Analyse der Bundestagswahl 2013 mit Twitter

Analysis of the German Parliament Elections 2013 with Twitter

Bachelor-Thesis von Uli Fahrer

Mai 2014

TECHNISCHE UNIVERSITÄT DARMSTADT

Analyse der Bundestagswahl 2013 mit Twitter
Analysis of the German Parliament Elections 2013 with Twitter

vorgelegte Bachelor-Thesis von Uli Fahrer

1. Gutachten: Prof. Dr. Chris Biemann
2. Gutachten: Steffen Remus

Tag der Einreichung:

Erklärung zur Bachelor-Thesis

Hiermit versichere ich, die vorliegende Bachelor-Thesis ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 7. Mai 2014

_____

(U. Fahrer)

Zusammenfassung

Diese Bachelor-Thesis beschreibt eine Methode, die auf Wortkollokationen basiert und zwei Schlüsselwörter anhand ihrer stark assoziierten Wörter kontrastiert. Das Konzept wird verwendet, um zu untersuchen, wie Ereignisse der realen Welt in Twitter reflektiert werden.

Ein selbst entwickeltes HTML-Dashboard ermöglicht es, die Daten der Analyse zu visualisieren und zu erforschen.

Basierend auf einer Studie zur deutschen Bundestagswahl werden zwei kontrastive Kollokationsanalysen durchgeführt, welche Gemeinsamkeiten und Unterschiede zwischen zwei Politikern aufzeigen. Um die Ergebnisse mit der Tagespresse zu vergleichen, wird die Plattform „Wörter des Tages“ der Universität Leipzig genutzt, welche tagesaktuell relevante Begriffe aus unterschiedlichen Tageszeitungen zeigt.

Die Ergebnisse zeigen, dass die Wörter im Schnitt der Analyse ein Indikator für zentrale politische Ereignisse sind. Außerdem reflektiert der Twitterstream Ereignisse der realen Welt und ist in hoher Übereinstimmung mit der Tagespresse.

Abstract

This bachelor thesis describes an approach based on word co-occurrence that contrasts two separate keywords regarding their strongly associated words. This approach is used to investigate how real-world events are reflected in Twitter.

Furthermore an HTML Dashboard is presented that allows real-time interaction to visualize and explore the data for analyses.

Based on a case study on the German election, two contrastive analyses are performed, that show differences and commonalities between two politicians. In order to compare the results to the daily press, the ”Wörter des Tages” platform of the University of Leipzig is used, which shows terms that are particularly relevant for a day with respect to different daily newspapers.

Results show that the words in the overlap of the analysis are an indicator for key political events. Moreover the Twitter stream reflects real-world events well, and is in high accordance with the daily press.

Inhaltsverzeichnis

1 Einführung

Die in dieser Arbeit vorgestellte Analyse, ermöglicht es dem Benutzer, Politiker und Parteien zu kontrastieren, um deren Unterschiede und Gemeinsamkeiten zu erfahren. Dieser Vergleich basiert auf Twitter-Nachrichten der Bundestagswahl 2013. Die Analyse wird mit einer Webanwendung visualisiert und enthält Wörter, die mit den Politikern oder den Parteien in Zusammenhang stehen. Dieses Einführungskapitel motiviert die Arbeit hinsichtlich unterschiedlicher Aspekte (Abschnitt 1.1) und enthält eine kurze Einführung in den Mikrobloggingdienst Twitter (Abschnitt 1.2). Der Abschnitt 1.3 gibt abschließend einen Überblick der Folgekapitel.

1.1 Motivation

Das Internet hat sich im Laufe der Jahre im Bezug auf die zwischenmenschliche Interaktion verändert. Soziale Online-Dienste und Mikroblogging Plattformen wie Twitter waren wichtige Faktoren in dieser Entwicklung. Aktuelle Proteste in der Ukraine oder der bekannte Arabische Frühling haben gezeigt, dass Twitter ein Werkzeug ist mit dem die Welt mit aktuellen Informationen über den Protest versorgt wird. Der Dienst wird verwendet, um die Aufmerksamkeit der internationalen Gesellschaft auf sich zu ziehen. Das wirft die Frage auf, wie Ereignisse der realen Welt in Twitter reflektiert werden.

Diese Forschungsfrage wird mit Hilfe einer Studie zur deutschen Bundestagswahl 2013 in Twitter untersucht. Der erfolgreiche Einsatz von sozialen Medien in der US-Präsidentschaftswahlkampagne von Barack Obama hat Twitter als integralen Bestandteil für den politischen Wahlkampf etabliert. Dies haben auch deutsche Politiker im Wahlkampf 2009 erkannt, so dass drei Viertel aller von Meckel und Stanoevska-Slabeva [29] identifizierten politischen Twitter-Accounts im Jahr 2009 eröffnet wurden.

In diesem Zusammenhang stellt diese Arbeit eine neuartige Methode vor, die es ermöglicht zwei Politiker oder Parteien zu vergleichen. Untersucht werden Gemeinsamkeiten und Unterschiede der Parteien oder Politiker. Diese kontrastive Analyse basiert auf den aggregierten Twitter-Nachrichten der Wahl und kann mit Hilfe einer Webanwendung durchgeführt und visualisiert werden. Die Webanwendung wurde eigens für diese Arbeit entwickelt und wird in Kapitel 6 vorgestellt.

Reflektiert werden Wörter, welche die Twitter-Nutzer häufig mit den gegebenen Politikern oder Parteien assoziieren, wodurch ein Spiegel der öffentlichen Meinung entsteht. Mit dieser Analyse lassen sich Beziehungen zwischen Politikern besser verstehen und erkunden. Außerdem erhält der Bürger einen transparenteren Einblick in politische Vorgänge.

Diese Art der Analyse ist nicht nur für Journalisten interessant, sondern auch für politisch interessierte Bürger, Wähler und Politiker selbst. Denkbar ist auch eine Verwendung im schulischen Bereich, bei dem Schüler den Verlauf der Wahl explorativ erkunden können.

Die Arbeit befasst sich mit den folgenden Forschungsfragen:

(1) Wie reflektiert Twitter Ereignisse der realen Welt?

(2) Welche Wörter sind typischerweise mit einem bestimmten Politiker oder einer Partei in Twitter assoziiert?

(3) Wie können benutzergenerierte Inhalte intuitiv visualisiert werden, so dass eine kontrastive Untersuchung von zwei Politikern oder Parteien möglich ist?

1.2 Einführung in Twitter

Twitter[1] (englisch für *Gezwitscher*) ist ein im Jahre 2006 gegründeter sozialer Mikrobloggingdienst im Internet. Auf Twitter kann jeder Benutzer auf 140 Zeichen beschränkte Kurznachrichten *twittern*. Diese sogenannten *Tweets* sind öffentlich auf der Website oder durch eine externe Anwendung einsehbar und werden auf der *Timeline* des Nutzers gespeichert. Diese Timeline ist demnach eine Sammlung aller verfassten Twitter-Nachrichten.

Es ist möglich anderen Benutzern zu folgen, da jeder Tweet der verfolgten Benutzer *(Followers)* in der eigenen Timeline angezeigt wird. Außerdem können Nutzer Tweets an ihre Follower weiterleiten. Diese *Retweets* sind mit dem Präfix „RT“ gekennzeichnet.

Mit Hilfe von Benutzeradressierungen (z.B. @angela_merkel) können andere Twitter-Accounts innerhalb des Tweets referenziert werden. Dadurch kann die Nachricht direkt an bestimmte Benutzer gerichtet werden, auch wenn diese dem Autor des Tweets nicht folgen.

In Tweets können auch sogenannte Hashtags (z.B. #BTW2013) eingefügt werden. Diese Hashtags dienen als automatischer Mechanismus, um Nachrichten unter einem gemeinsamen *Tag* zu kategorisieren. Die Bezeichnung leitet sich von dem Doppelkreuz „#“ ab, mit dem der Begriff markiert wird. Ein Hashtag kann aus Buchstaben und Ziffern bestehen, darf jedoch keine Satz- oder Leerzeichen enthalten. Diesen Hashtagkategorien kann gefolgt werden. Weiterhin ist es möglich, die Hashtag-Timeline einzusehen.

Twitter-Profile sind in der Regel öffentlich. Eine Ausnahme bilden Profile, die als geschützt markiert sind. Jeder Benutzer hat die Möglichkeit sein Profil zu schützen und kann dadurch bestimmen, welche Benutzer ihm folgen dürfen. Mit einer expliziten Erlaubnis kann die entsprechende Timeline eingesehen werden.

Des Weiteren können den Tweets optional geografische Positionsinformationen hinzugefügt werden. Dadurch enthält der Tweet die Position, an welcher der Benutzer den Tweet veröffentlicht hat. Diese Position wird als *Global Positioning System (GPS)* Koordinate gespeichert.

Trotz der einfachen Plattform ist Twitter ein sehr weit verbreiteter Dienst und wird sowohl aus politischen, als auch kommerziellen Gründen von Regierungen und Unternehmen verwendet. Die ursprüngliche Idee hinter dem Konzept Mikroblogging war das Veröffentlichen und Verbreiten von kleinen Statusupdates der Nutzer. Inzwischen enthalten die Veröffentlichungen jedoch die unterschiedlichsten Themen und Informationen, wie zum Beispiel Links zu Webseiten.

1.3 Übersicht

Diese Arbeit ist wie folgt organisiert: Kapitel 2 präsentiert Arbeiten, die sich mit der Reflexion der Politik in Twitter beschäftigen. Kapitel 3 gibt eine Übersicht über die einzelnen Komponenten des Softwaresystems, wie diese interagieren und welche Technik verwendet wird. Das Kapitel 4 beschreibt die Methodik zum Aggregieren der Twitter-Daten der Bundestagswahl 2013. Kapitel 5 widmet sich den Informationen, welche aus dem Korpus extrahiert werden und beschreibt die kontrastive Analyse. Das Kapitel 6 enthält Details zur Visualisierung der Analyse durch die Webanwendung und zeigt, wie die Ergebnisse erforscht werden können. Die Analyse wird in Kapitel 7 evaluiert. Außerdem werden exemplarisch zwei kontrastive Analysen prominenter Politiker vorgestellt und diskutiert.

Zuletzt werden einige Verbesserungen und Erweiterungen in Kapitel 8 gegeben und die Arbeit mit der Konklusion in Kapitel 9 abgeschlossen.

[1] `https://twitter.com/`

2 Verwandte Arbeiten

Während der Bundestagswahl wurden viele Plattformen zur Beobachtung der Social-Media Aktivitäten der Bundestagswahl 2013 veröffentlicht. Abschnitt 2.1 stellt einige dieser Instrumente vor und beschreibt ihre Funktion. Abschließend werden die vorgestellten Arbeiten in Abschnitt 2.2 zusammengefasst und in Hinblick auf diese Arbeit bewertet.

Wortassoziationen

Es wurde viel Forschung im Bereich von Wortassoziationen betrieben. Lexikalische Kollokationen[2] sind ein wichtiger Indikator für diese Assoziationen und waren damit die Motivation für einige grundlegenden Wortassoziationsmaße, wie Log-Likelihood (Dunning [13]), PMI (Church und Hanks [9]) und Dice (Dice [12]) .

Nach Chaudhari et al. [8] lassen sich Wortassoziationsmaße in drei Kategorien gliedern: *i*) Wortassoziationsmaße, die zusätzlich zur individuellen Unigrammfrequenz auch auf der Kollokationsfrequenz der Wörter basieren (Rapp [38], Dunning [13]), *ii*) Ähnlichkeitsverteilungsbasierte Assoziationsmaße, die ein Wort über die Verteilung anderer angrenzender Wörter beschreiben (Agirre et al. [2]) und *iii*) wissensbasierte Maße, die Wissensquellen wie Thesauri oder semantische Netzwerke nutzen (Milne und Witten [30], Gabrilovich und Markovitch [20]).

Diese Arbeit verwendet Wortassoziationsmaße, die sowohl auf der Unigrammfrequenz, als auch auf der Kollokationsfrequenz selbst basieren. Einen Überblick und Diskussionen über verschiedene Wortassoziationsmaße bieten Manning und Schütze [28]. Obwohl die mathematischen Eigenschaften dieser Wortassoziationsmaße ausführend diskutiert wurden, sind die Methoden zur qualitativen Evaluation nicht weit fortgeschritten. Evert und Krenn [15] untersuchen Methoden, um diese qualitative Evaluation durchzuführen. In diesem Zusammenhang zeigen diese, warum das normalerweise durchgeführte Bewerten der „n-besten“ durch ein bestimmtes Wortassoziationsmaß identifizierten Kandidaten nicht passend für eine qualitative Untersuchung ist.

Social-Media Analyse

Erste Untersuchungen von Twitter und politischen Wahlen wurden für die Vereinigten Staaten durchgeführt und stellen immer noch ein aktuelles Thema dar (Conway et al. [10], Wang et al. [46]). Inzwischen existieren auch einige Studien für andere Länder, wie beispielsweise Schweden (Larsson und Moe [26]).

Tumasjan et al. [44] haben für die deutsche Bundestagswahl 2009 eine Studie durchgeführt, um in einem ersten Schritt herauszufinden, ob Twitter als Forum für politische Deliberationen genutzt wird. In einem zweiten Schritt untersuchten diese, ob die Twitter-Nachrichten das politische Meinungsbild der Bevölkerung reflektieren. Die Analyse basiert auf einer Stimmungsbildanalyse der Tweets und umfasst 100.000 Tweets, die eine Partei oder einen Politiker referenzieren. Evaluiert wird die Analyse mit Hilfe von Wahlprogrammen und der Presse. Dazu erzeugen Tumasjan et al. basierend auf den einzelnen Wortfrequenzen mehrdimensionale Profile für die jeweiligen Politiker. Diese Profile sind Sterndiagramme und eignen sich besonders für Evaluationen mit festgelegten Kriterien und mehreren Serien. Als Kategorien werden Begriffe wie „zukunftsorientiert“, „Geld“, „Arbeit“, „Erfolg“ und „positive Emotionen“ gewählt.

