,000	4,429
13	2012-01-25-halmai-en.xml	2,000	2,000	4,903
14	2012-02-08-elsenhans-en.xml	3,742	1,247	1,891
15	2012-09-05-jahanbegloo-en.xml	2,646	2,646	6,675
16	2012-11-21-holmes-en.xml	2,000	2,000	5,792
17	2013-02-08-wallerstein-en.xml	2,000	2,000	4,096
18	2013-02-19-leggewie-en.xml	2,000	2,000	5,089
19	2013-02-26-james-en.xml	2,000	2,000	5,367
20	2013-05-03-muller-en.xml	2,000	2,000	4,226
21	2013-06-14-pomerantsev-en.xml	2,236	2,236	6,440
22	2013-07-29-gole-en.xml	2,000	2,000	4,187
23	2013-08-13-krastev-en.xml	2,236	2,236	4,813
24	2013-08-20-leggewienanz-en.xml	2,000	2,000	4,336
25	2013-09-11-deniztekin-en.xml	2,236	2,236	5,033
26	2013-11-08-vidanava-en.xml	2,000	2,000	4,897
27	2013-11-22-offe-en.xml	2,000	2,000	4,267
28	2013-12-12-margetts-en.xml	2,646	2,646	4,930

29	2013-12-12-pogonyi-en.xml	2,000	2,000	4,370
30	1117-2007-07-06-lapin1-en.xml	2,000	2,000	6,368
31	1193-2007-11-02-boulbina-en.xml	2,236	2,236	4,800
32	1270-2008-04-09-miklosi-en.xml	2,000	2,000	6,292
33	1344-2008-08-07-seymour-en.xml	2,000	2,000	4,944
34	2100-2011-09-27-scruton-en.xml	2,000	2,000	6,152
35	211-2002-12-20-verene-en.xml	2,000	2,000	5,021
36	2163-2012-01-11-ohlheiser-en.xml	2,000	2,000	4,514
37	2200-2012-03-20-mondediploo-en.xml	2,646	2,646	8,644
38	223-2003-01-31-des-en.xml	2,646	1,323	2,912
39	2447-2013-04-12-sanchez-en.xml	2,000	2,000	5,855
40	2495-2013-06-25-zhurzhenko-en.xml	2,449	2,449	7,149
41	2517-2013-08-13-osteuropa-en.xml	2,000	2,000	5,966
42	256-2003-02-11-kaplinski-en.xml	2,646	1,323	2,168
43	266-2003-02-16-mangasassen-en.xml	2,449	2,449	7,163
44	2666-2014-04-03-knausgard-en.xml	2,236	2,236	5,871
45	294-2003-03-04-ursic-en.xml	1,732	1,732	4,121
46	335-2003-05-15-henard-en.xml	2,449	2,449	4,878
47	414-2003-10-20-bogdanovic-en.xml	2,236	2,236	4,873
48	441-2003-11-28-abraham-en.xml	2,449	2,449	5,258
49	479-2004-03-03-senyener-en.xml	2,236	2,236	6,437
50	480-2004-03-04-cakmak-en.xml	2,000	2,000	5,987
51	505-2004-04-05-uys-en.xml	2,000	2,000	4,820
52	540-2004-06-21-peters-en.xml	2,000	2,000	7,066
53	62-2001-04-01-mistry-en.xml	2,449	2,449	5,639
54	661-2005-07-14-revista-en.xml	2,236	2,236	7,226
55	772-2006-02-01-boutang-en.xml	2,449	2,449	5,123
56	785-2006-02-16-sambrook-en.xml	2,449	2,449	5,221
57	80-2001-11-14-blecher-en.xml	2,000	2,000	5,847
58	904-2006-08-17-eder-en.xml	2,449	2,449	4,720

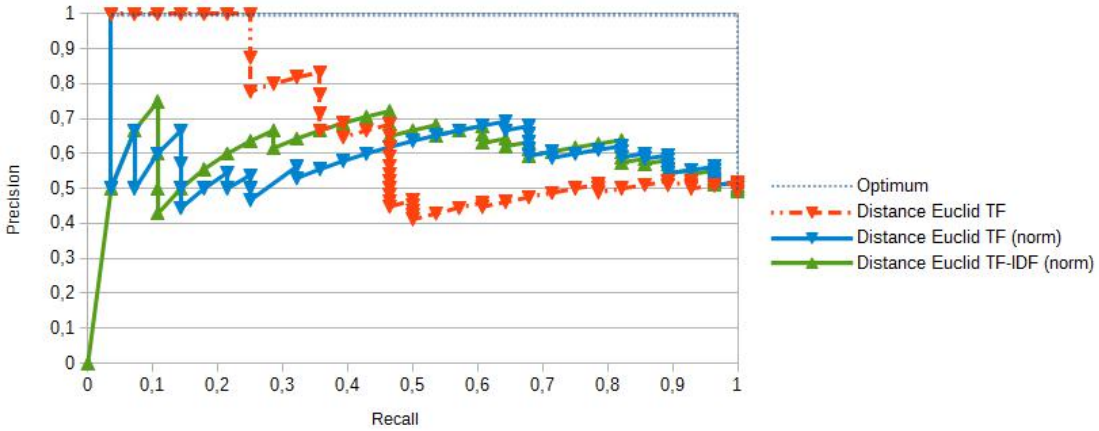
Parameter	Wert	Durchschnitt Precision	
Abstract	0	Optimal	
Title	0	Euclid TF	68,73%
Subheadings	1	Euclid TF (norm)	86,90%
Body	0	Euclid TF-IDF	85,52%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	4,796	4,796	8,876
3	2008-05-02-wennerhag-en.xml	3,606	NaN	NaN
4	2008-11-21-leggewiewelzer-en.xml	5,196	2,598	4,980
5	2009-04-21-fraser-en.xml	3,606	NaN	NaN
6	2009-07-14-biscione-en.xml	7,000	2,333	4,762
7	2009-09-09-kavaliauskas-en.xml	5,292	1,764	4,659
8	2010-09-14-ditchev-en.xml	3,606	NaN	NaN
9	2011-07-11-bluhdorn-en.xml	4,690	2,345	4,282
10	2011-11-02-G1000-en.xml	4,472	4,472	8,030
11	2011-11-10-sierakowski-en.xml	3,606	NaN	NaN
12	2011-12-19-amirpur-en.xml	5,292	2,646	4,611
13	2012-01-25-halmai-en.xml	5,916	1,972	3,226
14	2012-02-08-elsenhans-en.xml	5,292	1,764	3,163
15	2012-09-05-jahanbegloo-en.xml	3,606	NaN	NaN
16	2012-11-21-holmes-en.xml	5,292	5,292	11,775
17	2013-02-08-wallerstein-en.xml	3,606	NaN	NaN
18	2013-02-19-leggewie-en.xml	7,483	1,871	3,752
19	2013-02-26-james-en.xml	3,606	NaN	NaN
20	2013-05-03-muller-en.xml	4,583	4,583	8,876
21	2013-06-14-pomerantsev-en.xml	3,606	NaN	NaN
22	2013-07-29-gole-en.xml	5,831	2,915	6,137
23	2013-08-13-krastev-en.xml	5,916	2,958	5,700
24	2013-08-20-leggewienanz-en.xml	5,916	2,958	4,485
25	2013-09-11-deniztekin-en.xml	3,606	NaN	NaN
26	2013-11-08-vidanava-en.xml	3,606	NaN	NaN
27	2013-11-22-offe-en.xml	7,141	1,785	2,104
28	2013-12-12-margetts-en.xml	5,292	1,764	2,993

