FAKULTÄT FÜR !NFORMATIK

Klassifizierung von Web-Dokumenten

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

im Rahmen des Studiums

Software Engineering & Internet Computing

ausgeführt von

Bernhard Wachter
Matrikelnummer 0225650

am:
Institut für Softwaretechnik und Interaktive Systeme

Betreuung:
Betreuer/Betreuerin: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber

Wien, 18.07.2008

(Unterschrift Verfasser/in)

(Unterschrift Betreuer/in)

Dank

An dieser Stelle möchte ich mich bei allen Freunden, die mich bei der Erstellung dieser Arbeit auf verschiedene Art und Weise unterstützt haben, bedanken. Besonderer Dank gebührt auch meinem Betreuer Herrn Prof. Dipl.-Ing. Dr.techn. Andreas Rauber, der mir in allen Phasen meiner Arbeit mit Rat und Tat zur Seite gestanden ist.

Ich widme die vorliegende Arbeit meinen Eltern, die mir das Studium ermöglicht haben. Danke!

Abstract

Web archiving is the process of collecting and preserving web documents. The massive archives are rapidly growing and contain sensitive data. To prevent abuse it is important to identify sensitive data and restrict access to it. This also allows use cases where sensitive data are used for analysis without revealing them. The purpose of the genre-analysis is to classify a web-document based on its form and its style, independently of the underlying topic. The aim of this paper is to extend this method for usage within a web archive. This extension will allow distinguishing private from public elements within a web-document. Traditional approaches only allow operating on document-level. But especially web-documents often contain multiple genres within a single document. Therefore an approach is developed which allows the recognition of text segments and genre transitions. Based on this paragraph splitter a classifier for differing private from public elements of a web document is developed. This system may operate on document-level as well as on paragraph-level.

Kurzfassung

Bei der Web-Archivierung werden Web-Dokumente gesammelt und dauerhaft abgelegt. Die entstehen Archive wachsen rasant und enthalten auch sensitive Daten. Um Missbrauch vorzubeugen müssen sensitive Daten identifiziert und gegen unbefugte Zugriffe gesichert werden. Dadurch werden Anwendungsfälle denkbar in welchen auf Basis von sensitiven Daten Auswertungen vorgenommen werden ohne dabei die Daten selbst preis zugeben. Bei der Genre-Analyse werden Web-Dokumente aufgrund ihrer Form sowie des Stils einer Seite unabhängig vom eigentlichen Thema klassifiziert. Zielsetzung dieser Arbeit ist es, diese Methode dahingehend zu erweitern, dass es einem Archivierungssystem von Web-Dokumenten möglich ist, private und öffentliche Elemente von Web-Dokumenten unterscheiden zu können. Bisherige Ansätze in diesem Bereich agieren ausschließlich auf Dokumentenebene. Web-Dokumente enthalten jedoch häufig mehrere unterschiedliche Genres. Diese Arbeit entwickelt einen Ansatz zur Erkennung von zusammengehörenden Textsegmenten, welcher Absätze und gegebenenfalls damit verbundene Genre-Übergänge erkennt. Darauf aufbauend wird ein Ansatz für die Klassifizierung von privaten und öffentlichen Elementen von Web-Dokumenten auf Dokumenten- und Absatzebene vorgestellt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Das World Wide Web entstand 1989 als Hypertext-System und wird im allgemeinen Sprachgebrauch oft mit dem Internet gleichgesetzt. Während die Darstellung zuerst auf reinen Text beschränkt war, haben sich durch die Entwicklung von Browsern, welche verschiedenste Dateiformate verarbeiten und darstellen können, sehr schnell auch Web-Dokumente mit multimedialen Inhalten verbreitet.

Die Verbreitung des World Wide Web hat in vielfacher Weise den Umgang mit Informationen verändert. Das World Wide Web bietet heutzutage für nahezu alle Bereiche Informationen in gigantischem Umfang. Aufgrund seiner Bedeutung und der kurzen durchschnittlichen Lebensdauer von Web-Dokumenten ergibt sich die Notwendigkeit einer Archivierung. Aufgrund dieser Notwendigkeit entwickelte sich die Disziplin der Web-Archivierung.

Unter dem Begriff der Web-Archivierung versteht man das Sammeln sowie die dauerhafte Speicherung von Netzpublikationen mit dem Zweck, auch zukünftig einen Blick in die Vergangenheit zu ermöglichen. Die Web-Archivierung verfolgt das Ziel, einen möglichst großen Ausschnitt der im Internet vorhandenen Web-Präsenzen in systematischer Form abzubilden. Hierfür nötig ist eine Sammlungspolitik, ein Auswahlverfahren sowie ein festzulegende Häufigkeit der Archivierung. Nach Möglichkeit soll ein archiviertes Web-Dokument mit allen multimedialen Funktion (z.B. samt Stylesheets, Skripten, Bildern und Videos) auf Dauer erhalten werden.

Die größte internationale Einrichtung zur Web-Archivierung ist das Internet Archive in San Francisco (USA), welches sich als Archiv des gesamten World Wide Web versteht [22]. Staatliche Archive und Bibliotheken in vielen Ländern unternehmen bereits Anstrengungen zur Sicherung des World Wide Web - Bereichs der jeweiligen

länderspezfischen Top-Level-Domain [1]. So sind in Österreich, Deutschland und der Schweiz bereits Gesetze beschlossen bzw. in Begutachtung welche die rechtliche Basis für die Webarchivierung darstellen und die Sammlung von Online-Publikationen regeln. Entsprechende Pilotprojekte sind geplant bzw. wurden bereits initiiert. [21] [19] [20]. In Großbritannien wurde bereits 2004 ein entsprechendes Pilotprojekt durch das 'UK Web Archiving Consortium' gestartet [2].

Die durch diese Bewegung entstandenen Archive gigantischen Umfangs müssen natürlich auch vor Missbrauch geschützt werden. Hierbei ergeben sich nicht nur rechtliche Probleme, beispielsweise durch Copyright - Verletzungen, sondern auch in Hinsicht auf den Datenschutz sehr bedenkliche Anwendungsmöglichkeiten. Daher wird dem Schutz der Privatsphäre von Autoren eine hohe Bedeutung beigemessen.

Um Einsichtsmöglichkeiten für unterschiedliche Benutzergruppen zu beschränken ist es notwendig Web-Dokumente in 'öffentliche' und 'private' Elemente unterteilen zu können. Nur somit kann sichergestellt werden, dass beispielsweise ein Zeitungsartikel von Kommentaren oder eine Produktseite eines Webshops von Rezensionen getrennt werden kann.

Beispielsweise werden Standpunkte archiviert welche der Autor nicht mehr vertritt, sei dies in Form eines Kommentars, eines Blog Eintrags etc. Ein weiteres häufiges Problem ist, dass Inhalte archiviert werden welche gegen geltendes Recht verstoßen. Klassische Vertreter hierfür sind Verleumdungen, üble Nachreden, Ehrenbeleidigungen und ähnliches. Zugriffsmöglichkeiten für solche Inhalte sollen beschränkt werden können.

Zudem kann eine derartige Klassifizierung als weitere Unterstützung für Suchfunktionen, ähnlich den Ergebnissen der Genre-Analyse[2], dienen.

Formal betrachtet sind die Klassen 'öffentlich' und 'privat' jeweils ein Set von Genres. *Santini* definiert ein Genre als Typ eines Dokumentes mit gleichen lexikalischen, syntaktischen und Layout - Merkmalen sowie einer gleichen kommunikativen Intention [28]. Trotz der etwas vagen Definition wird klar, dass es sich bei einem Genre nicht um das Thema, sondern vielmehr um die Struktur und den Typ eines Dokumentes handelt. Somit können Texte mit verschiedenen Themen das gleiche Genre haben und Texte mit dem gleichen Thema können wiederum unterschiedlichen Genres angehören.

[1] Jeder Name einer Domain im Internet besteht aus einer Folge von durch Punkten getrennten Zeichen. 'Top-Level-Domain' bezeichnet dabei den letzten Namen dieser Folge und stellt die höchste Ebene der Namensauflösung dar.

[2] Bei der Genre-Analyse werden Web-Dokumente aufgrund ihrer Form sowie des Stils einer Seite unabhängig vom eigentlichen Thema klassifiziert. Ein Web-Dokument zu einem Thema kann beispielsweise ein wissenschaftlicher Artikel, eine Sammlung von kurzen Antworten auf spezielle Fragen oder aber auch eine Sammlung von Links zu themenverwandten Seiten sein.

Die besondere Herausforderung hierbei liegt bei der Findung von Merkmalen, welche zuverlässig private von öffentlichen Elementen trennen können. Eine weitere Schwierigkeit stellt die höhere erforderliche Granularität dar. Bei der klassischen Genre-Analyse wird lediglich auf Dokumentenebene klassifiziert. Eine Hierarchie in einem Webarchivierungs System kann jedoch mehrere Ebenen besitzen. So besteht eine WebSite aus mehreren Dokumenten, welche wiederum aus mehreren Absätzen bestehen. Die tiefste Stufe in dieser Hierarchie wäre schließlich die Wortebene. Eine solche Hierarchie erlaubt es die Zugriffsbeschränkungen zu parametrisieren. So können für bestimmte Benutzergruppen von vornherein WebSites welche in irgendeiner Form sensitive Daten enthalten gesperrt werden. Für andere Benutzergruppen wiederum können die Zugriffsmöglichkeiten so beschränkt werden, dass lediglich Absätze bzw. sogar einzelne Wörter mit sensitiven Inhalten entsprechend gesperrt bzw. ausgeblendet werden. Hierdurch ergibt sich die Schwierigkeit, ein Web-Dokument, in welchem üblicherweise mehrere Genres vertreten sind, möglichst präzise in zusammengehörende Absätze zu trennen.

1.1 Zielsetzung der Arbeit

Ziel dieser Arbeit ist es einen Ansatz für die Klassifizierung von privaten und öffentlichen Elementen von Web-Dokumenten auf Dokumenten- und Absatzebene zu erarbeiten. Dafür werden verschiedene Merkmale von Web-Dokumenten untersucht und jene, welche für die Klassifizierung geeignet sind, ausgewählt und anschließend mittels eines adäquaten Klassifizierungsalgorithmus ausgewertet. Weiters soll dieser Ansatz auf Dokumenten- und Absatzebene implementiert und evaluiert werden.

1.2 Gliederung der Arbeit

Die vorliegende Arbeit gliedert sich in die Abschnitte Analyse, Entwurf, Implementierung, Evaluierung und Zusammenfassung. In der nachfolgenden Analyse folgt eine detaillierte Problemstellung sowie eine Analyse der hierfür existierenden Lösungsansätze. Im nächsten Abschnitt Entwurf werden die theoretischen Grundlagen von verwendeten Merkmalen sowie jene des maschinellen Lernens und der Variablenselektion behandelt. In den darauf folgenden zwei Abschnitten wird schließlich die eigentliche Implementierung vorgestellt und evaluiert. Abschließend folgt eine Zusammenfassung.

2. Analyse

Dieser Abschnitt soll die Anforderungen an das Ziel-System aufzeigen und daraus ergebend eine formale Problemstellung ableiten. Weiters sollen die Möglichkeiten existierender Ansätze aufgezeigt werden beziehungsweise eine entsprechende Abgrenzung erfolgen.

2.1 Anforderungen

Die Zielsetzung des nachfolgend erarbeiteten Ansatzes ist ein Klassifikator welcher sowohl auf Dokumenten- als auch auf Absatzebene möglichst zuverlässig private und öffentliche Elemente von Web-Dokumenten trennen kann. Dafür erforderlich ist auch eine möglichst präzise Trennung von zusammengehörenden Absätzen.

Zielsetzung dieses Ansatzes ist es einen Recall[1] von privaten Elementen von 90% oder mehr zu erreichen, wobei zugleich die Precision[2] nicht unter 85% sinken soll. Aufgrund des möglichen Einsatz-Zwecks eines solchen Systems als Komponente in einem Web-Archivierungssystem liegt das Hauptaugenmerk auf der Maximierung des Recalls.

Um die Genre-Sets für die Ergebnis-Klassen bilden zu können, ist zuerst die Frage zu klären, welche Menge an Genres verwendet werden soll um das Web als Ganzes abzubilden. Der hier vorgestellte Ansatz basiert auf der Genre Menge des von *Santini* verwendeten Testkorpus "7 web genre collection"[29], welcher zugleich für die

[1] Recall ist ein Maß zur Beurteilung der Güte einer Treffermenge beim Information Retrieval. Der Recall gibt die Vollständigkeit eines Suchergebnisses an und beschreibt den Anteil relevanter Elemente aus der Gesamtmenge welche sich in der Ergebnismenge befinden.

[2] Precision ist ein Maß zur Beurteilung der Güte einer Treffermenge beim Information Retrieval. Die Precision gibt die Genauigkeit eines Suchergebnisses an und beschreibt mit dem Anteil relevanter Elemente in der Ergebnismenge die Genauigkeit eines Suchergebnisses.

Trainings-Instanzen als auch für Evaluierung auf Dokumentenebene verwendet wird (Abschnitt 5.2).

Als private Elemente sind alle Textelemente eines Web-Dokuments definiert, welche dem Genre 'Blog' oder aber dem Genre 'Personal Homepage' angehören (Tabelle 2.1). Die Differenz zwischen der Gesamtmenge an Genres und der eben definierten Menge an privaten Genres ist wiederum als öffentliche Klasse definiert. Die grundsätzliche Annahme auf welcher diese Einteilung basiert ist jene, dass die beiden der Klasse "privat" zugeordneten Genres möglichst viele Charakteristiken von "privaten" Elementen verkörpern, während zugleich die Gesamtanzahl an Genres überschaubar bleibt. Die relativ grobe Genre-Unterteilung des Webs soll sich zudem positiv auf die Güte des zu erstellenden Klassifikators auswirken.

Genre	Klasse
Blog	**Privat**
Personal Homepage	**Privat**
E-Shop	Öffentlich
FAQ	Öffentlich
Listings	Öffentlich
Newspaper Frontpage	Öffentlich
Search Page	Öffentlich

Tabelle 2.1: Auflistung der verwendeten Genre-Menge samt Klassenzugehörigkeiten.

Der nachfolgend beschriebene Ansatz wird für die englische Sprache implementiert. Allerdings werden die zur Klassifizierung verwendeten Merkmale auch unter dem Aspekt der Sprachunabhängigkeit ausgewählt, wodurch der vorgestellte Ansatz auch für andere Sprachen adaptiert werden kann.

2.2 Verwandte Arbeiten

Santini zeigt ein grundlegendes Problem bei der automatischen Merkmalsberechnung von Web-Dokumenten aufgrund deren Eigenheiten auf [29]. Tippfehler, Grammatik Fehler, die Verwendung von exotischen Namen, HTMl Tags oder Codestücke sind hierbei einige Beispiele welche eine entsprechende Vorverarbeitung unumgänglich machen. Doch selbst bei einer kompletten Entfernung von HTML Tags kann es zu Einflüssen kommen, welche das Ergebnis von Natural Language Processing (NLP) Tools stören. So können beispielsweise unerwartete Zusammensetzungen von Satzzeichen entstehen, bedingt durch das Zusammenfügen von nicht zusammenhängenden Textpassagen. Besondere Merkmale die von *Santini* verwendet werden, sind einerseits sogenannte "POS-Trigrams" und andererseits "Linguistic-Facets". Unter einem POS-Trigram versteht mann hierbei eine Zusammensetzung von drei aufeinander folgenden Part-Of-Speech (POS) Tags welche genre-typische Sprachmuster abbilden.

Das Merkmal Linguistic-Facet stellt hierbei ein aggregiertes Merkmal dar, welches auf Basis verschiedener genre-typischer Merkmale berechnet wird. Hierbei werden auch komplexe Muster als Merkmale verwendet. So wird beispielsweise das Merkmal "Postmodifier of a nominal" berechnet welches bei dem Satz "This mechanism is used for creating a connection between the two components." anschlagen würde da das Nomen "connection" im Nachhinein noch 'modifiziert' wird durch "between". Während Linguistic-Facets weitgehend unabhängig vom verwendeten Korpus sind, ist diese Unabhängigkeit bei den verwendeten 'POS-Trigrams' nur bedingt gegeben [29]. *Santini* erreicht bei der Evaluierung auf einem eigenen Korpus ('7-Web-Genre') je nach Genre F-Maß Werte zwischen 77 und 93 % während unter Verwendung des KI-04 Korpus (Meyer-zu-Eissen-web-page collection) F-Maß Werte zwischen 49 und 76 % erreicht werden.

Es konnte lediglich ein einziges System gefunden werden, welches eine Textkategorisierung auf Absatzebene erlaubt [15]. Dieses System erzielt Recall-Werte welche je nach verwendetem Korpus zwischen 66 % und 82 % liegen und mit nahezu identisch hohen Precision-Werten einhergehen. Jedoch wurde dieser Ansatz für die koreanische Sprache implementiert und evaluiert weshalb die Ergebniswerte nur bedingt vergleichbar sind. Die Funktionsweise dieses Systems ist nachfolgend abgebildet (Abbildung 2.1). Bei der Trennung auf Absatzebene wird hierbei das Dokument einer Komponente 'Passage Splitter' übergeben welche die Trennung von Absätzen vornimmt. Anschließend wird jeder Absatz für sich mittels der Komponente 'Text Classifier' klassifiziert. Abschließend werden die klassifizierten Absätze über die Komponente 'Category Merger' wieder zu einem gemeinsamen Dokument verschmolzen.

Grundsätzlich sind Ergebniswerte von klassischen Textkategorisierungs-Systemen nur bedingt vergleichbar mit den Ergebniswerten des hier vorgestellten Systems. Die Anzahl der in Frage kommenden Klassen eines Textkategorisierungs-Systems bewegt sich meist in einem Bereich von etwa 10 und mehr, im Gegensatz zum nachfolgend vorgestellten System welches diese zu zwei Klassen zusammenfasst.

Auch im Bereich der digitalen Bibliotheken wird neben der Archivierung auch der Kategorisierung von digitalen Objekten hohe Bedeutung beigemessen. *Kim und Ross* verwenden hierbei neben stilistischen, sprachbezogenen und semantischen Merkmalen auch Wissen der jeweiligen Domäne [39]. So wird beispielsweise auch ein Merkmal berechnet welches berücksichtigt welche anderen Arbeiten der jeweilige Autor verfasst hat. Weiters werden auch präsentationsbezogene Merkmale verwendet wobei bei diesem Ansatz lediglich die erste Seite einer Publikation verwendet wird. Der von *Kim und Ross* entwickelte Ansatz berücksichtigt hierbei beispielsweise das Ausmaß von Leerräumen und basiert auf der Annahme dass bestimmte Genres entsprechend

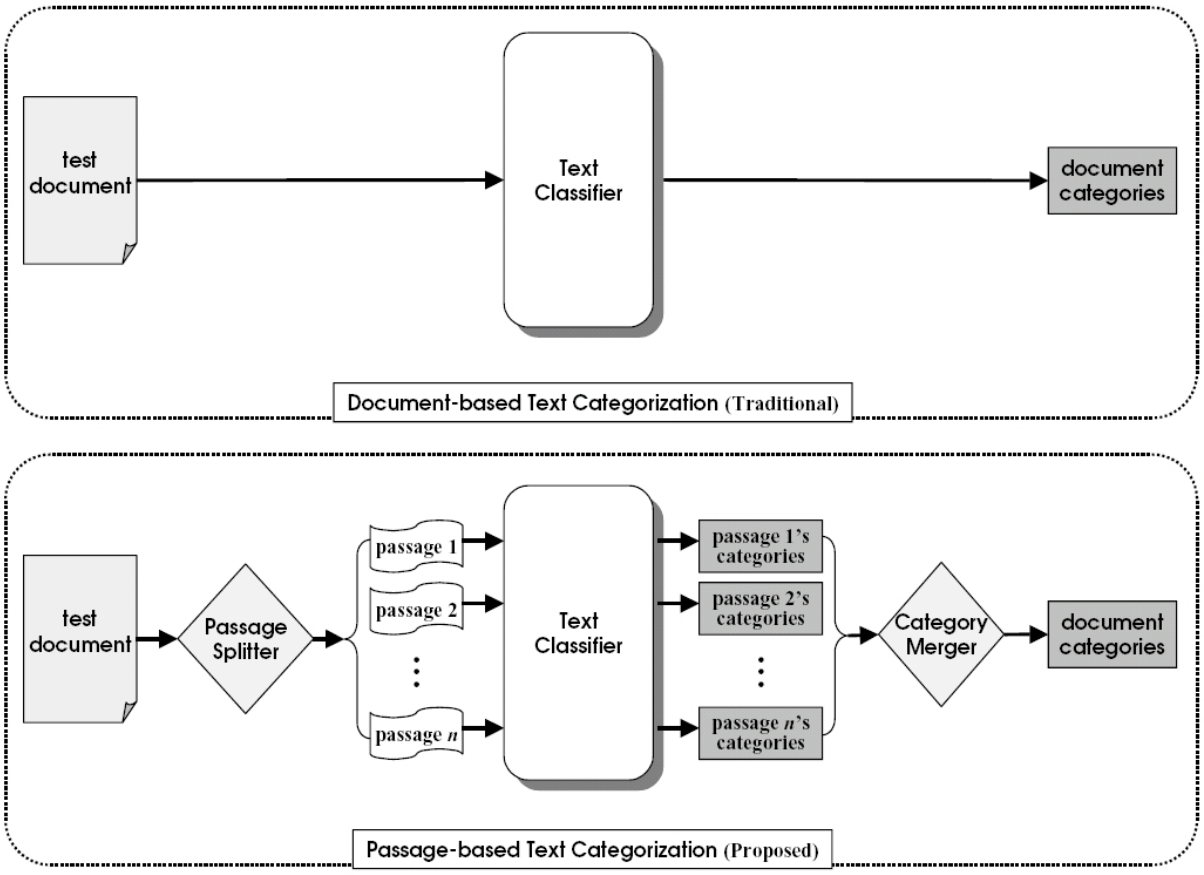

Abbildung 2.1: Funktionsweise eines Textkategorisierungs-Systems auf Dokumenten- und Absatzebene (Abbildung aus [15]).

eindeutige Formatierungen aufweisen [39]. So konnte ein Klassifikator auf Basis von präsentationsbezogenen Merkmalen der ersten Seite bei bestimmten Genres sehr hohe Erkennungsraten aufweisen. Für das Genre 'Formulare' konnte ein Recall - Wert von 88 % mit einer Precision von 69 % erreicht werden, während auch für das Genre 'Produktbeschreibungen' mittels dieses Klassifikators ein Recall von 80 % mit einer Precision von 62 % erreicht wurde.

Neben der eigentlichen Klassifikation existieren verwandte Arbeiten um den Zugriff auf sensible Daten einzuschränken. Neben klassischen Ansätzen im Bereich von Datenbanken existieren vor allem im Bereich des Data Mining interessante Arbeiten welche sich damit befassen Berechnungen auf sensitiven Daten durchzuführen ohne jedoch diese selbst preiszugeben.

Ein klassischer Ansatz um sensible Daten auf Datenbankebene vor unbefugten Zugriffen zu schützen ist neben der Beschränkung von Abfragemöglichkeiten das ge-

zielte 'Stören'. Hierbei werden Einzelwerte bzw. Ergebnisse welche anhand sensibler Daten berechnet wurden mit einem Rauschen versehen. Auch werden häufig die Werte von einzelnen Datensätzen vertauscht um einen Personenbezug unmöglich zu machen [24]. Zahlreiche Ansätze arbeiten zudem mit dynamischen Filtern. Somit wird die Entscheidung ob bzw. welche Daten gestört werden müssen dynamisch anhand entsprechender Richtlinien getroffen [10].

Insbesondere im Bereich des Data Minings gewinnt der Schutz von sensiblen Daten sehr stark an Bedeutung. Nicht zuletzt aufgrund der rasant steigenden Anzahl und Größe von Datenarchiven ergeben sich hierbei zahlreiche interessante Anwendungsfälle. Nicht zuletzt auch deshalb weil Data Mining Algorithmen in der Lage sind nach interessanten Informationen zu suchen ohne dass zuvor eine Hypothese aufgestellt wird [37]. Häufig ist eine Verknüpfung von Daten jedoch aufgrund deren Vertraulichkeit nicht möglich. Häufig entsteht somit ein Szenario in dem zwei Parteien ihre Datenbestände verknüpfen wollen ohne dabei ihre eigenen sensiblen Daten zu enthüllen.

Ein möglicher Ansatz für dieses Problem ist die Verwendung eines entsprechenden Protokolls welches einerseits die Vertraulichkeit der verwendeten Datensätze sicherstellt und andererseits trotzdem erlaubt diese zu aggregieren [27] [37]. *Lindell und Pinkas* haben ein Protokoll mit einem relativ geringen Overhead entwickelt, welches sensible Daten schützt [37].

Ein anderer Ansatz um möglichst präzise Modelle über aggregierten Daten zu erstellen ohne sensible Daten zu enthüllen verfolgt das Konzept Daten erst zu verzerren und anschließend deren Verteilung zu schätzen. Dieser Ansatz eignet sich für Anwendungsfälle in welchen anhand von Trainingsdaten ein Klassifikator erstellt werden soll. Zunächst werden die Ursprungsdaten verzerrt sodass die erzeugten Daten sich sowohl hinsichtlich ihrer Werte als auch ihrer Verteilung von den ursprünglichen Daten gravierend unterscheiden. Auf die genauen Werte von einzelnen Datensätzen kann somit nicht geschlossen werden, jedoch ist es möglich deren Verteilung zu rekonstruieren. *Agrawal und Srikant* haben gezeigt dass es möglich ist relativ genaue Modelle zu erstellen für Daten deren Störgröße gleich- oder normalverteilt ist [24].

2.3 Zusammenfassung

Zusammenfassend können die Anforderungen an das hier vorgestellte System wie folgt erfasst werden. Im Gegensatz zu klassischen Textkategorisierungs-Systemen werden lediglich zwei Klassen gebildet, welche jeweils als Set von Genres definiert werden. Zudem soll die Klassifizierung auch auf Absatzebene möglich sein, wofür wiederum eine möglichst präzise Trennung von Absätzen vorausgesetzt wird.

Klassische Textkategorisierungs-Systeme genügen diesen Ansprüchen hinsichtlich der geforderten Granularität nicht, zumal diese ausschließlich auf Dokumenten-Ebene agieren. Um in weiterer Folge Anwendungsfälle zu erstellen, in welchen die verfügbaren Daten für Auswertungen verwendet werden ohne sie hierbei preiszugeben, existieren einige interessante Ansätze aus dem Bereich des Data-Mining.

3. Entwurf

Dokumente werden anhand extrahierter Merkmale, welche verschiedene Eigenschaften eines Dokumentes repräsentieren, klassifiziert. Nachfolgend werden in diesem Abschnitt die verwendeten Merkmale vorgestellt. Weiters soll auch die Intention dargelegt werden, welche hinter der Erfassung des jeweiligen Merkmals steckt. Zudem werden die theoretischen Grundlagen des Maschinellen Lernens sowie der Variablenselektion erläutert.

3.1 Merkmale zur Klassifizierung

Die verwendeten und nachfolgend beschriebenen Merkmale zur Klassifizierung von Web-Dokumenten lassen sich in die Gruppen Textstatistik, Part-of-Speech und Präsentation unterteilen.

3.1.1 Textstatistik - Merkmale

Mittels Textstatistik-Merkmalen werden quantifizierbare Eigenschaften von Texten untersucht. Nachfolgend werden Merkmale auf Basis von Häufigkeiten einzelner Token-Kategorien sowie Merkmale auf Basis von Lesbarkeitsindizes vorgestellt.

3.1.1.1 Häufigkeit von Token-Kategorien

Die Verwendung der Häufigkeiten von Token-Kategorien[1] basiert auf der Annahme, dass unter anderem die relative Häufigkeit von Satzzeichen bei verschiedenen Genres unterschiedlich hoch ist. Ein Artikel enthält beispielsweise fast ausschließlich vollständige Sätze und demzufolge eine große Zahl an Kommas und Punkten. Während

[1] In der Computerlinguistik wird die Segmentierung eines Textes in Einheiten der Wortebene als Tokenisierung bezeichnet und die erzeugten Segmente daher als Tokens.

beispielsweise in einem Kommentar eines Lesers häufig relativ wenige Satzzeichen vorhanden sind und ein Geschäftsbericht hingegen sehr viele Symbole und Ziffern enthält.

Um diese Unterschiede in Merkmalen abbilden zu können, werden Elemente eines Textes, wie zum Beispiel die relative Häufigkeit von Buchstaben, Zahlen und einzelnen Satzzeichen, gezählt. Die Bedeutung einzelner Wörter und der inhaltliche Zusammenhang sind hierbei irrelevant, es werden lediglich die Zeichen an sich betrachtet [17]. In der Tabelle 3.1 sind die verwendeten Token-Kategorien dargestellt.

Stamatatos et al. zeigte, dass unter Mitbeachtung von Satzzeichen die Güte eines Klassifikators bei der Genre-Analyse deutlich verbessert werden kann. Insbesondere dann, wenn die Anzahl der Trainings-Daten relativ gering ist, nehmen die Häufigkeiten von Satzzeichen eine wichtige Rolle ein [34].