Die Ergebnisse zeigen, dass Twitter tatsächlich als Plattform für politische Deliberation genutzt wird und das die Stimmungsbilder auf Twitter mit politischen Programmen und Kandidatenprofilen korrelieren.

[2] Überzufällig häufige Wortpaare innerhalb eines Dokumentes

Eine weitere Arbeit in Bezug auf die Reflexion der Politik in Twitter stammt von Meckel und Stanoevska-Slabeva [29]. Diese analysieren die Verbindung zwischen 577 politischen Twitter-Accounts von Parteien und Politikern vor der Bundestagswahl 2009. Die Ergebnisse zeigen, dass die Verbindungen zwischen den untersuchten Twitter-Accounts keine Reflektion für die politischen Beziehungen der Parteien darstellt. Die Analyse fokussierte sich nur auf die Verbindungen der Benutzer und analysierte nicht den Inhalt der Twitter-Nachrichten.

Die Forschung von Kaczmirek et al. [25] konzentriert sich auf die Social-Media Kommunikation der Bundestagswahl 2013. Diese erzeugen für die Wahl unterschiedliche Datensätze, die aus Twitter und Facebook extrahiert wurden. Dadurch generieren sie insgesamt sechs unterschiedliche Korpora. Diese umfassen zum Zeitpunkt der Veröffentlichung eine Zeitspanne von Juli 2013 bis September 2013 für Twitter und für Facebook eine Zeitspanne von Januar 2009 bis Oktober 2013. Das Facebook-Korpus enthält Daten, die von Facebook-Pinnwänden der Kandidaten der Bundestagswahl extrahiert wurden. Die verbleibenden Korpora bestehen aus Daten von Twitter. Beispielsweise beinhaltet ein Korpus Tweets von Kandidaten der Bundestagswahl und ein weiteres Twitter-Nachrichten von Journalisten und Nachrichtenproduzenten. Die Autoren diskutieren außerdem mögliche Untersuchungen der erzeugten Korpora für Forscher und erörtern Probleme beim Aggregieren der Social-Media-Daten.

Blenn et al. [6] beschreiben einen neuartigen hybriden Klassifizierungsansatz für die Stimmungsbildanalyse im Bereich der Social-Media Anwendungen. In diesem Zusammenhang präsentieren diese eine Stimmungbildsanalyse mit Visualisierung von häufig mit Emoticons verwendeten Adjektiven basierend auf Twitter-Nachrichten. Adjektive, die mit traurigen Emoticons (z.B. „:(“ , „=(“, respektive) assoziiert sind, werden in Richtung der negativen Y-Achse dargestellt und analog, Adjektive, die mit fröhlichen Emoticons (z.B. „:)“ , „=)“, respektive) assoziiert sind, in Richtung der positiven Y-Achse. Sie nennen das Verfahren eine kombinierte Stimmungbildsanalyse. Die Autoren geben an, dass diese Technik nicht nur für die Stimmungsbildanalyse verwendet werden kann, sondern auf andere Domänen übertragbar ist. Dazu können die zwei Schlüsselwortmengen mit Wörtern für eine ausgewählte Studie ersetzt werden. Diese Methode liefert demnach eine Menge von Wörtern, die häufig mit den gegebenen Schlüsselwörtern auftreten und visualisiert diese in einer kombinierten Weise.

2.1 Analyseinstrumente der Bundestagswahl 2013

Dieser Abschnitt präsentiert Instrumente zur Analyse der Bundestagswahl 2013 und anderer politischer Vorgänge. Es werden jeweils Funktionen und Nutzen der Analyse vorgestellt. Es handelt sich jedoch nicht um eine vollständige Übersicht.

Twitterbarometer
Die von der Firma Buzzrank[3] entwickelte Plattform Twitterbarometer[4] misst die politische Stimmungslage unter den Twitter-Nutzern in Echtzeit. Diese erfasst Äußerungen, die mit Hashtags einer Partei markiert sind und bewertet den Tweet in Hinblick auf den Hashtag als positiv oder negativ. Dazu müssen Twitter-Nutzer die Hashtags mit einem „+“ oder „-“ kennzeichnen, um ihre Zustimmung oder Ablehnung auszudrücken. Möchte ein Twitter-Nutzer beispielsweise etwas Positives über die SPD sagen, verwendet dieser den Hashtag #spd+ in seinem Tweet. Die Plattform selbst stellt die jeweilig aktuelle Stimmung im Netz in einer Grafik dar.

Bundestwitter
Die Plattform Bundestwitter[5] sammelt und analysiert die Tweets der Mitglieder des deutschen Bundestags. Die Statistik präsentiert die aktivsten twitternden Politiker und deren Twitter-Profile. Außerdem werden die meist genutzen Hashtags in der Politik der letzten Monate dargestellt. Weiterhin ist es möglich, die letzten 200 Tweets mit Geodaten zu visualisieren. Die Seite liefert einen guten Eindruck über die Politiker und deren Twitteraktivität und wird in Echzeit aktualisiert.

politwi
Politwi[6] ist ein Forschungsprojekt der Hochschule Hof in Zusammenarbeit mit der Goethe-Universität Frankfurt. Analysiert werden die Top-Themen der Bundestagswahl 2013 mit Hilfe von Tweets. Der Fokus liegt besonders auf dem frühzeitigen Erkennen von Trends und basiert nicht nur auf der individuellen Frequenz der Wörter. Es können außerdem Trends der Vergangenheit mit ihrem zeitlichen Verlauf dargestellt werden.

2.2 Zusammenfassung und Reflektion der vorgestellten Arbeiten

Viele der vorgestellten Arbeiten führen eine Stimmungbildanalyse [27] durch. Damit sind die Ansätze auf Adjektive beschränkt, die häufig mit einer Menge von gegebenen Schlüsselwörtern auftreten. Basierend auf diesen Ergebnissen lässt sich ableiten, wie positiv oder negativ ein Politiker oder eine Partei von den Twitter-Nutzern wahrgenommen wird. Es lässt sich jedoch nicht ableiten, an welchen Wahlkampforten der Politiker aufgetreten ist, welche politischen Ereignisse mit dieser Person in Zusammenhang stehen oder welche Rolle der Politiker in einem Skandal spielt. Informationen werden zwar intuitiv und verständlich dargestellt, erhöhen aber nicht die Transparenz politischer Vorgänge. Außerdem sind die Analysen auf festgelegte Verfahren zum Sammeln der Daten beschränkt. Am Beispiel des Twitterbarometers ist dieses Verfahren das Erkennen von Hashtags mit dem „+“ und „-“ Suffix. Da jedoch nicht jeder Twitter-Nutzer in dieser Form twittert, repräsentieren die Ergebnisse nur einen Ausschnitt der Twitterstimmung.

[3] http://buzzrank.de/
[4] http://twitterbarometer.de/
[5] http://bundestwitter.de/
[6] http://politwi.de/

3 Überblick über das Softwaresystem

Dieses Kapitel gibt einen Überblick über die einzelnen Komponenten des Softwaresystems, das der in dieser Arbeit vorgestellten Analyse zugrunde liegt. Abschnitt 3.1 präsentiert die einzelnen Komponenten und erklärt, wie diese interagieren. Abschnitt 3.2 enthält Informationen über die verwendete Technologie und erörtert die Architektur der Komponenten.

3.1 Die Komponenten und deren Zusammenhang

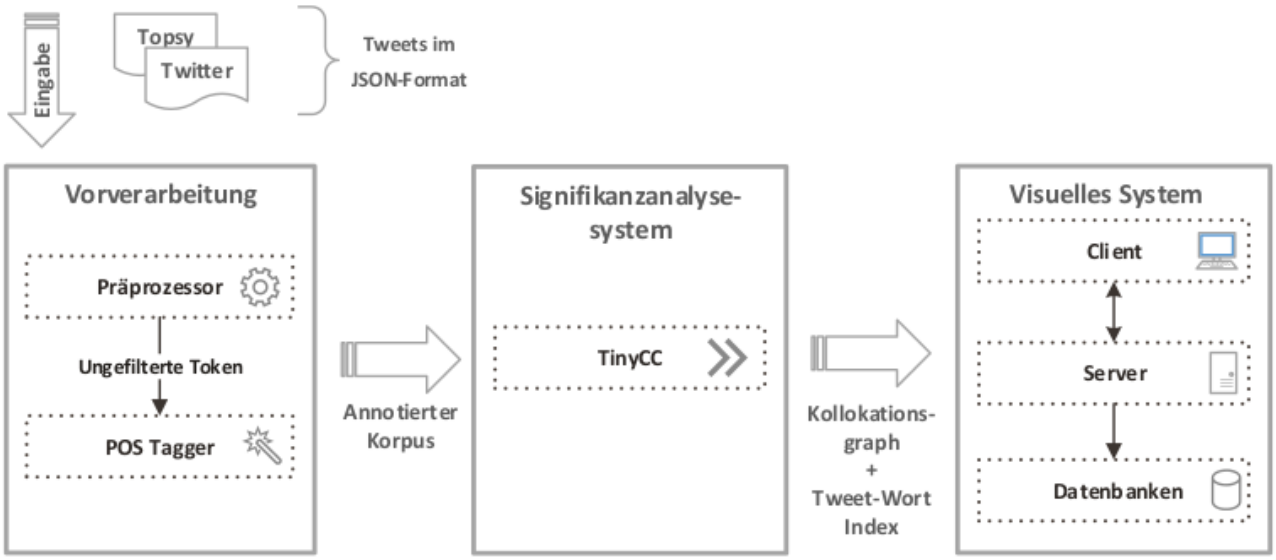


Abbildung 3.1: Übersicht über die Komponenten und den Datenfluss des Softwaresystems.

Abbildung 3.1 ist eine schematische Darstellung der im Softwaresystem vorhandenen Komponenten. Das System lässt sich in drei Teile gliedern: Der Präprozessor, der für das Annotieren des Korpus verantwortlich ist, die Signifikanzanalyse, welche aus dem annotierten Korpus Bigramme extrahiert und diese nach der Stärke ihrer Assoziation ordnet, sowie eine Komponente, die für die Visualisierung und Interaktion mit den Daten verantwortlich ist, um die Analyse durchzuführen.

Dieses Design hat die Einschränkung, dass der Struktur zur Laufzeit keine neuen Twitter-Daten hinzugefügt werden können. Die Daten müssen jeweils auf den gesamten Rohdaten berechnet und anschließend in die visuelle Komponente importiert werden. Bei großen Datenmengen nimmt diese Durchführung einige Zeit in Anspruch, was als Nachteil anzusehen ist. Verbesserungen dieses Designs werden in Kapitel 8 diskutiert.

Präprozessor

Der Präprozessor repräsentiert die erste Komponente im Softwaresystem und ist verantwortlich für das Annotieren des Korpus. Exakte Duplikate und Quasi-Duplikate haben einen enormen Einfluss auf die Ergebnisse der Analyse, so dass Retweets und exakte Duplikate vor der Verarbeitung mit einem Unix-Kommando entfernt werden.

Die Analyse der Rohdaten wird durch eine konfigurierbare Pipeline durchgeführt. Diese Pipeline enthält verschiedene Komponenten, die dem Korpus unterschiedliche Annotationen hinzufügen. Annotationen umfassen die Tokenisierung, Wortnormalisierung, Part-of-Speech (POS) Tagging, Spracherkennung und diverse Wortfilter. Eine detaillierte Beschreibung der Komponenten kann in Abschnitt 3.2 eingesehen werden. Der Präprozessor unterstützt drei unterschiedliche Eingabeformate: Das Twitter JSON-Format, das Topsy JSON-Format und Dateien, die zeilenweise Tweets enthalten.

Die Ausgabe des Präprozessors ist ein annotierter Korpus, der als Eingabe für das Signifikanzanalysesystem dient.

Signifikanzanalysesystem

Um Kollokationen zu finden und deren Signifikanz zu berechnen, werden Teile einer in der Universität Leipzig entwickelten Software[7] verwendet. Diese Kollokationen sind statistisch überzufällig häufige Wortkombinationen und die Grundlage für die in Abschnitt 5.4 vorgestellte kontrastive Analyse. Die Software erzeugt Textkorpora im Leipzig Corpus Collection (LLC) Format. Dieses von Quasthoff et al. [37] entwickelte Schema ist ein einfaches und flexibles Format, um einsprachige Sprachressourcen darzustellen und zu speichern.

Das LLC-Format wird erfolgreich im Wortschatz-Projekt[8] der Universität Leipzig eingesetzt, welches ermöglicht, 230 verschiedene einsprachige Wörterbücher zu durchsuchen. Neben Kollokationen umfasst das Format auch einen Index von Wörtern zu Tweets. Dieser wird genutzt, um Twitter-Nachrichten, die die identifizierten Kollokationen beinhalten, anzuzeigen. Dadurch ist es möglich, einen Einblick in den Kontext der Wörter zu erhalten. Die Identifikation von Kollokationen und die Berechnung der zugehörigen Signifikanz wird in Kapitel 5 näher beschrieben.