29	2013-12-12-pogonyi-en.xml	3,606	NaN	NaN
30	1117-2007-07-06-lapin1-en.xml	3,606	NaN	NaN
31	1193-2007-11-02-boulbina-en.xml	3,606	NaN	NaN
32	1270-2008-04-09-miklosi-en.xml	3,606	NaN	NaN
33	1344-2008-08-07-seymour-en.xml	4,899	4,899	10,872
34	2100-2011-09-27-scruton-en.xml	3,606	NaN	NaN
35	211-2002-12-20-verene-en.xml	3,606	NaN	NaN
36	2163-2012-01-11-ohlheiser-en.xml	3,606	NaN	NaN
37	2200-2012-03-20-mondediploo-en.xml	3,606	NaN	NaN
38	223-2003-01-31-des-en.xml	3,873	3,873	7,378
39	2447-2013-04-12-sanchez-en.xml	4,243	4,243	7,695
40	2495-2013-06-25-zhurzhenko-en.xml	6,782	2,261	4,765
41	2517-2013-08-13-osteuropa-en.xml	4,123	4,123	7,339
42	256-2003-02-11-kaplinski-en.xml	3,606	NaN	NaN
43	266-2003-02-16-mangasassen-en.xml	3,606	NaN	NaN
44	2666-2014-04-03-knausgard-en.xml	3,606	NaN	NaN
45	294-2003-03-04-ursic-en.xml	3,606	NaN	NaN
46	335-2003-05-15-henard-en.xml	3,606	NaN	NaN
47	414-2003-10-20-bogdanovic-en.xml	3,606	NaN	NaN
48	441-2003-11-28-abraham-en.xml	3,606	NaN	NaN
49	479-2004-03-03-senyener-en.xml	3,606	NaN	NaN
50	480-2004-03-04-cakmak-en.xml	3,606	NaN	NaN
51	505-2004-04-05-uys-en.xml	3,606	NaN	NaN
52	540-2004-06-21-peters-en.xml	7,211	1,803	2,572
53	62-2001-04-01-mistry-en.xml	6,325	3,162	6,488
54	661-2005-07-14-revista-en.xml	3,606	NaN	NaN
55	772-2006-02-01-boutang-en.xml	3,606	NaN	NaN
56	785-2006-02-16-sambrook-en.xml	3,606	NaN	NaN
57	80-2001-11-14-blecher-en.xml	3,606	NaN	NaN
58	904-2006-08-17-eder-en.xml	3,606	NaN	NaN

Parameter	Wert	Durchschnitt Precision	
Abstract	0	Optimal	
Title	0	Euclid TF	73,07%
Subheadings	0	Euclid TF (norm)	69,56%
Body	1	Euclid TF-IDF	70,44%

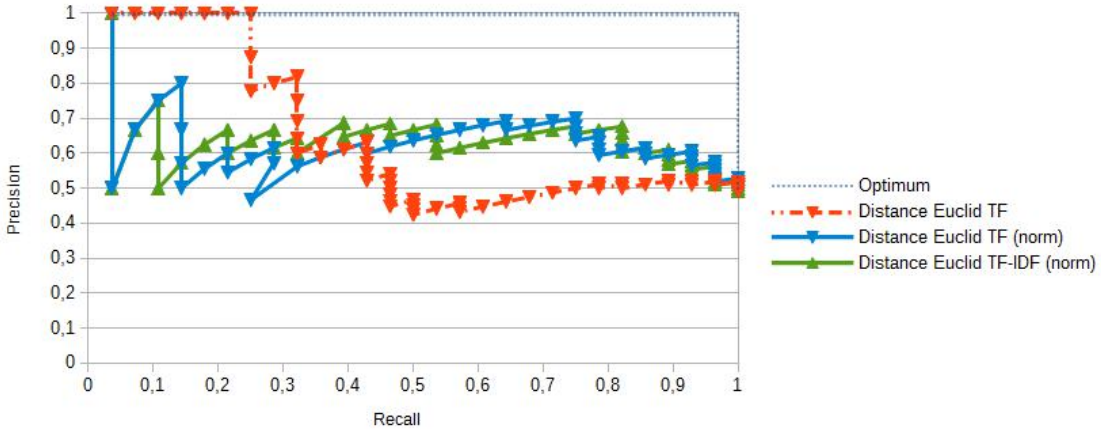


#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	171,706	3,733	5,192
3	2008-05-02-wennerhag-en.xml	177,924	2,170	2,707
4	2008-11-21-leggewiewelzer-en.xml	134,573	4,806	7,366
5	2009-04-21-fraser-en.xml	201,333	1,766	2,225
6	2009-07-14-biscione-en.xml	155,496	3,793	6,563
7	2009-09-09-kavaliauskas-en.xml	152,624	3,634	5,819
8	2010-09-14-ditchev-en.xml	135,440	8,465	13,269
9	2011-07-11-bluhdorn-en.xml	147,411	2,730	4,278
10	2011-11-02-G1000-en.xml	132,684	2,764	4,477
11	2011-11-10-sierakowski-en.xml	140,975	2,563	3,814
12	2011-12-19-amirpur-en.xml	159,031	3,001	5,559
13	2012-01-25-halmai-en.xml	209,929	1,578	2,010
14	2012-02-08-elsenhans-en.xml	161,608	5,213	7,404
15	2012-09-05-jahanbegloo-en.xml	147,374	4,466	7,260
16	2012-11-21-holmes-en.xml	141,517	2,573	4,056
17	2013-02-08-wallerstein-en.xml	155,090	5,170	7,022
18	2013-02-19-leggewie-en.xml	184,242	2,924	4,433
19	2013-02-26-james-en.xml	146,216	6,963	10,672
20	2013-05-03-muller-en.xml	154,360	3,150	4,558
21	2013-06-14-pomerantsev-en.xml	158,931	3,116	5,264
22	2013-07-29-gole-en.xml	176,193	2,098	3,778
23	2013-08-13-krastev-en.xml	139,162	6,627	9,883
24	2013-08-20-leggewienanz-en.xml	160,019	2,623	3,741
25	2013-09-11-deniztekin-en.xml	144,755	18,094	25,201
26	2013-11-08-vidanava-en.xml	145,024	8,531	12,718
27	2013-11-22-offe-en.xml	137,975	2,464	3,813
28	2013-12-12-margetts-en.xml	158,732	2,442	3,637