Textstatistik	Wort-Anzahl Silben-Anzahl	Wort-Länge
Token-Kategorien (Rel. Häufigkeiten)	Word-Tokens Symbol-Tokens Space-Tokens	Number-Tokens Punctuation-Tokens Control-Tokens
Lesbarkeitsindizes	Flesch Reading Ease	Flesch-Kincaid Grade Level

Tabelle 3.1: Textstatistik - Merkmale.

3.1.1.2 Lesbarkeitsindizes

Ein Lesbarkeitsindex ist definiert als Verfahren zur formalen Bestimmung der Lesbarkeit und der Verständlichkeit eines Textes [18]. Lesbarkeitsformeln sind grundsätzlich sprachspezifisch. Die ersten Lesbarkeitsformeln wurden für die englische Sprache entwickelt, es gibt aber bereits entsprechend adaptierte Formeln für andere Sprachen [30].

Flesch Reading Ease - Lesbarkeitsindex

Der Lesbarkeitsindex Flesch Reading Ease ist eine aus dem zu bewertenden Text berechnete Zahl zwischen 1 und 100. Je höher der Wert liegt, desto leichter verständlich ist der Text. Der Flesch Reading Ease ist in seiner Berechnung auf die englische Sprache abgestimmt und berechnet sich wie folgt [35]:

$$FRE = 206,835 - (1,015 * ASL) - (0,846 * ASW) \tag{3.1}$$

FRE ... Flesh Reading Ease - Lesbarkeitsindex
ASL ... Durchschnittliche Satzlänge
ASW ... Durchschnittliche Silbenanzahl pro Wort

Flesch-Kincaid Grade Level - Lesbarkeitsindex

Auch dieser Lesbarkeitsindex ist auf die englische Sprache und insbesondere auf das englische Schulsystem abgestimmt. Der Index versucht die Lesbarkeit auszudrücken indem er angibt wie viele Schuljahre der Leser absolviert haben muss um den Text verstehen zu können. Der Flesch-Kincaid Grade Level berechnet sich wie folgt:

$$FKGL = (0,39 * ASL) + (11,8 * ASW) - 15,59 \tag{3.2}$$

$FKGL$	...	Flesch Kincaid Grade Level - Lesbarkeitsindex
ASL	...	Durchschnittliche Satzlänge
ASW	...	Durchschnittliche Silbenanzahl pro Wort

Beim Flesch-Kincaid Grade Level hat die durchschnittliche Satzlänge im Vergleich zum Flesh Reading Ease einen größeren Einfluss. Bei beiden Indizes jedoch dominiert die durchschnittliche Silbenanzahl pro Wort, was auch einer der Gründe für die begrenzte Anwendbarkeit auf die deutsche Sprache darstellt [35].

3.1.2 POS - Merkmale

Bei der Part-of-Speech Analyse werden die im Text verwendeten Wörter in Bezug auf ihre Funktion im Satz sowie ihre Wortart untersucht. Die hierdurch gewonnen Informationen können genutzt werden um einen Überblick über den Sprachstil zu erhalten.

Für die Erstellung der entsprechenden Annotationen wird ein so genannter POS-Tagger verwendet, der jedes Wort eines Textes mit einer Markierung versieht, welche die Wortart enthält.

Häufig kann für ein einzelnes Wort nicht immer eindeutig die Wortart vorhergesagt werden. Beispielsweise kann das englische Wort ‚drink' sowohl als Verb als auch als Nomen verwendet werden. Um Mehrdeutigkeiten aufzulösen gibt es zwei mögliche Ansätze: Regelbasierte und stochastische POS–Tagger. Bei einem regelbasierten Tagger werden anhand eines Wörterbuchs zunächst für jedes Wort alle in Frage kommenden POS–Tags ausgesucht. Gibt es für ein Wort mehrere Möglichkeiten so wird über ein entsprechendes Regelwerk der zutreffende Tag gewählt. Bei stochastischen POS–Taggern wird die Auswahl des zutreffenden Tags über entsprechende Wahrscheinlichkeitsberechnungen bestimmt [17].

In der Tabelle 3.2 sind die verwendeten POS-Merkmale aufgelistet. Neben Wortarten werden hierbei auch Lookups als Merkmale verwendet. Lookups sind einfache Klassifikatoren, welche versuchen einzelne Wörter verschiedenen Klassen zuzuordnen.

Klasse	**Merkmale**	
Lookups	Orte	Adressen
(Rel. Häufigkeiten)	Zeit Datum	Gesellschafts-Namen
	Personen-Namen	Abkürzungen
	Währungen	Telefon- und Faxnummern
Part-Of-Speech	Nomen	Verben
(Rel. Häufigkeiten)	Adjektive	Adverben
	Modalverben	Konjunktionen
	Pronomen	Personalpronomen
	Artikel	Präpositionen
	Ausrufe	Listenelemente

Tabelle 3.2: POS - Merkmale.

Häufigkeit von Wortarten - Erkennung von Sprachstilen

Über die Häufigkeiten von Wortarten können Sprachstile erkannt werden, welche oft sehr stark mit bestimmten Genres korrelieren.

Beispielsweise lassen sich öffentliche Elemente eines Web-Dokuments sehr häufig mit einem Nominalstil[2] in Verbindung bringen. Der Nominalstil ist (häufig aufgrund der Sprachökonomie[3]) besonders in wissenschaftlichen Artikeln sowie in fachsprachlichen[4] Texten sehr verbreitet. Umgekehrt korrelieren private Elemente eines Web-Dokuments auch häufig mit einem Verbalstil[5]. So sind insbesondere erzählerische Texte meist in einem sehr ausgeprägten Verbalstil verfasst. Dieser Umstand ermöglicht es beispielsweise Rezensionen auf Produktseiten oder teilweise auch Kommentare zu einem Artikel relativ zielsicher anhand der Häufigkeiten von Wortarten identifizieren zu können.

Da die zu analysierenden Textfragmente oft relativ kurz sind, gestaltet es sich oft schwierig, Sprachstile nur anhand von Häufigkeiten von Wortarten zuverlässig erkennen zu können. Aus diesem Grund werden auch andere Merkmale herangezogen und kombiniert, um die Erkennungsrate zu steigern. Insbesondere private Elemente eines Web-Dokuments werden häufig durch erzählerische Texte repräsentiert, weshalb deren Erkennung von hoher Bedeutung ist. Abgesehen vom beschriebenen Verbalstil lassen sich erzählerische Texte häufig durch eine signifikant höhere Häufigkeit von Personalpronomen erkennen.

[2] Beim Nominalstil wird weitgehend auf die Verwendung von Verben verzichtet, während Nominalgruppen vorherrschen.

[3] Als Sprachökonomie bezeichnet man die Findung der rationalsten sprachlichen Lösung um kommunikative Ziele umzusetzen.

[4] Eine Fachsprache ist ein Jargon der in einem bestimmten Fachgebiet benutzt wird und inkludiert auch eine ausgeprägte Benutzung von Fachvokabular, welches außerhalb des Fachgebiets entweder ungebräuchlich ist oder aber eine andere Bedeutung besitzt.

[5] Der Verbalstil ist ein Sprachstil bei dem Verben relevante Aussageelemente bilden und Nomen vermieden werden bzw. verbal umschrieben werden (z.B. ”weil er arm ist” anstatt ”wegen Armut”).

3.1.3 Präsentationsbezogene Merkmale

Häufigkeit von HTML Tags

Das Layout einer Website wird über HTML Tags beschrieben. Daher bietet sich die Möglicheit, die Art und Weise der Textdarstellung unter Verwendung von HTML Tags zu analysieren. Um die Dimension der präsentationsbezogenen Merkmale einzugrenzen, werden verwandte HTML Tags in zusammengehörende Klassen gruppiert (Tabelle 3.3).

Somit können auf einfache Art und Weise anhand von relativen Häufigkeiten bestimmter HTML Tag - Klassen, charakteristische visuelle Elemente erkannt werden, welche wiederum eine Verbesserung des Klassifikators bewirken können. So lassen sich beispielsweise die für Kommentar- oder Rezensionsabschnitte sehr typischen Abtrennungen in Form von Separator-HTML Tags sehr zielsicher erkennen. Zudem lassen sich diese Bereiche in Kombination mit den typischen Eingabe-Möglichkeiten durch Formular-Elemente zu Beginn oder am Schluss recht deutlich identifizieren. Auch werden meist unterschiedliche Stile für die Überschrift, die Bezeichnung des Verfassers und für den eigentlichen Text verwendet. Dadurch entstehen mehrmals wiederkehrende Muster, welche einen relativ sicheren Indikator darstellen um Kommentare oder Rezensionen erkennen zu können.

Klasse	**HTML Tags**		
Links	<a>		
Überschriften	<h1>	<h2>	<h3>
	<h4>	<h5>	<h6>
	<h7>	<h8>	<h9>
Absätze	 		<blockquote>
	<p>	<hr>	
Textformattierungen		<font>	<b>
		<i>	<u>
	<tt>	<s>	<big>
	<small>	<strike>	<sub>
	<sup>		
Listen		<dl>	
Tabellen	<table>	<tr>	<td>
Graphiken	<img>		
Frames	<frameset>	<frame>	<iframe>
Formulare	<form>	<select>	<textarea>
	<input>		
Multimedia	<applet>	<object>	<embed>

Tabelle 3.3: Verwendete HTML Tag Klassen.

3.2 Dimensionsreduktion

Eine sehr hohe Dimension des Merkmalraums kann bei vielen hochentwickelten Algorithmend des maschinellen Lernens zu Problemen führen, sowohl in Hinsicht auf die Rechenzeit und des Speicherbedarfs als auch in Hinsicht auf die Klassifikationsrate [31]. Daher wird meist die Dimension des Merkmalraums reduziert (englisch: 'dimension reduction'). Ein weiterer Vorteil dieses Vorgehens liegt darin, dass das sogenannte 'Overfitting' tendenziell verringert wird. Als Overfitting wird das Phänomen bezeichnet dass ein Klassifikator derart trainiert wird, dass er zu abhängig von charakteristischen Merkmalen der Trainingsdaten ist und bei neuen unbekannten Daten wesentlich schlechtere Ergebnisse liefert [31].

Grundsätzlich wird bei der Dimensionsreduktion zwischen den nachfolgend angeführten Methoden 'Feature Selection' und 'Feature Extraction' unterschieden. Während bei der 'Feature Selection' eine Verringerung der Dimension durch das Weglassen von weniger wichtigen Merkmalen erfolgt, wird bei der Methode der 'Feature Extraction' ein Vektor von Merkmalen in einen Vektor mit einer niedrigeren Dimension transformiert.

3.2.1 Feature Selection

Die 'Feature Selection' ist ein Ansatz des maschinellen Lernens, wobei für einen Algorithmus nur eine Teilmenge der verfügbaren Merkmale verwendet wird. Die Feature Selection ist notwendig weil es meist unmöglich ist alle verfügbaren Merkmale mit einzubeziehen oder aber weil Differenzierungsprobleme auftreten.

Das Ziel der Feature Selection ist daher jene Merkmale zu eliminieren, welche nicht mit der Klasse korrelieren, oder aber redundant sind. Dadurch soll der Klassifikator robuster werden und die Güte verbessert werden. Weiters wirkt sich die Anwendung der Feature Selection zudem auch positiv auf die Performance des Lern-Algorithmus aus zumal die Dimension des Merkmalraums verringert wird. Zudem ist die Feature Selection auch hilfreich beim Verständnis über die Wichtigkeit einzelner Merkmale sowie deren Beziehungen zueinander.

Im Allgemeinen ist die Aufgabe der Feature Selection eine schwer zu lösende, viele Problemstellungen in diesem Bereich sind NP-vollständig. Ein typischer Prozess zur Findung einer optimalen Menge von Merkmalen gliedert sich in 4 Schritte. Zuerst werden basierend auf einer Suchstrategie potentielle Merkmal-Mengen gebildet, welche anschließend evaluiert werden und gegebenenfalls die bisherige Optimallösung ersetzen. Dies wird solange wiederholt bis ein zuvor definiertes Stop-Kriterium erreicht wird. Häufig werden minimale Anforderung an die Qualität oder aber eine maximale Anzahl an durchgeführten Iterationen als Stop-Kriterium verwendet. Im

abschließenden letzten Schritt wird die gefundene Lösung auf einem anderen Datenset evaluiert[38]. Die gebräuchlichsten Algorithmen gliedern sich in die nachfolgend erläuternden Ansätze.

Feature Selection Algorithmen lassen sich in zwei Kategorien einteilen: 'Feature Ranking' Algorithmen reihen Merkmale anhand einer Metrik während 'Feature Subset' Algorithmen mögliche Merkmalmengen untersuchen und die optimale Menge auswählen.

3.2.1.1 Feature Ranking

Aus dem Bereich der Informationstheorie existieren zahlreiche Funktionen um die Dimension des Merkmalraums zu verringern. Die wichtigsten Vertreter sind in der Tabelle 3.4 aufgeführt [31]. Um einen p-dimensionalen Merkmalraum auf q Dimensionen zu verringern ($q < p$) werden hierbei die q 'bestgereihten' Merkmale ausgewählt.

Wahrscheinlichkeiten werden hierbei als Ereignisse von Instanzen betrachtet. So ist beispielsweise $P(\bar{t}_k, c_i)$ die Wahrscheinlichkeit, dass, für eine zufällige Instanz x, das Merkmal t_k nicht in x vorkommt und x zur Klasse c_i gehört. Diese Wahrscheinlichkeit wird geschätzt durch das Zählen der Vorkommnisse in den Trainingsdaten.

Alle Funktionen sind lokal, d.h. für eine bestimmte Klasse c_i, definiert. Um einen global gültigen, d.h. klassenunabhängigen, Wert für ein Merkmal t_k zu erhalten existiert die übliche Methode der Summenbildung von Funktionen über alle Klassen.

$$f_{sum}(t_k) = \sum_{i=1}^{|C|} f(t_k, c_i) \tag{3.3}$$

Auch eine mit der Klassenwahrscheinlichkeit gewichtete Summe kann verwendet werden.

$$f_{wsum}(t_k) = \sum_{i=1}^{|C|} P(c_i) f(t_k, c_i) \tag{3.4}$$

Eine weitere übliche Methode um einen klassenunabhängigen Wert zu erhalten ist die Auswahl der maximalen lokalen Funktion.

$$f_{max}(t_k) = max_{i=1}^{|C|} f(t_k, c_i) \tag{3.5}$$

Diese Funktionen (Tabelle 3.4) basieren auf der Annahme dass die besten Merkmale einer Klasse c_i jene sind welche am ungleichmäßigsten verteilt sind in der Menge der positiven und negativen Datensätze. Das Ziel dieser Funktionen ist es zu messen wie unabhängig ein Merkmal t_k von einer Klasse c_i ist.

Bezeichnung	Notation	Funktion
DIA association factor	$z(t_k, c_i)$	$P(c_i\|t_k)$
Information gain	$IG(t_k, c_i)$	$\sum_{c\in c_i,\bar{c}_i} \sum_{t\in t_i,\bar{t}_i} P(t,c) log \frac{P(t_k,c_i)}{P(t_k)P(c_i)}$
Mutual Information	$MI(t_k, c_i)$	$log \frac{P(t_k,c_i)}{P(t_k)P(c_i)}$
Chi-square	$\chi^2(t_k, c_i)$	$\frac{\|T_r\|[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$
Relevancy score	$RS(t_k, c_i)$	$log \frac{P(t_k\|c_i)+d}{P(\bar{t}_k\|\bar{c}_i)+d}$
Odds Ratio	$OR(t_k, c_i)$	$\frac{P(t_k\|c_i)(1-P(t_k\|\bar{c}_i))}{(1-P(t_k\|c_i))P(t_k\|\bar{c}_i)}$

Tabelle 3.4: Feature Ranking - Funktionen aus der Informationstheorie.

Beispielsweise ist ein Merkmal t_k mit dem niedrigsten Wert für die Funktion $\chi^2(t_k, c_i)$ also jenes, welches die größte Unabhängigkeit von der Klasse c_i besitzt. Daher werden jene Merkmale ausgewählt für welche die Funktion $\chi^2(t_k, c_i)$ maximal ist [31].

3.2.1.2 Feature Subset Selection

Filter-Ansatz

Beim Filter-Ansatz startet der Algorithmus bei einer gegebenen Merkmal-Menge (meist die leere, die volle oder eine zufällig ausgewählte Menge) und wählt anhand einer bestimmten Suchstrategie Elemente aus dem Merkmalraum. Diese Elemente werden anschließend einzeln über ein vom Lernalgorithmus unabhängiges Maß evaluiert. Wird eine gefundene Lösung besser bewertet als jene zuvor, so nimmt die neu gefundene Lösung den Status der Optimallösung ein. Diese Suche wird durchgeführt bis ein definiertes Stop-Kriterium erreicht wurde.

Durch diese Vorgehensweise wird die Gewichtung und Auswahl der besten Merkmale unabhängig vom später verwendeten Lernalgorithmus durchgeführt. Dies führt zu einer sehr guten Performance, hat jedoch auch einige Nachteile. Einerseits werden redundante Merkmale verwendet, zumal verwandte Merkmale ähnlich stark gewichtet werden. Weiters werden Abhängigkeiten von Merkmalen nicht berücksichtigt, da Merkmale häufig nur in Kombination relevant sind und mit der Klasse korrelieren[38]. Ein weiteres großes Problem ist, dass die individuellen Stärken und Schwächen des verwendeten Lernalgorithmus bei der Auswahl der verwendeten

Merkmale nicht berücksichtigt werden. Insbesondere die Verwendung von redundanten Merkmalen kann je nach Ausprägung bei manchen Lernalgorithmen zu einer deutlich niedrigeren Güte führen.

Wrapper-Ansatz

Der Wrapper-Ansatz ist relativ ähnlich zum Filter-Ansatz mit dem signifikanten Unterschied dass die Evaluierung aller potentiellen Lösungs-Mengen über den eigentlichen Lernalgorithmus durchgeführt wird.

Dies führt zu einer Menge an ausgewählten Merkmalen welche optimal zum verwendeten Lernalgorithmus passt. Zudem ist die Güte des Klassifikators oft deutlich besser zumal auch Kombinationen von Merkmalen miteinbezogen werden, anstatt jedes Merkmal isoliert und unabhängig vom verwendeten Algorithmus zu bewerten[38]. Weiters können redundante Merkmale erkannt und entfernt werden, was sich je nach verwendetem Lern-Algorithmus positiv auswirken kann. Der gravierende Nachteil dieses Ansatzes ist der sehr hohe Berechnungsaufwand der meist eine maximale Anzahl an durchzuführenden Iterationen als Stop-Kriterium erfordert.

3.2.2 Feature Extraction

Im Gegensatz zur 'Feature Selection' werden bei der 'Feature Extraction' keine Merkmale verworfen. Stattdessen wird ein Merkmalvektor in einen Vektor mit einer geringeren Dimension transformiert mit dem Ziel des minimalen Informationsverlusts. Ein Nachteil dieses Verfahrens ist dass das Ergebnis der Transformation in Form eines Merkmalvektors nicht mehr interpretierbar ist. Das bekannteste Verfahren der Feature Extraction ist die Hauptkomponentenanalyse.

Hauptkomponentenanalyse

Die Hauptkomponentenanalyse ist ein Verfahren der multivariaten Statistik[6]. Das Ziel dieses Verfahrens ist es statistische Variablen durch eine geringere Zahl von möglichst aussagekräftigen Linearkombinationen (die 'Hauptkomponenten') anzunähern.

Eine Menge von n p-dimensionalen Merkmalvektoren kann als Menge von n Punkten im p-dimensionalen Raum R^p veranschaulicht werden. Ziel ist es nun, diese Datenpunkte in einen q-dimensionalen Unterraum R^q (q < p) zu transformieren wobei der Informationsverlust minimal bleiben soll. Die Varianz von Daten ist hierbei ein Maß für deren Informationsgehalt.

Hiefür wird eine 'Hauptachsentransformation' durchgeführt. Die Daten liegen als Punktwolke in einem p-dimensionalen kartesischen Koordinatensystem vor, in das

[6] Bei multivariaten Verfahren werden Variablen nicht isoliert untersucht, sondern das Zusammenwirken mehrerer Variablen zugleich.

nun ein neues q-dimensionales Koordinatensystem gelegt wird. Die erste Achse dieses neuen Koordinatensystem wird nun so rotiert dass die Varianz der Daten in dieser Richtung maximal wird. Die restlichen Achsen werden ebenfalls so rotiert dass die Varianz der Daten in deren Richtung maximal wird, wobei die Richtung der bereits gesetzten Achsen nicht verändert wird. Die k-te Achse bezeichnet hierbei die k-te Hauptkomponente. Die Suche nach den p Hauptkomponenten der Datenmenge entspricht der Suche nach p unkorrelierten Linearkombinationen der Vektoren mit maximaler Varianz [3] [14].

In der Abbildung 3.1 ist das Ergebnis einer Hauptkomponentenanalyse dargestellt bei der ein n-dimensionaler Merkmalraum durch 2 Linearkombinationen beschrieben wird.

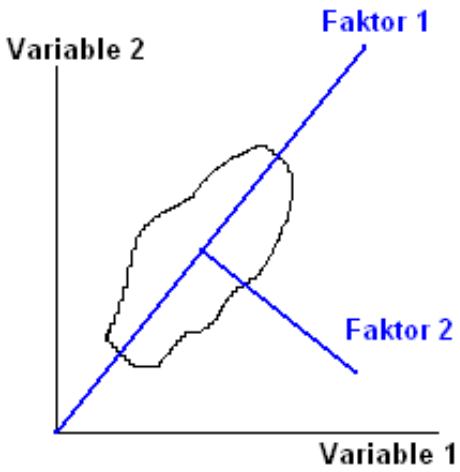

Abbildung 3.1: Veranschaulichung einer Hauptkomponentenanalyse bei der Reduktion auf 2 Dimensionen.

3.3 Klassifizierung

Eine Klassifikation ist eine systematische Vorgehensweise zur Sammlung von abstrakten Klassen, die zur Abgrenzung und Ordnung von Objekten verwendet werden. Die einzelnen Klassen werden mittels Klassifizierung, das heißt durch die Zuordnung von Objekten zu einer Klasse anhand bestimmter Merkmale, gewonnen. Bei automatisierten Klassifizierungen kann eine große Anzahl von Merkmalen der Objekte berücksichtigt werden.

3.3.1 Maschinelles Lernen

Maschinelles Lernen ist ein Oberbegriff für die künstliche Generierung von Wissen aus Erfahrung. Ein künstliches System lernt aus Beispielen und kann nach Abschluss

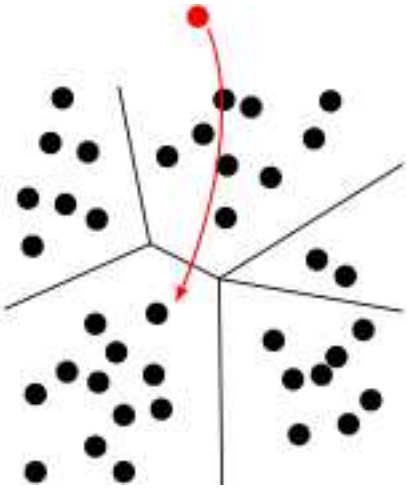

Abbildung 3.2: Beispiel einer Klassifikation eines 2-dimensionalen Merkmalraums in 5 Klassen und Klassifizierung eines Objektes.

der Lernphase verallgemeinern. Somit werden Gesetzmäßigkeiten in den Lerndaten erkannt. So kann das System auch unbekannte Daten beurteilen.

Verschiedene Algorithmen aus dem Bereich des maschinellen Lernens lassen sich in folgende drei Kategorien unterteilen:

3.3.1.1 Überwachtes Lernen

Beim überwachten Lernen (Supervised Learning) werden Beispiele in Form einer Trainings-Menge mit den zu verwendenden Merkmalen sowie deren Klassenzugehörigkeiten bereitgestellt. Hierbei fungieren die Merkmale als Lernmaterial währen die gegebene Klassenzugehörigkeit den ”Lehrer” verkörpert [25].

Um das Verfahren sinnvoll anwenden zu können, ist es nötig eine ausreichend große Anzahl an klassifizierten Objekten zur Verfügung zu haben, um einen Trainings- und einen Test-Korpus erstellen zu können.

Beispiele für diese Art des maschinellen Lernens sind Support-Vector-Maschinen und neuronale Netze.

3.3.1.2 Unüberwachtes Lernen

Beim unüberwachten Lernen (Unsupervised Learning) wird für eine Menge von Objekten ein Modell erzeugt, welches Vorhersagen ermöglicht. Es stehen zwar Beispiele zum Training des Klassifikators zur Verfügung, jedoch ist weder deren Klassenzugehörigkeit bekannt noch welche Klassen es überhaupt gibt. Es gibt jedoch Clustering-Verfahren, welche die Daten in mehrere Kategorien einteilen die sich durch charakteristische Muster voneinander unterscheiden [25].

Ein wichtiger Vertreter dieser Art des maschinellen Lernens ist der EM-Algorithmus, welcher iterativ die Parameter eines Modells so festlegt, dass es die betrachteten

Daten optimal erklärt. Der Algorithmus geht davon aus, dass einige Kategorien "beobachtbar" sind, andere wiederum nicht. Hierbei werden abwechselnd sowohl die Zugehörigkeit der Daten zu einer der Kategorien, als auch die Parameter welche die Kategorie ausmachen, geschätzt [25].

3.3.1.3 Bestärkendes Lernen

Bei der Variante des Bestärkenden Lernens (Reinforcement Learning) lernt ein Algorithmus durch Belohnung und Bestrafung eine Vorgehensweise, wie in potentiell aufzutretenden Situationen zu handeln ist um seinen Nutzen zu maximieren [25].

Bedingt durch die Belohnung oder Bestrafung wird die Klassifikation eines Beispiels bewertet, wodurch aber weniger Informationen für die Parametrisierung des verwendeten Algorithmus zur Verfügung stehen. Diese Art des maschinellen Lernens wird häufig für Agentensysteme verwendet, in welchen ein Agent selbständig lernen soll zu einem gegebenen Zustand die passende Aktion auszuwählen.

3.3.2 Klassifizierungs Algorithmen

3.3.2.1 k-Nearest Neighbor-Klassifikator

Der k-Nearest Neighbor-Klassifikator (k-NN) ist ein Verfahren bei dem die Klassenzuordnung durch Berücksichtigung seiner k nächsten Nachbarn vorgenommen wird. Bei diesem Algorithmus handelt es sich um einen "lazy learning" Algorithmus, das Lernen besteht lediglich aus dem Abspeichern der Trainingsbeispiele. Die eigentliche Berechnung erfolgt zur Laufzeit des Algorithmus.

Zur Klassenzuordnung werden die Distanzen zwischen der zu klassifizierenden Instanz und allen Trainingsinstanzen berechnet. Der Testinstanz wird jene Klasse zugeordnet der die meisten der k nächsten Nachbarn angehören. Die Zuordnung erfolgt also durch eine Mehrheitsentscheidung an der sich die k nächsten Instanzen beteiligen [5] [25].

Zur Distanzmessung sind hierbei verschiedene Abstandsmaße möglich, die gebräuchlichsten sind die Euklidische Distanz[7] sowie die Manhattan-Metrik[8].

Neben der Wahl einer geeigneten Metrik zur Abstandsmessung zweier Punkte ist die Wahl des Parameters k entscheidend für die Güte dieses Klassifikators. Wenn k zu klein gewählt wird können die Ergebnisse des Klassifikators durch ein "Rauschen" in den Trainingsdaten verschlechtert werden. Wenn k wiederum zu groß gewählt

[7] In einem n-dimensionalen euklidischen Vektorraum ist die Euklidische Distanz zweier Punkte definiert durch durch die euklidische Norm (Länge) des Differenzvektors[13].

[8] Die Manhattan-Metrik ist eine Metrik zur Distanzmessung zwischen zwei Punkten. Hierbei wird die Distanz definiert als die Summe der absoluten Differenzen der Vektor-Komponenten [25].

wird, führt dies dazu, dass auch Punkte in die Klassifikations-Entscheidung einbezogen werden, welche einen großen Abstand zur Testinstanz aufweisen. Insbesondere wenn die Menge an Trainings-Instanzen relativ klein oder aber ungleich verteilt ist, besteht diese Gefahr. Häufig wird daher eine gewichtete Funktion zur Distanzmessung verwendet, welche näher liegende Instanzen höher gewichtet als weiter entfernt liegende.

Eine große Schwäche des k-Nearest Neighbor-Klassifikators ist der Speicher- und Berechnungsaufwand welcher mit der Anzahl der Trainingsdaten steigt. Auch hochdimensionale Räume stellen hierbei ein Problem dar, zumal der Aufwand für die Distanzmessungen steigt.