Die Ausgabe des Signifikanzanalysesystems ist eine tabulatorgetrennte dateibasierte Darstellung der Kollokationen mit zugehöriger Signifikanz, inklusive der durch den Präprozessor erzeugten Annotationen.

Visuelles System

Das System verwendet eine *Client-Server-Architektur* und besteht aus zwei verschiedene Datenbanken, in denen die Ausgabedaten des Signifikanzanalysesystems gespeichert sind. Es ist als Webapplikation implementiert, damit die Anwendung der breiten Öffentlichkeit zugänglich ist. Dadurch kann die Anwendung in jedem modernen Browser ausgeführt werden, ohne der Notwendigkeit, zusätzliche Software installieren zu müssen.

Datenbanken

Das System enthält zwei verschiedene Datenbanken, die sich in ihrem Speicherformat unterscheiden. Zum einen wird eine Graphdatenbank verwendet, die gemeinsam mit den Annotationen des Präprozessors den Kollokationsgraphen speichert. Dieser Kollokationsgraph ist die grundlegende Struktur der kontrastiven Analyse und wird in Abschnitt 5.3 vorgestellt. Das Speicherformat der Datenbank ist speziell für Graphstrukturen optimiert und eignet sich damit für die der Analyse zugrundeliegenden Struktur.

Zum anderen verwendet die Webapplikation eine relationale Datenbank, die in Tabellen organisiert ist. Diese Datenbank speichert die Zuordnung von Wörtern zu den Tweets, um den Kontext der Wörter darzustellen. Die Webapplikation bietet außerdem eine Möglichkeit, die Ausgabedaten des Signifikanzanalysesystems in die Datenbanken zu importieren.

[7] http://wortschatz.uni-leipzig.de/~cbiemann/software/TinyCC2.html
[8] http://corpora.informatik.uni-leipzig.de/

Server

Der Server ist das *Daten-Backend* des Systems und interagiert mit den beiden Datenbanken. Dieser akzeptiert Anfragen der Clients und führt diese bei Bedarf aus. Jede Anfrage wird neu auf der Datenbank ausgeführt. Das bedeutet, es werden keine Datenbankinformationen durch den Server zwischengespeichert oder beim Start des Servers geladen. Der Server speichert außerdem keine Sitzung für die jeweiligen Clients. Diese übertragen bei jeder Anfrage an den Server alle nötigen Informationen. Die Sitzung beschreibt eine bestehende Verbindung zwischen dem Server und dem Client und speichert Informationen über die Verbindung.

Client

Die Benutzerschnittstelle der Webapplikation in Abbildung 3.2 besteht aus einer Menge von Steuerelementen, die als Parameter für das Ausführen neuer Analysen dienen. In der Mitte der Webseite wird die Visualisierung der Analyse dargestellt. Zusätzlich wird die Möglichkeit geboten, die Diagramme in verschiedene Formate zu exportieren und individuelle Wörter in der Visualisierung auszublenden.

Details zur Visualisierung und eine Übersicht und Erklärung der verschiedenen Parameter finden sich in Abschnitt 6.

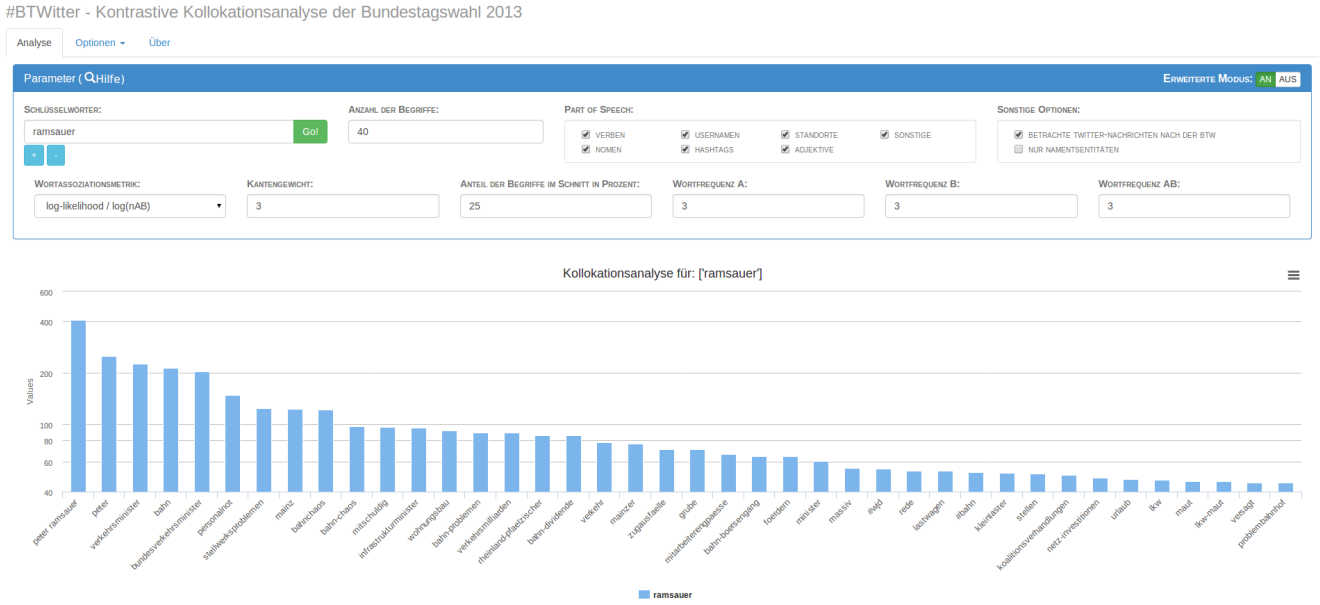


Abbildung 3.2: Die Benutzerschnittstelle der Webapplikation mit der Benutzer Analysen durchführen und betrachten können. Die Webapplikation unterstützt verschiedene Parameter, um das Erzeugen der Analyse zu beeinflussen.

3.2 Technologie und Architektur

Dieses Abschnitt beschreibt die Architektur der Komponenten und die verwendete Technologie, sowie Gründe für einzelne Designentscheidungen.

Präprozessor

Der Präprozessor ist in Python implementiert. Python ist eine interpretierte höhere Programmiersprache und unterstützt objektorientierte, aspektorientierte und funktionale Programmierparadigmen. Die Sprache besitzt ein dynamisches Typsystem und vor allem der funktionale Aspekt hat Vorteile, wenn große Datenmengen verarbeitet werden müssen.

Der Präprozessor unterstützt außerdem *Multiprocessing*. Im Gegensatz zum *Multithreading* werden die Daten in unterschiedlichen Prozessen verarbeitet, die sich keinen gemeinsamen Speicher teilen. Aufgrund des *Global Interpreter Lock*[9] kann nur ein *Thread* gleichzeitig Pythoncode ausführen. Deshalb muss Multiprocessing verwendet werden, um Rechenressourcen von Mehrkernrechnern zu nutzen. Dazu werden die Eingabedaten in Stapel unterteilt und an unterschiedliche *Worker* verteilt. Jeder Worker repräsentiert einen erzeugten Prozess der Anwendung.

Die Architektur des Präprozessors ist an das *Unstructured Information Management applications (UIMA) Framework*[10] angelehnt. UIMA ist ein von Apache[11] entwickeltes Softwaresystem, das es ermöglicht, große Mengen an unstrukturierten Informationen zu analysieren. Anwendungen, die auf dem UIMA Framework basieren, werden aus einzelnen Komponenten zusammengestellt, wobei jede dieser Komponenten eine bestimmte Aufgabe erfüllt [22].

Eine mögliche Abfolge von Komponenten lässt sich wie folgt definieren: Spracherkennung → Satzgrenzenbestimmung → Bestimmung der Entitäten (z.B. Personen, Orte und Organisationen) → Tokenisierung. Zwischen einigen Komponenten, wie dem Tokenisierer, existieren Abhängigkeiten. Dieser benötigt Satzgrenzen innerhalb des unstrukturierten Textes.

Der Präprozessor verwendet dieses komponentenbasierte Konzept, um die Twitter-Daten zu verarbeiten. Es bietet den Vorteil, dass einzelne Teile leicht ausgetauscht oder angepasst werden können. Den Komponenten wird ein sogenanntes *Common Analysis System (CAS)* übergeben, dem zusätzliche Informationen über den aktuellen Tweet hinzugefügt werden können. Außerdem besteht die Möglichkeit, bestehende Annotationen des CAS-Objektes für die Verarbeitung zu nutzen. Durch die Verwendung des CAS wird implizit ein einheitliches Eingabe- und Ausgabeformat zwischen den Komponenten definiert. Allgemein ist das CAS eine Repräsentation eines Artefakts, das analysiert werden soll. Das Artefakt für diesen Anwendungsfall ist ein Tweet, der im CAS-Objekt gespeichert wird und unveränderbar ist.

Es existieren jedoch Szenarien, in denen verschiedene Varianten eines Artefakts in unterschiedlichen Stufen der Verarbeitung benötigt werden. Dies ist in der Analyse eines Textes und seiner Übersetzung der Fall. Um diesen Anwendungsfall zu adressieren unterstützt das CAS unterschiedliche Sichten. Jede dieser Sichten enthält ein eigenes Artefakt, sowie eine eigene Menge an indizierten Features. Diese Features sind repräsentativ für erzeugte Annotationen und sind in einem Bereich des CAS gespeichert, der mit den Sichten geteilt wird. Einzig die Einträge des Index der individuellen Sicht sind einzigartig. Dadurch wird das redundante Speichern von Annotationen verhindert. Eine Komponente, die sich dieser Teilung des CAS bewusst ist, kann die entsprechende Repräsentation des Artefakts auslesen.

Die Implementierung des Präprozessors stellt die folgenden Komponenten zu Verfügung:

Collection Reader

Der *Collection Reader* dient dem Einlesen der Daten für die Pipeline. Jede Pipeline unterstützt nur einen Collection Reader. Dieser muss außerdem die erste Komponente in der Abfolge sein. Der Collec-

[9] http://docs.python.org/3/glossary.html#term-global-interpreter-lock
[10] http://uima.apache.org/
[11] http://apache.org/

tion Reader speichert den Tweet im CAS-Objekt und vergibt zusätzlich einen eindeutigen Identifikator für das Artefakt. Es kann zwischen drei verschiedenen Implementierungen des Collection Readers gewählt werden. Jede dieser Implementierungen unterstützt ein anderes Eingabeformat: Das Twitter JSON-Format, das Topsy JSON-Format und Dateien die zeilenweise Tweets enthalten. Da Twitter-Daten häufig HTML-Entitäten enthalten, konvertiert die Komponente diese außerdem in die entsprechende Textrepräsentation (z.B. „>“ anstelle von „>“).

Language Tagger

Der Spracherkenner[12] ist speziell für Kurznachrichtendienste wie Twitter optimiert und basiert auf einem *Infinity-Gram Modell* [34]. Dieser unterstützt 17 Sprachen und hat eine Genauigkeit von 99,1 Prozent. Der Spracherkenner fügt dem CAS-Objekt eine Annotation hinzu, welche die Sprache des Artefakts und damit die des Tweets enthält.

Tokenizer

Standard Tokenisierer sind nur für Zeitungen oder wissenschaftliche Arbeiten konzipiert und schneiden im Bereich Social-Media schlecht ab. Das Tokenisierermodul basiert auf dem Twitter-Tokenisierer von Gimpel et al. [21] und verwendet reguläre Ausdrücke. Diese erkennen Hashtags, „@“ Zeichen, Standardabkürzungen, Email-Adressen, Zeichenketten von Punktuation (z.B. „...“), Emoticons und Unicodezeichen (z.B. Musiknoten) als Token.

Normalizer

Ein häufig beobachtetes Phänomen im Bereich Social-Media, ist die Wiederholung eines Buchstaben innerhalb eines Wortes. Das Akronym *laughing out loud (lol)* findet sich in Twitter in verschiedenen Variationen, wie beispielsweise *lool, loool* oder *looool* und soll den Ausdruck der Emotion verstärken. Da diese Wörter grundsätzlich alle aus der Form *lol* abgeleitet sind, sollen die Variationen auf diese Grundform reduziert werden. Der implementierte Normalisierer führt eine einfache Reduzierung aller Buchstabenwiederholungen mit einer Frequenz von mehr als zwei durch. Diese Buchstabenwiederholungen werden auf zwei Buchstaben reduziert, um Konflikte bei Wörtern wie „hallo“ zu vermeiden. Der Normalisierer erzeugt jeweils eine *Error-Annotation* für die zu normalisierenden Token. Diese Annotation enthält ein Attribut, das eine Liste von Verbesserungen für dieses Token speichert. Einen regelbasierten Ansatz zur Normalisierung deutscher Twitter-Nachrichten präsentieren Sidarenka et al. [42].

Named Entity Tagger

Das Erkennen von Personen und Organisationen im Bereich natürlicher Sprachverarbeitung ist ein Klassifikationsproblem, das als *Named Entity Recognition (NER)* bezeichnet wird. Der NER Tagger basiert auf einer Wortliste, die aus Spitzenkandidaten der Wahl besteht. Wird ein Token als Person in der Liste identifiziert, erzeugt der Tagger eine neue *NER-Annotation* im CAS-Objekt. Nadeau und Sekine [31] bieten einen Überblick über Methoden im Bereich NER.