29	2013-12-12-pogonyi-en.xml	145,190	5,007	7,157
30	1117-2007-07-06-lapin1-en.xml	152,912	21,845	29,076
31	1193-2007-11-02-boulbina-en.xml	165,288	4,350	7,123
32	1270-2008-04-09-miklosi-en.xml	156,662	4,476	6,744
33	1344-2008-08-07-seymour-en.xml	225,672	2,564	4,674
34	2100-2011-09-27-scruton-en.xml	148,536	9,902	13,727
35	211-2002-12-20-verene-en.xml	170,438	2,506	3,966
36	2163-2012-01-11-ohlheiser-en.xml	151,387	4,883	8,243
37	2200-2012-03-20-mondediploo-en.xml	148,883	18,610	25,390
38	223-2003-01-31-des-en.xml	153,170	6,660	8,857
39	2447-2013-04-12-sanchez-en.xml	145,959	9,122	13,180
40	2495-2013-06-25-zhurzhenko-en.xml	190,544	2,071	5,320
41	2517-2013-08-13-osteuropa-en.xml	147,289	7,752	11,064
42	256-2003-02-11-kaplinski-en.xml	152,342	5,859	8,329
43	266-2003-02-16-mangasassen-en.xml	150,519	2,428	3,414
44	2666-2014-04-03-knausgard-en.xml	328,512	1,720	1,997
45	294-2003-03-04-ursic-en.xml	153,219	30,644	40,029
46	335-2003-05-15-henard-en.xml	142,292	9,486	14,747
47	414-2003-10-20-bogdanovic-en.xml	183,619	2,416	3,261
48	441-2003-11-28-abraham-en.xml	152,089	7,604	10,276
49	479-2004-03-03-senyener-en.xml	154,201	7,009	10,827
50	480-2004-03-04-cakmak-en.xml	151,911	5,425	7,978
51	505-2004-04-05-uys-en.xml	142,871	7,144	11,414
52	540-2004-06-21-peters-en.xml	190,234	2,238	2,747
53	62-2001-04-01-mistry-en.xml	221,820	2,641	4,757
54	661-2005-07-14-revista-en.xml	151,529	10,102	13,567
55	772-2006-02-01-boutang-en.xml	145,479	10,391	15,441
56	785-2006-02-16-sambrook-en.xml	145,921	4,707	7,130
57	80-2001-11-14-blecher-en.xml	147,604	4,473	6,466
58	904-2006-08-17-eder-en.xml	152,542	4,487	6,189

B.2 Hervorhebung der zusammenfassenden Anteile

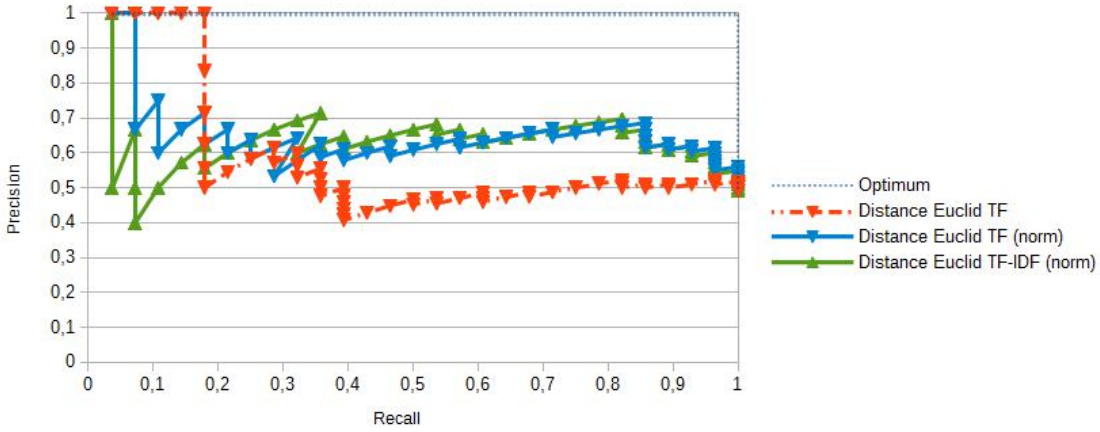
Parameter	Wert	Durchschnitt Precision	
Abstract	4	Optimal	
Title	3	Euclid TF	71,72%
Subheadings	2	Euclid TF (norm)	72,78%
Body	1	Euclid TF-IDF	73,68%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	184,225	3,232	4,727
3	2008-05-02-wennerhag-en.xml	197,924	1,960	2,507
4	2008-11-21-leggewiewelzer-en.xml	145,017	4,028	6,626
5	2009-04-21-fraser-en.xml	213,607	1,874	2,515
6	2009-07-14-biscione-en.xml	172,238	3,515	6,249
7	2009-09-09-kavaliauskas-en.xml	165,934	3,457	5,816
8	2010-09-14-ditchev-en.xml	145,685	6,070	10,049
9	2011-07-11-bluhdorn-en.xml	160,350	2,393	3,929
10	2011-11-02-G1000-en.xml	148,233	2,745	4,676
11	2011-11-10-sierakowski-en.xml	150,256	2,732	4,356
12	2011-12-19-amirpur-en.xml	175,417	2,741	5,147
13	2012-01-25-halmai-en.xml	229,521	1,561	2,039
14	2012-02-08-elsenhans-en.xml	173,853	4,346	6,500
15	2012-09-05-jahanbegloo-en.xml	162,468	4,062	6,918
16	2012-11-21-holmes-en.xml	151,911	2,337	3,871
17	2013-02-08-wallerstein-en.xml	168,187	4,805	6,828
18	2013-02-19-leggewie-en.xml	203,929	3,237	5,074
19	2013-02-26-james-en.xml	159,132	5,683	9,194
20	2013-05-03-muller-en.xml	168,039	3,429	5,230
21	2013-06-14-pomerantsev-en.xml	170,294	3,339	5,929
22	2013-07-29-gole-en.xml	194,895	2,030	3,743
23	2013-08-13-krastev-en.xml	151,199	5,214	8,461
24	2013-08-20-leggewienanz-en.xml	177,679	2,820	4,117
25	2013-09-11-deniztekin-en.xml	156,585	10,439	15,611
26	2013-11-08-vidanava-en.xml	156,199	6,791	10,910