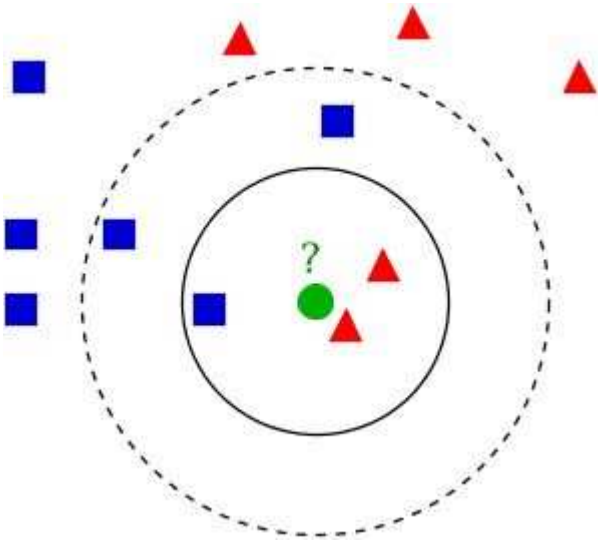

Abbildung 3.3: Beispiel einer k-NN Klassifikation. Die Testinstanz (Kreis) soll entweder zur Klasse der Quadrate oder zur Klasse der Dreiecke zugeordnet werden. Wenn k = 3 ist wird die Testinstanz zur Klasse der Dreiecke zugeordnet da sich im inneren Kreis 2 Dreiecke und 1 Quadrat befinden. Wenn k = 5 ist wird die Testinstanz zur Klasse der Quadrate zugeordnet da sich 3 Quadrate und 2 Dreiecke im äußeren Kreis befinden.

3.3.2.2 Support Vector Machine-Klassifikator

Eine Support Vector Machine ist ein Klassifikator, welche eine Menge derart in Klassen unterteilt, so dass um die Klassengrenzen herum ein möglichst breiter Bereich frei von Objekten bleibt.

Die Ausgangslage für die Erstellung einer Support Vector Machine ist eine Menge von Trainingsobjekten mit bekannten Klassenzugehörigkeiten. Jedes Objekt bzw. dessen Merkmale werden durch einen Vektor in einem Vektorraum repräsentiert. Aufgabe

der Support Vector Machine ist es nun, in diesen Raum eine mehrdimensionale Hyperebene[9] einzupassen, welche als Trennfläche zwischen den Klassen fungiert. Die Lage der Hyperebene ist dann optimal, wenn der Abstand zu den Klassen maximal ist. Ein entsprechend breiter leerer Rand soll sicherstellen, dass auch Objekte, welche nicht genau die selben Merkmalsausprägungen wie die Trainingsobjekte aufweisen, möglichst zuverlässig klassifiziert werden [25].

Eine Hyperebene kann nicht "verbogen" werden, so dass eine saubere Trennung mit einer Hyperebene nur dann möglich ist, wenn die Objekte linear trennbar sind (Abbildung 3.4). Dies ist in realen Anwendungsfällen im Allgemeinen nicht der Fall. Daher werden die Trainingsvektoren üblicherweise in einen höherdimensionalen Raum überführt. Diese Transformation ist jedoch sehr aufwändig.

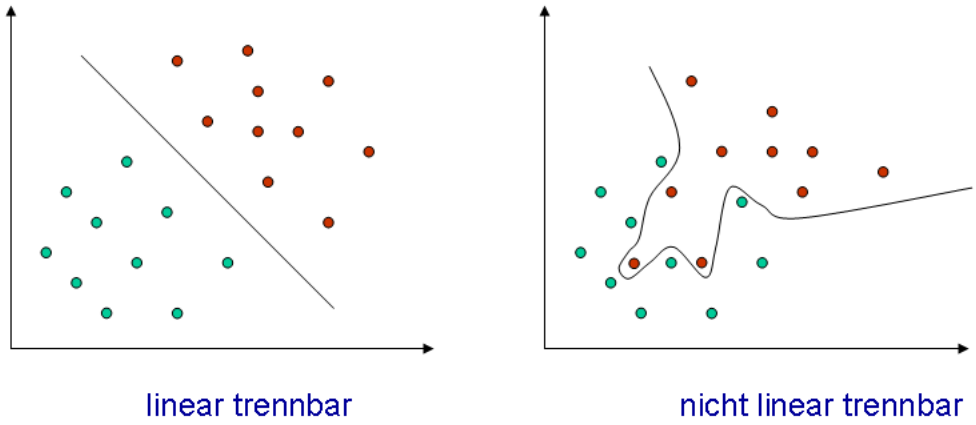

Abbildung 3.4: Lineare Trennbarkeit von Trainings - Objekten.

3.3.2.3 Bayes-Klassifikator

Ein Bayes - Klassifikator ist ein aus dem Bayestheorem[10] abgeleiteter Klassifikator. Er ordnet ein Objekt jener Klasse zu, der es mit der größten Wahrscheinlichkeit angehört.

Voraussetzung hierfür ist dass die Wahrscheinlichkeit, dass ein Punkt des Merkmalraums zu einer bestimmten Klasse gehört, bekannt ist. Folglich ist für jede Klasse eine Wahrscheinlichkeitsdichte nötig. In der Praxis sind diese Dichtefunktionen aber üblicherweise nicht bekannt, weshalb sie geschätzt werden müssen. Daher wird hinter jeder Klasse ein Typ von Wahrscheinlichkeitsverteilung vermutet, in der Regel eine Normalverteilung. Es wird versucht anhand der vorhandenen Daten die Parameter der zu erwartenden Verteilung abzuschätzen[25].

[9] Eine affine Hyperebene ist die Verallgemeinerung einer normalen Ebene im dreidimensionalen Raum auf Vektorräume[13].

[10] Das Bayestheorem (auch Satz von Bayes) ist ein Ergebnis der Wahrscheinlichkeitstheorie, benannt nach dem Mathematiker Thomas Bayes. Es gibt an, wie man mit bedingten Wahrscheinlichkeiten rechnet[3].

Naiver Bayes-Klassifikator

Aufgrund seiner schnellen Berechenbarkeit bei guter Erkennungsrate sehr beliebt ist auch der naive Bayes-Klassifikator. Mittels des naiven Bayes-Klassifikators ist es möglich, die Zugehörigkeit eines Objektes zu einer Klasse zu bestimmen. Er basiert auf dem Bayesschen Theorem. Ein naiver Bayes-Klassifikator ist ein sternfömiges Bayessches Netz [11].

Grundannahme ist hierbei, dass jedes Merkmal eines Objekts mit der zugehörigen Klasse korreliert. Obwohl dies in der Realität selten zutrifft, erzielen naive Bayes-Klassifikatoren bei praktischen Anwendungen häufig sehr gute Ergebnisse, solange die Attribute nicht zu stark miteinander korrelieren[25]. Dieser Umstand ist umso erstaunlicher zumal die Schätzung der zugrunde liegenden Wahrscheinlichkeitsverteilung oft relativ ungenau ist[26].

3.4 Zusammenfassung

Da aufgrund von Differenzierungsproblemen oder implementierungstechnischen Beschränkungen häufig nicht alle verfübaren Merkmale miteinbezogen werden können ist die so genannte Dimensionsreduktion (Abschnitt 3.2) meist unumgänglich. Bei dem Ansatz der 'Feature Selection' wird nur eine Teilmenge der verfügbaren Merkmale verwendet. Eine Methode der Feature Selection ist das 'Feature Ranking' bei dem die einzelnen Merkmalen nach ihrer Wichtigkeit gereiht werden. Für diese Bewertung existieren zahlreiche Funktionen aus der Informationstheorie, welche die Unabhänigkeit eines Merkmals von einer Klasse messen. Die Methode der 'Feature-Subset Selection' untersucht mögliche Merkmalmengen mit dem Ziel die optimale Menge auszuwählen. Während beim einfacheren Filter-Ansatz Merkmale isoliert und unabhängig vom verwendeten Lernalgorithmus bewertet werden, kaschiert der komplexere Wrapper-Ansatz diese Unzulänglichkeiten was jedoch zu einem enorm hohen Berechnungsaufwand führen kann. Ein grundlegend verschiedener Ansatz der Dimensionsreduktion ist die sogenannte 'Feature Extraction'. Hierbei wird ein Merkmalvektor in einen Vektor mit einer geringeren Dimension transformiert mit dem Ziel des geringstmöglichen Informationsverlusts (Abschnitt 3.2.2).

Grundlegend lassen sich Merkmale zur Klassifizierung von Web-Dokumenten in die Gruppen Textstatistik, Part-of-Speech und Präsentation unterteilen. Lesbarkeitsindizes (Abschnitt 3.1.1.2), welche im Allgemeinen auf Basis der Wort- und Silbenanzahl ermittelt werden, stellen hierbei einen besonders robusten Vertreter der Gattung Textstatistik-Merkmale dar. Bei der Part-of-Speech Analyse (Abschnitt 3.1.2)

[11] Ein Bayes'sches Netz ist ein gerichteter azyklischer Graph, in dem die Knoten Zufallsvariablen und die Kanten bedingte Abhängigkeiten zwischen den Variablen beschreiben.

werden die im Text verwendeten Wörter in Bezug auf ihre Funktion im Satz sowie ihre Wortart untersucht. Beiden Gruppen gemein ist, dass sie, im Gegensatz zu präsentationsbezogenen Merkmalen, auf jegliche Art von Texten unabhängig von der eigentlichen Darstellung angewandt werden können. Die einzige Voraussetzung ist das Vorhandensein von vollständigen Sätzen. Die verwendeten Merkmale ergeben einen 41-dimensionalen Merkmalvektor für jede zu klassifizierende Textpassage.

Die anschließende Zuordnung von Objekten zu Klassen auf Basis der gesammelten Merkmale erfolgt durch einen Klassifikator welcher aber zuerst anhand von Beispieldaten trainiert werden muss. Wenn selbige in klassifizierter Form vorliegen, kann durch einen Algorithmus aus der Kategorie "Überwachtes Lernen" (Abschnitt 3.3.1.1) ein Klassifikator ohne weiteres Zutun trainiert werden. Sind hingegen weder die Klassenzugehörigkeiten der Beispieldatensätze bekannt, noch welche Klassen überhaupt existieren, so müssen diese erst ermittelt werden. Diese Art des maschinellen Lernens wird als "Unüberwachtes Lernen" (Abschnitt 3.3.1.2) bezeichnet.

4. Implementierung

Dieser Abschnitt soll einen Überblick über die eigentliche Implementierung des vorgestellten Ansatzes geben. Hierbei werden die zentralen Bibliotheken, auf welchen die Implementierung aufbaut, kurz vorgestellt. Anschließend wird die Architektur des implementierten Ansatzes skizziert und ein grober Überblick über die Funktionsweise gegeben. Zudem sollen die Kern-Komponenten näher beschrieben werden und auf implementierungsspezifische Eigenheiten eingegangen werden. Abschließend wird das Ergebnis der durchgeführten Variablenselektion präsentiert.

4.1 Verwendete Bibliotheken

4.1.1 GATE

GATE ist eine Architektur, ein Framework und eine graphische Entwicklungsumgebung für die Verarbeitung von natürlicher Sprache und wird von der University of Sheffield seit 1995 entwickelt. Es wurde bereits vielfach in Projekten eingesetzt und die Liste von Nutzern umfasst zahlreiche Branchengrößen wie beispielsweise Hewlett Packard oder AT&T[32].

GATE enthält zahlreiche Tools zur Verarbeitung von natürlicher Sprache. Üblicherweise wird GATE über einem Korpus von Texten ausgeführt und erzeugt eine Menge von annotierten Texten. Die Architektur erlaubt die Komposition von einzelnen Komponenten zu so genannten "Processing Pipelines", welche ausführbar sind. Manche dieser Komponenten sind sprach- und domänenunabhängig wodurch keinerlei Adaptierungen nötig sind für die Verwendung in eigenen Applikationen[1]. Weiters werden standardmäßig zahlreiche Input Formate unterstützt [2] und ein umfangreiches

[1] Derzeit werden in GATE die Sprachen Englisch, Spanisch, Chinesisch, Arabisch, Französisch, Deutsch, Hindi, Rumänisch und Russisch unterstützt

[2] GATE kann derzeit folgende Input Formate verarbeiten: PlainText, HTML, Microsoft Word (Doc), PDF, PostgreSQL & Oracle (per JDBC Treiber).

Informations Extrahierungs System "Annie" (A Nearly-New Information Extraction System) mitgeliefert. Annie besteht aus einem "Tokenizer"[3], einem "Gazetteer"[4], einem "Sentence Splitter", einem "Part Of Speech Tagger', einem "Named Entities Transducer"[5] sowie einem "Coreference Tagger"[6].

4.1.2 WEKA

WEKA ist eine sehr populäre Implementierung von zahlreichen Machine Learning und Data-Mining[7]Algorithmen und wurde von der University of Waikato entwickelt. WEKA enthält eine Java-API sowie einige graphische Tools zur Verarbeitung und Visualisierung von Daten [12]. WEKA umfasst Implementierungen von zahlreichen Standard Data-Mining Aufgaben, wie beispielsweise Clustering, Klassifikation, Regression[8] sowie Variablenselektion.

Zur Evaluierung verschiedener Klassifizierungs-Algorithmen sowie zur Variablenselektion wurde der WEKA Explorer (Abbildungen 4.1 und 4.2) verwendet. Der Explorer dient als graphische Entwicklungsumgebung, welche sowohl Funktionen zur Datenaufbereitung als auch zum Clustering sowie zur Klassifizierung von Daten enthält. Weiters ist auch eine graphisches Interface zur Variablenselektion enthalten sowie die Möglichkeit die Verteilungen der Merkmale sowie deren Korrelationsmatrix (Abbildung 4.2) visuell darzustellen.

4.1.3 HTML Parser

HTML Parser ist eine sehr effiziente und robuste Java-Implementierung eines HTML Parsers. Ein besonderes Merkmal dieser Implementierung ist die Möglichkeit ein HTML Dokument in eine Plain Text-Repräsentation zu transformieren. Eine weitere Stärke dieses Parsers ist die Möglichkeit, auch von fehlerbehafteten Websites eine Darstellung der Elemente in Form eines Baumes zu generieren [6].

4.2 Architektur

In der nachfolgenden Abbildung 4.3 sind die Kern-Komponenten der Implementierung in Form eines Package-Diagramms dargestellt.

[3] Ein Tokenizer wird verwendet um eine Menge von Zeichen in einzelne Strings zu gruppieren, beispielsweise kann so ein Satz in einzelne Wörter zerlegt werden.

[4] Gazetteer ist die englische Bezeichnung für Ortsregister und besteht aus zahlreichen Wörterbüchern um geographische Informationen klassifizieren zu können.

[5] Aufgabe eines Named Entities Transducer ist die Klassifizierung von einzelnen Wörtern in gegebene Klassen, beispielsweise die Klassifizierung von Namen von Personen oder Firmen.

[6] Ein Coreference Tagger wird verwendet um erkennen zu können, ob sich mehrere Ausdrücke eines Satzes auf den selben Referent beziehen. Die Entscheidung ob zwei Ausdrücke sich auf ein- und den selben Referenten beziehen wird häufig per logistischer Regression getroffen.

[7] Unter Data-Mining versteht man die Anwendung von statistisch-mathematischen Methoden auf einen üblicherweise sehr großen Datenbestand mit dem Ziel der Mustererkennung.

[8] Die Regressionsanalyse ist ein statistisches Analyseverfahren zur Feststellung von Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen.

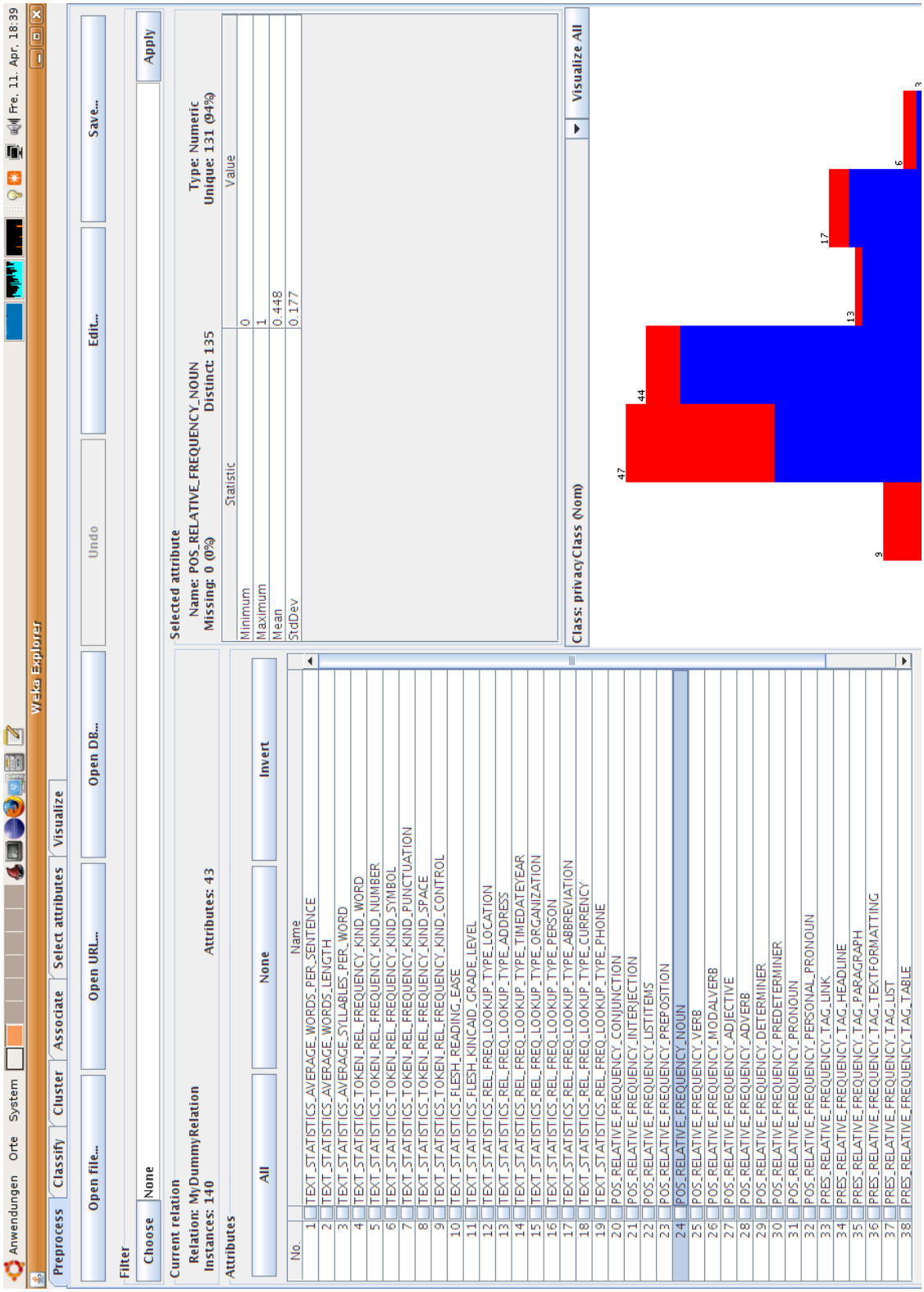

Abbildung 4.1: Weka Explorer - Darstellung der Verteilung des Merkmals 'Relative Häufigkeit von Nomen' (Klasse öffentlich - blau, Klasse privat - rot).

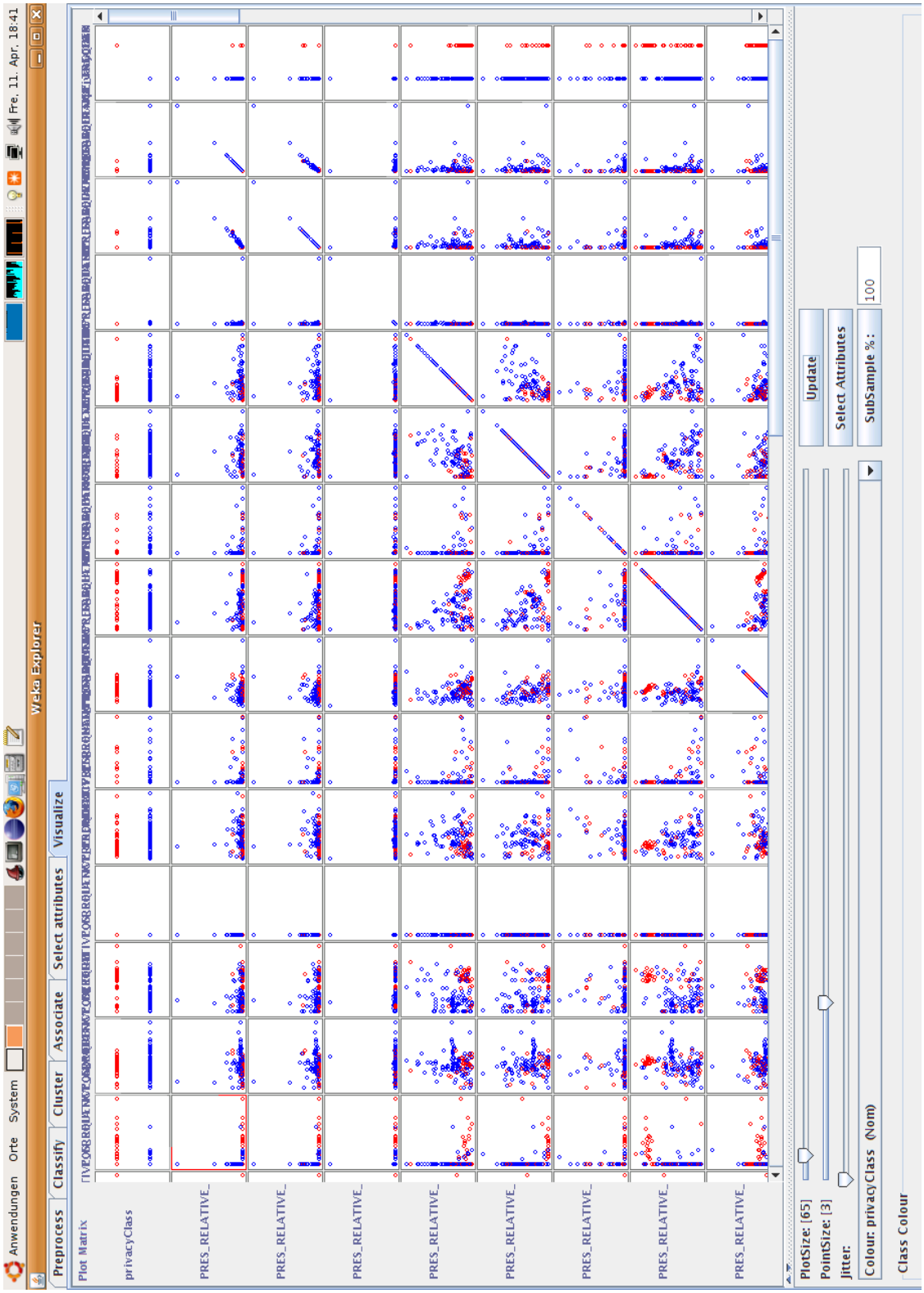

Abbildung 4.2: Weka Explorer - Visuelle Darstellung der Korrelationsmatrix.

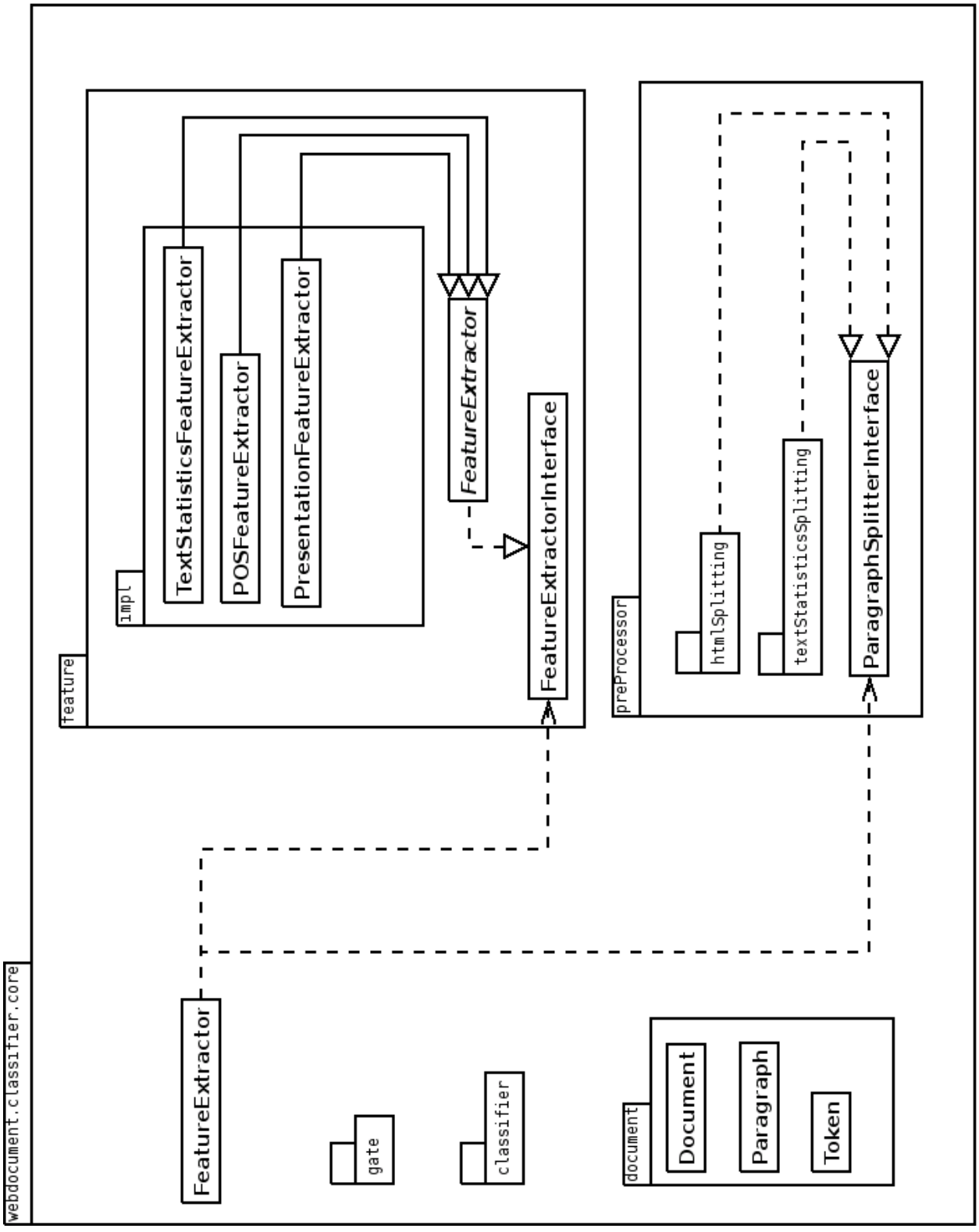

Abbildung 4.3: Package Diagramm der Kern-Komponenten.

Im Package [9] 'document' sind die verwendeten 'Model-Klassen'[10] enthalten. Hierbei stellt ein Dokument die zentrale Entität dar, welche wiederum multiple Absätze enthält. Jeder Absatz besteht aus mehreren Tokens. Ein Token ist hierbei die Entität welche einer logisch zusammenhängenden Zeichenfolge, beispielsweise einem Wort, Satzzeichen oder Symbol, entspricht.

Die Klasse 'FeatureExtractor' im zentralen Package 'webdocument.classifier.core' stellt hierbei die zentrale Komponente zur Extrahierung beziehungsweise der Berechnung von Merkmalen sowie der dazu nötigen Absatz-Trennung dar.

Die Implementierungen von Modulen zur Merkmalsberechnung befinden sich im Package 'webdocument.classifier.core.feature.impl'. Allen Implementierungen gemein ist die Erweiterung der abstrakten Basisklasse 'FeatureExtractor', welche grundlegende Hilfsfunktionen zur Berechnung von Merkmalen enthält und selbst das Interface 'FeatureExtractorInterface' implementiert. Über dieses Interface werden die einzelnen Implementierungen von der zentralen Komponente 'FeatureExtractor' angesteuert.

Implementierungen zur Absatz-Trennung wurden im Package 'webdocument.classifier-.core.preProcessor' platziert. Die Steuerung der Module erfolgt per Interface 'ParagraphSplitterInterface' über die zentrale Komponente 'FeatureExtractor', äquivalent zur Steuerung der Module zur Merkmalsberechnung.

Das Package 'webdocument.classifier.gate' enthält einen Wrapper für das Gate-Framework (Abschnitt 4.1.1) und wird von Implementierungen der Komponenten zur Merkmalsberechnung verwendet. Nahezu alle Textstatistik- sowie Part-Of-Speech Merkmale (Abschnitt 3.1.1 und 3.1.2) werden auf Basis von GATE berechnet.

Ein weiterer Wrapper ist im Package 'webdocument.classifier.classifier' enthalten, welcher Klassifizierungs-Funktionalitäten auf Basis von WEKA (Abschnitt 4.1.2) zur Verfügung stellt.

4.3 Funktionsweise

Die grundlegende Funktionsweise der Implementierung ist im nachfolgenden Activity-Diagramm skizziert (Abbildung 4.4).

Sofern die Klassifizierung auf Absatz-Ebene stattfindet, muss direkt nach der Initiierung des Klassifizierungs-Prozess eine Absatz-Trennung erfolgen. Das zuständige Modul wird über einen Konfigurations-Eintrag ermittelt. Anschließend erfolgt die Berechnung der für die Klassifizierung notwendigen Merkmale. Diese Berechnung

[9] Ein Package ist eine Sammlung von zusammengehörenden Java-Klassen.
[10] Model-Klassen repräsentieren Datenspeicher und entsprechen den Entitäten des Domänenmodells.

erfolgt für alle getrennten Absätze beziehungsweise das Dokument im Gesamten (abhängig von der gewünschten Granularitätsstufe). Die ermittelten Merkmale werden an die Klassifizierungs-Komponente übergeben, welche abschließend die gewünschte Klasse des Dokuments beziehungsweise jene der einzelnen Absätze ermittelt.