Stopword Tagger

Diese Komponente dient dem Erkennen und Annotieren von Stoppwörtern. Stoppwörter sind Funktionswörter (im linguistischen Sinn), die meist nur eine grammatikalische Funktion haben und damit keinen Aufschluss über den Inhalt geben. Im Deutschen übliche Stoppwörter sind bestimmte Artikel, unbestimmte Artikel, Konjunktionen und häufige Präpositionen. Der Tagger benötigt einen bereits tokenisierten Tweet im CAS-Objekt.

Punctuation Tagger

Diese Komponente dient dem Erkennen und Annotieren von Zeichensetzung. Die Funktionalität ist analog zum *Stopword Tagger* spezifiziert.

[12] https://github.com/shuyo/ldig

CAS-Consumer

Der *CAS-Consumer* befindet sich meist am Ende einer Pipeline und hat die Aufgabe das CAS-Objekt in einem gewünschten Format zu serialisieren. Der Präprozessor stellt zwei verschiedene CAS-Consumer zu Verfügung. Die erste Variante erzeugt die Ausgabe im Format für das Signifikanzanalysesysteme und die zweite Variante serialisiert die Annotationen in der *Extensible Markup Language (XML)* (Listing 3.1).

Der POS Tagger ist nicht Teil dieser Pipeline, sondern wird als externe Anwendung gestartet. POS Tagging beschreibt das Zuordnen von Worten und Satzzeichen zu Wortarten unter Berücksichtigung des Kontext. Forscher benutzen häufig den von Schmid [41] entwickelten *TreeTagger* (Sidarenka et al. [42], Neunerdt et al. [32]). Dieser basiert auf einem *Markov Modell* [45] und verwendet das *STTS Tagset* [40], das üblicherweise für NLP-Methoden genutzt wird. Rehbein [39] präsentiert hingegen einen speziellen POS Tagger für Twitter, der das STTS Tagset um weitere twitterspezifische Tags erweitert.

Als POS Tagger für die Vorverarbeitung wird das nicht überwachte POS Tagging-System von Biemann [4] verwendet, um Tokens mit Wortklassen zu annotieren. Diese Wortklassen sind repräsentativ für Wortarten. Im Gegensatz zum aktuellen Stand der Technik bestimmt der Algorithmus die Anzahl dieser Klassen. Das Verfahren bietet den Vorteil, dass die Wortklassen nicht auf herkömmliche Wortarten beschränkt sind. Es hat sich herausgestellt, dass einige der erzeugten Wortklassen Städte innerhalb Deutschland repräsentieren. Diese Wortklasse ist vor allem interessant für eine Betrachtung der Wahlkampforte von Politikern.

Der Workflow des Präprozessors lässt sich mit drei verschiedenen Arten von Komponenten beschreiben: Dem *Collection Reader*, der *Analyse-Engine* und dem *CAS-Consumer*. Die *Analyse-Engine* repräsentiert hierbei die einzelnen Komponenten zum Analysieren der Daten. Abbildung 3.3 zeigt die Verarbeitung im Präprozessor mit den vorgestellten Komponenten.

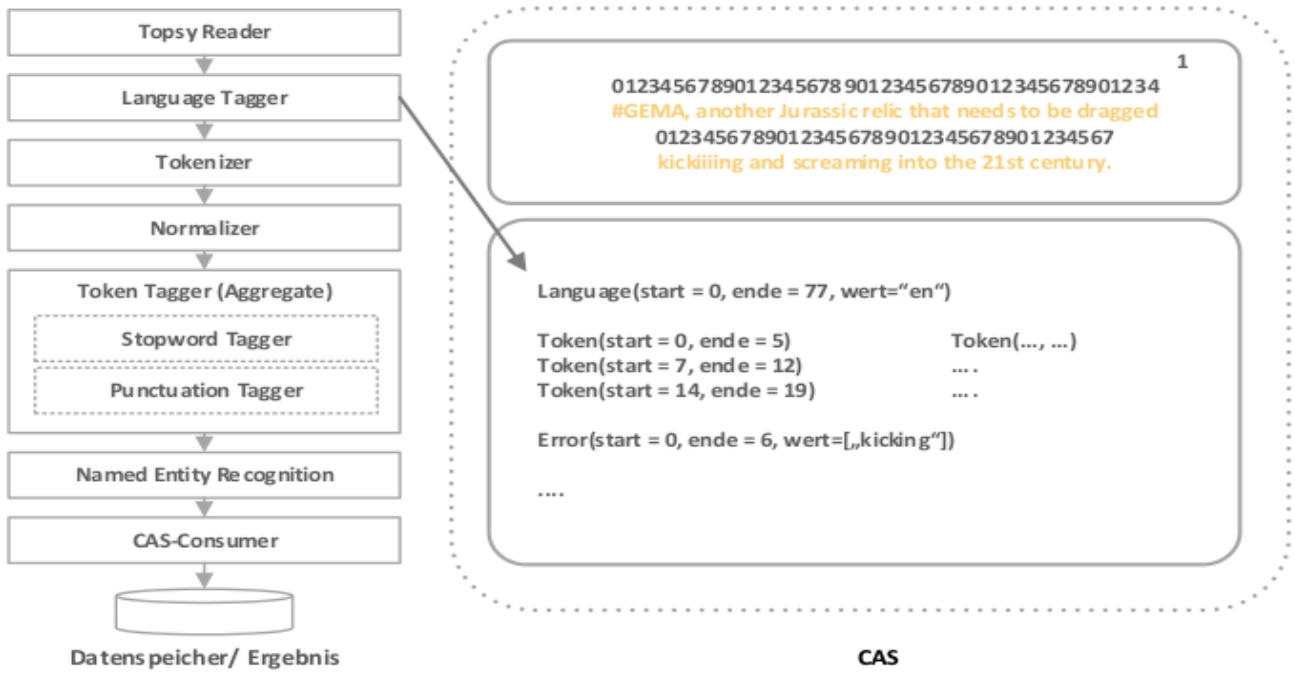


Abbildung 3.3: Präprozessor-Pipeline mit Annotationen anhand eines Beispieltweets. Die Abbildung ist inspiriert von der Vorlesung „Natural Language Processing and the Web" an der Technischen Universität Darmstadt[13].

[13] http://lt.informatik.tu-darmstadt.de/de/teaching/lectures-and-classes/natural-language-processing-and-the-web/

Der *Topsy Reader* speichert den Tweet als Artefakt im CAS-Objekt und vergibt einen eindeutigen Identifikator, der als Nummer repräsentiert wird. Die nächste Komponente in der Pipeline ist der *Language Tagger*. Dieser analysiert die Sprache des Tweets und fügt dem CAS-Objekt eine *Language-Annotation* hinzu. Die Language-Annotation referenziert mit den Attributen „Start“ und „Ende“ den Zeichenindexoffset innerhalb des Tweets und enthält im Attribut „Wert“ die Sprache des Tweets.

Dieses Vorgehen zum Annotieren der Daten wird auch *Standoff Markup* genannt und verändert im Gegensatz zum *Inline Markup* nicht das zu analysierende Dokument. Die Annotationen werden durch einen Zeichenindexoffset mit dem Artefakt verknüpft. Aus diesem Grund ist das Artefakt unveränderbar, denn eine Veränderung des Dokumentes beschädigt die Annotationen. Mit diesem Konzept lassen sich außerdem auch überlappende Annotationen modellieren.

Nach dem *Language Tagger* folgt der *Tokenizer*. Dieser fügt dem CAS-Objekt *Token-Annotationen* hinzu, wobei die Attribute „Start“ und „Ende“ genau angeben, welcher Text innerhalb des Tweets referenziert wird.

Der *Normalizer* erweitert die Menge der Annotationen um eine *Error-Annotation*, deren Wert das korrigierte Token enthält. Nach dem Normalizer folgt ein aggregiertes Modul, das den Stopword und den Punctuation Tagger enthält und wie eine normale Komponente verwendet werden kann. Jeder Annotator hat eine bestimmte Aufgabe und eine damit verbundene Verantwortlichkeit. Modularisierung ist wichtig für die Wiederverwendbarkeit der Komponenten. Wenn die Modularisierung jedoch zu feingranular wird, bekommt die Pipeline eine unnötige Komplexität. Dieses Problem wird durch sogenannte aggregierte Module gelöst, die sich wie normale Komponenten verhalten und eine Menge von individuellen Komponenten kapseln können. Der *Token Tagger* ist ein solches aggregiertes Modul und fügt dem CAS-Objekt *Stopword und Punctuation Annontationen* hinzu.

Die letzte Komponente der Analyse-Engine ist die NER. Diese entscheidet anhand des Tokens und einer Liste von Entitäten, ob es sich bei dem Token um eine Namensentität handelt und erzeugt eine *NER-Annotation*.

```
<dokument id="927">
	<text>@13DX110 pssst - das Wissen die #Piraten- doch nicht</text>
	<annotationen>
		<annotation start="0" ende="52" typ="Language" wert="de" \>
		<annotation start="0" ende="7" typ="Token" \>
		<annotation start="9" ende="13" typ="Token" \>
		<annotation start="15" ende="15" typ="Token" \>
		<annotation start="17" ende="19" typ="Token" \>
		<annotation start="21" ende="26" typ="Token" \>
		<annotation start="28" ende="30" typ="Token" \>
		<annotation start="32" ende="39" typ="Token" \>
		<annotation start="40" ende="40" typ="Token" \>
		<annotation start="42" ende="45" typ="Token" \>
		<annotation start="47" ende="51" typ="Token" \>
		<annotation start="9" ende="13" typ="Error" wert="psst" \>
		<annotation start="17" ende="19" typ="Stopword" \>
		<annotation start="28" ende="30" typ="Stopword" \>
		<annotation start="42" ende="45" typ="Stopword" \>
		<annotation start="47" ende="51" typ="Stopword" \>
		<annotation start="15" ende="15" typ="Punctuation" \>
		<annotation start="40" ende="40" typ="Punctuation" \>
	</annotationen>
</dokument>
```

Listing 3.1: Standoff Markup für den Tweet: „@13DX110 pssst - das Wissen die #Piraten- doch nicht“.

Am Ende der Pipeline steht der CAS-Consumer. Dieser wird genutzt, um die unterschiedlichen Eingabeformate für die folgenden Verarbeitungschritte zu generieren. Zwischenergebnisse und das annotierte Korpus werden in tabulatorgetrennten Dateien gespeichert. Das bietet den Vorteil der Verwendung von Kommandozeilenprogrammen, die Funktionen wie Sortieren, Suchen und Ersetzen, sowie Ausschneiden einzelner Spalten ausführen. Außerdem kann der Inhalt dieser Dateien im Gegensatz zu Datenbanken einfach gelesen werden.

Erzeugt wird eine zeilenweise nummerierte Datei mit den gefilterten Token für das Signifikanzanalysesystem, eine Datei mit den ungefilterten Token für den POS Tagger, sowie eine zeilenweise nummerierte Datei mit den unveränderten Tweets.

Der *Consumer* erzeugt außerdem eine Statistik über die verarbeiteten Daten. Diese Statistik enthält die Anzahl der Tweets, die Anzahl der Tokens, die Anzahl eindeutigen Hashtags, die Anzahl der eindeutigen Usernamen und die Anzahl der nicht deutschen Tweets.

Das Listing 3.1 zeigt beispielhaft für einen Tweet aus dem Korpus die vom Präprozessor erzeugten Annotationen in XML.

Signifikanzanalysesystem

Die an der Universität Leipzig entwickelte TinyCC-Software besteht aus einer Menge von Perl-Scripten und Unix-Kommandos. Die Software wird verwendet, da diese bereits alle benötigten Informationen berechnet und in einem geeigneten Format serialisiert. TinyCC teilt einen unstrukturierten Text in Sätze und erzeugt verschiedene tabulatorgetrennte Dateien:

- Sätze
- Wortliste mit Frequenz
- Index Wort-Satz
- Liste mit Quelldateien
- Index Quelldateien-Satz
- Signifikanz Nachbar Kollokationen
- Signifikanz Satz basierter Kollokationen

Der Software werden bereits tokenisierte Tweets aus dem Präprozessor übergeben und es wird ein Tokenisierer verwendet, der Wörter anhand der Leerzeichen trennt.

Außerdem erkennt das Signifikanzanalysesystem sogenannte Mehrworteinheiten, welche in der Analyse als ein gemeinsames Token betrachtet werden. Verwendet werden diese Mehrworteinheiten für die Vor-und Nachnamen von Politikern der Bundestagswahl.

Als Signifikanztest wird das von Dunning [13] eingeführte Log-Likelihood Maß verwendet. Von den tabulatorgetrennten Dateien werden der Wort-Satz Index und die satzbasierten Kollokationen mit zugehöriger Signifikanz genutzt. Nach Angaben der Autoren skaliert TinyCC auf mindestens 50 Millionen Sätzen (750 Millionen Wörter). Dies entspricht in etwa dem 71-fachen des aus einzigartigen Tweets bestehenden Korpus.

Visuelles System

Das visuelle System ist eine Kombination aus verschiedenen Technologien, um die kontrastive Analyse durchzuführen und zu visualisieren.

Datenbanken

Die Ausgabedaten des Signifikanzanalysesystems, welche die Kollokationen und deren Signifikanz enthalten, werden in eine Neo4J[14] Graphdatenbank importiert. Dieser Vorgang wird durch die Webapplikation unterstützt. Während des Importvorgangs kann außerdem eine Untergrenze für die Signi-

[14] `http://neo4j.org/`

fikanz importierter Relationen angegeben werden. Signifikanzwerte, die unter diese Schwelle fallen, werden beim Importieren verworfen, um die Menge an Relationen zu reduzieren. Um die Abfragen mit der Datenbank zu beschleunigen, wird zusätzlich ein Index für die Wörter in den Knoten des Graphen erzeugt, da die Anfragen der Clients diese enthalten.