27	2013-11-22-offe-en.xml	148,805	2,325	3,760
28	2013-12-12-margetts-en.xml	174,565	2,359	3,616
29	2013-12-12-pogonyi-en.xml	159,944	3,999	5,984
30	1117-2007-07-06-lapin1-en.xml	165,251	11,017	15,969
31	1193-2007-11-02-boulbina-en.xml	182,732	4,351	7,467
32	1270-2008-04-09-miklosi-en.xml	168,567	4,322	6,825
33	1344-2008-08-07-seymour-en.xml	238,122	2,533	4,700
34	2100-2011-09-27-scruton-en.xml	160,801	8,463	12,486
35	211-2002-12-20-verene-en.xml	185,324	2,346	3,831
36	2163-2012-01-11-ohlheiser-en.xml	163,438	4,191	7,391
37	2200-2012-03-20-mondediploo-en.xml	159,634	19,954	28,814
38	223-2003-01-31-des-en.xml	167,815	4,093	5,756
39	2447-2013-04-12-sanchez-en.xml	158,701	8,353	12,810
40	2495-2013-06-25-zhurzhenko-en.xml	207,593	2,097	5,408
41	2517-2013-08-13-osteuropa-en.xml	157,886	8,310	12,441
42	256-2003-02-11-kaplinski-en.xml	164,994	6,346	9,386
43	266-2003-02-16-mangasassen-en.xml	161,858	2,611	4,025
44	2666-2014-04-03-knausgard-en.xml	338,167	1,771	2,098
45	294-2003-03-04-ursic-en.xml	163,905	32,781	45,361
46	335-2003-05-15-henard-en.xml	153,366	8,072	13,221
47	414-2003-10-20-bogdanovic-en.xml	200,142	2,199	3,096
48	441-2003-11-28-abraham-en.xml	163,851	7,124	10,229
49	479-2004-03-03-senyener-en.xml	166,111	6,389	10,646
50	480-2004-03-04-cakmak-en.xml	167,269	5,227	8,484
51	505-2004-04-05-uys-en.xml	155,023	5,537	9,205
52	540-2004-06-21-peters-en.xml	219,324	2,215	2,744
53	62-2001-04-01-mistry-en.xml	243,758	2,902	5,099
54	661-2005-07-14-revista-en.xml	162,395	10,826	15,365
55	772-2006-02-01-boutang-en.xml	158,183	6,327	9,893
56	785-2006-02-16-sambrook-en.xml	161,762	4,622	7,526
57	80-2001-11-14-blecher-en.xml	158,085	4,790	7,312
58	904-2006-08-17-eder-en.xml	169,873	3,693	5,400

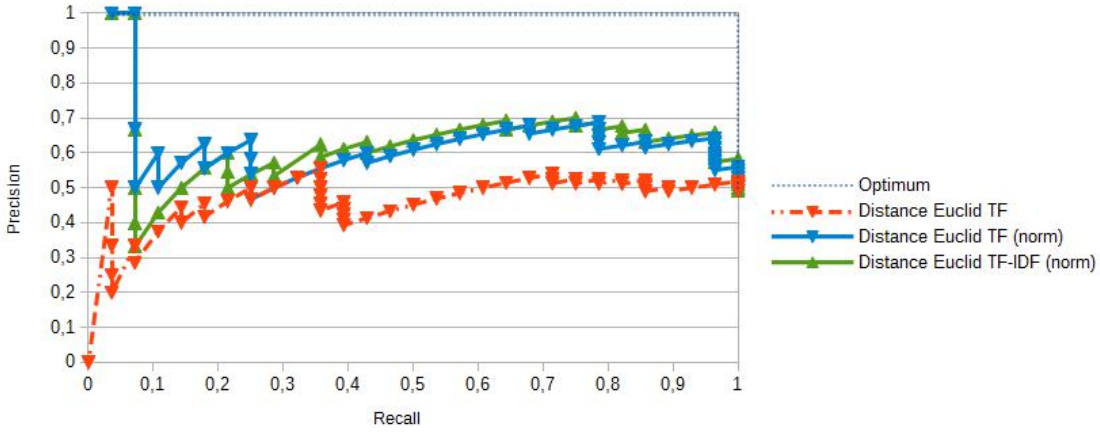
Parameter	Wert	Durchschnitt Precision	
Abstract	10	Optimal	
Title	8	Euclid TF	66,43%
Subheadings	5	Euclid TF (norm)	74,77%
Body	1	Euclid TF-IDF	73,46%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid 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