4.4 Komponenten

4.4.1 Absatz-Trennung

4.4.1.1 Absatz-Trennung auf Basis von HTML Mustern

Die grundlegende Annahme dieses Ansatzes ist jene, dass bestimmte HTML Tags[11] bzw. bestimmte aus HTML Tags bestehende Muster als zuverlässige Indikatoren bei der Trennung von Absätzen dienen.

Beispielsweise sind "Style-Tags" relativ sichere Indikatoren um Absätze zu trennen. Sehr häufig können auch mehrfach vorkommende Muster von HTML Tags als Delimiter verwendet werden. Kommentare oder Rezensionen sind meist durch einen gemeinsamen Stil für die Überschrift sowie für den Autor und gegebenenfalls durch eine Bewertungsgraphik[12] bzw. ein adäquates Skript erkennbar.

Im nachfolgenden Beispiel (Abbildung 4.5) wird eine Produkt-Website von amazon.com gezeigt. In der Abbildung wird der Übergang von der Produktbeschreibung zu den zugehörigen Rezensionen dargestellt. Dabei ist folgendes Muster erkennbar: Jede Rezension beginnt mit einer Bewertungsgraphik gefolgt von einer Überschrift, welche mit einem eigenen "Style" versehen wurde. Anschließend folgt der Autor, welcher ebenfalls mit einem eigenen "Style" ausgegeben wird. Abschließend folgt noch eine weitere Bewertungsgraphik für die Kategorie "Fun".

Funktionsweise

Die grundsätzliche Funktionsweise der Absatz-Trennung auf Basis von HTML Mustern lässt sich wie folgt darstellen. Im ersten Schritt wird nach Indikatoren gesucht um potentielle Absätze zu erkennen, insbesondere nach einer Kombination aus einem Absatz - Tag bzw. einem Zeilsprung in einer Tabelle und einem Stilwechsel der Schrift. Beim Auffinden eines solchen Indikators wird eine vorläufige Absatz-Trennung vorgenommen.

Dieses Vorgehen führt zu einer sehr hohen Zahl an Absatz-Trennungen weshalb in einem weiteren Schritt eine Überprüfung stattfindet ob eine getrennte Textpassage

[11] HTML Tags sind strukturierende Markierungen, welche Textbereichen eine Bedeutung zuordnen.

[12] Eine Bewertungsgraphik zeigt an, wie die Besucher einer Website den Beitrag eines Benutzers auf einer festgelegten Skala bewertet haben.

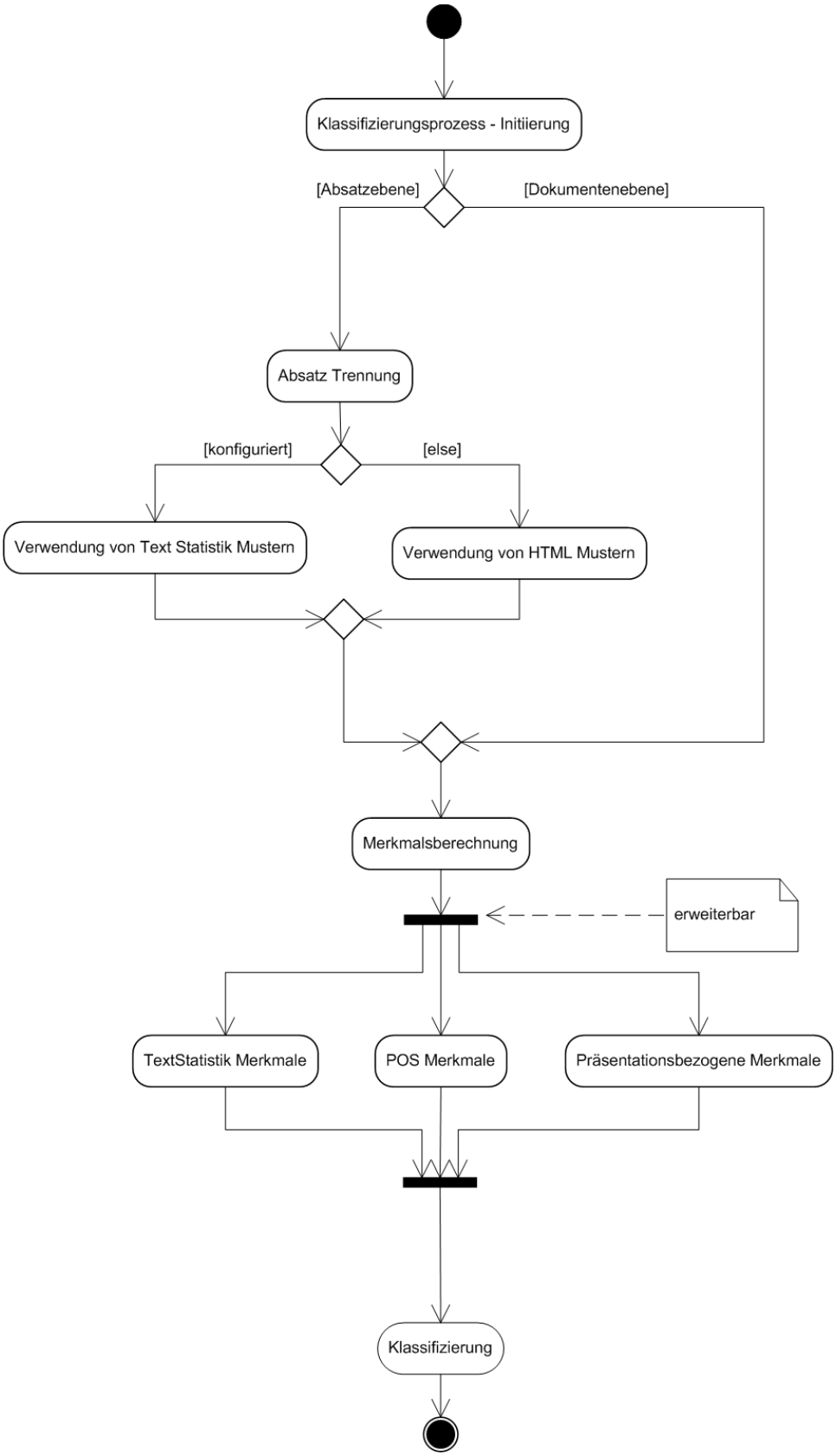

Abbildung 4.4: Activity Diagramm des Klassifizierungs-Prozesses.

tatsächlich als eigener Absatz zu werten ist oder die Trennung wieder rückgänig gemacht werden sollte. Diese Entscheidung wird über die Textgröße des vermeintlichen Absatzes getroffen. Bei Experimenten mit den verwendeten Testkorpora (Abschnitt 5.2) wurden hierbei Werte zwischen 1 und 5 als minimale Anzahl von Sätzen für einen Absatz versucht. Eine Anforderung von mindestens drei Sätzen für einen Absatz hat sich hierbei bewährt.

In der Abbildung 4.6 ist ein Beispiel ersichtlich, an dem die Funktionsweise illustriert wird. Hierbei handelt es sich um eine Kundenrezension zu einer Produktbeschreibung. Hierbei werden Tabellen zur Strukturierung verwendet. Eine Rezension besteht in diesem Fall aus einer Bewertung der Rezension, einer Überschrift sowie einer Angabe des Autors. Am Ende jedes dieser Felder erfolgt ein Absatz sowie eine Änderung der Schrift weshalb eine (vorläufige) Absatz-Trennung vorgenommen wird. Im nächsten Schritt werden die jeweils enthaltenenen Texte analysiert. Die Bewertung der Rezension enthält hierbei den Text '243 of 278 people found the following review helpful:'. Die Überschrift enthält den Text 'Behind the burqa, March 24, 2007'. Der Text des Autor-Feldes enthält den Text 'By Amanda Richards "Modest to the extreme" (Georgetwon, Washington D.C.)'. Alle diese Felder enthalten jeweils nur einen einzigen Satz. Daher werden die einzelnen Absatz-Trennungen wieder rückgängig gemacht und sie werden mit dem nachfolgenden Absatz, welcher den eigentlichen Rezensionstext enthält, vereint.

4.4.1.2 Absatz-Trennung auf Basis von syntaktischen Mustern

Die Annahme dieses Ansatzes ist jene dass sich Textelemente, welche verschiedenen Genres angehören, auf Basis von syntaktischen Mustern erkennen lassen. So ist beispielsweise der Übergang eines Artikels zu einem Kommentar oder der Übergang einer Produktbeschreibung zu einer Rezension häufig von einer signifikanten Veränderung des Sprachstils gekennzeichnet. Insbesondere Wechsel zwischen Nominal- und Verbalstil lassen sich recht einfach über die relativen Häufigkeiten der Wortarten erkennen und sind zugleich ein sehr zuverlässiger Indikator um eine Absatztrennung vorzunehmen. Auch eine signifikante Veränderung in der Häufigkeit von Personalpronomen stellt einen relativ sicheren Indikator für eine Absatz-Trennung dar. Ein weiterer relativ sicherer Indikator ist eine deutliche Veränderung der Lesbarkeit. Da Lesbarkeitsindizes im Allgemeinen sowohl die Satzlänge als auch die Silbenanzahl als Berechnungskriterium verwenden, sind sie ein relativ sicheres Mittel zur Erkennung von Stiländerungen.

Das nachfolgende Beispiel (Abbildung 4.7) zeigt eine Produktseite von amazon.com. In dieser Abbildung ist der Übergang zwischen der Produktbeschreibung und einer zugehörigen Rezension gekennzeichnet. Dieser Übergang soll erkannt werden damit eine Trennung zwischen diesen Absätzen erfolgen kann.

Abbildung 4.5: Anwendungsmöglichkeiten der Absatz-Trennung auf Basis von HTML Mustern.

Die Merkmale der Absätze bzw. jene der entsprechenden Sätze sind in der Tabelle 4.1 aufgelistet. Weiters ist an diesem Beispiel deutlich ersichtlich, dass ein Wechsel vom Nominalstil zum Verbalstil erfolgt. Die relative Häufigkeit von Nomen sinkt von durchschnittlich 33,98 % in der Produktbeschreibung auf einen Wert von 14,26 % in der Rezension. Gleichzeitig steigt die relative Häufigkeit von Verben von durchschnittlich 9,58 % auf 20,68 %. Weiters wird ersichtlich, dass in der Produktbeschreibung die relative Häufigkeit von Personalpronomen lediglich bei 1,85 % liegt. Im Abschnitt der Rezension hingegen ist eine Steigerung auf 9,99 % zu beobachten. Ein weiterer Indikator, welcher in diesem Beispiel anschlägt, ist der Lesbarkeitsindex. Der verwendete 'Flesch Readability Index' steigt von einem Durschnittswert im Abschnitt der Produktbeschreibung von 59,17 auf 83,34.

Die in diesem Beispiel gezeigten Abweichungen von Merkmalen zwischen privaten und öffentlichen Elementen sind zweifellos als Extremwerte zu betrachten, lassen sich jedoch verallgemeinern, wenn auch in abgeschwächter Form. Zu erklären sind diese wie folgt. Häufig werden Rezensionen, Kommentare oder ähnliche Kurztexte mit der Absicht verfasst "Wärme" oder "Greifbarkeit" auszustrahlen, während hingegen sachliche Texte meist wenig anschaulich und oft sogar "hölzern" wirken. Dieser Umstand führt dazu, dass öffentliche Elemente, welche meist die Charakte-

Abbildung 4.6: Beispiel der Absatz-Trennung auf Basis von HTML Mustern. Die drei dargestellten Elemente werden nach der Trennung mit dem nachfolgenden Absatz vereint zumal sie jeweils nur einen einzigen Satz enthalten.

ristiken von Sachtexten vereinen und gewöhnlich in einem Nominalstil verfasst sind. Jedoch gibt es auch Ausnahmen von dieser Regel. Beispielsweise werden Texte mit Handlungsanweisungen (technische Anleitungen oder ähnliches) gelegentlich in einem Verbalstil verfasst um anschaulicher zu wirken, beziehungsweise um den Grad der Komplexität zu senken.

Die deutliche Steigerung der Lesbarkeit der hier gezeigten Rezension ist damit zu erklären, dass in den meisten nicht-professionell verfassten Texte eine Tendenz zu kurzen Sätzen besteht. Zudem ist die durschnittliche Silbenanzahl meist deutlich niedriger wenn keinerlei fachspezifische Begriffe oder Bezeichnungen verwendet werden. Da der hier verwendete Lesbarkeitsindex auf genau diesen beiden Faktoren basiert, ist die Abweichung dadurch zu erklären.

4.4.2 Merkmalsberechnung

Die Merkmalsberechnung lässt sich in drei Komponenten gliedern. Es existiert eine Komponente für Textstatistik-bezogene Merkmale (Abschnitt 3.1.1) welche auf Basis von GATE (Abschnitt 4.1.1) implementiert wurde. Dieser Komponente obliegt die Berechnung der nachfolgend aufgelisteten Merkmalsklassen 'Textstatistik', 'Token-Kategorien', 'Lesbarkeitsindizes' sowie 'Lookups'. Weiters existiert eine Komponente

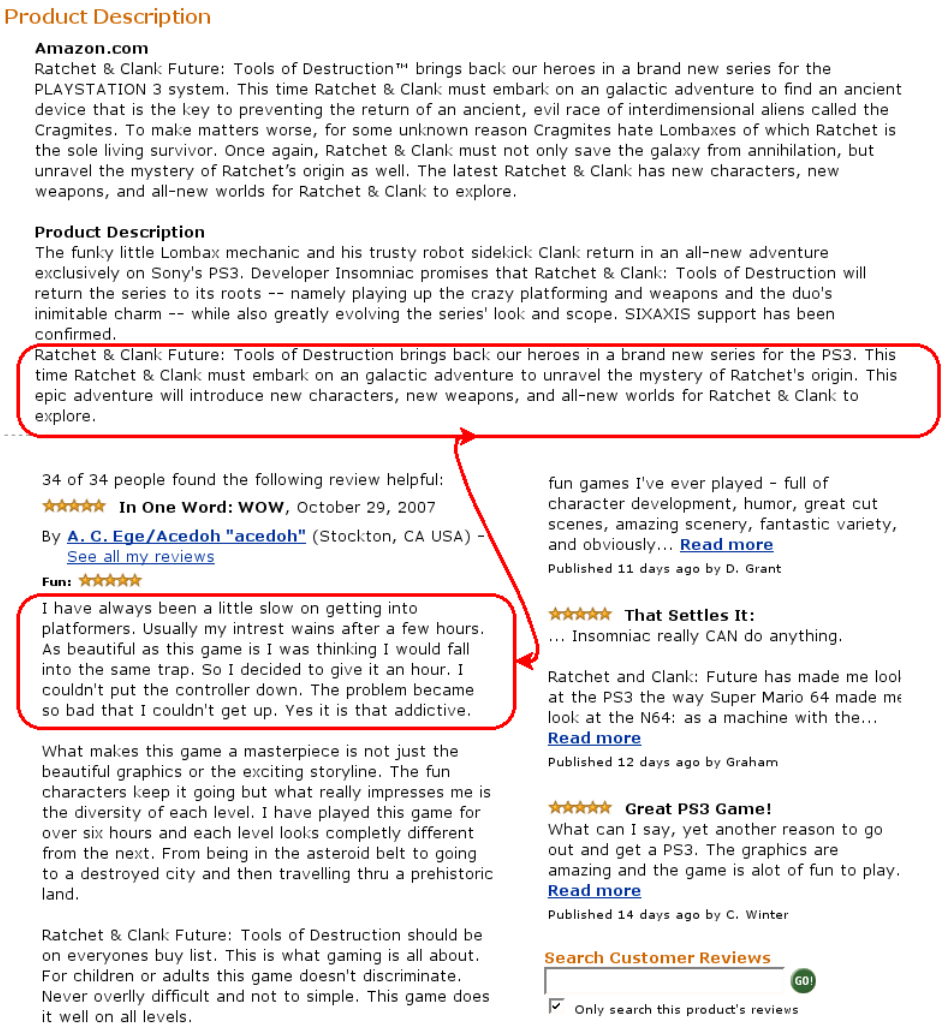

Abbildung 4.7: Anwendungsmöglichkeiten der Absatz-Trennung auf Basis von syntaktischen Mustern.

zur Berechnung der Merkmale aus der Klasse 'Part-Of-Speech' (Abschnitt 3.1.2) welche ebenfalls auf Basis von GATE implementiert wurde. Die letzte der drei Komponenten zur Merkmalsberechnung zeichnet sich für präsentationsbezogene Merkmale (Abschnitt 3.1.3) zuständig und wurde auf Basis von HTML Parser (Abschnitt 4.1.3) implementiert.

Alle implementierten und verwendeten Merkmale sind in der Tabelle 4.2 aufgelistet.

Textstatistik

Für jeden Satz werden die Anzahl der Wörter, sowie deren Länge und Silbenanzahl als absolute Werte gespeichert. Dies ermöglicht eine Aggregation dieser Werte für unterschiedliche Hierarchiestufen einer Website. Diese Vorgehensweise ist auch eine zwingende Voraussetzung um für jeden beliebigen Bereich eines Dokuments Lesbarkeitsindizes berechnen zu können.

Satz	Flesh Readability Index	Nomen (Relative Häufigkeit)	Verben (Relative Häufigkeit)	Personal-pronmen (Relative Häufigkeit)
Merkmale der Produktbeschreibung				
Ratchet & Clank Future: Tools of Destruction brings back our heroes in a brand new series for the PS3.	76,23	33,33 %	5,55 %	5,55 %
This time Ratchet & Clank must embark on an galactic adventure to unravel the mystery of Ratchet's origin.	42,87	35,29 %	17,64 %	0,00 %
This epic adventure will introduce new characters, new weapons, and all-new worlds for Ratchet & Clank to explore.	58,42	33,33 %	5,55 %	0,00 %
Merkmale der Rezension				
I have always been a little slow on getting into platformers.	72,62	9,09 %	18,18 %	9,09 %
Usually my intrest wains after a few hours.	71,82	25,00 %	12,50 %	12,50 %
As beautiful as this game is I was thinking I would fall into the same trap.	79,56	12,50 %	31,25 %	12,50 %
So I decided to give it an hour.	92,96	12,50 %	25,00 %	12,50 %
I couldn't put the controller down.	87,95	14,28 %	28,56 %	14,28 %
The problem became so bad that I couldn't get up.	95,16	9,09 %	27,27 %	9,09 %
Yes it is that addictive.	83,32	20,00 %	20,00 %	0,00 %

Tabelle 4.1: Absatztrennung auf Basis von syntaktischen Mustern - Beispiel eines Überganges von einer Produktbeschreibung zu einer Rezension.

Token-Kategorien

Die Kategorien einzelner Tokens werden für jeden Satz als relative Häufigkeiten gespeichert um eine Aggregation dieser Werte für unterschiedliche Hierarchiestufen einer Website zu ermöglichen. Neben der Kategorien 'Word-Tokens' für Wörter sowie 'Number-Tokens' für Zahlen werden auch Symbole über die Kategorie 'Symbol-Tokens' erkannt.

Gerade bei Zeitungsartikeln, insbesondere aus der Rubrik Wirtschaft, kann hierbei häufig eine erhöhte Häufigkeit von Zahlen und Symbolen festgestellt werden. Bedingt durch die zahlreiche Verwendung von Währungssymbolen oder Prozentzeichen in Kombination mit den eigentlichen Zahlenwerten kann so eine annähernd gleiche Häufigkeit von Number-Tokens und Symbol-Tokens zustande kommen. In der nachfolgenden Tabelle 4.3 ist ein Beispiel ersichtlich, wobei die ersten drei Sätze aus dem

Klasse	**Merkmale**	
Textstatistik	Wort-Anzahl	Wort-Länge
	Silben-Anzahl	
Token-Kategorien	Word-Tokens	Number-Tokens
(Rel. Häufigkeiten)	Symbol-Tokens	Punctuation-Tokens
	Space-Tokens	Control-Tokens
Lesbarkeitsindizes	Flesch Reading Ease	Flesch-Kincaid Grade Level
Lookups	Orte	Adressen
(Rel. Häufigkeiten)	Zeit Datum	Gesellschafts-Namen
	Personen-Namen	Abkürzungen
	Währungen	Telefon- und Faxnummern
Part-Of-Speech	Nomen	Verben
(Rel. Häufigkeiten)	Adjektive	Adverben
	Modalverben	Konjunktionen
	Pronomen	Personalpronomen
	Artikel	Präpositionen
	Ausrufe	Listenelemente
Präsentation	Links	Überschriften
(Rel. Häufigkeiten)	Absätze	Textformattierungen
	Listen	Tabellen
	Graphiken	Frames
	Formulare	Multimedia

Tabelle 4.2: Implementierte und verwendete Merkmale.

eigentlichen Artikel stammen und eine entsprechend höhere Häufigkeit an Symbolen und Zahlen aufweisen. Im zweiten Satz wird eine Prozentzahl verwendet, im dritten hingegen ein Währungsbetrag. Beiden Sätzen gemein ist eine jeweils (gleichmäßig) erhöhte Häufigkeit von Number- und Symbol Tokens.

Teils werden bei User Kommentaren aus Bequemlichkeit auf Satzzeichen verzichtet oder aber nur sehr 'sparsam' verwendet. Auch dies zeichnet sich bei dem verwendeten Beispiel ab (Tabelle 4.3). Die ersten drei Sätze, welche aus dem eigentlichen Artikel stammen, weisen eine durchschnittliche relative Häufigkeit von 4,00 % an Satzzeichen auf. Die zwei nachfolgenden Sätze, welche User - Kommentare darstellen, weisen deutlich weniger Satzzeichen auf. So weist der erste Kommentar einen Wert von 3,00 % auf, während der zweite Kommentar gänzlich auf die Verwendung von Satzzeichen verzichtet.

Die Häufigkeit von 'Space Tokens' gibt Auskunft über die Anzahl der Leerzeichen, während hingegen 'Control Tokens' Zeilensprünge repräsentieren.

Lesbarkeitsindizes

Die beiden implementierten Lesbarkeitsindizes 'Flesch Reading Ease' und 'Flesch-Kincaid Grade Level' werden jeweils auf Basis der durchschnittlichen Wortanzahl

Text	WT	NT	ST	PT	SPT	CT
New "green" taxes on household waste will cause more toxic pollution as ratepayers dodge the levy by burning their own rubbish, it was warned last night.	0.47	0.00	0.04	0.04	0.45	0.00
Around 40 % of homes will opt to build their own garden bonfires if Labour's bin taxes are introduced, research suggests.	0.46	0.02	0.02	0.05	0.45	0.00
The tax could see some middle-class homes paying up to £100 a year extra to have their bins emptied.	0.48	0.03	0.03	0.03	0.43	0.00
Hey, I already do this as my bins are only emptied once every two weeks	0.48	0.00	0.00	0.03	0.45	0.00
I have to burn the excess as there's no alternative method of disposal	0.50	0.00	0.00	0.00	0.47	0.00

Tabelle 4.3: Beispiel für Token-Kategorie Merkmale. Legende: WT=Word Tokens, NT=Number Tokens, ST=Symbol Tokens, PT=Punctuation Tokens, SPT=Space Tokens, CT=Control Tokens

pro Satz sowie der durchschnittlichen Silbenanzahl pro Wort berechnet. Diese Informationen stehen über die zuvor erwähnten Textstatistik Merkmale zur Verfügung, wodurch für beliebige Ausschnitte eines Dokuments ein Lesbarkeitsindex berechnet werden kann.

Beide Indizes werden auf einen Wertebereich zwischen 0.00 und 1.00 normalisiert. Der 'Flesch Reading Ease' Index kann einen Wertebereich von 0 bis 121 einnehmen, wobei der theoretische Bestwert 121 nur erreicht werden kann wenn jeder Satz aus einem einzigen einsilbigen Wort besteht. Der 'Flesch-Kincaid Grade Level' Index kann im theoretischen Fall von aus jeweils einem einsilbigen Wort bestehenden Sätzen einen Bestwert von -3.4 annehmen (im Gegensatz zum 'Flesch Reading Ease' Index ist hier ein niedriger Wert besser und lässt sich dahingehend interpretieren als dass weniger Schuljahre nötig sind um den Text zu verstehen). Die obere Grenze wird jedoch durch die durchschnittliche Wortanzahl pro Satz definiert, weshalb auch der Wert 12, welcher dem höchsten amerikanischen Schulgrad entspricht, überschritten werden kann. Daher wurde als obere Schranke der Wert 15 gewählt. Bei der Normalisierung wurde der Wertebereich invertiert um sicherzustellen dass bei allen normierten Indizes 0.0 dem schlechtest möglichen Wert entspricht während 1.0 das Optimum darstellt.

Häufig sind beispielsweise User - Kommentare oder Blog Einträge aufgrund der meist kürzeren Sätze und der weitgehenden Vermeidung von Fremdwörtern bzw. von vielsilbigen Wörtern einfacher verständlich. Diese einfachere Verständlichkeit soll durch die Verwendung von Lesbarkeitsindizes gemessen werden. In der Tabelle 4.4 ist ein Beispiel gelistet, wobei die ersten drei Sätze aus einem Artikel stammen während

die anderen drei zugehörige Kommentare repräsentieren. In der Tabelle sind für die einzelnen Sätze jeweils die berechneten 'Flesch Reading Ease' und 'Flesch-Kincaid Grade Level' Lesbarkeitsindizes mit den Original- sowie den normalisierten Werten dargestellt. Für die ersten drei Sätze aus dem Artikel werden als normalisierte Lesbarkeitsindizes die Werte 0.62 (FREN) bzw. 0.56 (FKGN) berechnet. Für die anderen drei Sätze, welche zugehörige Kommentare repräsentieren, ergeben sich hingegen die deutlich höheren Werte 0.78 (FREN) bzw. 0.87 (FKGN).

Text	**FRE**	**FREN**	**FKG**	**FKGN**
New "green" taxes on household waste will cause more toxic pollution as ratepayers dodge the levy by burning their own rubbish, it was warned last night.	63.3	0.52	10.9	0.41
Around 40 % of homes will opt to build their own garden bonfires if Labour's bin taxes are introduced, research suggests.	72.7	0.60	8.3	0.55
The tax could see some middle-class homes paying up to £100 a year extra to have their bins emptied.	94.0	0.78	4.9	0.73
Where does a large amount of waste come from?	103.7	0.85	1.0	0.95
Who is a major donor to the Labour party?	84.9	0.70	3.7	0.79
The Government could tackle the problem at source?	93.0	0.77	2.3	0.88

Tabelle 4.4: Beispiel für Lesbarkeitsindizes, Darstellung der Merkmale mit originalen und normalisierten Werten. Legende: FRE=Flesch Reading Ease, FREN=Flesch Reading Ease (Normiert), FKG=Flesch-Kincaid Grade Level, FKGN=Flesch-Kincaid Grade Level (Normiert)

Lookups

Lookups sind einfache Klassifikatoren, welche anhand eines Wörterbuches oder Mustern, versuchen einzelne Wörter verschiedenen Klassen zuzuordnen. Hierbei wurden die in Gate (Abschnitt 4.1.1) implementierten Lookup-Klassen als Merkmale verwendet, wobei die berechneten Werte deren relativen Häufigkeiten entsprechen.

Die Annahme welche hinter der Verwendung dieser Merkmale steckt ist jene, dass sich starke Ausprägungen bestimmter Wort-Klassen in manchen Fällen als guter Diskriminator zwischen privaten und öffentlichen Elementen erweisen. Obwohl zwar auch auf privaten Websiten Kontaktdaten zu finden sind, lassen zahlreich vorhandene Kontaktdaten auf eine öffentliche Website schließen. So können beispielsweise klassische Firmenprofil - Seiten recht zielsicher erkannt werden wenn die Anzahl der Personen-Namen Lookups mit der Anzahl der erkannten Telefon- und Faxnummern nahezu identisch ist. Auch ein für Firmen meist verpflichtendes Impressum lässt sich recht zielsicher erkennen durch die Kombination aus einem Gesellschaftsnamen, dem

Personennamen eins Verantwortlichen sowie einer Adresse und gegebenenfalls einer Telefon- oder Faxnummer.

Ein weiter Grund für die Notwendigkeit dieses Merkmals ist die Tatsache dass eine solche Klassifizierung auf Wortebene Voraussetzung ist um bestimmte, besonders sensitive Elemente eines Satzes zu sperren bzw. auszublenden (wenngleich diese Granularitätsstufe in dieser Arbeit nicht näher behandelt wird). Ein möglicher Anwendungsfall wäre die Anonymisierung von personenbezogenen Daten in Dokumenten. So könnten beispielsweise Namen von Personen, deren Geburtsdatum, Anschrift und sonstige sensitive Daten 'geschwärzt' werden.

Part-Of-Speech

Auf Basis des Part-Of-Speech Taggers von Gate (Abschnitt 4.1.1) wurde für jede Wortart ein Merkmal implementiert welches die jeweilige relative Häufigkeit angibt. Bemerkenswert ist hierbei dass Nummern und Buchstaben welche als Aufzählungszeichen fungieren erkannt werden und hierfür vom POS-Tagger eine eigene Wortart existiert (in der Merkmal Tabelle 4.2 als 'Listenelemente' bezeichnet).