Folgende Gründe waren ausschlaggebend, eine Graphdatenbank für das Speichern der Kollokationen und deren Annotationen gegenüber einer relationalen Datenbank zu bevorzugen:

(1) Darstellung von Graphen und Netzwerken mit Performance.

(2) Schnelle Traversierung anstelle von langsamen SQL-Queries mit vielen Operationen, um Tabellen zu verknüpfen.

(3) Skalierbarkeit: Mehrere Milliarden von Knoten, Relationen und Attributen auf einer einzelnen Java Virtual Machine (JVM).

Zum Speichern der Zuordnung der Wörter zu den einzelnen Tweets wird zusätzlich die relationale Datenbank Postgre-SQL[15] verwendet. Das Speicherformat eignet sich besonders gut für diese Art von Index. Sowohl die Graphdatenbank, als auch die relationale Datenbank werden nach dem Import der Daten nicht mehr geändert.

Server

Der Server ist in Python implementiert und nutzt das Flask[16] Framework. Flask ist ein in Python geschriebenes Mikro-Framework zur Entwicklung von Webanwendungen. Der größte Unterschied zu Web-Frameworks wie Django oder Play ist, dass Flask keine Komponenten zur Verfügung stellt, für die bereits Lösungen existieren, sondern das Integrieren bestehender Bibliotheken erlaubt. Damit bleibt der Kern des Frameworks einfach und erweiterbar.

Client

Der Client ist in Javascript implementiert, als *Frontend-Framework* wird Bootstrap[17] verwendet. Bootstrap ist eine Sammlung von Hilfsmitteln für die Gestaltung von Webanwendungen und enthält Vorlagen für Buttons, Tabellen und andere Navigationselemente. Für die Visualisierung wird die Javascript Bibliothek Highcharts[18] eingesetzt, die das Erstellen von interaktiven Diagrammen ermöglicht.

[15] http://postgresql.org/
[16] http://flask.pocoo.org/
[17] http://getbootstrap.com/
[18] http://highcharts.com/

4 Twitter Datenerfassung

Dieses Kapitel beschreibt den Vorgang zum Erfassen von Twitter-Daten für die Studie zur Bundestagswahl 2013. Diese Daten wurden sowohl von Twitter, als auch von Topsy[19] über einen Zeitraum vom 2. August 2013 bis zum 16. Oktober 2013 gesammelt. Insgesamt konnten so 10.524.367 deutschsprachige Tweets gesammelt werden. Die Datenerfassung lässt sich in drei Phasen unterteilen. Abschnitt 4.1 umfasst das Identifizieren von Suchbegriffen, die verwendet werden, um Twitter und Topsy zu durchsuchen. Der Abschnitt 4.2 beschreibt die Vorgehensweise zur Durchführung der Datenerfassung und erläutert die dafür verwendeten Mechanismen. Abschnitt 4.3 präsentiert verschiedene Statistiken über die gesammelten Daten und dient der Evaluation der Durchführung.

4.1 Vorbereitung

Für das Sammeln der Tweets von Twitter und Topsy müssen Suchbegriffe definiert werden, um den Umfang des Korpus festzulegen. Enthält ein Tweet mindestens einen dieser Suchbegriffe, wird der jeweilige Tweet in das Korpus aufgenommen. Die Suchbegriffe umfassen die sechs im Bundestag vertretenen Parteien (CDU/CSU, SPD, FDP, Bündnis 90/Die Grünen und Die Linke), deren Spitzenkandidaten und relevante Wahlkampfthemen aus deutschen Zeitungen. Eine vollständige Liste der Suchbegriffe kann im Anhang A eingesehen werden.

Das Vorgehen zur Datenerfassung unterscheidet sich zu Kaczmirek et al. [25] hinsichtlich der Suchbegriffe und Zusammenstellung des Korpus. Die Autoren aggregieren mit verschiedenen themenbezogenen Suchbegriffen unterschiedliche Korpora. Unter Berücksichtigung der Verfasser der Tweets, erzeugen Kaczmirek et al. insgesamt sechs Korpora. Neben Twitter-Daten verwenden diese auch Daten von Facebook. Diese Unterscheidung wird für die Datenerfassung nicht getroffen, sondern aus allen gesammelten Daten wird ein Korpus erzeugt.

4.2 Implementierung der Datenerfassung und Durchführung

Für die Datenerfassung wurde ein Twitter-Crawler entwickelt. Dies ist ein Programm, das auf dem *Tweepy-Modul*[20] basiert und die Plattform Twitter automatisch nach Nachrichten durchsucht. Das Tweepy-Modul bietet eine Schnittstelle für die *Twitter Search* und *Streaming API* und liefert Twitter-Nachrichten zu festgelegten Suchbegriffen. Der Crawler selbst serialisiert diese Nachrichten.

Die von Twitter bereitgestellten APIs unterscheiden sich wesentlich in ihrer Funktion und Verwendung und sollen nachfolgend vorgestellt werden.

Search API

Die Search API fokussiert auf Relevanz und nicht auf Vollständigkeit der Ergebnisse. Außerdem wird im Gegensatz zur Streaming API eine größere Menge an Operatoren unterstützt, um Ergebnisse zu filtern. Beispielsweise kann mit Emoticons die Konnotation der Tweets festgelegt werden. Ein trauriger Emoticon (z.B. „:(“) steht in diesem Zusammenhang für einen negativ konnotierten Tweet. Die Obergrenze an möglichen Suchbegriffen ist nicht dokumentiert. Es wird jedoch empfohlen, nicht mehr als zehn Begriffe zu verwenden. Wird eine Anfrage zu komplex formuliert, antwortet die API mit einem

[19] `http://topsy.com/`
[20] https://github.com/tweepy/tweepy

speziellen Fehlercode. Außerdem ist es Möglich, Tweets zu sammeln, die maximal vor einer Woche getwittert wurden.

Streaming API

Die Möglichkeit, Ergebnisse mit der Streaming API zu filtern sind eingeschränkt. Dafür konzentriert sich die API auf Vollständigkeit der Ergebnisse. Es ist möglich, bis zu 400 Suchbegriffe zu spezifizieren, um den aktuellen Twitterstream zu durchsuchen. Die Ergebnisse enthalten keine Twitter-Nachrichten der Vergangenheit. Das bedeutet, die durch die Streaming API zurückgelieferten Tweets sind aktuell in Twitter verfasste Nachrichten.

Zum Zeitpunkt dieser Arbeit hatte der Wahlkampf der Bundestagswahl 2013 bereits begonnen. Aus diesem Grund wird in einem ersten Schritt die Search API genutzt, um die Twitter-Nachrichten der vergangenen Woche zu aggregieren. Die Suche basiert auf den zuvor definierten Schlüsselwörtern. Zusätzlich wird der von der API angebotene Sprachparameter auf Deutsch eingestellt. Damit kann sichergestellt werden, dass nur deutsche Twitter-Nachrichten bezüglich der übergebenen Begriffe gesammelt werden.

Die von der API zurückgelieferten Twitter-Nachrichten sind im *JavaScript Object Notation (JSON)* Format kodiert, welches mit dem Crawler serialisiert wird. JSON ist eine kompakte, für Menschen leicht lesbare Textform und wird häufig für den Austausch von Daten zwischen Anwendungen verwendet. Tabelle 4.1 zeigt die wichtigsten Attribute der durch JSON dargestellten Tweets.

Nachdem die Twitter-Daten der vergangenen Woche gesammelt wurden, wird in einem zweiten Schritt die Streaming API verwendet. Dadurch kann eine aktuelle und möglichst vollständige Menge an Tweets bezogen auf die festgelegten Suchbegriffe aggregiert werden.

Attribut	**Bedeutung**
id	Eindeutige Integer-Repräsentation für die Identität des Users
created_at	Koordinierte Weltzeit *(Universal Time Coordinated)*, zu welcher der Tweet verfasst wurde
coordinates	Geografische Position des Nutzers
screen_name	Nutzername, der anderen Nutzern angezeigt wird
name	Name des Nutzers
text	Twitter-Nachricht
lang	Eine BCP 47 Sprachkennung [36]
retweet_count	Anzahl der Retweets
in_reply_to_screen_name	Orginalautor des Tweets, wenn es ein Retweet ist

Tabelle 4.1: Die wichtigsten Attribute der durch JSON dargestellten Twitter-Nachrichten.

4.3 Auswertung und Fehleranalyse

Abbildung 4.1 zeigt die Häufigkeit der Tweets verteilt über die Dauer der Datenerfassung. Die schwarze Kurve beschreibt die Daten, welche mit Hilfe der Twitter-API gesammelt wurden. Am Tag des Kanzler-TV-Duells, dem 2. September, wurden 56.603 Twitter-Nachrichten gesammelt. Das entspricht weniger als die Hälfte der von Twitter[21] für diesen Tag gemessenen 173.000 Tweets.

Das Problem ist auf die gewählten Suchbegriffe zurückzuführen. Viele Tweets, die in Zusammenhang mit öffentlichen Auftritten von Politikern stehen, enthalten spezielle Hashtags, welche in den Suchbegriffen fehlen. Außerdem wurde die Datenerfassung mit der Twitter-API nach der Bundestagswahl am 22. September beendet, weshalb keine Tweets nach diesem Datum gesammelt werden konnten.

Dadurch lassen sich zwar Tweets vor der Bundestagswahl betrachten, es können jedoch keine Vergleiche mit Tweets nach der Bundestagswahl durchgeführt werden. Dieser Vergleich könnte für die Analyse von Interesse sein und neue Erkenntnisse offenbaren.

Um die genannten Probleme zu adressieren, wird der Online-Dienst Topsy verwendet, welcher zertifizierter Twitter-Partner ist. Topsy stellt einen großen, aus mehreren Milliarden Tweets bestehenden Index zur Verfügung, der bis in das Jahr 2006 zurückreicht. Dieser Dienst hat gegenüber der Twitter-API den Vorteil, dass Twitter-Nachrichten gesammelt werden können, die mehr als eine Woche in der Vergangenheit liegen. Die Suchbegriffe werden um bundestagswahlspezifische Hashtags (Anhang A) erweitert und für das Durchsuchen von Topsy verwendet.

Die blaue Kurve beschreibt die durch Topsy gesammelten Tweets. Neben dem schon genannten Kanzler TV-Duell ist im Vergleich zu den durch Twitter gesammelten Daten auch der Wahltag am 22. September charakteristisch. Diese Erkenntnis ist in Übereinstimmung mit der von Twitter durchgeführten Analyse, so dass sich die Anpassung der Suchbegriffe als erfolgreich beschreiben lässt.

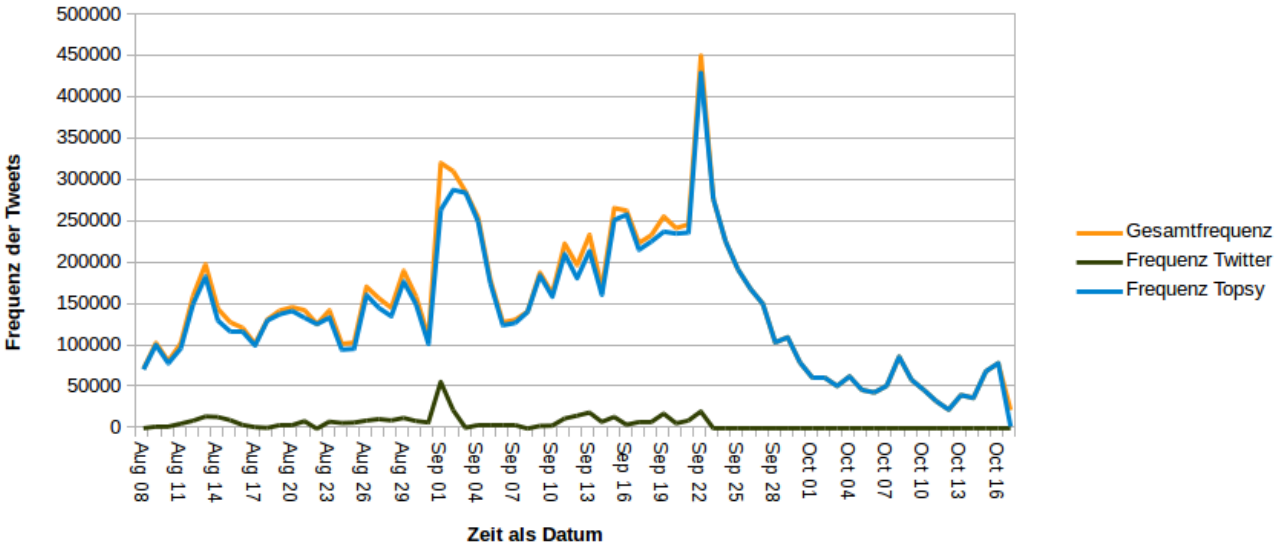


Abbildung 4.1: Häufigkeitsverteilung der gesammelten Twitter-Daten zwischen dem 2. August 2013 und 16. Oktober 2013. Die beiden orangen Spitzen korrelieren mit dem TV-Duell und dem Wahltag am 22 September.

[21] https://blog.twitter.com/2013/the-german-election-on-twitter

Das Korpus für die Studie besteht sowohl aus Daten der Twitter-API, als auch aus den neu gewonnenen Daten von Topsy. Tabelle 4.2 ist eine Statistik der ungefilterten Daten zwischen dem 2. August 2013 und 16. Oktober 2013.