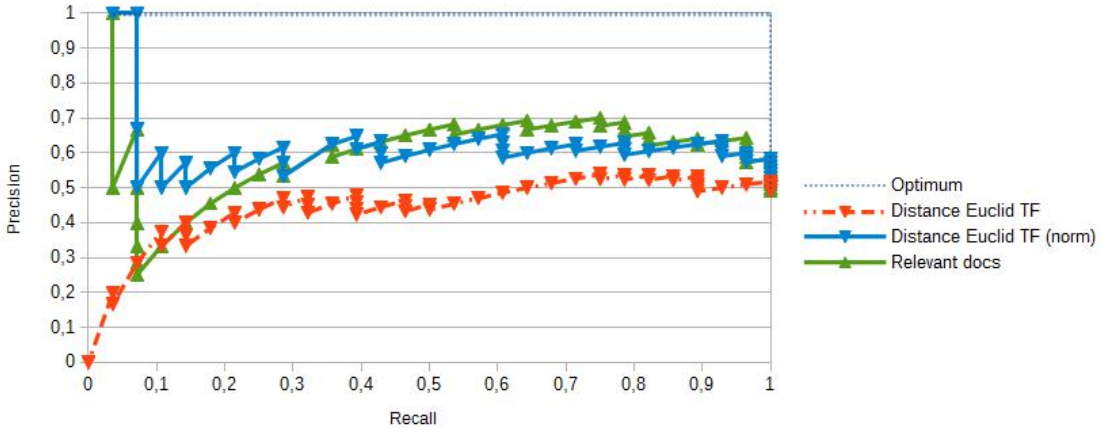
Parameter	Wert	Durchschnitt Precision	
Abstract	20	Optimal	
Title	10	Euclid TF	54,65%
Subheadings	5	Euclid TF (norm)	72,33%
Body	1	Euclid TF-IDF	72,75%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	255,456	2,580	4,367
3	2008-05-02-wennerhag-en.xml	301,516	1,753	2,476
4	2008-11-21-leggewiewelzer-en.xml	220,851	3,248	6,227
5	2009-04-21-fraser-en.xml	305,254	2,678	4,313
6	2009-07-14-biscione-en.xml	270,137	3,335	6,116
7	2009-09-09-kavaliauskas-en.xml	236,335	3,527	6,769
8	2010-09-14-ditchev-en.xml	214,567	3,832	6,939
9	2011-07-11-bluhdorn-en.xml	231,927	2,128	3,922
10	2011-11-02-G1000-en.xml	243,916	2,323	4,256
11	2011-11-10-sierakowski-en.xml	215,659	3,478	6,696
12	2011-12-19-amirpur-en.xml	256,673	2,760	5,464
13	2012-01-25-halmai-en.xml	320,180	1,703	2,451
14	2012-02-08-elsenhans-en.xml	249,493	2,901	4,861
15	2012-09-05-jahanbegloo-en.xml	243,010	3,857	7,381
16	2012-11-21-holmes-en.xml	215,353	2,154	4,084
17	2013-02-08-wallerstein-en.xml	240,703	3,253	5,310
18	2013-02-19-leggewie-en.xml	288,175	3,167	5,391
19	2013-02-26-james-en.xml	232,237	4,074	7,436
20	2013-05-03-muller-en.xml	243,335	2,765	4,981
21	2013-06-14-pomerantsev-en.xml	246,382	2,899	6,068
22	2013-07-29-gole-en.xml	276,919	2,067	4,064
23	2013-08-13-krastev-en.xml	225,812	3,273	6,314
24	2013-08-20-leggewienanz-en.xml	266,141	4,032	6,275
25	2013-09-11-deniztekin-en.xml	228,449	6,012	11,115
26	2013-11-08-vidanava-en.xml	226,952	4,126	7,934
27	2013-11-22-offe-en.xml	210,528	2,193	4,023
28	2013-12-12-margetts-en.xml	251,865	2,799	4,708

29	2013-12-12-pogonyi-en.xml	246,546	3,121	5,245
30	1117-2007-07-06-lapin1-en.xml	232,867	4,955	9,295
31	1193-2007-11-02-boulbina-en.xml	281,931	3,661	7,136
32	1270-2008-04-09-miklosi-en.xml	238,092	4,329	7,958
33	1344-2008-08-07-seymour-en.xml	300,063	2,655	5,186
34	2100-2011-09-27-scruton-en.xml	232,826	4,565	8,070
35	211-2002-12-20-verene-en.xml	255,605	2,166	3,898
36	2163-2012-01-11-ohlheiser-en.xml	236,502	3,331	6,681
37	2200-2012-03-20-mondediploo-en.xml	211,592	21,159	36,263
38	223-2003-01-31-des-en.xml	258,159	2,506	4,178
39	2447-2013-04-12-sanchez-en.xml	235,317	5,739	10,405
40	2495-2013-06-25-zhurzhenko-en.xml	279,834	2,499	6,209
41	2517-2013-08-13-osteuropa-en.xml	208,708	9,938	17,231
42	256-2003-02-11-kaplinski-en.xml	233,630	4,768	8,119
43	266-2003-02-16-mangasassen-en.xml	239,522	3,863	8,063
44	2666-2014-04-03-knausgard-en.xml	391,504	2,050	2,764
45	294-2003-03-04-ursic-en.xml	214,898	21,490	35,306
46	335-2003-05-15-henard-en.xml	214,271	6,122	11,760
47	414-2003-10-20-bogdanovic-en.xml	283,533	1,942	3,122
48	441-2003-11-28-abraham-en.xml	227,280	7,576	13,255
49	479-2004-03-03-senyener-en.xml	233,673	5,564	11,657
50	480-2004-03-04-cakmak-en.xml	268,890	5,602	11,844
51	505-2004-04-05-uys-en.xml	224,738	3,746	6,800
52	540-2004-06-21-peters-en.xml	364,251	2,118	2,818
53	62-2001-04-01-mistry-en.xml	352,780	2,556	4,143
54	661-2005-07-14-revista-en.xml	214,210	14,281	23,995
55	772-2006-02-01-boutang-en.xml	238,514	3,727	6,718
56	785-2006-02-16-sambrook-en.xml	255,730	3,602	7,044
57	80-2001-11-14-blecher-en.xml	219,868	6,663	11,929
58	904-2006-08-17-eder-en.xml	270,710	2,880	4,822

Parameter	Wert	Durchschnitt Precision	
Abstract	30	Optimal	
Title	20	Euclid TF	50,53%
Subheadings	5	Euclid TF (norm)	71,92%
Body	1	Euclid TF-IDF	69,82%