Ähnlich verhält es sich mit sogenannten 'Ausrufen', welche ebenfalls eine eigene Wortart darstellt (in der Merkmal Tabelle 4.2 als 'Ausrufe' bezeichnet). Hierbei werden bestimmte Wörter wie zum Beispiel "yes", "no" oder "well" abhängig vom Kontext als sogenannte Ausrufe bzw. Zwischenrufe gewertet. Beispielsweise wird das Wort "No" im Satz "No! I just don't agree." als Aufruf bzw. Zwischenruf gewertet. Im Satz 'I have to say no.' hingegen ist dies nicht der Fall. Eine erhöhte Häufigkeit von 'Ausrufen' (bzw. überhaupt ein Wert > 0.0) kann auf ein Gespräch bzw. eine Diskussion privater Natur hinweisen.

Die Merkmale der Gruppe 'Part-Of-Speech' werden vorwiegend dazu genutzt um Sprachstile zu erkennen. Durch eine hohe Häufigkeit von Nomen bzw. Verben lassen sich beispielsweise recht einfach erkennen ob Elemente eines Dokuments in einem Nominal- oder Verbalstil verfasst wurden, welche wiederum typisch für öffentliche bzw. private Elemente sind. In der Tabelle 4.5 ist eine Produktbeschreibung und ein darauf bezogener Kommentar mit den Häufigkeiten der Wortarten Nomen, Verben und Personalpronomen dargestellt. Während die relative Häufigkeit von Nomen in der Produktbeschreibung bei 0,33 liegt, beträgt sie in der Rezension lediglich 0,15. Im Gegenzug liegt die relative Häufigkeit von Verben in der Produktbeschreibung bei 0,09, in der Rezension hingegen bei 0,23. Währen die Produktbeschreibung in einem Nominalstil verfasst wurde, zeigt sich beim Kommentar eine Tendenz zum Verbalstil. Auch die Häufigkeit der Personalpronomen ist bei einem sachlichen Text meist deutlich geringer als in beispielsweise einer Rezension oder einem Blog Eintrag welche häufig in einem erzählerischen Stil verfasst sind. So liegt die Häufigkeit

von Personalpronomen in der Produktbeschreibung bei 0,02 während hingegen beim Kommentar ein Wert von 0,10 vorliegt.

Text	**Nomen**	**Verben**	**Personal-pronmen**
Ratchet & Clank Future: Tools of Destruction brings back our heroes in a brand new series for the PS3.	0,33	0.05	0,05
This time Ratchet & Clank must embark on an galactic adventure to unravel the mystery of Ratchet's origin.	0,35	0,17	0,00
This epic adventure will introduce new characters, new weapons, and all-new worlds for Ratchet & Clank to explore.	0,33	0,05	0,00
I have always been a little slow on getting into platformers.	0,09	0,18	0,09
Usually my intrest wains after a few hours.	0,25	0,13	0,13
As beautiful as this game is I was thinking I would fall into the same trap.	0,13	0,31	0,13
So I decided to give it an hour.	0,13	0,25	0,13
I couldn't put the controller down.	0,14	0,29	0,14
The problem became so bad that I couldn't get up.	0,09	0,27	0,09
Yes it is that addictive.	0,20	0,20	0,00

Tabelle 4.5: Part-Of-Speech Merkmalwerte einer Produktbeschreibung und einer zugehörigen Rezension.

Präsentation

Präsentationsmerkmale werden als relative Häufigkeiten von Tag-Klassen berechnet (Abschnitt 3.1.3). Die Idee hinter diesem Merkmal ist durch erhöhte Werte von Tag-Klassen charakteristische visuelle Elemente erkennen zu können.

Insbesondere Kommentare, Rezensionen, Blog-Einträge oder Beiträge in Foren weisen häufig gemeinsame charakteristische Häufigkeiten von verschiedenen Tag-Klassen auf. In der Abbildung 4.8 ist ein Beispiel mit Rezensionen einer Produktseite samt HTML Code dargestellt. Die relative Häufigkeit der Tabellen-Tag Klasse für eine Rezension beträgt in diesem Beispiel 0,30, jene für die Absatz-Tag Klasse 023. Die relative Häufigkeit der Klasse für Formattierungen nimmt den Höchstwert von 0,38 ein während die Häufigkeit für die Bilder-Tag Klasse bei lediglich 0,07 liegt.

Sehr typisch für derartige Abschnitte ist der hohe Wert von Tabellen-Tags welcher mit einem hohen Wert für die Klassen Absatz sowie Formattierungen einhergeht. So werden meist Tabellen zur Strukturierung von Abschnitten verwendet. Umso kürzer hierbei die eigentlichen Inhalte sind, desto höher ist die relative Häufigkeit der

Tag-Klasse zur Strukturierung. Die hohen Häufigkeiten für Formattierungen und Absätzen sind durch die meist zahlreichen verschiedenen Felder zu erklären. So gibt es nahezu immer eigene Felder mit unterschiedlichen Formattierungen für die Darstellung des Autors, der Überschrift und des eigentlichen Textes. Je nach Anwendung kann die jeweilige Anzahl von Feldern per Abschnitt auch höher sein. Trotzdem korrelieren die Werte der Tabellen-Tag Klasse meist mit jenen der Formattierungen.

Closed-Class Word Sets (Verworfen)

Sogenannte 'Closed-Class Word Sets' sind vordefinierte, geschlossene Klassen von Wörtern. Üblicherweise enthalten derartige Klassen sehr wenige und eindeutig zuzuordnende Elemente. Beispielsweise lassen sich Online-Shops relativ eindeutig über dieses Merkmal erkennen, zumal die Häufigkeit von Copyright-Hinweisen oder bestimmten, meist sehr plakativen Schlagwörtern signifikant höher ist als bei anderen Genres.

In hochentwickelten Ansätzen wird zudem auch häufig neben dem Auftreten bestimmter Wort-Klassen deren Verteilung berücksichtigt. Hierbei ist es oft entscheidend ob sich die Vorkomnisse gleichmäßig über den gesamten Text erstrecken oder aber auf wenige Häufungspunkte konzentrieren.

Die relativen Häufigkeiten von Wörtern, welche einer bestimmten Klasse zugeordnet werden können, korrelieren meist sehr stark mit dem der Klasse zugewiesenen Genre. Aufgrund des hohen Stellenwerts und der einfachen Verwendung sowie Adaptierbarkeit dieses Merkmals, sind 'Closed-Class Word Sets' ein in der Genre Analyse sehr beliebtes Merkmal.

Dieses Merkmal wurde testweise implementiert. Hierbei wurden ein Wörterbuch für die in den Testkorpora vorkommenden Wörter erstellt. Hierbei wurde der '7 web genre collection' Korpus (Abschnitt 5.2.1) verwendet. Anschließend wurden die 'tf-idf' [13] Werte ermittelt. Die Auswahl der relevanten Elemente der einzelnen Klassen erfolgt per Feature-Ranking (Abschnitt 3.2.1.1) mittels WEKA (Abschnitt 4.1.2) unter Verwendung der 'Information Gain' - Funktion. Die verwendeten Parameter sind in der Tabelle 4.6 dargestellt. Die Definition der einzelnen Word Sets erfolgte hierbei durch die Auswahl der für die Klasse jeweils bestgereihten Wörter.

Die Durchführung der Tests erfolgte mit Word Sets mit einer Anzahl von 10 bzw. 20 bzw. 30 Elementen unter Verwendung eines naiven Bayes-Klassifikators. Die verwendeten Parameter sind in der Tabelle 4.6 abgebildet. Testweise wurde anschließend

[13] tf-idf (Term Frequency - Inverse Document Frequency) ist eine Gewichtungsmethode für Schlüsselwörter beim Information Retrieval. Die Termfrequenz (Term Frequency) gibt hierbei einen Hinweis auf die Bedeutung des Terms für ein Dokument. Die inverse Dokumentenhäufigkeit (Inverse Document Frequency) wiederum misst die Bedeutung eines Terms für die Gesamtmenge der betrachteten Dokumente.

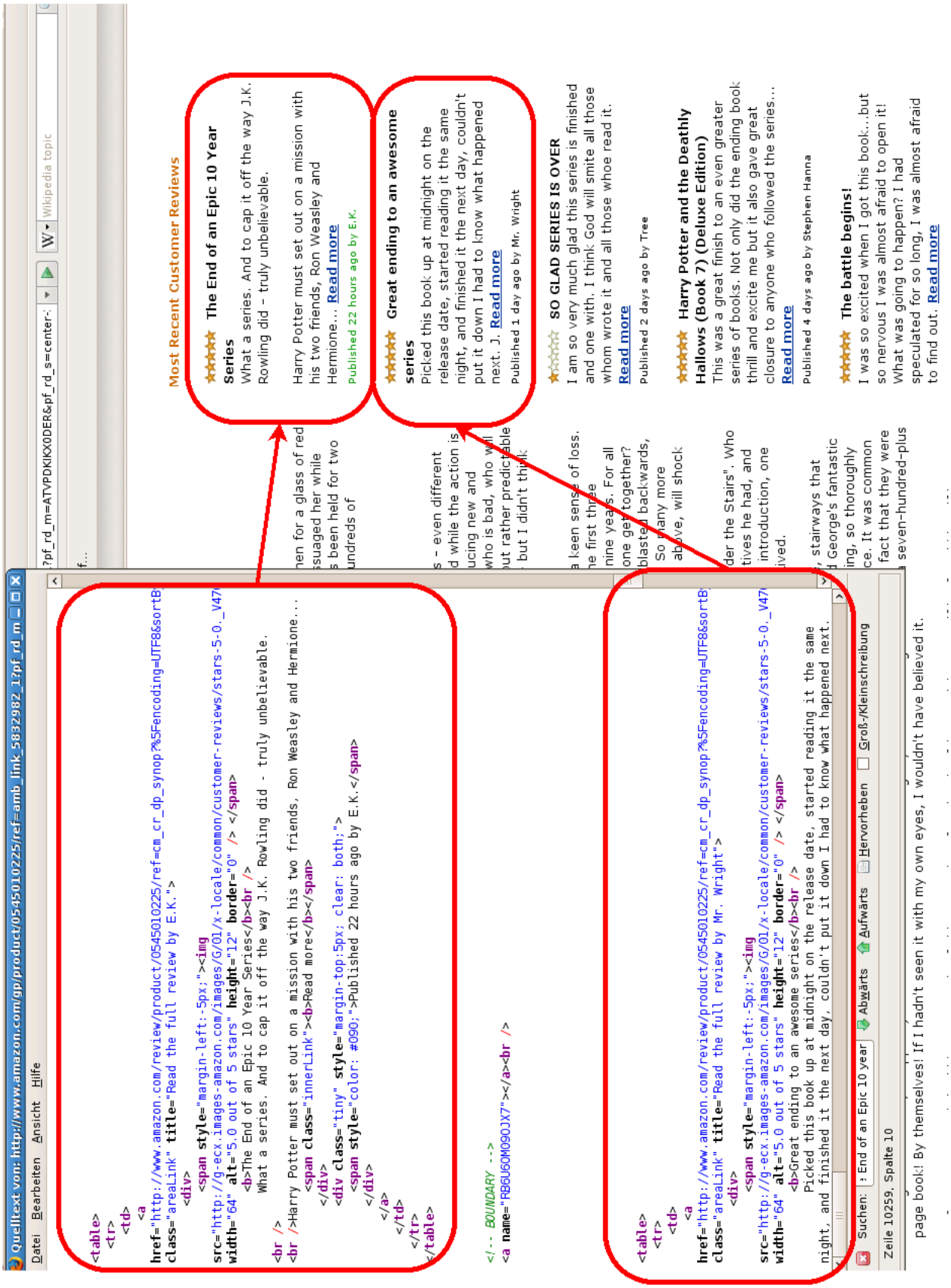

Abbildung 4.8: Präsentationsmerkmale - Beispiel von wiederkehrenden Mustern.

Evaluator	weka.attributeSelection.InfoGainAttributeEval
Search	weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Klassifikator	weka.classifiers.bayes.NaiveBayes -D

Tabelle 4.6: WEKA Parameter zur Auswahl der Closed-Class Word Set - Einträge und Klassifikation unter Verwendung selbiger.

eine Klassifikation mit den 7 Genre - Klassen des Korpus durchgeführt. Hierbei kam es zu einer, wenig überraschenden, signifikant höheren Erkennungsrate. Die besten Ergebnisse wurden mit Set-Größen von 20 Wörtern erzielt. Hierbei konnte der zuvor sehr schwache Recall Wert von knapp 50 % auf über 70 % gesteigert werden.

Anschließend wurde eine Klassifikation mit den beiden Klassen 'privat' und 'öffentlich' durchgeführt. Hierbei konnten ohne Verwendung von 'Closed-Class Word Sets' Recall Werte im bereich von knapp 75 % erzielt werden. Durch die Miteinbeziehung dieses Merkmals konnte leider keine spürbare Verbesserung erzielt werden. Im Gegenteil, der Klassifikator arbeitete sogar etwas effizienter (Recall und Precision verschlechterten sich unter Einbeziehung der Word Sets um etwa 0,5 %) ohne die zusätzlichen Merkmale.

Eine mögliche Erklärung hierfür ist jene, dass die für diese beiden Klassen relevanten Wörter bereits durch andere Merkmale (insbesondere durch die relative Häufigkeit von Personalpronomen welche für die Genres 'Personal Homepage' und 'Blog' jeweils bei den bestgereihten Einträgen der Word Sets lagen) abgedeckt werden. 'Closed-Class Word Sets' sind zur Bestimmung von einzelnen Genres sehr gute Diskriminatoren. In diesem Fall wurde ein Genre-Problem aber auf ein 2-Klassen Problem reduziert, wodurch die Entscheidungsgrenze ('Decision Boundary') wohl zu komplex wurde.

4.4.3 Klassifikator

Die Klassifizierung erfolgte mittels WEKA-Implementierungen eines naiven Bayes-Klassifikators (Abschnitt 3.3.2.3), eines k-Nearest Neighbor-Klassifikators (Abschnitt 3.3.2.1) sowie eines Support Vector Machine-Klassifikators (Abschnitt 3.3.2.2). In den nachfolgenden Abbildungen 4.9 und 4.10 sind Beispiele von visualisierten Klassenzuordnungen von Elementen eines Web-Dokuments dargestellt. Private Elemente sind hierbei rot dargestelt, während öffentliche Elemente grün dargestellt werden. Die Wahrscheinlichkeit, welche der Klassifikator für die Klassenzugehörigkeit berechnet, ist hierbei ausschlaggebend für die Stärke des Farbtons.

4.5 Dimensionsreduktion

Die Dimensionsreduktion erfolgte anhand des WEKA - Explorers (Abschnitt 4.1.2).

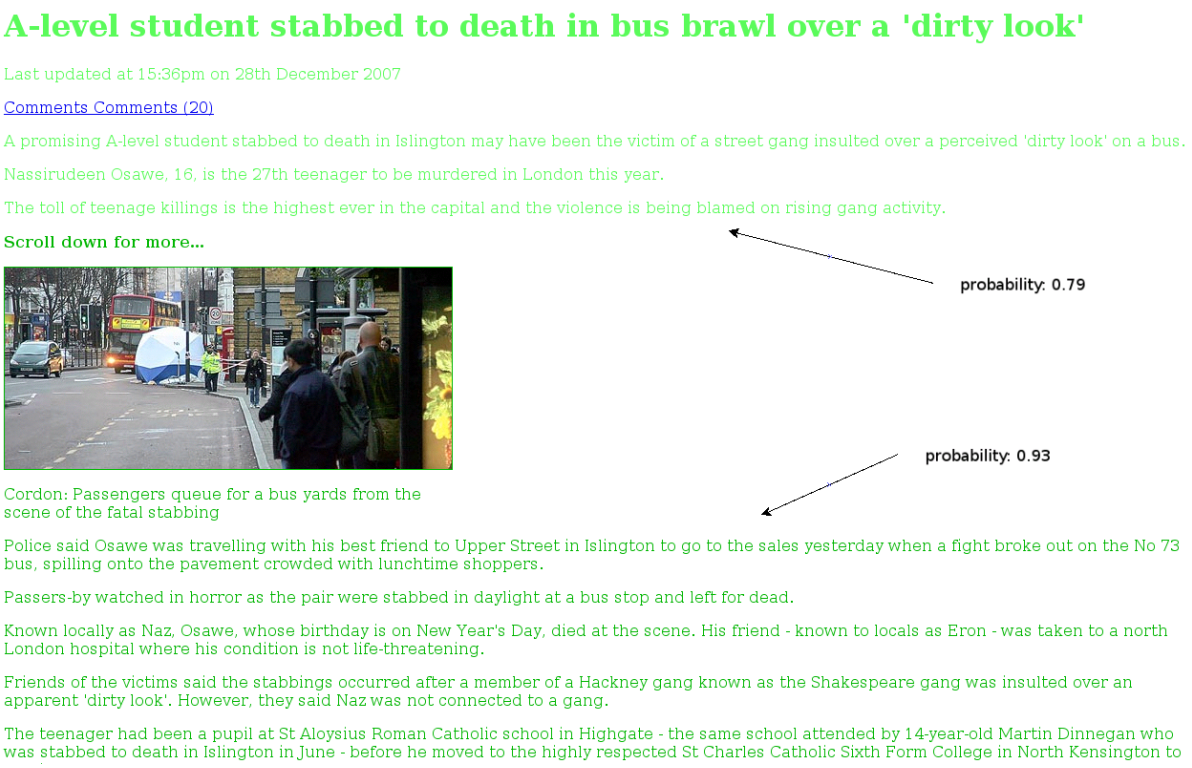

Abbildung 4.9: Visualisierung der Klassenzuordnungen von Elementen eines Web-Dokuments. Öffentliche Elemente sind grün dargestellt, private rot. Die Stärke des Farbtons gibt die Wahrscheinlichkeit an, welche der Klassifikator für die Klassenzugehörigkeit berechnet.

Zur Ermittlung der Merkmalmenge für den naiven Bayes-Klassifikator wurde ein Wrapper-Ansatz (Abschnitt 3.2.1.2) verwendet. Die verwendeten Parameter sind in der Tabelle 4.7 abgebildet. Nachfolgend (Tabelle 4.8) sind die Ergebnisse der Dimensionsreduktion aufgelistet.

Evaluator	weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.bayes.NaiveBayes -F 10 -T 0.01 -R 1 –
Search	weka.attributeSelection.GreedyStepwise -R -T -1.7976931348623157E308 -N -1

Tabelle 4.7: WEKA Parameter zur Dimensionsreduktion des naiven Bayes Klassifikators (Wrapper Ansatz unter Verwendung eines naiven Bayes-Klassifikators).

Für die Dimensionsreduktion des SVM-Klassifikators wurde ein ein 'Feature Ranking' Ansatz verwendet, wobei zur Evaluierung der Attribute die in Weka implementierte 'SVMAttributeEval' - Funktion verwendet wurde. Diese Funktion evaluiert Attribute unter Verwendung des SVM - Klassifikators. Hierbei erhält ein Attribut als Rang die Wurzel des vom SVM - Klassifikator zugewiesenen Gewichts. Die hier-

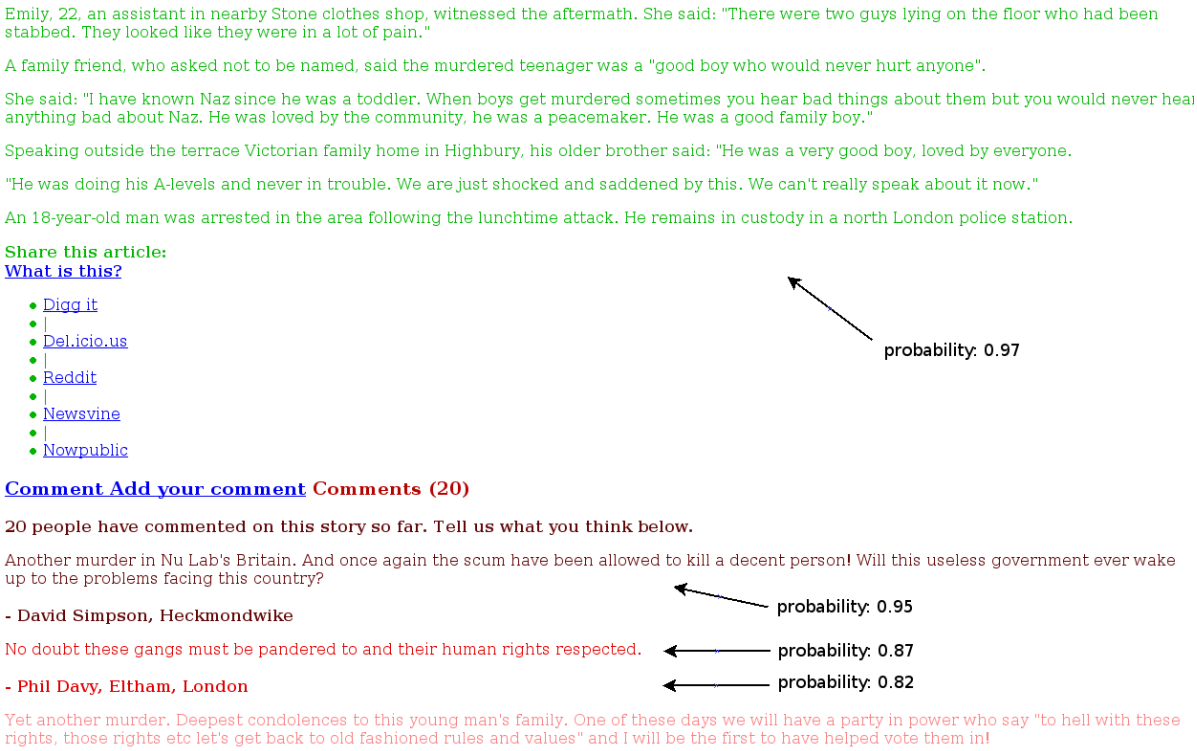

Abbildung 4.10: Visualisierung der Klassenzuordnungen von Elementen eines Web-Dokuments. Öffentliche Elemente sind grün dargestellt, private rot. Die Stärke des Farbtons gibt die Wahrscheinlichkeit an, welche der Klassifikator für die Klassenzugehörigkeit berechnet.

für verwendeten Parameter sind in der Tabelle 4.9 abgebildet. Die Ergebnisse sind in der Tabelle 4.10 abgebildet.

Für die Dimensionsreduktion des k-NN-Klassifikators wurde ein ein 'Feature Ranking' Ansatz verwendet, wobei zur Evaluierung der Attribute die 'InformationGain' - Funktion verwendet wurde. Durch diese Funktion wird ermittelt, wieviel Information für die Unterscheidung zwischen verschiedenen Klassen gewonnen wird oder verloren geht, wenn dieses Merkmal weggelassen wird. Die hierfür verwendeten Parameter sind in der Tabelle 4.11 abgebildet. Die Ergebnisse sind in der Tabelle 4.12 abgebildet.

Interpretation der Ergebnisse

Durch die Korrelationsmatrix (ein Ausschnitt ist in der Abbildung 4.2 ersichtlich) lassen sich recht schnell jene Merkmale erkennen, welche direkt mit der Ziel-Klasse korrelieren. Durch die relative Häufigkeit von Nomen und Verben (Abbildungen 4.11 und 4.12) zeichnet sich recht deutlich ein ausgeprägter Nominalstil bei öffentlichen Elementen ab. Bei privaten Elementen hingegen wird eine stärkere Ausprägung des

Merkmal
*Abkürzungen Lookups (Relative Häufigkeit)**
*Multimedia-Tags (Relative Häufigkeit)**
*Absatz-Tags (Relative Häufigkeit)**
*Listenelemente (Relative Häufigkeit)**
*Orte Lookups (Relative Häufigkeit)**
*Währungen-Lookups (Relative Häufigkeit)**
Adjektive (Relative Häufigkeit)
Adressen Lookups (Relative Häufigkeit)
Adverben (Relative Häufigkeit)
Artikel (Relative Häufigkeit)
Ausrufe (Relative Häufigkeit)
Control-Tokens (Relative Häufigkeit)
Durchschnittliche Silbenanzahl pro Wort
Durchschnittliche Wortanzahl pro Satz
Durchschnittliche Wortlänge
Flesch Reading Ease - Lesbarkeitsindex
Flesch-Kincaid Grade Level - Lesbarkeitsindex
Formular-Tags (Relative Häufigkeit)
Frame-Tags (Relative Häufigkeit)
Gesellschafts-Namen Lookups (Relative Häufigkeit)
Graphik-Tags (Relative Häufigkeit)
Konjunktionen (Relative Häufigkeit)
Link-Tags (Relative Häufigkeit)
Listen-Tags (Relative Häufigkeit)
Modalverben (Relative Häufigkeit)
Nomen (Relative Häufigkeit)
Number-Tokens (Relative Häufigkeit)
Personalpronomen (Relative Häufigkeit)
Personen-Namen Lookups (Relative Häufigkeit)
Präpositionen (Relative Häufigkeit)
Pronomen (Relative Häufigkeit)
Punctuation-Tokens (Relative Häufigkeit)
Space-Tokens (Relative Häufigkeit)
Symbol-Tokens (Relative Häufigkeit)
Tabellen-Tags (Relative Häufigkeit)
Telefon- und Faxnummern Lookups (Relative Häufigkeit)
Textformattierungs-Tags (Relative Häufigkeit)
Überschriften-Tags (Relative Häufigkeit)
Verben (Relative Häufigkeit)
Word-Tokens (Relative Häufigkeit)
Zeit Datum Lookups (Relative Häufigkeit)

Tabelle 4.8: Dimensionsreduktion - Naiver Bayes-Klassifikator Ergebnisse. Die kursiv dargestellten Merkmale welche mit einem Stern markiert sind wurden hierbei nicht berücksichtigt für die optimale Merkmalmenge.

Verbalstils deutlich. Weiters wird der Sprachstil sehr deutlich durch die relative

Evaluator	weka.attributeSelection.SVMAttributeEval -X 1 -Y 0 -Z 0 -P 1.0E-25 -T 1.0E-10 -C 1.0 -N 2
Search	weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Tabelle 4.9: WEKA Parameter zur Dimensionsreduktion des SVM Klassifikators (Feature Ranking Ansatz unter Verwendung des SVM-Klassifikators).

Häufigkeit von Präpositionen, Adverben und Personalpronomen beschrieben. Insbesondere Personalpronomen zeigen sich als enorm wichtig, so nehmen sie doch den unbestrittenen (bei einer relativ geringen Abweichung) ersten Rang in der Ergebnisliste ein.

Ein weiteres sehr wichtiges Merkmal ist die Lesbarkeit, respektive die durchschnittliche Wort- und Silbenanzahl. In Abbildung 4.13 ist die Korrelation des Flesch Lesbarkeitsindex mit der Ziel-Klasse ersichtlich. Hierbei ist recht deutlich ersichtlich dass private Elemente im Vergleich zu öffentlichen eine deutlich höhere Lesbarkeit aufweisen.

Interessanterweise eignen sich sowohl die Häufigkeit von 'Number-Tokens' im Allgemeinen als auch jene von erkannten Telefon- oder Faxnummern sehr gut um öffentliche von privaten Elementen unterscheiden zu können.

Auf der Präsentationsebene zeigen sich Tabellen-Tags als verlässlichstes Merkmal zur Klassifikation. Dieser Umstand ist durch die häufige Verwendung von Tabellen zur Strukturierung von Kommentaren bzw. User-Einträgen zu erklären. Von daher ist die relative Häufigkeit von Tags, welche zur Erstellung von Tabellen dienen, in privaten Elementen häufig signifikant höher als in öffentlichen Elementen.

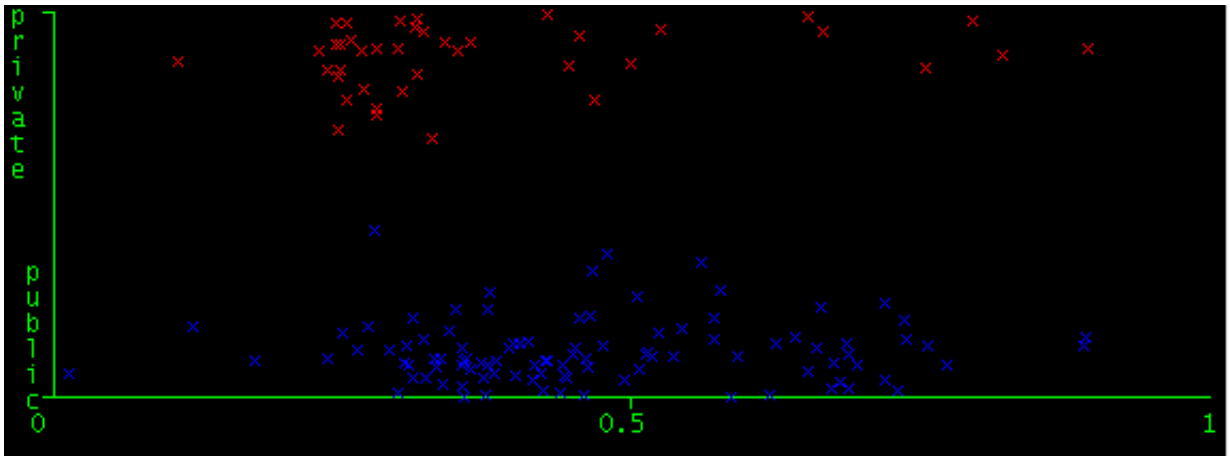

Abbildung 4.11: Dimensionsreduktion - Merkmal Relative Häufigkeit von Nomen.