Von den insgesamt 697.912 einzigartigen Tweets sind 21.553 Tweets nicht deutsch, trotz deutschem Sprachparameter. Gründe hierfür sind zum einen, dass Tweets zu kurz sind um ihre Sprache präzise erkennen zu können und URLs, Usernamen und Hashtags nicht hilfreich bei der Klassifizierung sind. Außerdem werden häufig Akronyme, Abkürzungen und in der Länge veränderte Worte wie „looool“ verwendet, was die Klassifizierung zusätzlich erschwert. Um die verbleibenden nicht deutschen Tweets zu entfernen, wird ein zusätzlicher Sprachfilter im Präprozessor verwendet. Damit konnten drei Prozent nicht deutsche Tweets identifiziert und entfernt werden.

Die Nutzernamenverweise in Tabelle 4.2 sind nicht als Autoren der Tweets zu verstehen, sondern beschreiben die Benutzeradressierungen (z.B. @peersteinbrück) innerhalb des Tweets. Mehr als die Hälfte aller Tweets in den Daten (54 Prozent), enthalten solch ein „@“ Zeichen. Diese Beobachtung ist in Übereinstimmung mit Honeycutt und Herring [23], welche herausfanden, dass die überwiegende Mehrheit der „@“ Zeichen genutzt wird, um einen Tweet direkt an einen bestimmten Nutzer zu richten.

Viele Twitter-Nutzer teilen Beiträge auch mit ihren Followers. Diese Retweets enthalten häufig Informationen, die der Nutzer interessant findet, wie zum Beispiel Links zu anderen Webseiten. Bezüglich der gesammelten Daten enthalten 39 Prozent aller Tweets einen solchen Link.

Die Anzahl der einzigartigen Links ist jedoch nicht repräsentativ, da in Twitter häufig sogenannte Kurzlinks verwendet werden. Kurzlinks sind ein Alias für beliebige Links existierender Webseiten und bestehen aus wenigen Buchstaben oder Zahlen. Vor allem in Twitter ist die Nutzung sehr verbreitet, da nur eine begrenzte Anzahl von Zeichen erlaubt ist. Dadurch können viele Kurzlinks die gleiche Webseite referenzieren und werden trotzdem von der Statistik als einzigartig repräsentiert.

Diese Studie hat einige Einschränkungen. Zum einen sind die gesammelten Daten limitiert auf Tweets, welche die zuvor festgelegten Suchbegriffe enthalten. Daher könnten Antworten auf Nachrichten, die zu einer Diskussion gehören, nicht gesammelt worden sein. Denn diese müssen nicht notwendigerweise den gleichen Suchbegriff in der Antwort enthalten. Trotz der Tatsache, dass ein Beitrag zu einem bestimmten Thema oder einer bestimmten Diskussion zuzuordnen ist, verwendet nicht jeder Twitter-Nutzer den gleichen Hashtag für dieselbe Diskussion.

Außerdem enthält der Topsy-Index nur relevante Tweets und ist keine vollständige Liste aller Tweets bezüglich eines Suchbegriffes. Diese Relevanz wird durch die Anzahl der Retweets und dem vergangenen Einfluss des Autors auf Twitter berechnet. Somit werden Tweets von Nachrichtendiensten bevorzugt, welche häufig über politische Themen twittern.

Merkmal	**Ausprägung**
Anzahl der Tweets	10.524.367
Anzahl der einzigartigen Tweets	697.912
Anzahl der Token aus einzigartigen Tweets	12.341.005
Anzahl einzigartiger Hashtags	52.614
Anzahl einzigartiger Nutzernamenverweise	37.899
Anzahl einzigartiger Links	274.669

Tabelle 4.2: Statistik über die ungefilterten Daten.

5 Signifikanzanalyse

Dieses Kapitel beschreibt den Hauptaspekt der Arbeit. Abschnitt 5.1 umfasst eine allgemeine Einführung in Kollokationen und motiviert die Verwendung von Kollokationen in der natürlichen Sprachverarbeitung. Der Abschnitt 5.2 befasst sich mit dem Finden von aussagekräftigen Kollokationen durch verschiedene Metriken und differenziert diese gegeneinander. In Abschnitt 5.3 wird der Kollokationsgraph eingeführt, auf dem die in 5.4 vorgestellte kontrastive Kollokationsanalyse basiert.

5.1 Definition und Einführung in den Begriff der Kollokation

Der Begriff der Kollokation wurde von Firth [18] eingeführt, welcher diesen als Ausdruck von zwei oder mehreren lexikalischen Einheiten (z.B. Wörter) in einer übergeordneten Einheit, wie zum Beispiel einem Satz, beschreibt (*"Collocations of a given word are statements of the habitual or customary places of that word.", Firth [18]*).

Nach dieser Definition ist eine Kollokation ein sprachliches Mittel, um etwas auf herkömmliche, dem Sprachgebrauch angepasste Weise, auszudrücken. Kollokationen umfassen unter anderem Nominalphrasen (NP), wie beispielsweise „starker Tee". Diese Kollokation meint jedoch nicht, dass der Tee physikalisch stark ist und schwere Dinge heben kann, sondern ist ein herkömmliches Sprachmittel, um die Intensität des Teegeschmacks zu charakterisieren. Aussage wie, „intensiver Tee" oder „kräftiger Tee" sind weniger gebräuchlich.

Die gesamte Bedeutung des Ausdrucks „starker Tee" lässt sich nicht aus der Bedeutung der einzelnen Wörter ableiten [3]. Diese Form der Kollokation wird auch als phraseologischer Mehrwortausdruck bezeichnet und lässt sich normalerweise nicht wortweise in eine andere Sprache übersetzen.

Man kann den Begriff der Kollokation hinsichtlich zweier unterschiedlicher Aspekte betrachten. Zum einen als vorgefertigte lexikalische Einheit, mit der, wie oben angeführt, etwas auf konventionelle Weise ausgedrückt werden kann (*"[...] a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.", Sinclair [43]*).

Zum anderen lässt sich der Begriff operationalisieren und als statistisch ausgeprägtes Auftreten von Kollokationen innerhalb des Korpus auffassen. Vor allem im Bereich der natürlichen Sprachverarbeitung spielen Kollokationen und deren empirische Quantität eine große Rolle.

Kollokationen werden daher als statistisch überzufällig häufige Wortpaare betrachtet, die innerhalb einer festgelegten übergeordneten Einheit beobachtet werden können. Eine Einheit der Größe ± 2 entspricht beispielsweise Wortbigrammen. Im Bereich Text Mining unterschiedet man neben dieser Nachbarschaftskollokation auch die Satzkollokation, bei der das Auftreten innerhalb des Satzes beobachtet wird. Für die Umsetzung dieser Arbeit werden Kollokationen auf Satzebene observiert. Dadurch kann für ein Wort jedes gemeinsame Auftreten mit anderen Wörtern innerhalb eines Tweets beobachtet werden. Eine zentrale Annahme hierbei ist, dass die Nähe der Wörter zueinander ein Indikator für die Beziehung zwischen diesen ist.

5.2 Maße für Wortassoziationen aus der Informationstheorie und Statistik

Um Kollokationen zu finden, muss bestimmt werden, ob das gemeinsame Auftreten von Wortpaaren zufälliger Natur ist, oder ob zwischen diesen beiden Wörtern eine statistische Abhängigkeit existiert. Für diesen Anwendungsfall können verschiedene Maße aus der Informationstheorie und Statsitik verwendet werden. Mit Hilfe dieser Maße werden aussagekräftige Kollokationen aus dem Korpus extrahiert und gemäß der Stärke ihrer Assoziation geordnet. Nachfolgend werden drei dieser Maße vorgestellt und deren Vor-und Nachteile beschrieben. Eine detailliertere Übersicht und Beschreibung kann in Manning und Schütze [28] eingesehen werden.

Bigrammfrequenz

Die denkbar einfachste Möglichkeit, um Kollokationen in einem Korpus zu finden, ist die Frequenz von Wortbigrammen zu betrachten. Die Frequenz des Bigramms $w_1 w_2$ ist gegeben als n_{AB}.

Tabelle 5.1 zeigt die observierte Frequenz der häufigsten Bigramme im Brown Korpus [19]. Der Brown Korpus wurde 1961 an der Brown Universität erstellt und enthält etwa eine Million Wörter, extrahiert aus verschiedenen Textgenres *(News, Fiction, Adventure, Government)*.

w_1	w_2	n_{AB}	Rang
of	the	9774	1
in	the	6156	2
to	the	3525	3
on	the	2491	4
and	the	2307	5
for	the	1862	6
to	be	1718	7
at	the	1677	8
with	the	1545	9
of	a	1502	10

Tabelle 5.1: Die zehn häufigsten Bigramme extrahiert aus dem Brown Korpus.

Das Problem bei diesem Vorgehen ist, dass viele der dadurch bestimmten Wörter Stoppwörter sind und keine aussagekräftigen Kollokationen bilden. Justeson und Katz [24] präsentiert eine simple Heuristik, welche diese häufigen Wortpaare entfernt und die Ergebnisse verbessert. Dieser Ansatz basiert auf einem POS Filter, der bestimmte Phrasenmuster erkennt und alle anderen Kollokationen entfernt.

Pointwise Mutal Informationen

Pointwise Mutal Information (PMI) [9] ist ein Maß, um aussagekräftige Kollokationen zu bestimmen und entstammt der Informationstheorie und Statistik. Allgemein beschreibt PMI, wie viel ein Wort über ein anderes Wort aussagt. Diese wage Aussage wird nachfolgend konkretisiert.

Gegeben zwei Zufallsvariablen X und Y und zwei mögliche Ergebnisse $X = x$ und $Y = y$, beschreibt PMI, ob die Ereignisse x und y unabhängig sind (PMI=0), positiv korreliert (PMI>0) oder negativ korreliert sind (PMI<0).

Mit den beiden Ereignissen x und y, welche das Auftreten zweier Wörter beschreiben, lässt sich PMI wie folgt definieren:

$$
\begin{aligned}
pmi(x,y) &= log_2 \frac{P(xy)}{P(x) \cdot P(y)} \\
&= log_2 \frac{P(x|y)}{P(x)} \\
&= log_2 \frac{P(y|x)}{P(y)}
\end{aligned}
\tag{5.1}
$$

Angenommen für die beiden Wörter $w_1 = Merkel$ und $w_2 = Berlin$ wird das Ergebnis $pmi(w_1, w_2) = 15.3$ erhalten. PMI gibt demnach an, dass der Informationsgehalt über das Auftreten des Wortes „Merkel“ an der Position i mit dem Auftreten des Wortes „Berlin“ an der Position $i+1$ um 15.3 Bits ansteigt. Dies gilt auch für den umgekehrten Fall, denn nach Gleichung 5.1 ist das Maß symmetrisch. Das bedeutet, es gilt:

$$pmi(x,y) = pmi(y,x)$$

PMI ist jedoch in vielen Fällen kein gutes Maß für aussagekräftige Kollokationen, wie die nachfolgende Diskussion zeigen soll. Der Nenner $P(x) \cdot P(y)$ ist der Erwartungswert für $P(x,y)$ unter der Annahme, das x und y unabhängig sind. Der Wert für $pmi(x,y)$ wird groß, wenn $P(x|y)$ konstant bleibt und $p(x)$ kleiner wird. Eine Folge davon ist, dass Wörter die nur einmal im Korpus enthalten sind, jedoch zusammen auftreten, einen großen PMI zugeordnet bekommen. Das bedeutet, ein großer PMI muss nicht zwangsläufig ein Indikator für die starke Assoziation von Bigrammen sein.

Ein weiterer Nachteil ist die Abhängigkeit zwischen den Wörtern. Angenommen zwei Wörter sind statistisch perfekt abhängig. Das bedeutet, wenn das eine Wort auftritt, tritt das andere Wort auch auf. Für die Gleichung 5.1 folgt:

$$pmi(x,y) = log_2 \frac{1}{P(y)}$$

Je seltener das Wort ist, desto größer ist der zugehörige PMI Wert. Ein großer PMI muss demnach nicht zwangsläufig eine große Abhängigkeit zwischen den Wörtern beschreiben, sondern kann vielmehr seltenere Wörter extrahieren. Eine Möglichkeit, um dieses Problem zu beheben ist einen Schwellwert für die Wortfrequenz festzulegen. Dieser Ansatz ist jedoch keine Lösung für das dem PMI zugrundeliegende Problem.

Tabelle 5.2 ist ein Vergleich von extrahierten Bigrammen mit dem PMI Maß aus dem Twitterkorpus. Betrachtet wird außerdem der Einfluss eines Schwellwert für die Frequenz der Wörter auf PMI. Die Spalten n_A und n_B beschreiben zusätzlich zur Bigrammfrequenz n_{AB} die individuelle Wortfrequenz der Wörter w_1 und w_2.

Wird PMI ohne den Schwellwert verwendet, werden besonders seltene Kollokationen gefunden. Diese Kollokationen sind jedoch nicht aussagekräftig für die Bundestagswahl.

Das Ergebnis für den Frequenzschwellwert enthält zwar noch Kollokationen die nicht aussagekräftig für die Bundestagswahl sind (z.B. „justin – bieber“), beinhaltet im Vergleich aber auch bundestagswahlspezifische Kollokationen (z.B. „krankenversicherer – fdp-gesundheitsminster“, „buergerinnenhand – energiewende-charta“).

Die Kollokation „leidig – sabine leidig“ zeigt eine Mehrworteinheit, welche in der Signifikanzanalyse als Einheit betrachtet wird.