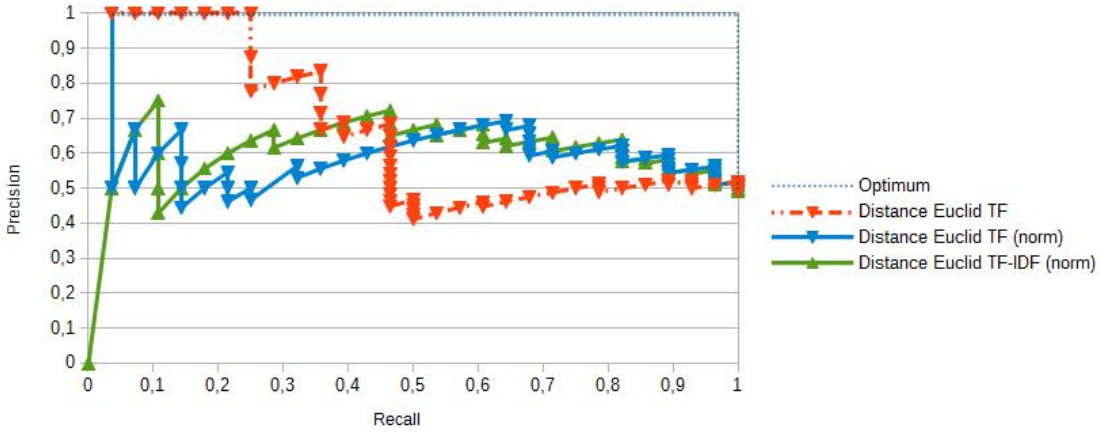


#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	314,353	2,437	4,340
3	2008-05-02-wennerhag-en.xml	384,619	1,733	2,528
4	2008-11-21-leggewiewelzer-en.xml	288,193	3,275	6,446
5	2009-04-21-fraser-en.xml	386,316	3,389	5,729
6	2009-07-14-biscione-en.xml	345,013	3,350	6,198
7	2009-09-09-kavaliauskas-en.xml	292,359	3,797	7,664
8	2010-09-14-ditchev-en.xml	272,725	3,588	6,617
9	2011-07-11-bluhdorn-en.xml	291,393	2,096	3,977
10	2011-11-02-G1000-en.xml	318,959	2,363	4,404
11	2011-11-10-sierakowski-en.xml	271,936	3,316	6,752
12	2011-12-19-amirpur-en.xml	320,782	2,839	5,718
13	2012-01-25-halmai-en.xml	386,206	1,857	2,811
14	2012-02-08-elsenhans-en.xml	318,884	2,531	4,331
15	2012-09-05-jahanbegloo-en.xml	310,313	3,739	7,452
16	2012-11-21-holmes-en.xml	269,438	2,245	4,456
17	2013-02-08-wallerstein-en.xml	301,294	2,897	4,942
18	2013-02-19-leggewie-en.xml	349,550	3,149	5,574
19	2013-02-26-james-en.xml	293,656	3,375	6,433
20	2013-05-03-muller-en.xml	303,763	2,574	4,897
21	2013-06-14-pomerantsev-en.xml	312,384	2,499	5,517
22	2013-07-29-gole-en.xml	336,250	2,183	4,393
23	2013-08-13-krastev-en.xml	287,386	2,903	5,820
24	2013-08-20-leggewienanz-en.xml	334,352	4,399	7,025
25	2013-09-11-deniztekin-en.xml	290,188	5,003	9,848
26	2013-11-08-vidanava-en.xml	286,613	3,822	7,702
27	2013-11-22-offe-en.xml	261,461	2,254	4,325
28	2013-12-12-margetts-en.xml	315,335	3,153	5,470

29	2013-12-12-pogonyi-en.xml	316,552	2,904	5,050
30	1117-2007-07-06-lapin1-en.xml	287,553	4,292	8,799
31	1193-2007-11-02-boulbina-en.xml	358,978	3,355	6,776
32	1270-2008-04-09-miklosi-en.xml	296,054	4,555	8,926
33	1344-2008-08-07-seymour-en.xml	349,482	2,841	5,676
34	2100-2011-09-27-scruton-en.xml	291,836	4,110	7,757
35	211-2002-12-20-verene-en.xml	312,336	2,110	3,973
36	2163-2012-01-11-ohlheiser-en.xml	296,164	3,255	6,798
37	2200-2012-03-20-mondediploo-en.xml	254,501	12,725	23,889
38	223-2003-01-31-des-en.xml	335,538	2,193	3,813
39	2447-2013-04-12-sanchez-en.xml	297,412	4,876	9,350
40	2495-2013-06-25-zhurzhenko-en.xml	336,552	2,759	6,748
41	2517-2013-08-13-osteuropa-en.xml	249,718	11,891	22,130
42	256-2003-02-11-kaplinski-en.xml	291,964	4,231	7,548
43	266-2003-02-16-mangasassen-en.xml	307,491	3,494	7,921
44	2666-2014-04-03-knausgard-en.xml	437,396	2,290	3,370
45	294-2003-03-04-ursic-en.xml	255,306	12,765	22,592
46	335-2003-05-15-henard-en.xml	265,917	5,909	11,960
47	414-2003-10-20-bogdanovic-en.xml	352,038	1,893	3,190
48	441-2003-11-28-abraham-en.xml	280,029	7,001	13,164
49	479-2004-03-03-senyener-en.xml	290,832	5,193	11,794
50	480-2004-03-04-cakmak-en.xml	350,117	5,226	11,751
51	505-2004-04-05-uys-en.xml	281,508	3,519	6,571
52	540-2004-06-21-peters-en.xml	463,895	2,000	2,786
53	62-2001-04-01-mistry-en.xml	436,273	2,451	3,874
54	661-2005-07-14-revista-en.xml	255,980	12,799	23,433
55	772-2006-02-01-boutang-en.xml	307,033	3,266	6,118
56	785-2006-02-16-sambrook-en.xml	330,058	3,268	6,696
57	80-2001-11-14-blecher-en.xml	272,180	6,805	12,945
58	904-2006-08-17-eder-en.xml	353,474	2,851	4,929

B.3 Hervorhebung des Textanteils

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

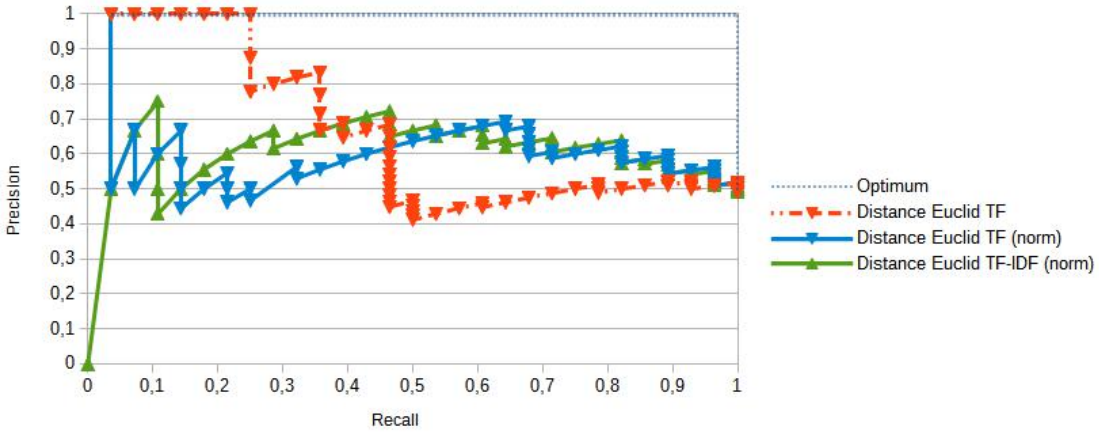


#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	1720,259	3,740	5,210
3	2008-05-02-wennerhag-en.xml	1784,126	2,163	2,701
4	2008-11-21-leggewiewelzer-en.xml	1348,188	4,781	7,343
5	2009-04-21-fraser-en.xml	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