4.6 Zusammenfassung

In diesem Abschnitt erfolgte ein grober Überblick über die Implementierung eines Systems, welches in der Lage ist auf Dokumenten- und Absatzebene private von

Rang	Merkmal
1	Personalpronomen (Relative Häufigkeit)
2	Präpositionen (Relative Häufigkeit)
4	Number-Tokens (Relative Häufigkeit)
6	Gesellschafts-Namen Lookups (Relative Häufigkeit)
6.5	Telefon- und Faxnummern Lookups (Relative Häufigkeit)
6.5	Durchschnittliche Wortlänge
7	Flesch Reading Ease - Lesbarkeitsindex
7.5	Adverben (Relative Häufigkeit)
9.5	Word-Tokens (Relative Häufigkeit)
10	Tabellen-Tags (Relative Häufigkeit)
10.5	Konjunktionen (Relative Häufigkeit)
15	Durchschnittliche Silbenanzahl pro Wort
15	Artikel (Relative Häufigkeit)
15.5	Punctuation-Tokens (Relative Häufigkeit)
16.5	Zeit Datum Lookups (Relative Häufigkeit)
17.5	Textformattierungs-Tags (Relative Häufigkeit)
18.5	Control-Tokens (Relative Häufigkeit)
19	Adressen Lookups (Relative Häufigkeit)
20	Graphik-Tags (Relative Häufigkeit)
21.5	Link-Tags (Relative Häufigkeit)
22	Absatz-Tags (Relative Häufigkeit)
22	Durchschnittliche Wortanzahl pro Satz
22.5	Pronomen (Relative Häufigkeit)
22.5	Nomen (Relative Häufigkeit)
24	Präpositionen (Relative Häufigkeit)
26	Flesch-Kincaid Grade Level - Lesbarkeitsindex
26	Adjektive (Relative Häufigkeit)
27.5	Listenelemente (Relative Häufigkeit)
28	Orte Lookups (Relative Häufigkeit)
28.5	Listen-Tags (Relative Häufigkeit)
28.5	Verben (Relative Häufigkeit)
29	Formular-Tags (Relative Häufigkeit)
32	Modalverben (Relative Häufigkeit)
32.5	Space-Tokens (Relative Häufigkeit)
35	Überschriften-Tags (Relative Häufigkeit)
35.5	Ausrufe (Relative Häufigkeit)
35.5	Personen-Namen Lookups (Relative Häufigkeit)
38	Währungen-Lookups (Relative Häufigkeit)
38	Symbol-Tokens (Relative Häufigkeit)
39.5	Frame-Tags (Relative Häufigkeit)
40	Multimedia-Tags (Relative Häufigkeit)
40.5	Abkürzungen Lookups (Relative Häufigkeit)

Tabelle 4.10: Absteigend sortierte Auflistung der Ergebnisse der SVM - Dimensionsreduktion.

öffentlichen Elementen eines Web-Dokuments zu unterscheiden. Die Umsetzung er-

Evaluator	weka.attributeSelection.InfoGainAttributeEval
Search	weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Tabelle 4.11: WEKA Parameter zur Dimensionsreduktion des k-NN Klassifikators (Feature Ranking Ansatz).

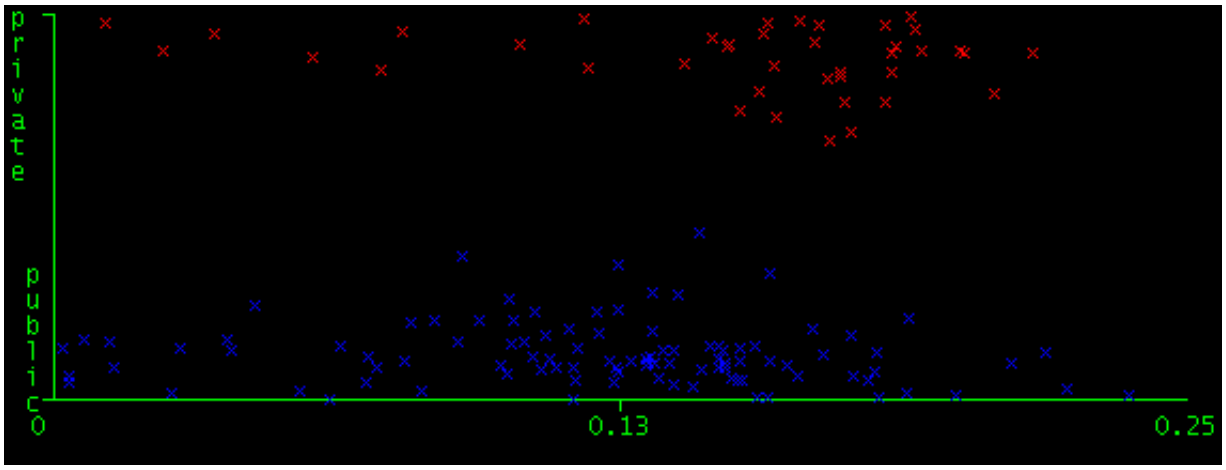

Abbildung 4.12: Dimensionsreduktion - Merkmal Relative Häufigkeit von Verben.

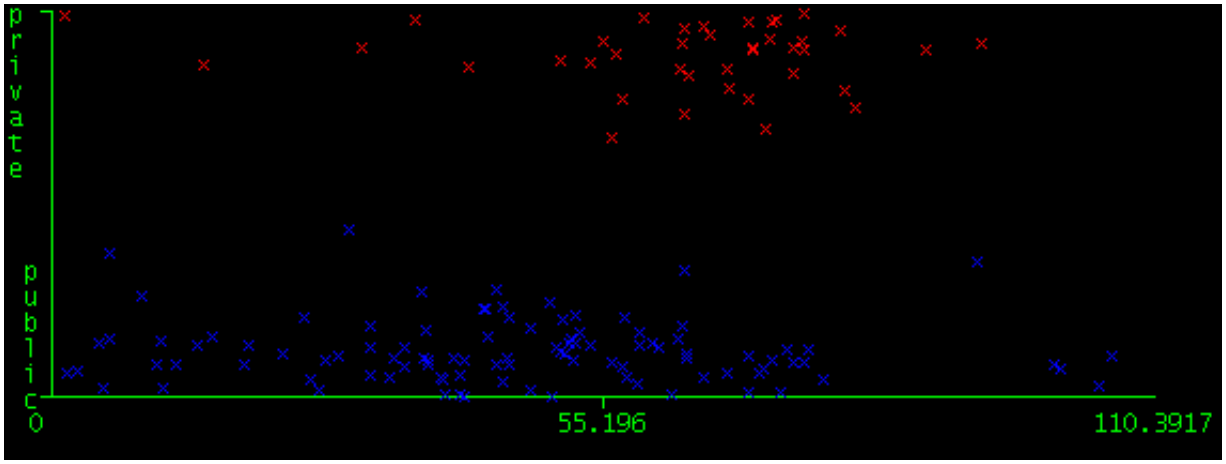

Abbildung 4.13: Dimensionsreduktion - Merkmal Flesch Lesbarkeitsindex.

folgte auf Basis von Gate (Abschnitt 4.1.1), einem Framework zur Verarbeitung von natürlicher Sprache und WEKA (Abschnitt 4.1.2), einer Implementierung von unterschiedlichen Machine Learning Algorithmen. Die grundsätzliche Funktionsweise dieses Systems ist jene, dass zuerst eine Trennung von Absätzen erfolgt, welche anschließend klassifiziert werden.

Ein implementierter Ansatz zur Trennung von Absätzen versucht hierbei wiederkehrende, aus HTML Tags bestehende Muster zu erkennen. Diese Muster werden hierbei als Indikatoren für die Trennung von Absätzen verwendet. Ein weiterer implementierter Ansatz basiert auf der Idee, Genre-Wechsel auf Basis von syntaktischen Mustern zu erkennen. Insbesondere der Wechsel zwischen Nominal- und Verbalstil hat sich hierbei als verlässliches Muster bewährt.

Rang	Merkmal
0	Personalpronomen (Relative Häufigkeit)
0	Nomen (Relative Häufigkeit)
0	Number-Tokens (Relative Häufigkeit)
0	Verben (Relative Häufigkeit)
0	Flesch Reading Ease - Lesbarkeitsindex
0	Adverben (Relative Häufigkeit)
0	Word-Tokens (Relative Häufigkeit)
0	Textformattierungs-Tags (Relative Häufigkeit)
0	Konjunktionen (Relative Häufigkeit)
0	Pronomen (Relative Häufigkeit)
0	Artikel (Relative Häufigkeit)
0	Telefon- und Faxnummern Lookups (Relative Häufigkeit)
0	Durchschnittliche Wortlänge
0.0556	Punctuation-Tokens (Relative Häufigkeit)
0.0719	Flesch-Kincaid Grade Level - Lesbarkeitsindex
0.0802	Tabellen-Tags (Relative Häufigkeit)
0.0887	Control-Tokens (Relative Häufigkeit)
0.09	Adressen Lookups (Relative Häufigkeit)
0.0931	Graphik-Tags (Relative Häufigkeit)
0.0946	Link-Tags (Relative Häufigkeit)
0.0957	Gesellschafts-Namen Lookups (Relative Häufigkeit)
0.1181	Durchschnittliche Wortanzahl pro Satz
0.1365	Absatz-Tags (Relative Häufigkeit)
0.1421	Präpositionen (Relative Häufigkeit)
0.1434	Präpositionen (Relative Häufigkeit)
0.1556	Durchschnittliche Silbenanzahl pro Wort
0.1582	Adjektive (Relative Häufigkeit)
0.1674	Listenelemente (Relative Häufigkeit)
0.1755	Orte Lookups (Relative Häufigkeit)
0.1885	Listen-Tags (Relative Häufigkeit)
0.1906	Zeit Datum Lookups (Relative Häufigkeit)
0.2025	Formular-Tags (Relative Häufigkeit)
0.2025	Modalverben (Relative Häufigkeit)
0.2191	Space-Tokens (Relative Häufigkeit)
0.2240	Überschriften-Tags (Relative Häufigkeit)
0.2385	Ausrufe (Relative Häufigkeit)
0.2412	Personen-Namen Lookups (Relative Häufigkeit)
0.2502	Währungen-Lookups (Relative Häufigkeit)
0.3113	Symbol-Tokens (Relative Häufigkeit)
0.3921	Frame-Tags (Relative Häufigkeit)
0.4297	Multimedia-Tags (Relative Häufigkeit)
0.4670	Abkürzungen Lookups (Relative Häufigkeit)

Tabelle 4.12: Absteigend sortierte Auflistung der Ergebnisse der k-NN - Dimensionsreduktion.

5. Evaluierung

Dieser Abschnitt behandelt die durchgeführte Evaluierung. Nachfolgend werden die Maße zur Beurteilung sowie die verwendeten Korpora vorgestellt. Weiters wird die Vorgehensweise der Evaluierung beschrieben und die erzielten Ergebnisse dargestellt. Abschließend folgt eine Interpretation der ermittelten Ergebnisse. Die Evaluierung wurde sowohl auf Dokumenten- als auch auf Absatzebene durchgeführt. Das Hauptaugenmerk dieser Arbeit liegt auf der Klassifizierung auf Absatzebene.

5.1 Verwendete Maße zur Beurteilung

Recall (auch Sensitivität), Precision (auch Relevanz) und das F-Maß sind die gebräuchlichsten Maße zur Beurteilung der Güte eines Information Retrieval Systems. Alle drei Maße können Werte zwischen Null und Eins (beziehungsweise 0% bis 100%) annehmen und hängen voneinander ab, weshalb sie auch gemeinsam betrachtet werden sollten.

Der Recall ist das Maß für die Vollständigkeit des Suchergebnisses. Auf Dokumentenebene beschreibt er den Anteil der gefundenen privaten Dokumente im Verhältnis zu deren Gesamtanzahl. Auf Absatzebene beschreibt er das Verhältnis zwischen der Anzahl der gefundenen privaten Elemente im Verhältnis zur Anzahl aller vorhandenen privaten Elemente. Die Messung erfolgt dabei auf Zeichenbasis.

Die Precision beschreibt die Genauigkeit eines Suchergebnisses und beschreibt die Fähigkeit, nicht relevante Elemente auszuscheiden. Auf Dokumentenebene beschreibt sie das Verhältnis der Anzahl an gefundenen privaten Dokumente zur Gesamtanzahl aller gefundenen Dokumente. Auf Absatzebene beschreibt sie das Verhältnis zwischen der Anzahl an gefundenen privaten Elemente im Verhältnis zur Gesamtanzahl aller vom System als privat klassifizierten Elemente. Auch bei der Precision erfolgt die Messung dabei auf Zeichenbasis.

Sinnvoll ist jedoch nur eine Betrachtung beider Maße, da der Recall sich leicht auf das Maximum 1 setzen lässt indem die vollständige Menge als Suchergebnis verwendet wird. In diesem Fall wäre die Precision jedoch wiederum sehr niedrig. Umgekehrt lässt sich die Precision nahezu beliebig erhöhen wenn nur sehr wenige Elemente in die Ergebnismenge übernommen werden.

Das F-Maß kombiniert Precision und Recall mittels des gewichteten harmonischen Mittels.

5.2 Verwendete Testkorpora

Die verwendeten Testkorpora sind in der Tabelle 5.1 aufgelistet.

Korpus	Beschreibung	Größe (Dokumente)	Größe (Zeichen)
7 web genre collection	Web-Dokumente 7 verschiedener Genres	1.400	
amazon.com Top 100 Books of 2007	Produktbeschreibungen mit Kundenrezensionen	50	863.422
amazon.com Video Games	Produktbeschreibungen mit Kundenrezensionen	20	370.365
Daily Mail	Zeitungsartikel mit Leserkommentaren	40	458.181
The Times	Zeitungsartikel mit Leserkommentaren	40	481.130

Tabelle 5.1: Auflistung der Testkorpora.

5.2.1 7 web genre collection

Die Evaluierung auf Dokumenten - Ebene wurde anhand des ”7 web genre collection” Korpus von *Santini* durchgeführt [29]. Der Korpus besteht aus 7 Genres mit jeweils 200 Web-Dokumenten. Die 7 Genres sind in der Folge kurz beschrieben. Nachfolgend ist ein Beispiel eines Web-Dokuments aus der Kategorie ”Personal Homepage” (Abbildung 5.1) sowie aus der Kategorie ”Listings” (Abbildung 5.2) abgebildet.

Genre Blog (Klasse privat)

Blogs (Kurzform für ”Web Logs”) sind meist lange Texte welche in tägliche Berichte unterteilt sind, ähnlich den Einträgen eines Tagebuchs. Blogs werden üblicherweise dazu verwendet um die Gedanken des Autors nieder zu schreiben oder aber Erlebnisse zu schildern. Meist besteht auch die Möglichkeit dass Besucher Kommentare zu einzelnen Berichten verfassen können.

Genre Personal Homepage (Klasse privat)

Die in diesem Genre enthaltenen Elemente sind definiert als Web-Dokumente welche von einer natürlichen Person erstellt und gewartet werden. Meist dient eine derartige Website zur Selbstdarstellung und umfasst persönliche Daten sowie Informationen über Interessengebiete der jeweiligen Person. Häufiger Bestandteil einer "persönlichen Homepage" ist auch ein Fotoalbum.

Genre E-Shop (Klasse öffentlich)

Web-Dokumente von E-Shops sind meist sehr interaktive Dokumente. Ein Charakteristikum ist ein hoher Anteil an Programmlogik, meist in Form von Skripten. Zudem ist ein E-Shop häufig als Liste von Produkten mit Preisen organisiert. Produktseiten weisen zudem häufig eine nahezu identische Struktur, geprägt von Produktabbildungen sowie kurzen Beschreibungen und Preisinformationen, auf. Häufig enthalten die verwendeten Texte auch stilistisch ähnliche Floskeln mit dem Ziel Besucher zum Kauf zu bewegen

Genre FAQ (Klasse öffentlich)

FAQs (Kurzform für "Frequently Asked Questions") haben große Ähnlichkeit mit dem Abschnitt "Problembeseitigung" in Handbüchern. FAQs existieren im Web allerdings nicht nur für technische Betriebsanleitungen. Nahezu jede Website hat einen FAQ-Bereich, in dem häufig auftretende Fragen samt Antworten aufgelistet sind. In der Strukturierung sind jedoch deutliche Unterschiede erkennbar. Beispielsweise kann ein FAQ ein einzelnes Dokument mit mehreren Fragen und unmittelbar darauf folgenden Antworten sein. Eine weitere Möglichkeit ist es jede Frage mit der Antwort in einem eigenen Dokument zu platzieren. Eine weitere gebräuchliche Art der Strukturierung ist ein zentrales Dokument mit aufgelisteten Fragen, welche wiederum als Link zu den einzelnen Antwort-Dokumenten modelliert werden[29].

Genre Listings (Klasse öffentlich)

Eine Liste ist eine Form von Verzeichnisstruktur zur Übermittlung von Informationen oder Instruktionen. Charakteristisch sind die Aufzählungszeichen in Form von aufsteigenden Nummern oder graphischen Symbolen. Beispiele für Vertreter der Kategorie Listings sind Sitemaps[1], Inhaltsverzeichnisse oder Checklisten.

Genre Newspaper Frontpage (Klasse öffentlich)

Dieses Genre enthält Startseiten von Onlinepräsenzen verschiedener Zeitungen. Charakteristisch ist das Vorhandensein von Inhalten in Form von mehreren Artikeln welche das Gegenstück zur Titelseite der Papier-Variante darstellt. Weiters enthält die Startseite meist eine mehr oder wenige komplexe Navigationsmöglichkeit.

[1] Eine Sitemap ist die hierarchisch strukturierte Darstellung aller einzelnen Dokumente einer Website.

Search Page (Klasse öffentlich)

Charakteristisch für Suchmasken ist üblicherweise die kurze textuelle Beschreibung sowie ein zentrales Eingabefeld für die Abfrage. Je nach Komplexität der zu bedienenden Suchmaschine sind auch häufig zahlreiche weitere Eingabefelder enthalten.

5.2.2 amazon.com Top 100 Books of 2007

Dieser Korpus enthält 50 Produktseiten der Kategorie "Top 100 Books of 2007" von amazon.com und dient zur Evaluierung auf Absatz-Ebene zumal Produktseiten auch Rezensionen von Kunden enthalten. Diese von Amazon erstellte Liste enthält vorwiegend Romane, vereinzelt sind jedoch auch Biographien zu finden. Für den Test Korpus wurden die ersten 50 Dokumente herangezogen (Tabelle A.1). Nachfolgend ist ein Beispiel einer Produktseite (Abbildung 5.3) sowie einer zugehörigen Kundenrezension (Abbildung 5.4) abgebildet.

Dieser Korpus besitzt die Eigenheit, dass überdurchschnittlich viele professionelle Rezensionen enthalten sind, welche einerseits kaum von der eigentlichen Produktbeschreibung unterschieden werden können und zum Anderen auch keine privaten Elemente darstellen. Daher steht dieser Korpus in zwei Varianten zur Verfügung. Die erste Variante behandelt alle Rezensionen als private Elemente, während bei der zweiten hingegen eine Fallunterscheidung stattfand und einige Rezensionen als öffentlich annotiert wurden zumal sie in die Kategorie "Professionelle Rezension" fallen und nicht als Rezension eines Normallesers betrachtet werden können.

Die Annotierung der privaten und öffentlichen Elemente der einzelnen Web-Dokumente erfolgte manuell. Um möglichst einfach zu parsende Seiten zu erhalten, erfolgte der Download per Identifikation als GoogleBot[2].

5.2.3 amazon.com Video Games

Dieser Korpus enthält 20 Produktseiten der Kategorie "Video Games" von amazon.com und dient als weiterer Testkorpus zur Evaluierung auf Absatz-Ebene. Hierbei wurden von den Bestsellern 2007 die jeweils 10 meistverkauften Spiele der Plattformen 'Playstation 3' und 'XBox 360' ausgewählt (Tabelle A.2). Die Kategorie "Video Games" wurde insbesondere deshalb gewählt weil die Sprache der Rezensionen sich teils doch sehr deutlich von jenen des "amazon.com Top 100 Books of 2007"-Korpus unterscheidet.

Die Annotierung der privaten und öffentlichen Elemente der einzelnen Web-Dokumente erfolgte manuell. Um möglichst einfach zu parsende Seiten zu erhalten, erfolgte der Download per Identifikation als GoogleBot.

[2] Googlebot ist der Webcrawler der Suchmaschine Google. Dabei handelt es sich um ein Programm welches Web-Dokumente herunterlädt und diese für die Websuche von Google indiziert.

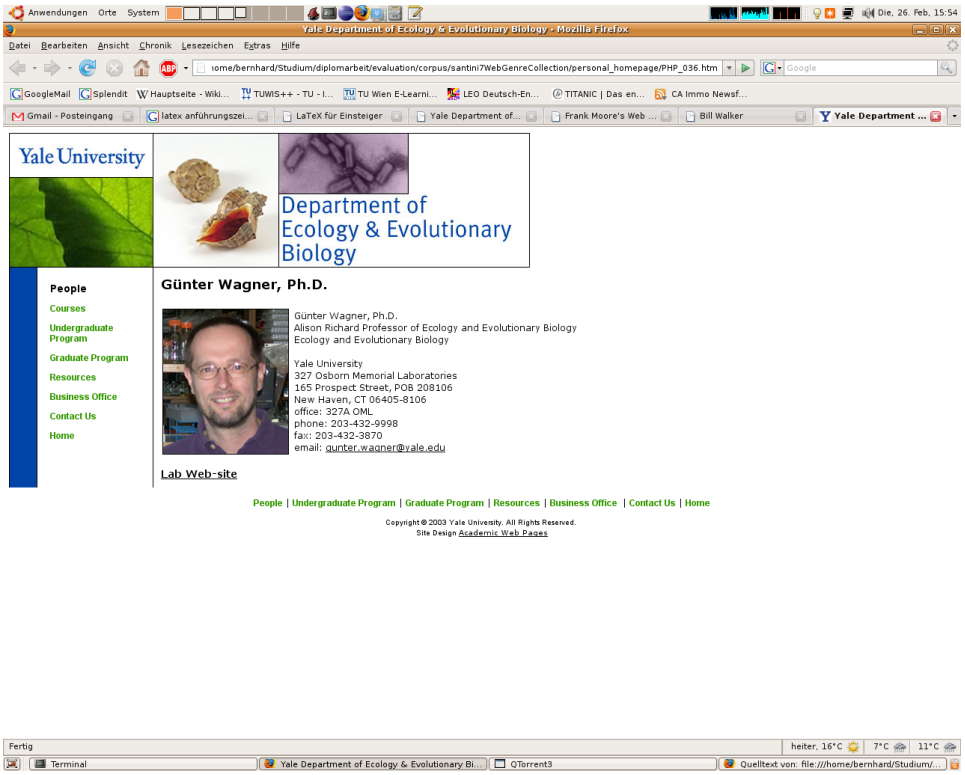

Abbildung 5.1: ”7 web genre collection” - Beispiel eines Web-Dokuments aus dem Genre ”Personal Homepage”.

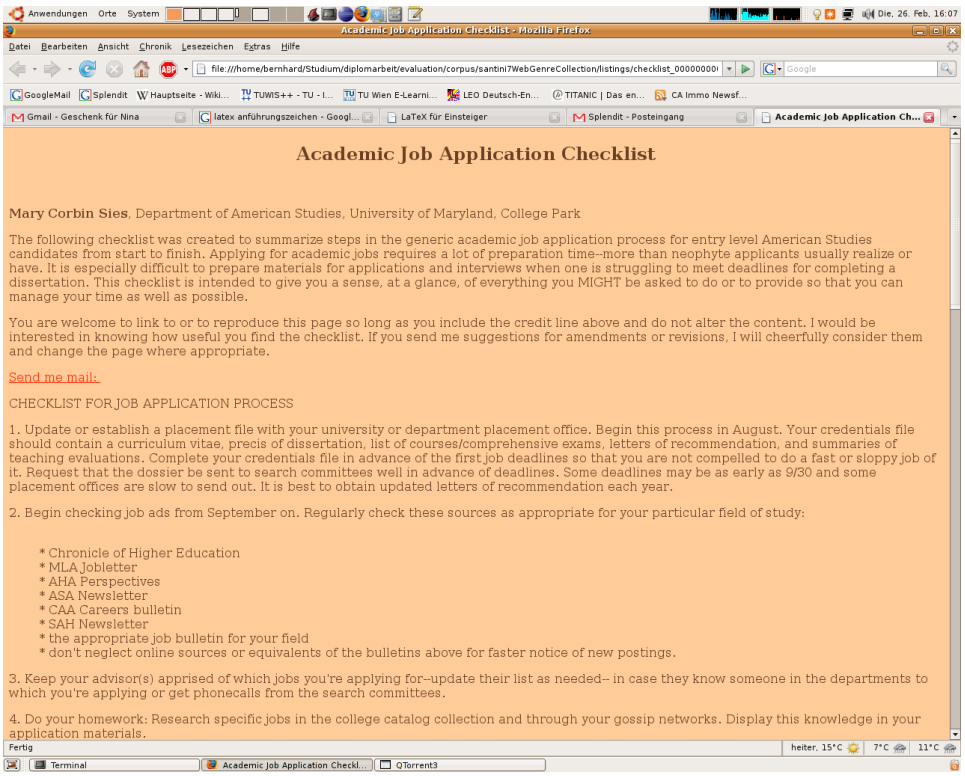

Abbildung 5.2: ”7 web genre collection” - Beispiel eines Web-Dokuments aus dem Genre ”Listings”.

Abbildung 5.3: ”amazon.com Top 100 Books of 2007” - Beispiel einer Produktseite.

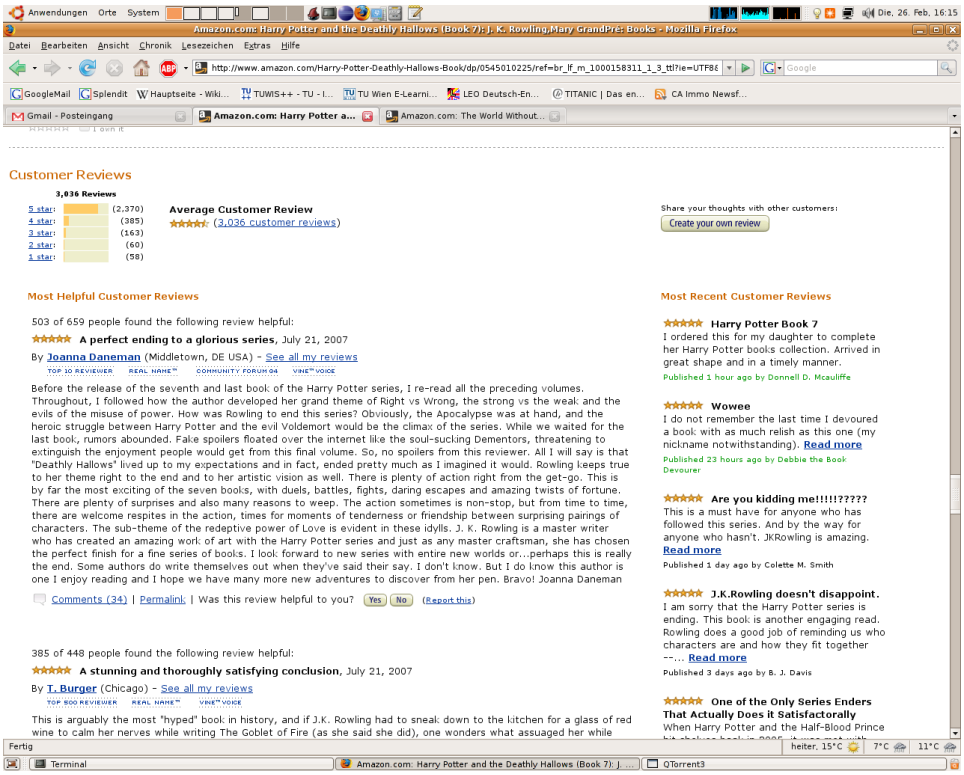

Abbildung 5.4: ”amazon.com Top 100 Books of 2007” - Kundenrezensionen einer Produktseite.

5.2.4 Daily Mail

Die Daily Mail ist eine britische Zeitung mit einer Auflage von 2 Millionen Druckexemplaren. Dieser Korpus enthält 40 Artikel bzw. Kurzmeldungen der Onlinepräsenz der Daily Mail welche über einen Zeitraum von einer Woche aus der Kategorie "Headlines" zusammengetragen wurden (Tabelle A.3). Nachfolgend ist ein Beispiel eines Artikels (Abbildung 5.5) sowie eines zugehörigen Leserkommentars (Abbildung 5.6) abgebildet.

Die Annotierung der privaten und öffentlichen Elemente der einzelnen Web-Dokumente erfolgte manuell. Um möglichst einfach zu parsende Seiten zu erhalten, erfolgte der Download per Identifikation als GoogleBot.

5.2.5 The Times

The Times ist eine nationale Tageszeitung in Großbritannien. Dieser Korpus enthält 40 Artikel der Onlinepräsenz der Times welche über einen Zeitraum von einer Woche aus der Kategorie "News" zusammengetragen wurden (Tabelle A.4).