PMI				PMI mit Frequenzschwellwert			
Bigramm w_1w_2	n_A	n_B	n_{AB}	**Bigramm w_1w_2**	n_A	n_B	n_{AB}
ichwaehlesienicht ichwerde	12	2	2	leidig sabine leidig	64	61	59
justinbieber marília	8	2	2	krankenversicherer fdp-gesundheitsminister	66	63	61
icky ptang	8	2	2	buergerinnenhand energiewende-charta	68	65	61
developers krabbenburger	7	2	2	spionen fdp-cryptoparty	70	64	58
badeinsel intex	6	4	2	zweistelligen milliardenbetrag	69	62	55
gammeligen untermieter	4	2	2	nervenzusammenbruch kanzler-amt	89	86	83
europapokaal europapokal	4	2	2	costa concordia	82	72	64
europapokaal fesselt	4	2	2	baerbel baerbel hoehn	84	61	60
entgiften smoothies	4	2	2	stadtbildes moscheen	84	84	83
entgiften 3-tage	4	2	2	justin bieber	78	76	62

Tabelle 5.2: Zehn Bigramme extrahiert aus dem Twitterkorpus der Bundestagswahl 2013 und geordnet mit PMI. PMI wird sowohl mit und ohne Frequenzschwellwert betrachtet. Es gilt: $n_A > 60 \wedge n_B > 60$.

Log-Likelihood Ratio

Das von Dunning [13] eingeführte Log-Likelihood Maß wird häufig in der natürlichen Sprachverarbeitung verwendet, um die Stärke von Assoziationen, speziell lexikalischen Assoziationen, zu bestimmen. Wenn zwei Wörter eines häufigen Bigramms selbst sehr häufig beobachten werden können, ist die Wahrscheinlichkeit hoch, dass beide Wörter häufig auftreten, auch wenn sie keine aussagekräftige Kollokation bilden. Dieses Problem adressiert das Log-Likelihood Maß, welches zu den Hypothesentests zählt [33]. Für zwei Wörter soll bestimmt werden, ob diese statistisch überzufällig häufig auftreten. Statistisch überzufällig bedeutet, dass die beiden Wörter häufiger beobachtet werden können, als man dies basierend auf ihren individuellen Frequenzen erwarten würde. Für das Finden von Kollokationen lassen sich nach Dunning [13] die folgenden beiden Hypothesen für die Auftrittshäufigkeit eines Bigramms w_1w_2 formulieren:

$$\begin{aligned} &\text{Hypothese } H_0: \quad P(w_2|w_1) = p = P(w_2|\neg w_1) \\ &\text{Hypothese } H_1: \quad P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1) \end{aligned} \tag{5.2}$$

Die erste Hypothese H_0 beschreibt, dass das Auftreten des Wortes w_2 unabhängig von dem zuvor aufgetretenen Wort w_1 ist. Die zweite Hypothese H_1 stellt eine Formalisierung der Abhängigkeit zwischen w_1 und w_2 dar und ist damit ein guter Indikator für aussagekräftige Kollokationen.

Für die Berechnung von p, p_1 und p_2 wird die *Maximum-Likelihood-Schätzungen* verwendet. Damit ist das Log-Likelihood Maß gegeben als:

$$-2 \cdot log\lambda = 2 \cdot log \frac{L(H_0)}{L(H_1)} \tag{5.3}$$

Hierbei ist $L(H_i)$ für $i \in \{0,1\}$ die Likelihood Wahrscheinlichkeit von H_i und kann mit der Annahme einer Binominialverteilung berechnet werden. Damit folgt für den binominialen Fall schließlich:

$$-2 \cdot \log\lambda = 2 \cdot \log \frac{L(H_0)}{L(H_1)} = 2 \cdot \begin{bmatrix} n \cdot \log n - n_A \log n_A - n_B \log n_B + n_{AB} \log n_{AB} \\ +(n - n_A - n_B + n_{AB}) \log(n - n_A - n_B + n_{AB}) \\ +(n_A - n_{AB}) \log(n_A - n_{AB}) \\ +(n_B - n_{AB}) \log(n_B - n_{AB}) \\ -(n - n_A) \log(n - n_A) - (n - n_B) \log(n - n_B) \end{bmatrix} \tag{5.4}$$

Ein Vorteil des Log-Likelihood Ratio Tests gegenüber anderen Hypothesentests ist die intuitive Interpretation. Der Wert ist eine Zahl, die angibt welche Hypothese wahrscheinlicher ist. Dadurch ist das Ergebnis beispielsweise einfacher zu deuten, als das des t-Tests, bei dem das Ergebnis in einer Tabelle nachgeschlagen werden muss.

$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{n_{AB}}$	**Log-Likelihood**	**Rang**
angela	angela merkel	7273	7454.74	1
peer	peer steinbrueck	5940	6331.7	2
angela	merkel	7439	4113.28	3
peer	steinbrueck	6228	3208.49	4
stimme	#twitterwahlen	2457	2819.49	5
grosse	koalition	3131	2729.68	6
NUMBER	uhr	6303	2705.75	7
steinbrueck	merkel	8812	2600.39	8
#cdu	#csu	10182	2499.56	9
#piratenpartei	#forum	1774	2489.09	10

Tabelle 5.3: Zehn Bigramme extrahiert aus dem Twitterkorpus der Bundestagswahl 2013 und geordnet mit Log-Likelihood.

Tabelle 5.3 zeigt die ersten zehn mit dem Log-Likelihood Maß geordneten Kollokationen. Diese Kollokationen sind aus dem gesamten Korpus extrahiert, der den Zeitraum nach der Bundestagswahl einschließt. Aus diesem Grund sind auch die Wörter des Bigramms „grosse – koalition“ sehr stark assoziiert. Die beiden Spitzenkandidaten „Angela Merkel“ und „Peer Steinbrück“ sind die am stärksten assoziierten Wortpaare, gefolgt von der Kollokation „stimme – #twitterwahlen“, die am Tag der Wahl von sehr vielen Benutzern verwendet wurde, um über Twitter ihre Präferenz für die jeweilige Partei abzugeben. Das Wort *_NUMBER_* ist ein Platzhalter für observierte Zahlen, die während der Signifikanzanalyse durch dieses Wort ersetzt werden.

5.3 Definition des Wortkollokationsgraphen

Kollokationen eines Korpus lassen sich grafisch in einem sogenannten Kollokationsgraphen darstellen. Werden Wörter eines Korpus als Knoten betrachtet und die Anzahl des Auftretens zweier Wörter als Kanten interpretiert, ergibt sich der Wortkollokationsgraph eines Korpus aus der Menge aller Wortkollokationen.

Für die Darstellung dieses Wortkollokationsgraphen wird ein ungerichteter, gewichteter Graph $G = (V, E)$ verwendet. Die Knoten V und Kanten E sind wie folgt definiert:

$$E \subseteq V^2 \times \mathbb{N} \times \{\text{Vorher}, \text{Nachher}\} \times \mathbb{Z}^4$$
$$V \subseteq K \times \mathbb{N}^2 \times \{0, 1\}$$

Knoten V:

Jeder Knoten $v = (\text{Identifikator}, \text{Wortfrequenz}, \text{Wortklasse}, \text{Namensentität}) \in V$ ist ein 4-Tupel. Dieses 4-Tupel besteht aus einem einzigartigen Identifikator, der das Wort repräsentiert, der Frequenz des Wortes im Korpus, einer Zahl welche die Zugehörigkeit des Wortes zu einer Wortklasse modelliert, sowie einer Zahl aus der Menge {0, 1}. Die Zahl 1 gibt an, dass es sich bei dem Wort um eine Namensentität handelt, wohingegen die Zahl 0 das Gegenteil modelliert. Die Menge K repräsentiert die Menge aller Wörter im Twitterkorpus.

Kanten E:

Jede Kante $e = (\text{Knoten}_1, \text{Knoten}_2, \text{Kantengewicht}, \text{Relation}, \text{Likelihood}, \text{Likelihood}_{norm}, \text{Pmi}, \text{Lmi}) \in E$ ist ein 8-Tupel. Dieses 8-Tupel enthält die durch die Kante verbundenen Knoten, das Kantengewicht, einen Relationsnamen und vier unterschiedliche Signifikanzmaße. Das Kantengewicht gibt an, wie häufig eine Kollokation im Korpus observiert wurde. Die Signifikanzmaße werden genutzt, um die Kollokationen für die kontrastive Analyse zu ordnen. Die Relation beschreibt, ob eine Kollokation repräsentiert durch zwei verbundene Knoten, vor oder nach der Bundestagswahl 2013 beobachtet wurde. So wird ermöglicht, dass Twitter-Nachrichten nach der Bundestagswahl in die Ergebnissen aufgenommen oder ausgeschlossen werden können.

Es ist möglich, diese Graphstruktur zu ersetzen, um weitere Annotationen hinzuzufügen. Abbildung 5.1 zeigt einen Ausschnitt des Kollokationsgraphen für den Tweet „Merkel und Steinbrück live #tvduell“. Für die Knoten 1 und 2, welche die Wörter „Merkel“ und „live“ repräsentieren, sind die Eigenschaften der Kante exemplarisch aufgetragen. Der Relationsname „Vorher“ beschreibt in diesem Zusammenhang, dass die beiden Wörter mit den zugehörigen Eigenschaften vor der Bundestagswahl observiert wurden und das Kantengewicht beschreibt die Häufigkeit dieser Observation. Der große Wert des Log-Likelihood Maßes für die Wörter „Merkel“ und „live“ zeigt eine statistische Abhängigkeit, womit sich die beiden Wörter als besonders stark assoziiert beschreiben lassen.

Die hier vorgestellte Struktur ist die Grundlage der in Abschnitt 5.4 eingeführten Analyse und muss auf dem gesamten Korpus vorberechnet werden. Der Kollokationsgraph selbst ist in einer Graphdatenbank persistiert.

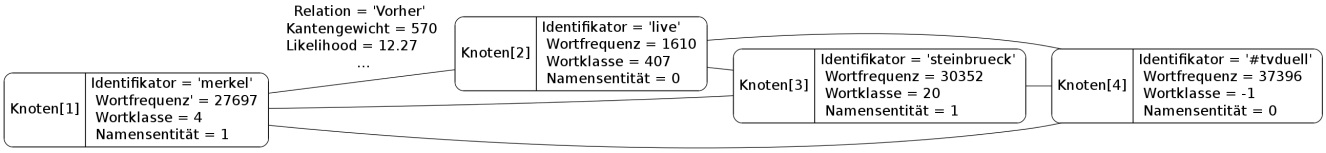


Abbildung 5.1: Ausschnitt des Kollokationsgraphen für den Tweet: „Merkel und Steinbrück live #tvduell“.

5.4 Kontrastive Kollokationsanalyse

Die vorgestellte Analyse verwendet das Konzept der Kollokationen, betrachtet diese jedoch im Gegensatz zu bisherigen Ansätzen nicht als Einheit. Für ein beliebiges, aber fest gewähltes Wort *A* werden alle Wörter *B* betrachtet, in denen *A* als Teil der Kollokation „A – B“ vorkommt. Durch dieses Vorgehen lässt sich eine Menge bilden, die solche Wörter enthält, die besonders stark mit dem Wort *A* assoziiert sind. Dieser Ansatz wird auch in anderen Forschungsgebieten verwendet, um Wortbedeutungen zu erhalten. Mehr Informationen zu diesem Thema bietet Bordag [7].

Um aussagekräftige Kollokationen zu finden und zu ordnen, werden die in Kapitel 5.2 vorgestellten Signifikanzmaße genutzt. Tabelle 5.4 zeigt die zehn am stärksten assoziierten Begriffe mit der Bundeskanzlerin, Angela Merkel. Diese Begriffe wurden mit Hilfe von PMI und einer normierten Variation des Log-Likelihood Maß geordnet. Für PMI wird außerdem gezeigt, wie sich der Frequenzschwellwert auf die Ergebnisse auswirkt. Die Spalte f(Wort) beschreibt die Frequenz des Wortes. Mit PMI werden solche Wörter bestimmt, die besonders selten sind. Durch einen Wortfrequenzschwellwert von 60 kann eine teilweise Übereinstimmung mit den Wörtern, die durch das Log-Likelihood Maß gefunden werden, erreicht werden.

Durch die hier vorgestellte Betrachtung von Kollokationen ist es möglich, zwei Schlüsselwörter durch ihre besonders stark assoziierten Wörter zu kontrastieren. Dieses Methode wird kontrastive Kollokationsanalyse genannt und nachfolgend vorstellt.

Log-Likelihood		*PMI*		*PMI mit Schwellwert*	
Wort	**f(Wort)**	**Wort**	**f(Wort)**	**Wort**	**f(Wort)**
angela merkel	7273	deutschrap	2	christlich-vegetarische	61
angela	7439	entwickelten	2	angela merkel	7273
steinbrueck	8812	fueherschaft	2	moscheen	83
#tvduell	6141	radiohoerer	2	stadtbildes	83
tv-duell	1717	parolieren	2	volksprofessor	80
bundeskanzlerin	1026	steinschleuder	2	gruenen-anhaenger	67
kanzlerin	1444	unbequemeren	2	angela	7439
frau	1610	juso-landesdelegierten	2	bundeskanzlerin	1026
raab	655	gewinkt	2	tv-duells	81
vertrauen	536	angela merkel	2	fernsehduell	102

Tabelle 5.4: Die zehn am stärksten mit der Bundeskanzlerin Angela Merkel assoziierten Begriffe. Die Begriffe sind geordnet mit Log-Likelihood, PMI und PMI mit einem Frequenzschwellwert von f(Wort) > 60.