B.4 Hervorhebung von Headings

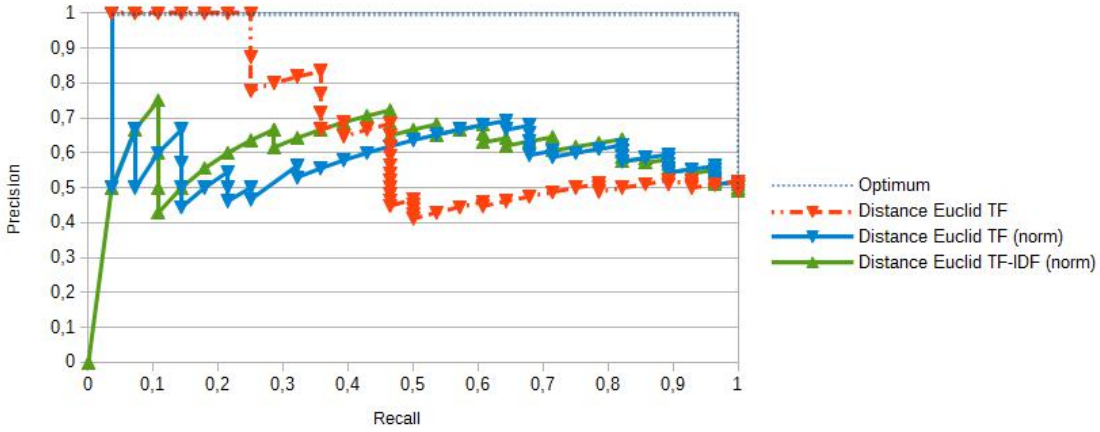
Parameter	Wert	Durchschnitt Precision	
Abstract	1	Optimal	
Title	5	Euclid TF	70,57%
Subheadings	5	Euclid TF (norm)	72,21%
Body	1	Euclid TF-IDF	74,19%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	183,456	3,276	4,821
3	2008-05-02-wennerhag-en.xml	190,887	2,098	2,762
4	2008-11-21-leggewiewelzer-en.xml	147,017	3,770	6,159
5	2009-04-21-fraser-en.xml	210,131	1,843	2,462
6	2009-07-14-biscione-en.xml	171,869	3,997	7,328
7	2009-09-09-kavaliauskas-en.xml	167,812	3,496	6,106
8	2010-09-14-ditchev-en.xml	146,335	8,130	13,390
9	2011-07-11-bluhdorn-en.xml	162,490	2,462	4,019
10	2011-11-02-G1000-en.xml	146,840	2,997	5,098
11	2011-11-10-sierakowski-en.xml	150,688	2,740	4,349
12	2011-12-19-amirpur-en.xml	177,969	2,579	4,741
13	2012-01-25-halmai-en.xml	230,239	1,535	1,997
14	2012-02-08-elsenhans-en.xml	177,983	3,423	5,103
15	2012-09-05-jahanbegloo-en.xml	161,468	4,140	7,067
16	2012-11-21-holmes-en.xml	156,077	2,517	4,209
17	2013-02-08-wallerstein-en.xml	167,619	4,789	6,850
18	2013-02-19-leggewie-en.xml	210,264	2,767	4,451
19	2013-02-26-james-en.xml	158,811	5,882	9,556
20	2013-05-03-muller-en.xml	167,666	3,422	5,205
21	2013-06-14-pomerantsev-en.xml	169,041	3,315	5,820
22	2013-07-29-gole-en.xml	195,954	2,041	3,769
23	2013-08-13-krastev-en.xml	154,376	4,980	8,060
24	2013-08-20-leggewienanz-en.xml	177,432	2,688	3,976
25	2013-09-11-deniztekin-en.xml	156,684	11,192	16,661
26	2013-11-08-vidanava-en.xml	156,349	9,197	14,499

27	2013-11-22-offe-en.xml	153,678	2,650	4,210
28	2013-12-12-margetts-en.xml	178,093	2,095	3,187
29	2013-12-12-pogonyi-en.xml	157,721	4,381	6,621
30	1117-2007-07-06-lapin1-en.xml	164,521	18,280	26,287
31	1193-2007-11-02-boulbina-en.xml	177,423	4,549	7,658
32	1270-2008-04-09-miklosi-en.xml	167,917	4,664	7,370
33	1344-2008-08-07-seymour-en.xml	239,056	2,543	4,741
34	2100-2011-09-27-scruton-en.xml	160,187	10,012	14,875
35	211-2002-12-20-verene-en.xml	183,448	2,446	4,004
36	2163-2012-01-11-ohlheiser-en.xml	161,521	4,895	8,514
37	2200-2012-03-20-mondediploo-en.xml	160,459	20,057	29,490
38	223-2003-01-31-des-en.xml	166,676	4,630	6,534
39	2447-2013-04-12-sanchez-en.xml	158,357	7,541	11,558
40	2495-2013-06-25-zhurzhenko-en.xml	215,397	2,013	5,254
41	2517-2013-08-13-osteuropa-en.xml	159,154	7,579	11,474
42	256-2003-02-11-kaplinski-en.xml	165,185	5,696	8,458
43	266-2003-02-16-mangasassen-en.xml	161,645	2,607	3,941
44	2666-2014-04-03-knausgard-en.xml	335,031	1,754	2,091
45	294-2003-03-04-ursic-en.xml	164,381	32,876	46,053
46	335-2003-05-15-henard-en.xml	154,305	9,077	14,745
47	414-2003-10-20-bogdanovic-en.xml	196,647	2,341	3,305
48	441-2003-11-28-abraham-en.xml	164,024	6,561	9,461
49	479-2004-03-03-senyener-en.xml	165,934	7,215	11,756
50	480-2004-03-04-cakmak-en.xml	163,786	5,648	8,816
51	505-2004-04-05-uys-en.xml	154,561	7,025	11,664
52	540-2004-06-21-peters-en.xml	213,839	2,299	2,925
53	62-2001-04-01-mistry-en.xml	240,624	2,865	5,204
54	661-2005-07-14-revista-en.xml	162,960	10,864	15,673
55	772-2006-02-01-boutang-en.xml	157,569	7,503	11,730
56	785-2006-02-16-sambrook-en.xml	158,423	4,951	7,906
57	80-2001-11-14-blecher-en.xml	158,808	4,812	7,381
58	904-2006-08-17-eder-en.xml	164,791	4,454	6,523