Die Annotierung der privaten und öffentlichen Elemente der einzelnen Web-Dokumente erfolgte manuell. Um möglichst einfach zu parsende Seiten zu erhalten, erfolgte der Download per Identifikation als GoogleBot.

5.3 Evaluierung auf Dokumentenebene

5.3.1 Durchführung

Die Evaluierung auf Dokumentenebene wurde anhand des "7 web genre collection" Korpus durchgeführt. Die Evaluierung erfolgte mittels einer 10-fachen Kreuzvalidierung[3]

Klassifikator

Bei der Evaluierung mittels des naiven Bayes Klassifikator (Abschnitt 3.3.2.3) erwies sich die Variante mit Diskretisierung als optimale. Die verwendeten Parameter sind in der Tabelle 5.2 abgebildet. Während nominale Attribute eine festgelegte Wertemenge besitzen, haben numerische Attribute keine explizit festgelegte Wertemenge. Beim Vorgang der Diskretisierung werden einzelne numerische Werte zu einem Intervall zusammengefasst. Ein naiver Bayes Klassifikator arbeitet häufig besser, wenn numerische Attribute diskretisiert werden.

[3] Die Kreuzvalidierung ist ein Testverfahren zur Messung der Güte. Bei der k-fachen Kreuzvalidierung wird eine Datenmenge in k Teilmengen aufgeteilt. Nun werden k Durchläufe gestartet wobei die jeweils k-te Teilmenge als Testmenge verwendet wird und die verbleibenden k-1 Teilmengen als Trainingsmengen fungieren.

Abbildung 5.5: ”Daily Mail” - Beispiel eines Artikels.

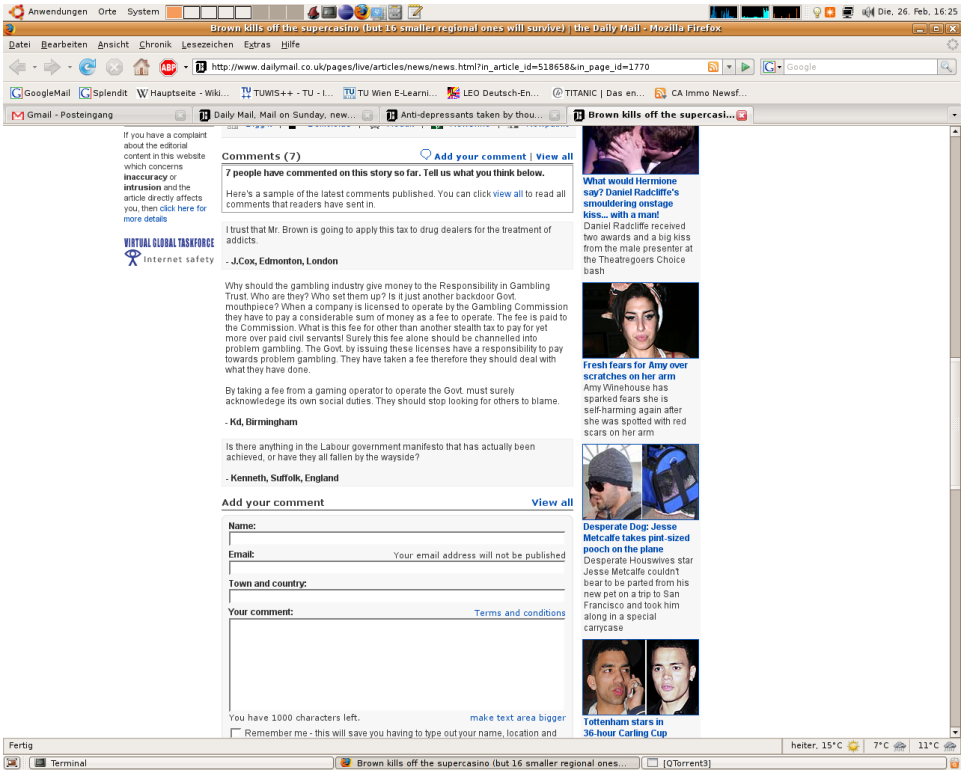

Abbildung 5.6: ”Daily Mail” - Kommentare zu einem Artikel.

Naiver Bayes Kl.	weka.classifiers.bayes.NaiveBayes -D
k-Nearest Neighbor Kl.	weka.classifiers.lazy.IBk -K 7 -W 0
SVM Klassifikator (Polynomineller Kernel)	weka.classifiers.functions.SMO -C 30.0 -E 1.0 -G 0.01 -A 250007 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
SVM Klassifikator (RBF - Kernel)	weka.classifiers.functions.SMO -C 80.0 -E 1.0 -G 0.01 -A 250007 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1

Tabelle 5.2: Bei der Evaluierung verwendete WEKA Parameter für die Klassifikatoren.

Für die Evaluierung mittels des k-Nearest Neighbor Klassifikator (Abschnitt 3.3.2.1) wurden k - Parameterwerte zwischen 1 und 20 untersucht. Der Parameter k definiert hierbei die Anzahl der k nächsten Nachbarn welche für die Klassenzuordnung verwendet werden. Die besten Ergebnisse konnten hierbei mit einem Parameter k von 7 erzielt werden. Die verwendeten Parameter sind in der Tabelle 5.2 abgebildet.

Für die Evaluierung mittels des SVM - Klassifikator (Abschnitt 3.3.2.2) wurden ein RBF Kernel sowie ein polynomineller Kernel untersucht. Für die beiden Kernel wurden unterschiedliche Kostenparameter C mit Werten zwischen 1 und 100 versucht. Der Kostenparameter C definiert hierbei die Ausgeglichenheit zwischen dem Erlauben von Fehlern während des Trainings und dem Erzwingen von starren Grenzen. Im Optimalfall soll somit ein leicht nachgiebige Grenze erzeugt werden welche einige Fehler beim Trainieren erlaubt um somit einen möglichst allgemein gültigen Klassifikator zu erstellen. Der Kostenparameter C ist definiert als Kosten für einen falsch klassifizierten Punkt im Vektorraum und kann bei falscher Wahl (bei einem zu hohen Wert) in einem sogenannten 'Overfitting' [4] resultieren.

Bei der Verwendung eines polynominellen Kernels wurde das optimale Resultat bei einem relativ niedrigen Kostenparameter von 30 erzielt, während hingegen bei der Verwendung des RBF - Kernels ein höherer Kostenparameter von 80 die besten Ergebnisse erzielte. Mittels des polynominellen Kernels konnten geringfügig bessere Ergebnisse erzielt werden, weshalb dieser zum Einsatz kam. Die optimalen Parameter für beide Kernel sind in der Tabelle 5.2 dargestellt.

5.3.2 Ergebnisse

In der Tabelle 5.3 sind die Gesamtergebnisse abgebildet. Diese Tabelle stellt zugleich den Bezug zur eigentlichen Aufgabenstellung her zumal hier die Recall- und Precision Werte für private Dokumente gezeigt werden. Unter Verwendung des naiven Bayes-Klassifikator konnten hierbei 85,75 % aller privaten Dokumente erkannt werden. 73,45 % aller als privat erkannten Dokumente waren auch tatsächlich privat. Mittels

[4] Als Overfitting wird das Phänomen bezeichnet dass ein Klassifikator derart trainiert wird, dass er zu abhängig von charakteristischen Merkmalen der Trainingsdaten ist und bei neuen unbekannten Daten wesentlich schlechtere Ergebnisse liefert [31]

des Support Vector Machine-Klassifikator konnten ähnlich gute Werte erzielt werden. Hierbei konnte ein Recall von 81,50 % mit einer Precision von 71,96 % erzielt werden. Unter Verwendung des k-Nearest Neighbour - Klassifikator wurde ein Recall von 78 % mit einer Precision von 67,09 % erzielt.

In den Tabellen 5.4, 5.5 und 5.6 sind die entsprechenden Gesamt - Konfusionsmatrizen dargestellt. Bei einer Korpusgröße von 1.400 Dokumenten konnten mittels des naiven Bayes-Klassifikator 343 von 400 privaten Dokumenten erkannt werden, 124 von 1.000 öffentlichen Dokumenten wurden hingegen fälschlicherweise als privat eingestuft. Unter Verwendung des k-Nearest Neighbour - Klassifikator konnten 312 von 400 privaten Dokumenten erkannt werden, von 1.000 öffentlichen Dokumenten wurden hingegen 153 als privat klassifiziert. 326 von 400 privaten Dokumenten konnten auf Basis des Support Vector Machine-Klassifikator erkannt werden, während zugleich 127 öffentliche Dokumente fälschlicherweise als privat eingestuft wurden.

Weiters sind in den Tabellen 5.4, 5.5 und 5.6 auch Konfusionsmatrizen der einzelnen privaten Genres ersichtlich. Da der Korpus auf fünf öffentlichen und zwei privaten Genres besteht beträgt die Gesamtanzahl an Dokumenten hierbei jeweils 1.200. Durch diese Konfusionsmatrizen werden die Einzelergebnisse der Genres ersichtlich. Unter Verwendung des naiven Bayes-Klassifikator konnten für das Genre 'Blog' 192 von 200 Dokumenten korrekt als private Dokumente klassifiziert werden, dies entspricht einem Recall von 96 %. Für das Genre 'Personal Homepage' konnten hingegen nur 151 von 200 privaten Dokumenten korrekt erkannt werden, dies enstpricht einem Recall von 75,5 %.

Klassifikator	Recall	Precision	F-Maß
NB	**85,75 %**	**73,45 %**	**79,06 %**
k-NN	78,00 %	67,09 %	72,13 %
SVM	81,50 %	71,96 %	76,43 %

Tabelle 5.3: Evaluierung auf Dokumentenebene - Gesamtergebnisse.

		Tatsächliche Klasse	
		Privat	Öffentlich
Gesamt	Privat	343	124
	Öffentlich	57	876
Genre 'Blog'	Privat	192	124
	Öffentlich	8	876
Genre 'Personal Homepage'	Privat	151	124
	Öffentlich	49	876

Tabelle 5.4: Evaluierung auf Dokumentenebene - Konfusionsmatrix (Naiver Bayes-Klassifikator).

		Tatsächliche Klasse	
		Privat	Öffentlich
Gesamt	Privat	312	153
	Öffentlich	88	847
Genre 'Blog'	Privat	182	153
	Öffentlich	18	847
Genre 'Personal Homepage'	Privat	130	153
	Öffentlich	70	847

Tabelle 5.5: Evaluierung auf Dokumentenebene - Konfusionsmatrix (k-Nearest Neighbor-Klassifikator).

		Tatsächliche Klasse	
		Privat	Öffentlich
Gesamt	Privat	326	127
	Öffentlich	74	873
Genre 'Blog'	Privat	187	127
	Öffentlich	13	873
Genre 'Personal Homepage'	Privat	139	127
	Öffentlich	61	873

Tabelle 5.6: Evaluierung auf Dokumentenebene - Konfusionsmatrix (Support Vector Machine-Klassifikator).

5.3.3 Interpretation der Ergebnisse

Unter Verwendung des naiven Bayes-Klassifikator konnten 85,75 % aller privaten Dokumente korrekt erkannt werden. 73,45 % aller als privat klassifizierten Dokumente waren hierbei auch tatsächlich privat. Während das Genre 'Blog' mit sehr guten Werten klassifiziert werden konnte, zeigen sich jedoch beim Grenre 'Personal Homepage' Schwächen. Der Recall Wert beträgt hier lediglich 75,5 %. Diese Schwäche kann allerdings etwas relativiert werden zumal dieses Genre relativ "vielschichtig" ist und relativ viele Sub-Genres vereint. Bei der Evaluierung von *Santini* zeigt sich ein ähnliches Bild, auch dort sinkt die Erkennungsrate bei diesem Genre auf 79 %, während für die restlichen Korpora Werte im Bereich von 90 % erreicht werden [29].

Für die im Vergleich zu *Santini* doch spürbar schlechteren Werte kommen mehrere Faktoren in Frage. Einerseits basiert die Implementierung von *Santini* auf einem NLP - Parser der Firma Connexor [5] weshalb auch Merkmale verwendet werden, welche in Gate nicht umgesetzt sind. Ein Beispiel hierfür ist die Erkennung von Aktiv- oder Passivsätzen. Auch kann der verwendete HTML - Parser eine sehr entscheidende Rolle spielen. Gerade bei komplexen Web-Dokumenten mit einem hohen Anteil an Skripten kann ein Fehler beim Parsen gravierenden Einfluss auf die Anordnung der Elemente haben. Unter Verwendung einer anderen (höher entwickelten) HTML

[5] Connexor, http://www.connexor.eu/

Rendering Engine könnten je nach Schwierigkeit der zu parsenden Dokumente bessere Ergebnisse erzielt werden. Eine mögliche Alternative wäre hierbei die Verwendung der 'Gecko Rendering Engine' welche von zahlreichen Internet-Browsern wie beispielsweise 'Mozilla Firefox' verwendet wird.

Weiters muß darauf hingewiesen werden dass *Santini* POS-Trigram Merkmale verwendet [6] welche nur bedingt Korpus - unabhängig sind [29].

5.4 Evaluierung auf Absatzebene

5.4.1 Durchführung

Die Evaluierung auf Absatzebene wurde auf Basis der Korpora "amazon.com Top 100 Books of 2007", "amazon.com Video Games", "Daily Mail" und "The Times" durchgeführt. Die zuvor manuell durchgeführte Annotierung von öffentlichen und privaten Elementen stellt die Voraussetzung für eine Evaluierung auf Absatzebene dar.

Um möglichst repräsentative Ergebnisse für die Evaluierung auf Absatzebene zu erreichen wurde die Evaluierung mit dem selben Klassifikator wie zuvor auf Dokumentenebene durchgeführt (Abschnitt 5.3.1). Das heißt, dem verwendeten Klassifikator wurden keine Testdaten aus einem der für die Evaluierung auf Absatzebene verwendeten Korpora zur Verfügung gestellt.

5.4.2 Ergebnisse

Bedingt durch den Umstand, dass die Trennung der Absätze nicht immer korrekt erfolgt, wurden alle Ergebnisse auf Basis der Zeichenanzahl gemessen. Eine möglichst korrekte Trennung der Absätze stellt eine zwingende Voraussetzung für eine korrekte Klassifizierung auf Absatzebene dar. Daher wurden sämtliche Messungen mit allen implementierten Modulen zur Trennung von Absätzen durchgeführt.

Die Maße zur Beurteilung des Klassifikators auf Absatzebene wurden sowohl unter Verwendung des Moduls zur Absatz-Trennung auf Basis von HTML Mustern (Tabellen 5.7 & 5.8) durchgeführt als auch unter Verwendung der Absatz-Trennung auf Basis von Textstatistik-Mustern(Tabellen 5.9 & 5.10) ermittelt. Weiters stehen die Ergebnisse als Gesamtergebnisse sowie als Einzelergebnisse auf Basis der einzelnen Korpora zur Verfügung. Sämtliche Ergebnisse wurden zudem mittels der Implementierung eines naiven Bayes - Klassifikators (Abschnitt 3.3.2.3), jener eines k-Nearest Neighbor-Klassifikators (Abschnitt 3.3.2.1) und jener eines Support Vector Machine-Klassifikators (Abschnitt 3.3.2.2) ermittelt. Hierbei wurden dieselben

[6] Unter einem POS-Trigram versteht mann hierbei eine Zusammensetzung von drei aufeinander folgenden Part-Of-Speech (POS) Tags welche genre-typische Sprachmuster abbilden.

Parameter verwendet wie bei der Evaluierung auf Dokumentenebene. Die Parameter sind in der Tabelle 5.2 abgebildet.

Die Messung für den Korpus "amazon.com Top 100 Books of 2007" wurde für beide Varianten durchgeführt. Bei der mit einem Stern markierten Variante wurden professionelle Rezensionen als öffentlich annotiert (Abschnitt 5.2.2).

Klassifikator	Recall	Precision	F-Maß
NB	**91,42 %**	**79,82 %**	**85,15 %**
k-NN	88,34 %	78,16 %	82,87 %
SVM	90,30 %	79,39 %	84,41 %

Tabelle 5.7: Evaluierung auf Absatzebene - Gesamtergebnisse (Absatz-Trennung auf Basis von HTML Mustern).

Klassifikator	Korpus	Recall	Precision	F-Maß
NB	**amazon.com Top 100 Books of 2007**	**84,06 %**	**77,91 %**	**80,87 %**
NB	**amazon.com Top 100 Books of 2007***	**85,75 %**	**78,27 %**	**81,84 %**
NB	**amazon.com Video Games**	**96,02 %**	**87,66 %**	**91,65 %**
NB	**Daily Mail**	**98,55 %**	**77,87 %**	**87,00 %**
NB	**The Times**	**92,75 %**	**77,41 %**	**84,39 %**
k-NN	amazon.com Top 100 Books of 2007	82,67 %	76,58 %	79,51 %
k-NN	amazon.com Top 100 Books of 2007*	84,60 %	77,09 %	80,67 %
k-NN	amazon.com Video Games	92,52 %	86,68 %	89,51 %
k-NN	Daily Mail	91,95 %	74,58 %	82,36 %
k-NN	The Times	89,98 %	75,86 %	82,32 %
SVM	amazon.com Top 100 Books of 2007	83,74 %	77,61 %	80,56 %
SVM	amazon.com Top 100 Books of 2007*	85,26 %	78,05 %	81,50 %
SVM	amazon.com Video Games	93,91 %	87,11 %	90,38 %
SVM	Daily Mail	97,55 %	77,51 %	86,38 %
SVM	The Times	90,95 %	76,37 %	83,02 %

Tabelle 5.8: Evaluierung auf Absatzebene - Einzelergebnisse (Absatz-Trennung auf Basis von HTML Mustern).

Klassifikator	Recall	Precision	F-Maß
NB	**61,07 %**	**59,22 %**	**59,75 %**
k-NN	57,82 %	57,45 %	57,31 %
SVM	59,80 %	58,44 %	58,76 %

Tabelle 5.9: Evaluierung auf Absatzebene - Gesamtergebnisse (Absatz-Trennung auf Basis von Textstatistik-Mustern).

5.4.3 Interpretation der Ergebnisse

Unter Verwendung des naiven Bayes - Klassifikator wurden sowohl auf Dokumenten- als auch auf Absatzebene die besten Ergebnisse erzielt. Bei der Absatz-Trennung hat

Klassifikator	Korpus	Recall	Precision	F-Maß
NB	**amazon.com Top 100 Books of 2007**	**45,32 %**	**52,37 %**	**48,59 %**
NB	**amazon.com Top 100 Books of 2007***	**48,71 %**	**57,95 %**	**52,93 %**
NB	**amazon.com Video Games**	**74,96 %**	**60,16 %**	**66,75 %**
NB	**Daily Mail**	**76,52 %**	**64,40 %**	**69,94 %**
NB	**The Times**	**59,86 %**	**61,20 %**	**60,52 %**
k-NN	amazon.com Top 100 Books of 2007	43,74 %	51,31 %	47,22 %
k-NN	amazon.com Top 100 Books of 2007*	47,80 %	57,29 %	52,12 %
k-NN	amazon.com Video Games	69,66 %	58,12 %	63,37 %
k-NN	Daily Mail	71,77 %	62,14 %	66,61 %
k-NN	The Times	56,15 %	58,40 %	57,25 %
SVM	amazon.com Top 100 Books of 2007	45,01 %	52,17 %	48,32 %
SVM	amazon.com Top 100 Books of 2007*	48,23 %	57,78 %	52,58 %
SVM	amazon.com Video Games	71,62 %	58,96 %	64,68 %
SVM	Daily Mail	74,97 %	63,51 %	68,76 %
SVM	The Times	59,19 %	59,76 %	59,47 %

Tabelle 5.10: Evaluierung auf Absatzebene - Einzelergebnisse (Absatz-Trennung auf Basis von Textstatistik-Mustern).

sich die Implementierung auf Basis von HTML - Mustern bewährt. Unter Zuhilfenahme dieser Komponente konnte ein Recall Wert für private Elemente von 91,42 % erreicht werden wobei eine Precision von 79,82 % erzielt wurde (Tabelle 5.7). In Bezug auf die ursprüngliche Aufgabenstellung lässt sich hierbei folgende Aussage ableiten. Auf Absatzebene konnten, auf Zeichenbasis gemessen, 91,42 % aller privaten Elemente entdeckt werden wobei 79,38 % aller als privat klassifizierten Elemente auch tatsächlich privater Natur waren. Daraus folgt dass 8,62 % aller sensitiven Informationen nicht entdeckt werden konnten und zudem 20,18 % aller als sensitiv eingestuften Informationen in Wahrheit nicht-sensitive Informationen waren, welche fälschlicherweise als sensitiv klassifiziert wurden.

Die größten Probleme traten bei dem Korpus 'amazon.com Top 100 Books of 2007' auf. Da professionelle Rezensionen, welche einerseits kaum von einer Produktbeschreibung zu unterscheiden sind und andererseits auch keine Rezensionen im eigentlichen Sinne darstellen, eine nahe liegende Vermutung für die relativ schlechten Ergebnisse waren, wurde dieser Korpus in einer zweiten Variante annotiert. Hierbei wurden professionelle Rezensionen als öffentlich deklariert, bei diesem Korpus zeigte sich zwar eine spürbare Verbesserung, wenn gleich nicht in dem erhofften Ausmaß. Die Ursache hierfür dürfte schlicht und einfach am Schreibstil dieser Gruppe von Rezensionisten liegen, welcher häufig sehr professionell wirkt und wenige erzählerische Elemente enthält. Als Kontrast zu Bücher-Rezensionen stellen sich hierbei Rezensionen zu Videospielen dar, bei welchen professionell wirkende Rezensionen die Ausnahme sind und der Sprachstil einer Rezension sich sehr deutlich von jenem der

Produktbeschreibung unterscheidet. So ist dieser Korpus der absolute Spitzenreiter, wobei F-Maß Werte von über 90 % erreicht werden. Die beiden Zeitungs-Korpora pendeln sich hierbei in der Mitte ein, bei einem F-Maß von etwa 85 %. Nachfolgend ist ein Beispiel einer 'typischen', korrekt klassifizierten Rezension.

> I like books that teach me something, and there is a lot to learn in Splendid Suns. Previously, I didn't know much about the political turmoil in Afghanistan and the various factions vying for power. I knew women had an appalling time living under the Taliban regime, but I didn't realize how horrible conditions really were. The childbirth section will fill you with horror. I also learned of the natural beauty of Afghanistan and her fascinating history.

Die nachfolgende Rezension ist ein Beispiel einer 'professionellen' Rezension welche als öffentlich (und somit nicht-sensitiv) klassifiziert wurde.

> With his second novel, Khaled Hosseini proves beyond a shadow of doubt that "The Kite Runner" was no flash in the Afghan pan. Once again set in Afghanistan, the story twists and turns its way through the turmoil and chaos that ensued following the fall of the monarchy in 1973, but focuses mainly on the lives of two women, thrown together by fate.
>
> The story starts decades before the Taliban came into power in 1996, and ends after the era of Taliban rule. The main character begins life as a "harami" - the illegitimate daughter of a wealthy man and one of his housekeepers.
>
> ...

Die Ergebnisse bei der Verwendung einer Absatz-Trennung auf Basis von Textstatistik-Mustern waren deutlich schlechter als jene auf Basis von HTML Mustern. Wenngleich sich aus HTML Tags bestehende Muster sehr bewährt haben, liegt die Vermutung nahe, dass dass noch Optimierungspotential in der Implementierung der Absatz-Trennung auf Basis von Textstatistik-Mustern liegt, zumal die Leistungsdifferenz derzeit doch sehr groß ist.

Ein weiterer interessante Erkenntnis war, dass deutlich schlechtere Ergebnisse erzielt werden, wenn der Klassifikator mit den Instanzen aus den eigentlichen Korpora zur Evaluierung auf Absatzebene trainiert wird. Dieser Versuch erfolgt auf Basis eines Gesamt-Korpus ("amazon.com Top 100 Books of 2007", "amazon.com Video Games", "Daily Mail" und "The Times") mittels einer 10-fachen Kreuzvalidierung.

Sowohl Recall- als auch Precision Werte waren bei diesem Szenario etwa 20 % niedriger als bei der Verwendung des Klassifikators, welcher anhand des Santini-Korpus auf Dokumentenebene trainiert wurde. Ein Störfaktor der hierbei auf jeden Fall mitverantwortlich ist sind Unzulänglichkeiten des HTML - Parsers. Allen verwendeten Korpora gemein ist, dass sie intensiven Gebrauch von Java-Script machen. Dies führt dazu dass teilweise Texte auseinandergerissen bzw. nicht korrekt zusammengefügt werden. Ein weiterer gravierender Störfaktor der bei einem Test in Form einer Plain-Text Transformation ermittelt werden konnte ist, dass vereinzelt Script - Code in beträchtlichem Umfang als Text übernommen wird.

5.5 Zusammenfassung

Die Zielsetzung wurde dahingehend formuliert als dass zumindest 90 % aller sensitiven Informationen von Web-Dokumenten auf Dokumenten- und Absatzebene entdeckt werden. Zudem soll die Fehlerquote von als sensitiv klassifizierten Elemente nicht mehr als 15 % betragen.

Die Evaluierung auf Dokumentenebene wurde auf Basis des '7 web genre collection' - Korpus durchgeführt welcher 7 Genres mit jeweils 200 Web-Dokumenten enthält. Hierbei konnten 85,75 % aller sensitiven Dokumente erkannt werden. 26,55 % aller als sensitiv erkannten Dokumente wurden hierbei fälschlicherweise als solche klassifiziert und enthielten nicht-sensitive Informationen.

Für die Evaluierung auf Absatzebene wurden als Test-Korpora Produktseiten von 'amazon.com' samt Kundenrezensionen sowie Artikel samt Leserkommentaren zweier Zeitungen ausgewählt. Der verwendete Klassifikator wurde hierbei ohne Daten aus einem dieser Korpora trainiert. Die besten Ergebnisse wurden hierbei unter Verwendung einer Absatz-Trennung auf Basis von HTML-Mustern erzielt. 91,42 % aller sensitiven Elemente wurden erkannt. Lediglich 8,58 % aller sensitiven Elemente konnte nicht erkannt werden. 79,82 % aller als sensitiv eingestuften Elemente waren hierbei tatsächlich relevant. 20,18 % aller als sensitiv klassifizierten Elemente wurden hingegen fälschlicherweise als solche eingestuft.

Die meisten Probleme traten bei der Erkennung von Kundenrezensionen zu Büchern auf. Hierbei hatte der Klassifikator Probleme Rezensionen zu erkennen, welche in einem zu professionell wirkenden Sprachstil verfasst wurden und zugleich wenige erzählende Elemente enthielten. Bei diesem Korpus sank das F-Maß auf 81,51 %. Als Gegenbeispiel zeigten sich hierbei Kundenrezensionen zu Videospielen. Bei diesem Korpus hebt sich der Sprachstil der Kundenrezensionen doch deutlich von jenem der Produktbeschreibung ab. Bei diesem Korpus konnte ein F-Maß von 90,13 % erreicht werden, dies stellt zugleich den höchsten erzielten Wert dar.

6. Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Ansatz zur Klassifizierung von privaten und öffentlichen Elementen von Web-Dokumenten Auf Dokumenten- und Absatzebene präsentiert. Klassische Textkategorisierungs-Systeme operieren ausschließlich auf Dokumentenebene. Als minimale Zielsetzung, um in einem Web-Archivierungssystem verwendet werden zu können, wurde hierbei ein minimaler Recall von 90 % für private Daten definiert. Durch diese Anforderung wird sichergestellt dass mindestens 90 % aller sensitiven Informationen auch als solche erkannt werden. Die Klassen 'öffentlich' und 'privat' wurden jeweils als ein Set von Genres definiert.

Grundlegend lassen sich Merkmale zur Klassifizierung von Web-Dokumenten in die Gruppen Textstatistik, Part-of-Speech und Präsentation unterteilen. Lesbarkeitsindizes, welche im Allgemeinen auf Basis der Wort- und Silbenanzahl ermittelt werden, stellen hierbei einen besonders robusten Vertreter der Gattung Textstatistik-Merkmale dar. Bei der Part-of-Speech Analyse werden die im Text verwendeten Wörter in Bezug auf ihre Funktion im Satz sowie ihre Wortart untersucht. Beispielsweise können über die Häufigkeiten von Wortarten Sprachstile erkannt werden, welche wiederum häufig mit bestimmten Genres korrelieren. Beiden Merkmal-Gruppen gemein ist dass sie, im Gegensatz zu präsentationsbezogenen Merkmalen, auf jegliche Art von Texten unabhängig von der eigentlichen Darstellung angewandt werden können.

Voraussetzung für eine Klassifizierung auf Absatzebene ist eine möglichst präzise Trennung von zusammengehörenden Textsegmenten. Für dieses Problem wurden zwei mögliche Ansätze vorgestellt. Ein Ansatz beruht auf der Annahme, dass bestimmte HTML Tags bzw. bestimmte aus HTML Tags bestehende Muster als zuverlässige Indikatoren bei der Trennung von Absätzen dienen. Ein anderer Ansatz

basiert auf der Annahme dass sich Textelemente, welche verschiedenen Genres angehören, auf Basis von syntaktischen Mustern erkennen lassen.