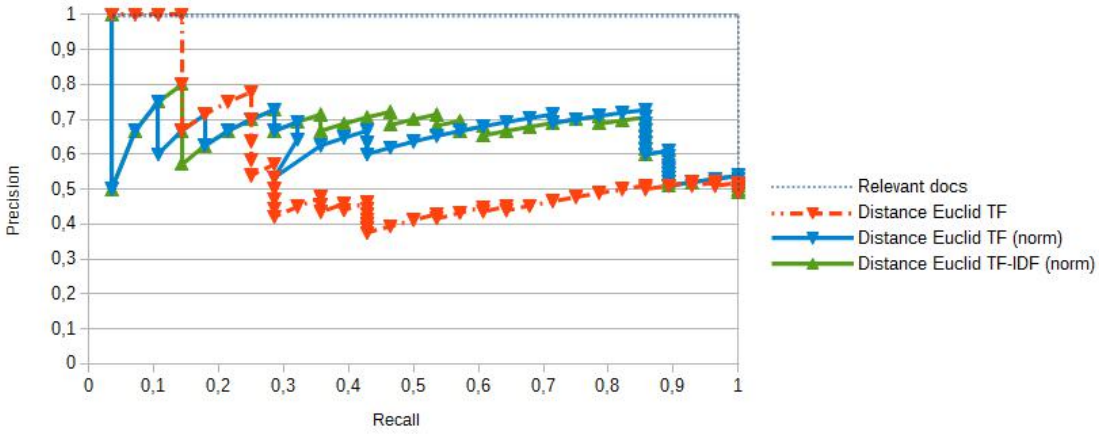
Parameter	Wert	Durchschnitt Precision	
Abstract	1	Optimal	
Title	10	Euclid TF	64,15%
Subheadings	10	Euclid TF (norm)	75,42%
Body	1	Euclid TF-IDF	76,68%



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid TF-IDF (norm)
1	2013-02-01-krastev-en.xml	0,000	0,000	0,000
2	2001-11-27-rosenberg-en.xml	197,411	2,991	4,638
3	2008-05-02-wennerhag-en.xml	202,529	2,110	2,942
4	2008-11-21-leggewiewelzer-en.xml	164,040	3,038	5,184
5	2009-04-21-fraser-en.xml	219,784	1,928	2,714
6	2009-07-14-biscione-en.xml	192,559	3,320	6,370
7	2009-09-09-kavaliauskas-en.xml	186,604	3,521	6,740
8	2010-09-14-ditchev-en.xml	158,836	7,942	13,595
9	2011-07-11-bluhdorn-en.xml	180,203	2,371	3,980
10	2011-11-02-G1000-en.xml	162,764	3,071	5,418
11	2011-11-10-sierakowski-en.xml	162,040	2,946	4,936
12	2011-12-19-amirpur-en.xml	198,879	2,368	4,308
13	2012-01-25-halmai-en.xml	250,789	1,520	2,033
14	2012-02-08-elsenhans-en.xml	200,519	2,445	3,761
15	2012-09-05-jahanbegloo-en.xml	176,187	4,004	7,121
16	2012-11-21-holmes-en.xml	174,499	2,604	4,637
17	2013-02-08-wallerstein-en.xml	180,724	4,518	6,782
18	2013-02-19-leggewie-en.xml	239,731	2,374	4,023
19	2013-02-26-james-en.xml	172,151	5,380	9,173
20	2013-05-03-muller-en.xml	182,666	3,321	5,275
21	2013-06-14-pomerantsev-en.xml	180,375	3,537	6,428
22	2013-07-29-gole-en.xml	216,583	2,043	3,852
23	2013-08-13-krastev-en.xml	174,648	4,260	7,291
24	2013-08-20-leggewienanz-en.xml	197,198	2,777	4,228
25	2013-09-11-deniztekin-en.xml	169,823	8,938	14,073
26	2013-11-08-vidanava-en.xml	168,879	9,934	16,365
27	2013-11-22-offe-en.xml	176,528	2,802	4,436
28	2013-12-12-margetts-en.xml	199,855	1,903	2,940

29	2013-12-12-pogonyi-en.xml	170,649	4,162	6,611
30	1117-2007-07-06-lapin1-en.xml	176,839	14,737	22,700
31	1193-2007-11-02-boulbina-en.xml	189,153	4,850	8,346
32	1270-2008-04-09-miklosi-en.xml	180,003	5,000	8,282
33	1344-2008-08-07-seymour-en.xml	253,571	2,561	4,877
34	2100-2011-09-27-scruton-en.xml	172,699	10,794	17,102
35	211-2002-12-20-verene-en.xml	196,311	2,454	4,146
36	2163-2012-01-11-ohlheiser-en.xml	172,522	5,228	9,322
37	2200-2012-03-20-mondediploo-en.xml	173,485	17,348	27,589
38	223-2003-01-31-des-en.xml	181,248	3,940	5,878
39	2447-2013-04-12-sanchez-en.xml	171,951	6,613	10,692
40	2495-2013-06-25-zhurzhenko-en.xml	243,549	1,996	5,229
41	2517-2013-08-13-osteuropa-en.xml	172,641	6,640	10,615
42	256-2003-02-11-kaplinski-en.xml	179,126	4,593	7,096
43	266-2003-02-16-mangasassen-en.xml	174,281	2,811	4,533
44	2666-2014-04-03-knausgard-en.xml	341,388	1,787	2,203
45	294-2003-03-04-ursic-en.xml	176,383	17,638	26,296
46	335-2003-05-15-henard-en.xml	167,765	7,626	12,842
47	414-2003-10-20-bogdanovic-en.xml	209,213	2,351	3,465
48	441-2003-11-28-abraham-en.xml	177,085	5,903	9,018
49	479-2004-03-03-senyener-en.xml	178,645	7,767	13,231
50	480-2004-03-04-cakmak-en.xml	175,986	6,068	9,940
51	505-2004-04-05-uys-en.xml	167,194	7,600	13,055
52	540-2004-06-21-peters-en.xml	237,628	2,425	3,228
53	62-2001-04-01-mistry-en.xml	260,883	2,867	5,299
54	661-2005-07-14-revista-en.xml	175,488	11,699	18,140
55	772-2006-02-01-boutang-en.xml	170,860	6,572	10,737
56	785-2006-02-16-sambrook-en.xml	171,356	5,355	8,910
57	80-2001-11-14-blecher-en.xml	171,275	5,190	8,412
58	904-2006-08-17-eder-en.xml	177,090	4,786	7,383

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



#	Filenename	Distance Euclid TF	Distance Euclid TF (norm)	Distance Euclid 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