Bei der Evaluierung auf Dokumentenebene konnten 85,75 % aller Dokumente mit sensitivem Inhalt erkannt werden. Bei der Evaluierung auf Absatzebene wurden als Test-Korpora Produktseiten von 'amazon.com' samt Kundenrezensionen sowie Artikel samt Leserkommentaren zweier Zeitungen ausgewählt. Der verwendete Klassifikator wurde hierbei ohne Instanzen aus einem dieser Korpora trainiert. Die besten Ergebnisse konnten bei einer Absatz-Trennung auf Basis von HTML-Mustern erzielt werden. Hierbei wurden 91,42 % aller sensitiven Daten, gemessen auf Zeichen-Ebene, erkannt. Lediglich 8,58 % aller sensitiven Daten wurden nicht korrekt als solche klassifiziert. 79,82 % aller als sensitiv eingestuften Daten waren hierbei tatsächlich relevant. Die erzielten Ergebnisse werden auch in einem Beitrag des diesjährigen 'International Web Archiving Workshop' publiziert [1].

Als zukünftige Erweiterung wäre auch eine Klassifizierung auf Wortebene möglich. Diese Erweiterung wäre aber domänenspezifisch und kaum in einer generischen Variante umsetzbar. So wäre beispielsweise ein Anwendungsfall denkbar, bei dem Kommentare eines Zeitungsartikels, als Teil der 'Netzkultur', ebenfalls als öffentlich klassifiziert werden. Im Gegenzug allerdings könnten die Verfasser, als auch Referenzen zu anderen Verfassern von Kommentaren erkannt und gegebenenfalls ausgeblendet werden. In der nachfolgenden Abbildung 6.1 ist dies beispielhaft dargestellt. Personen im eigentlichen Artikel werden hierbei nicht ausgeblendet. Bei den Kommentaren hingegen wird unterschieden, ob Personenbezeichnungen sich auf eine im Artikel vorkommende Person beziehen oder aber auf Verfasser eines Kommentars, welche durch 'schwarze Balken' ausgeblendet werden.

Rice: Mideast peace deal still within reach

Updated 3h 23m ago | Comments 392 | Recommend 10 E-mail | Save | Print | RSS

Enlarge By Omar Rashidi, Palestinian Press Office

Secretary of State Condoleezza Rice meets with Palestinian Presiden Mahmoud Abbas on Sunday n Ramallah, West Bank.

JERUSALEM (AP) — U.S. Secretary of State Condoleezza Rice on Sunday said a year-end goal for an Israeli-Palestinian peace deal is still achievable, even though both sides question whether the target is realistic.

Mixx it
Other ways to share:
Yahoo! Buzz
Digg
Newsvine
Reddit
Facebook
What's this?

Rice made the comments after a meeting with Palestinian President Mahmoud Abbas, who has sounded increasingly pessimistic about reaching an agreement with the Israelis. Abbas accuses Israel of undermining talks by continuing to build Jewish settlements on lands the Palestinians claim for a future state, and refusing to remove hundreds of military checkpoints that dot the West Bank.

At a news conference with the Palestinian leader, Rice urged Israel not to prejudice a final deal — a reference to the settlement construction. And in unusually pointed criticism, Rice suggested the Israeli government could do more to improve life for West Bank residents.

She said Israeli gestures in the West Bank must have a "real effect" on the lives of people there. "We are trying to look not just at quantity, but also quality of improvements," she said.

Abbas and Israeli Prime Minister Ehud Olmert relaunched peace talks at a U.S.-hosted conference last November and set a December 2008 target for a peace deal.

FIND MORE STORIES IN: President Bush | White House | Islamic | Palestinians | Cabinet | West Bank | Jewish | Israelis | Gaza Strip | Hamas | Rice | Palestinian President Mahmoud Abbas | Israeli-Palestinian | Israeli Prime Minister Ehud Olmert | Jenin | Israeli Foreign Minister Tzipi Livni | U.S.-hosted | Israeli Defense Minister Ehud Barak | Palestinian Prime Minister Ahmed Qureia

Their talks are to be based on the U.S.-backed "road map," a peace plan that sets out a phased process leading to the formation of an independent Palestinian state. As interim measures, Israel is supposed to halt settlement activity and take steps to improve the freedom of movement for Palestinians, while the Palestinians are supposed to dismantle militant groups. Neither side has fully met these obligations.

Rice said carrying out the roadmap is "very painstaking work," but noted that U.S. President George W. Bush believes "the time has come for the establishment of the Palestinian state."

"That is why we are working so hard on the roadmap simultaneously with the negitiations. And we continue to believe that it is an achievable goal to have an agreement between the Palestinians and the Israelis by the end of the year and by the end of President Bush's term," she said.

Rice arrived on her latest peace mission on Saturday night, and spent Sunday in a series of meetings with Israeli and Palestinian leaders. With no concrete signs of progress, Rice is seeking to breathe new life into peace talks before a visit to the region later this month by Bush, who is joining Israel's 60th anniversary celebrations.

Rice said that during talks with Israeli Defense Minister Ehud Barak, there was an "extensive discussion" of the checkpoints.

Israel maintains hundreds of roadblocks and checkpoints throughout the West Bank, saying they are needed to protect

Comments: (392) Showing: Newest first

[redacted] wrote: 1m ago
[redacted] wrote: 32m ago
[redacted] wrote: 1m ago
The politics here is only what is seen with carnal eyes. To see the true war going on here, your spiritual eyes must be open. This is a heavenly as well as earthly battle. Satan wants Israel destroyed to try and thwart the prophecies of God. It will not happen. The scriptures say there will be no peace until Christ's Kingdom comes with Him to earth. So, don't sweat these things. God has things well in hand. All this talk of peace in the Middle East will lead to a seven year peace treaty between the Antichrist and Israel and then then the Antichrist will break that halfway through. 3 1/2 years of false peace is all that will eventually come of this. And Bush will not be part of it.

++++++++++++++++++++++++++

Again [redacted], did the finger of God write UN181? And did the murdering, thieving, lying terrorists that founded Israel act as instruments of God?
---------- ---------- ------
Again, I'll tell you. God turned the hearts of men to write UN 181.

Based on this, Israel has clearly overstepped itself by occupying the lands in West Bank and territories (like Golan Heights) seized in 1967.

Thus, Israel needs to clear out of those areas and start minding its own business instead of squeezing the Palestinians out of the region entirely.

Recommend | Report Abuse

Abbildung 6.1: Abbildung einer möglichen Klassifizierung auf Wortebene. Private Elemente werden hierbei kontextabhängig durch 'schwarze Balken' ausgeblendet.

A. Appendix

A.1 amazon.com Top 100 Books of 2007 - Dokumentenliste

Titel	**Autor**	**Version**
A Thousand Splendid Suns (Hardcover)	Khaled Hosseini	19.11.2007
The Brief Wondrous Life of Oscar Wao (Hardcover)	Junot Diaz	19.11.2007
Harry Potter and the Deathly Hallows (Book 7) (Hardcover)	J. K. Rowling (Author), Mary GrandPré (Illustrator)	19.11.2007
The World Without Us (Hardcover)	Alan Weisman	19.11.2007
The God of Animals: A Novel (Hardcover)	Aryn Kyle	19.11.2007
Schulz and Peanuts: A Biography (Hardcover)	David Michaelis	19.11.2007
The Beautiful Things That Heaven Bears (Hardcover)	Dinaw Mengestu	19.11.2007
A Long Way Gone: Memoirs of a Boy Soldier (Hardcover)	Ishmael Beah	19.11.2007
Better: A Surgeon's Notes on Performance (Hardcover)	Atul Gawande	19.11.2007
The Year of Living Biblically: One Man's Humble Quest to Follow the Bible as Literally as Possible (Hardcover)	A. J. Jacobs	19.11.2007
The Great Man: A Novel (Hardcover)	Kate Christensen	19.11.2007
Call Me by Your Name: A Novel (Hardcover)	Andre Aciman	19.11.2007
fortgesetzt auf nachfolgender Seite		

Tabelle A.1 – Fortsetzung

Titel	Autor	Version
God Is Not Great: How Religion Poisons Everything (Hardcover)	Christopher Hitchens	19.11.2007
The Age of Turbulence: Adventures in a New World (Hardcover)	Alan Greenspan	19.11.2007
The Name of the Wind (The Kingkiller Chronicle, Day 1) (Hardcover)	Patrick Rothfuss	19.11.2007
Loving Frank: A Novel (Hardcover)	Nancy Horan	19.11.2007
No One Belongs Here More Than You: Stories (Hardcover)	Miranda July	19.11.2007
Animal, Vegetable, Miracle: A Year of Food Life (Hardcover)	Barbara Kingsolver, Camille Kingsolver, Steven L. Hopp	19.11.2007
Vietnam Zippos: American Soldiers' Engravings and Stories (1965-1973) (Hardcover)	Sherry Buchanan	19.11.2007
Einstein: His Life and Universe (Hardcover)	Walter Isaacson	19.11.2007
The Rest Is Noise: Listening to the Twentieth Century (Hardcover)	Alex Ross	19.11.2007
The Spellman Files: A Novel (Hardcover)	Lisa Lutz	19.11.2007
The Gathering (Man Booker Prize) (Paperback)	Anne Enright	19.11.2007
The Invention of Hugo Cabret (Hardcover)	Brian Selznick	19.11.2007
Run (Hardcover)	Ann Patchett	19.11.2007
Musicophilia: Tales of Music and the Brain (Hardcover)	Oliver Sacks	19.11.2007
The Zookeeper's Wife: A War Story (Hardcover)	Diane Ackerman	19.11.2007
The Day of Battle: The War in Sicily and Italy, 1943-1944 (The Liberation Trilogy) (Hardcover)	Rick Atkinson	19.11.2007
The Nine: Inside the Secret World of the Supreme Court (Hardcover)	Jeffrey Toobin	19.11.2007
The Art of Simple Food: Notes, Lessons, and Recipes from a Delicious Revolution (Hardcover)	Alice Waters	19.11.2007
The Reluctant Fundamentalist (Hardcover)	Mohsin Hamid	19.11.2007
About Alice (Hardcover)	Calvin Trillin	19.11.2007
The Black Swan: The Impact of the Highly Improbable (Hardcover)	Nassim Nicholas Taleb	19.11.2007
1776: The Illustrated Edition (Hardcover)	David McCullough	19.11.2007

fortgesetzt auf nachfolgender Seite

Tabelle A.1 – Fortsetzung

Titel	**Autor**	**Version**
The War: An Intimate History, 1941-1945 (Hardcover)	Geoffrey C. Ward, Ken Burns	19.11.2007
The Face of Death (Hardcover)	Cody Mcfadyen	19.11.2007
Out Stealing Horses: A Novel (Hardcover)	Per Petterson	19.11.2007
On Chesil Beach: A Novel (Hardcover)	Ian McEwan	19.11.2007
Planet Earth: As You've Never Seen It Before (Hardcover)	Alastair Fothergill, David Attenborough	19.11.2007
The Abstinence Teacher (Hardcover)	Tom Perrotta	19.11.2007
Crashing Through: A True Story of Risk, Adventure, and the Man Who Dared to See (Hardcover)	Robert Kurson	19.11.2007
Fire in the Blood (Hardcover)	Irene Nemirovsky, Sandra Smith	19.11.2007
The Maytrees: A Novel (Hardcover)	Annie Dillard	19.11.2007
Blackwater: The Rise of the World's Most Powerful Mercenary Army (Hardcover)	Jeremy Scahill	19.11.2007
The River Cottage Meat Book (Hardcover)	Hugh Fearnley-Whittingstall	19.11.2007
Spook Country (Hardcover)	William Gibson	19.11.2007
Knuffle Bunny Too: A Case of Mistaken Identity (Hardcover)	Mo Willems	19.11.2007
Armed America: Portraits of Gun Owners in Their Homes (Hardcover)	Kyle Cassidy	19.11.2007
The Young Man and the Sea: Recipes and Crispy Fish Tales from Esca (Hardcover)	David Pasternack, Ed Levine, Christopher Hirsheimer	19.11.2007
Someone Knows My Name: A Novel (Hardcover)	Lawrence Hill	19.11.2007

Tabelle A.1: amazon.com Top 100 Books of 2007 Korpus
- Auflistung der enthaltenen Dokumente

A.2 amazon.com Video Games - Dokumentenliste

Titel	**Plattform**	**Version**
Ratchet & Clank Future: Tools of Destruction	Playstation 3	15.12.2007
Elder Scrolls IV: Oblivion: Game of the Year Edition	Playstation 3	15.12.2007
The Orange Box	Playstation 3	15.12.2007
Rock Band Special Edition	Playstation 3	15.12.2007
Call of Duty 4: Modern Warfare	Playstation 3	15.12.2007
Uncharted: Drake's Fortune	Playstation 3	15.12.2007
Heavenly Sword	Playstation 3	15.12.2007
Assassin's Creed	Playstation 3	15.12.2007
Guitar Hero III: Legends of Rock Bundle	Playstation 3	15.12.2007
WarHawk Bundle with Bluetooth Headset	Playstation 3	15.12.2007
BioShock	Xbox 360	15.12.2007
The Orange Box	Xbox 360	15.12.2007
Rock Band Special Edition	Xbox 360	15.12.2007
Call of Duty 4: Modern Warfare	Xbox 360	15.12.2007
Halo 3	Xbox 360	15.12.2007
Mass Effect	Xbox 360	15.12.2007
Forza Motorsport 2	Xbox 360	15.12.2007
Guitar Hero III: Legends of Rock Bundle	Xbox 360	15.12.2007
Assassin's Creed	Xbox 360	15.12.2007
Skate	Xbox 360	15.12.2007

Tabelle A.2: amazon.com Video Games Korpus - Auflistung der enthaltenen Dokumente

A.3 Daily Mail - Dokumentenliste

Titel	Autor	Version
Sweaty Murat was breathless and excited during Maddie police quiz	Rebecca Camber	15.12.2007
Prince Harry's polo coach leaves his wife after affair with stable girl	Katie Nicholl and Richard Creasy	15.12.2007
Classic children's names conveying style and elegance top this year's favourites	Polly Dunbar	15.12.2007
First picture of Kelly, the all-American mother who reveals John Darwin was the 'creepiest and most frightening man I ever met'	Sharon Churcher	15.12.2007
Bank of England governor was warned of Northern Rock collapse two years ago	Simon Walters	15.12.2007
Boy-next-door Leon snatches X Factor glory from red-hot favourite Rhydian	Cath Sherwood	15.12.2007
Wardrobe malfunction for Posh as the Spice Girls hit the UK	Bevan Hurley	15.12.2007
Electricity firm breaks into house to change meter - even though homeowner is NOT a customer	Jason Lewis	15.12.2007
Home Secretary plotted to undermine police claims for a pay rise in order to subsidise War on Terror	Christopher Leake	15.12.2007
Labour's bin tax will lead to more bonfires and pollution, say experts	Joanathan Oliver	15.12.2007
Cameron's pint-sized Rasputin - by the blonde from Tory HQ	Simon Walters	16.12.2007
Brown facing a spring revolt as MP compares him to defeated Anthony Eden	Jonathan Oliver	16.12.2007
Fraudsters pose as MPs to steal personal bank details	James Slack	16.12.2007
Labour donor David Abrahams banned from taking beauty queen to club - because of her views on Holocaust	Simon Walters	16.12.2007
Paul Weller declares: 'I enjoyed my midlife drug spree'	k.A.	16.12.2007
Parky's last show after 26 years is a two-hour celebrity marathon	k.A.	16.12.2007

fortgesetzt auf nachfolgender Seite

Tabelle A.3 – Fortsetzung

Titel	Autor	Version
William swaps the nightclub for a quiet meal with Camilla - and there's no sign of Charles or Harry	k.A.	16.12.2007
Richard and Judy's glamorous daughter and some very rude gestures...	k.A.	16.12.2007
Kate Moss' bed-head hair looks worse for wear	k.A.	16.12.2007
Leona flaunts her curves as she enjoys a romantic Caribbean break with her boyfriend	k.A.	28.12.2007
Wife killed by mentally ill husband told police days earlier she feared for her life	Arthur Martin	28.12.2007
A shaggy hog story: Boris, the extinct 'sheep-pig', gains a new trotter-hold in UK	Lucy Ballinger	28.12.2007
War hero's daughter facing arrest for tackling yobs who 'trashed war memorial'	Luke Salkeld	28.12.2007
Politically correct 'non jobs' cost the taxpayer £600million a year	k.A.	28.12.2007
A-level student stabbed to death in bus brawl over a 'dirty look'	k.A.	28.12.2007
Two potholders drown in 'flash flood' cave usually used by schoolchildren on adventure trips	Chris Brooke	28.12.2007
Mother, 34, marries foster son she adopted with husband as a refugee from war-torn Kosovo	k.A.	28.12.2007
Madeleine: Murat's alibi in doubt after two new witnesses claim they saw him on the night she disappeared	k.A.	28.12.2007
Pakistan's police shoot rioters and 23 die as Bhutto lies buried next to her father	k.A.	28.12.2007
Swimming pools, volleyball and massages... just a normal day at the office for Google staff	k.A.	29.12.2007
Benazir Bhutto 'died after hitting head on sunroof - NOT from bullet or shrapnel wounds'	k.A.	29.12.2007
Government chief adviser demands smoking ban in cars	k.A.	29.12.2007
'Doctors told us to abort our disabled baby - but our son is proof that we were right to say no'	Luke Salkeld	29.12.2007
		fortgesetzt auf nachfolgender Seite

Tabelle A.3 – Fortsetzung

Titel	**Autor**	**Version**
500,000 fewer Britons in work following influx of Eastern Europeans	James Slack	29.12.2007
Muslims should be proud to support England's cricket team, says first Islamic Minister	Jane Merrick	29.12.2007
More property gloom as UK house prices set to stall next year, warns Nationwide	k.A.	29.12.2007
Experts warn record numbers of people could go bankrupt in 2008 as financial fears deepen	k.A.	29.12.2007
Caught on camera: The moment 'model citizen' youth worker battered stranger with brick in drunken street attack	k.A.	29.12.2007
Detox diets are a waste of time, says government advisor	David Derbyshire	29.12.2007
Model behaviour: Kate Moss celebrates Christmas in style as alcohol bottles stack up AND remembers to recycle	k.A.	29.12.2007

Tabelle A.3: Daily Mail Korpus - Auflistung der enthaltenen Dokumente

A.4 The Times - Dokumentenliste

Titel	Autor	Version
When a bundle of joy can be a ball and chain	Sarah Vine	15.12.2007
Archbishop of Canterbury, Dr Rowan Williams, warns American church leaders to curb their pro-gay agenda	Ruth Gledhill	15.12.2007
I can't believe I'm saying this, but we need to learn from Tony Benn about how the State can change people's habits	Matthew Parris	15.12.2007
Give the talented poor a hand	Janice Turner	15.12.2007
Carbon stand-off puts climate talks at risk	Jonathan Leake	16.12.2007
How Facebook has become a very British way to stay in touch	Elizabeth Judge, Dan Sabbagh	16.12.2007
Tech groups' Microsoft challenge threatens to reignite browser wars	Rhys Blakely	16.12.2007
Wood power and rain tanks at heart of new eco-village	James Rossiter	16.12.2007
Google unveils rival to Wikipedia	Rhys Blakely	16.12.2007
'Tis the season to be joyful shoppers, but the young are not buying into the tradition	Peter Riddell	16.12.2007
Climate deal sealed in Bali	Times Online and agencies	16.12.2007
Oh please, you lard-butt British frumps have got off too lightly	Tad Safran	16.12.2007
Hollywood's A-list of overpaid stars	John Harlow	16.12.2007
Inside the Taliban's fallen town of fear	Stephen Grey	16.12.2007
Gordon Brown in 'crisis of morale'	Irwin Stelzer	16.12.2007
America's constitution produces a pure democracy we will never have	Simon Jenkins	16.12.2007
Labour's Scottish chief in new cash row	Jason Allardyce	16.12.2007
Motorists latest victims of missing data scandal	Marie Woolf	16.12.2007
Police plan pay protest march	Alan Schofield and Steven Swinford	16.12.2007
Charles Kennedy set to climb back on frontline wagon	Marie Woolf	16.12.2007
X Factor surprise	Leon Jackson and Amy Fallon	16.12.2007

fortgesetzt auf nachfolgender Seite

Tabelle A.4 – Fortsetzung

Titel	**Autor**	**Version**
Dawkins to preach atheism to US	Maurice Chittenden and Roger Waite	28.12.2007
What's smug and deserves to be decapitated?	Matthew Parris	28.12.2007
Ofcom to probe Catherine Tate Christmas special	Catherine Tate	28.12.2007
Bhutto 'died after hitting head on car roof', Pakistan government claims	Jenny Booth and agencies	28.12.2007
Daughter held over Christmas murders of six of her family	James Bone	28.12.2007
Polls put President ahead in Kenyan elections	Jason Allardyce	28.12.2007
Barack Obama urges voters to reject 'politics of fear' in face of resurgent Hillary Clinton Democratic presidential hopeful, Senator Barack Obama	Tim Reid	28.12.2007
'I have only now begun to mourn my wife's death. Now my heart is broken'	Sonia Verma	28.12.2007
Benazir Bhutto - world and readers' reactions	From The Times	28.12.2007
Leaders denounce senseless murder of a courageous woman and friend	Philip Webster	29.12.2007
News could be pivotal in the race to replace Bush as president	Tom Baldwin	29.12.2007
Why latest gadgets are already out of date	Chris Ayres	29.12.2007
Illegal film and TV downloaders could lose their links to the web	Irwin Stelzer	29.12.2007
Bad connection 'could unplug rural economy'	Valerie Elliott	29.12.2007
Childminder's 'babycam' brings issue of trust into sharp focus	Marie Tourres and Charles Bremner	29.12.2007
Well-behaved pupils given video games and executive perks	Alexandra Frean	29.12.2007
Amazon partners with fans' online record label	Jonathan Richards	29.12.2007
How online fraudsters helped themselves on Christmas Day	Jack Malvern	29.12.2007
A nation online	Hannah Fletcher	29.12.2007

Tabelle A.4: The Times Korpus - Auflistung der enthaltenen Dokumente

Literaturverzeichnis

[1] Andreas Rauber, Max Kaiser, Bernhard Wachter: *Ethical Issues in Web Archive Creation and Usage - Towards a Research Agenda.* In: *Proceedings of the 8th International Web Archiving Workshop, Aarhus, Denmark, September 18-19, 2008.*

[2] Archives, The National: *Information on web archiving Project Homepage.* http://www.nationalarchives.gov.uk/preservation/webarchive/future.htm. [Version vom 24.6.2008].

[3] Bosch, Karl: *Statistik-Taschenbuch.* Oldenbourg, 3. Auflage, Jänner 1998.

[4] Bouckaert, Remco R.: *A probabilistic line breaking algorithm.* In: *Proceedings of the Sixteenth Australian Joint Conference on Artificial Intelligence, Australia, December 3-5, 2003.*, Band 2903/2003, Seiten 390–401. Springer-Verlag, 2003.

[5] David Aha, Dennis Kibler: *Instance-based learning algorithms.* Machine Learning, 6:37–66, 1991.

[6] Derrick Oswald, Somik Raha, Ian Macfarlane David Walters: *HTML Parser Project Homepage.* http://htmlparser.sourceforge.net/. [Version vom 25.3.2008].

[7] Domingos, Pedro und Michael J. Pazzani: *On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.* Machine Learning, 29(2-3):103–130, 1997.

[8] Gerhard Marinell, Gabriele Steckel-Berger: *Einführung in die Bayes-Statistik.* Oldenbourg, 3. Auflage, Jänner 2001.

[9] Hall, Mark: *Correlation-based Feature Selection for Machine Learning.* Ph.D dissertation, New Zealand: Waikato University, Department of Computer Science., 1998.

[10] Han-Yuen Ong, Miri Ali: *Privacy preserving database access through dynamic privacy filters with stable data randomization.* In: *IEEE International Conference on Systems, Man and Cybernetics 2007*, Band 4654, Seiten 3333–3338. Springer-Verlag, Oktober 2007.

[11] Heinonen, Oskari: *Optimal multi-paragraph text segmentation by dynamic programming.* In: *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, Seiten 1484–1486, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[12] Holmes, G., A. Donkin und I.H. Witten: *WEKA: A Machine Learning Workbench.* In: *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994.

[13] Ilja N. Bronstein, Konstantin A. Semendjajew, Gerhard Musiol: *Taschenbuch der Mathematik.* Harri Deutsch Verlag, 5. Auflage, Juni 2001.

[14] Joachim Hartung, Bärbel Elpelt, Karl-Heinz Klösener: *Statistik: Lehr- und Handbuch der angewandten Statistik.* Oldenbourg, 14. Auflage, September 2005.

[15] Kim, Jinsuk und Myoung Ho Kim: *An Evaluation of Passage-Based Text Categorization.* Journal of Intelligent Information Systems, 23(1):47–65, 2004.

[16] Kohavi, Ron: *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*, Seiten 1137–1145, 1995.

[17] Lahn, Andrea: *Genre-Analyse von Web-Dokumenten.* Diplomarbeit, Bauhaus-Universität Weimar, Fakultät Medien, Mediensysteme, November 2006.

[18] McCallum, Douglas R. und James L. Peterson: *Computer-based readability indexes.* In: *ACM 82: Proceedings of the ACM '82 conference*, Seiten 44–48, New York, NY, USA, 1982. ACM.

[19] Nationalbibliothek, Deutsche: *Archivierung von Netzpublikationen Projekt Homepage.* http://www.d-nb.de/netzpub/index.htm. [Version vom 24.6.2008].

[20] Nationalbibliothek, Schweizerische: *Elektronische Publikationen Projekt Homepage.* http://www.nb.admin.ch/slb/themen/e-publikationen/index.html. [Version vom 24.6.2008].

[21] Nationalbibliothek Österreichische: *Webarchivierung Projekt Homepage.* http://www.onb.ac.at/about/webarchivierung.htm. [Version vom 24.6.2008].

[22] Organization, Internet Archive: *About the Internet Archive.* http://www.archive.org/about/about.php. [Version vom 2.1.2008].

[23] P. R. Krishnaiah, L. N. Kanal: *Handbook of Statistics 2: Classification, Pattern Recognition and Reduction of Dimensionality.* North-Holland Publishing, 2. Auflage, 1987.

[24] Rakesh Agrawal, Ramakrishnan Srikant: *Privacy-preserving data mining.* In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, Seiten 439–450. ACM Press, 2000.

[25] Richard O. Duda, Peter E. Hart, David G. Stork: *Pattern Classification.* Wiley and Sons, 2. Auflage, Jänner 2001.

[26] Rish, Irina: *IBM Research Report RC 22230 - An empirical study of the naive Bayes classifier.* http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf, November 2001. [Version vom 25.3.2008].

[27] Rong She, Ke Wang, Ada Waichee Fu Yabo Xu: *Computing Join Aggregates over Private Tables.* In: *Data Warehousing and Knowledge Discovery*, Band 4654, Seiten 77–88. ACM Press, 2007.

[28] Santini, Marina: *Identifying Genres on the Web.* Technical Report ITRI-03-06, 2003, ITRI, University of Brighton (UK), 2003.

[29] Santini, Marina: *Automatic Identification of Genre in Web Pages.* Doktorarbeit, University of Brighton, Brighton (UK), 2007.

[30] Schriver, Karen A.: *Readability formulas in the new millennium: what's the use?* ACM Journal of Computer Documentation (JCD), 24(3):138–140, 2000.

[31] Sebastiani, Fabrizio: *Machine learning in automated text categorization.* ACM Computing Surveys, 34(1):1–47, 2002.

[32] Sheffield, University of: *GATE, A General Architecture for Text Engineering.* http://gate.ac.uk/. [Version vom 28.2.2008].

[33] Sporleder, Caroline und Mirella Lapata: *Broad coverage paragraph segmentation across languages and domains.* ACM Transactions on Speech and Language Processing (TSLP), 3(2):1–35, 2006.

[34] Stamatatos, E., N. Fakotakis und G. Kokkinakis: *Text genre detection using common word frequencies.* In: *Proceedings of the 18th conference on Computational linguistics, Saarbrücken, Germany, July 31st to August 4th 2000*, Band 2, Seiten 808–814, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[35] Talburt, John: *The Flesch index: An easily programmable readability analysis algorithm.* In: *SIGDOC '85: Proceedings of the 4th annual international conference on Systems documentation*, Seiten 114–122, New York, NY, USA, 1985. ACM.

[36] Witten, Ian H., Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes und Sally Jo Cunningham: *Weka: Practical Machine Learning Tools and Techniques with Java Implementations.* In: Kasabov, Nikola und Kitty Ko (Herausgeber): *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, Seiten 192–196, 1999. Dunedin, New Zealand.

[37] Yehuda Lindell, Benny Pinkas: *Privacy Preserving Data Mining.* Lecture Notes in Computer Science, 1880:36–54, 2000.

[38] Yu, Huan Liu Lei: *Toward integrating feature selection algorithms for classification and clustering.* IEEE Transactions on Knowledge and Data Engineering, 17(4):491–502, April 2005.

[39] Yunhyong Kim, Seamus Ross: *Searching for Ground Truth: A Stepping Stone in Automating Genre Classification.* In: *Digital Libraries: Research and Development, Proceedings of the First International DELOS Conference, Pisa, Italy, February 13-14, 2007*, Band 4877/2007, Seiten 248–261. Springer-Verlag, 2007.