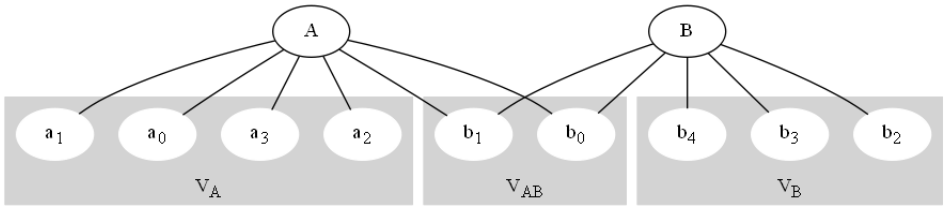


Abbildung 5.2: Schematische Abbildung des Kollokationsgraphen für die Schlüsselwörter A und B, sowie die zugehörigen Wortgruppen V_A, V_B, V_{AB}.

Die kontrastive Kollokationsanalyse ermöglicht es, sowohl Unterschiede als auch Gemeinsamkeiten zwischen zwei gegebenen Entitäten, wie beispielsweise Politikern, zu analysieren. Dazu werden basierend auf zwei gegebenen Schlüsselwörtern A und B drei Gruppen von Wörtern (V_{AB}, V_A, V_B) gebildet. Abbildung 5.2 zeigt für diese Schlüsselwörter den schematischen Kollokationsgraphen und skizziert die drei Wortgruppen.

Diese Gruppen sind in Gleichung 5.5 definiert. Die Definitionen basieren auf der Struktur des Kollokationsgraphen aus Abschnitt 5.3 und den folgenden Hilfsfunktionen:

$$\begin{aligned}
\forall\, e = (\text{Knoten}_1, \text{Knoten}_2, \text{Kantengewicht}, \text{Relation}, \text{Likelihood}, \text{Likelihood}_{norm}, \text{Pmi}, \text{Lmi}) \in E \\
start(e) &= \text{Knoten}_1 \\
ende(e) &= \text{Knoten}_2 \\
lmi(e) &= \text{Lmi} \\
pmi(e) &= \text{Pmi} \\
ll(e) &= \text{Likelihood} \\
ll_{norm}(e) &= \text{Likelihood}_{norm}
\end{aligned}$$

$$\begin{aligned}
\text{Seien } v_1, v_2 \in V : verbunden(v_1, v_2) \Leftrightarrow \exists\, e \in E : (start(e) = v_1 \wedge ende(e) = v_2) \vee \\
(start(e) = v_2 \wedge ende(e) = v_1)
\end{aligned}$$

$$\begin{aligned}
v \in V_{AB} &\iff (verbunden(v,A) \wedge verbunden(v,B)) \\
v \in V_A &\iff (verbunden(v,A) \wedge v \notin V_{AB}) \\
&\iff (verbunden(v,A) \wedge \neg verbunden(v,B)) \\
v \in V_B &\iff (verbunden(v,B) \wedge v \notin V_{AB}) \\
&\iff (verbunden(v,B) \wedge \neg verbunden(v,B))
\end{aligned} \tag{5.5}$$

$$\begin{aligned}
<_{pmi} &= \{(e_1, e_2) \in E^2 : pmi(e_1) < pmi(e_2)\} \\
<_{lmi} &= \{(e_1, e_2) \in E^2 : lmi(e_1) < lmi(e_2)\} \\
<_{ll} &= \{(e_1, e_2) \in E^2 : ll(e_1) < ll(e_2)\} \\
<_{ll_{norm}} &= \{(e_1, e_2) \in E^2 : ll_{norm}(e_1) < ll_{norm}(e_2)\}
\end{aligned}$$

V_A modelliert Wörter, die besonders stark mit dem Schlüsselwort A assoziiert sind, jedoch nicht mit dem Wort B. V_B ist analog modelliert. Die beiden Wortgruppen entsprechen damit dem Unterschied der Wörter A und B, welcher durch die jeweils stark assoziierten Wörter ausgedrückt wird.

V_{AB} hingegen beschreibt Wörter, die besonders stark mit dem Schlüsselwort A und B assoziiert sind und stellt damit die Gemeinsamkeiten der beiden Schlüsselwörter dar. Die Elemente von V_A, V_B und V_{AB} werden entsprechend dem ausgewähltem Signifikanzmaß nach $<_i$ mit $i \in \{pmi, lmi, ll, ll_{norm}\}$ geordnet.

Diese Gruppen werden in einem Tornadodiagramm visualisiert, um Gemeinsamkeiten und Unterschiede anschaulich darzustellen. Details zu dieser Visualisierung und Skalierung der Daten werden in Kapitel 6 beschrieben.

Die Technik wird am Beispiel der Schlüsselwörter „Rainer Brüderle“ und „Jürgen Trittin“ demonstriert. Das System ordnet die assoziierten Begriffe der beiden Schlüsselwörter anhand der Stärke ihrer Assoziation. Rainer Brüderle ist ein deutscher Politiker und Spitzenkandidat der FDP. Jürgen Trittin ist ebenfalls Politiker und Mitglied der Grünen.

Abbildung 5.3 zeigt die Visualisierung der kontrastiven Kollokationsanalyse zwischen den beiden Politikern. Zum ordnen der Ergebnisse wird die normalisierte Form des Log-Likelihood Maß verwendet und der Schwellwert für das Kantengewicht auf drei eingestellt. Ein minimales Kantengewicht von drei bedeutet, dass beide Wörter mindestens dreimal als Bigramm im Korpus enthalten sein müssen. Außerdem werden Usernamen ausgeblendet und Tweets nach der Bundestagswahl in die Ergebnisse der Analyse aufgenommen.

Die Wörter auf der linken Seite des Graphen sind Wörter, die besonders stark mit Rainer Brüderle assoziiert sind (V_A), wohingegen Wörter auf der rechten Seite stark mit Jürgen Trittin assoziiert sind (V_B). Der Schnitt in der Mitte des Graphen beschreibt Wörter, die sowohl mit Rainer Brüderle, als auch Jürgen Trittin assoziiert sind (V_{AB}). Die Ausprägung der Balken gibt jeweils die Stärke der Assoziation des gewählten Signifikanzmaßes an. Die Ausrichtung der Balken repräsentiert nicht das Vorzeichen des Signifikanzmaßes, sondern gibt die Zugehörigkeit des Wortes zum jeweiligen Schlüsselwort an. Es werden nur solche Wörter in das Ergebnis aufgenommen, deren Signifikanzmaß größer null ist.

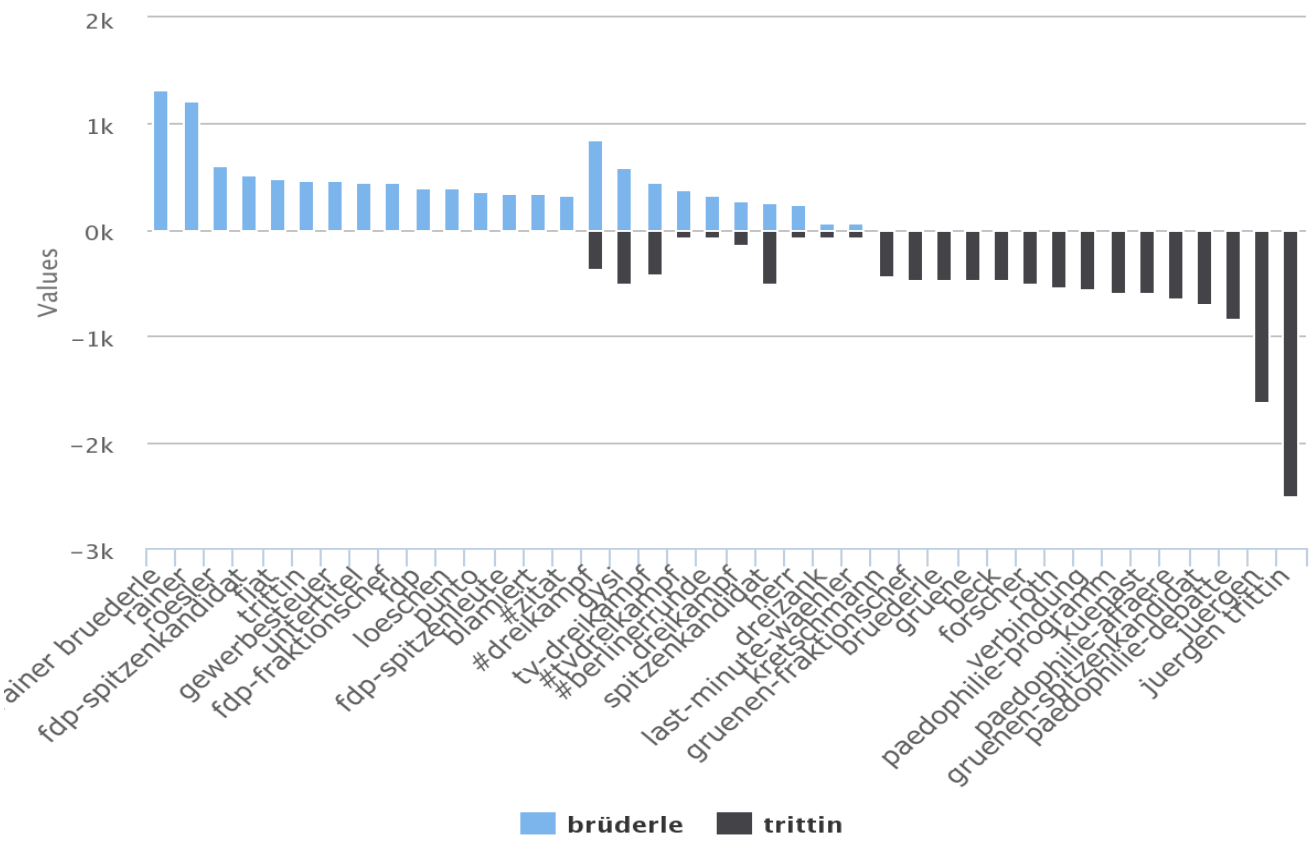


Abbildung 5.3: Kontrastive Analyse zwischen den Politikern Rainer Brüderle und Jürgen Trittin

Das hier vorgestellte Konzept lässt sich außerdem auf eine Gruppe von Schlüsselwörtern erweitern, wie exemplarisch in einer Stimmungsbildanalyse von Blenn et al. [6] gezeigt. Diese präsentieren ein Beispiel zweier fiktiver Teehersteller: *McArrow's orange-peppermint tea* und *DrBrew's strawberry-melon tea*. Für diese beiden Hersteller lassen sich Schlüsselwortgruppen A_{Gruppe} und B_{Gruppe} bilden, die jeweils die Produktpalette repräsentieren. Verwendet man die hier vorgestellte Technik, lassen sich Wörter ableiten, die stark mit diesen Gruppen assoziiert sind. Damit kann der Hersteller anhand seiner Produktpalette bewertet werden. Übertragen auf die Bundestagswahl könnten damit Parteien durch ihre Politiker charakterisiert und ferner zwei Parteien kontrastiert werden. Abbildung 5.4 beschreibt diesen Ansatz schematisch, die Umsetzung dieser Erweiterung liegt jedoch außerhalb des Umfangs dieser Arbeit.

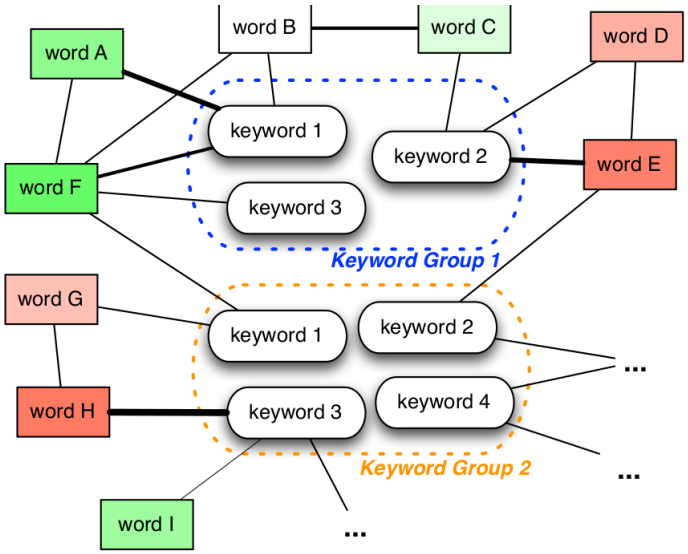


Abbildung 5.4: Kombinierte Simmungsbildanalyse durch die stark assoziierten Wörter mit den zwei Schlüsselwortgruppen A_{Gruppe} und B_{Gruppe}. Rot hinterlegte Wörter repräsentieren ein negatives Stimmungsbild, wohingegen grün hinterlegte Wörter ein positives Stimmungsbild darstellen. Die Grafik stammt aus der Arbeit von Blenn et al. [6].

6 Durchführung und Visualisierung der kontrastiven Analyse

Dieses Kapitel beschreibt die Möglichkeiten des visuellen Systems kontrastive Analysen zu erstellen und zu erforschen. Abschnitt 6.1 beschreibt das Visualisierungskonzept, welches verwendet wird, um die Analyse darzustellen. In Abschnitt 6.2 werden verschiedene Einstellungen präsentiert, mit denen semiprozedural das Erzeugen der Analyse beeinflusst werden kann. Außerdem wird gezeigt, wie der Benutzer die Ergebnisse explorativ erkunden kann.

6.1 Darstellung der kontrastiven Kollokationsanalyse

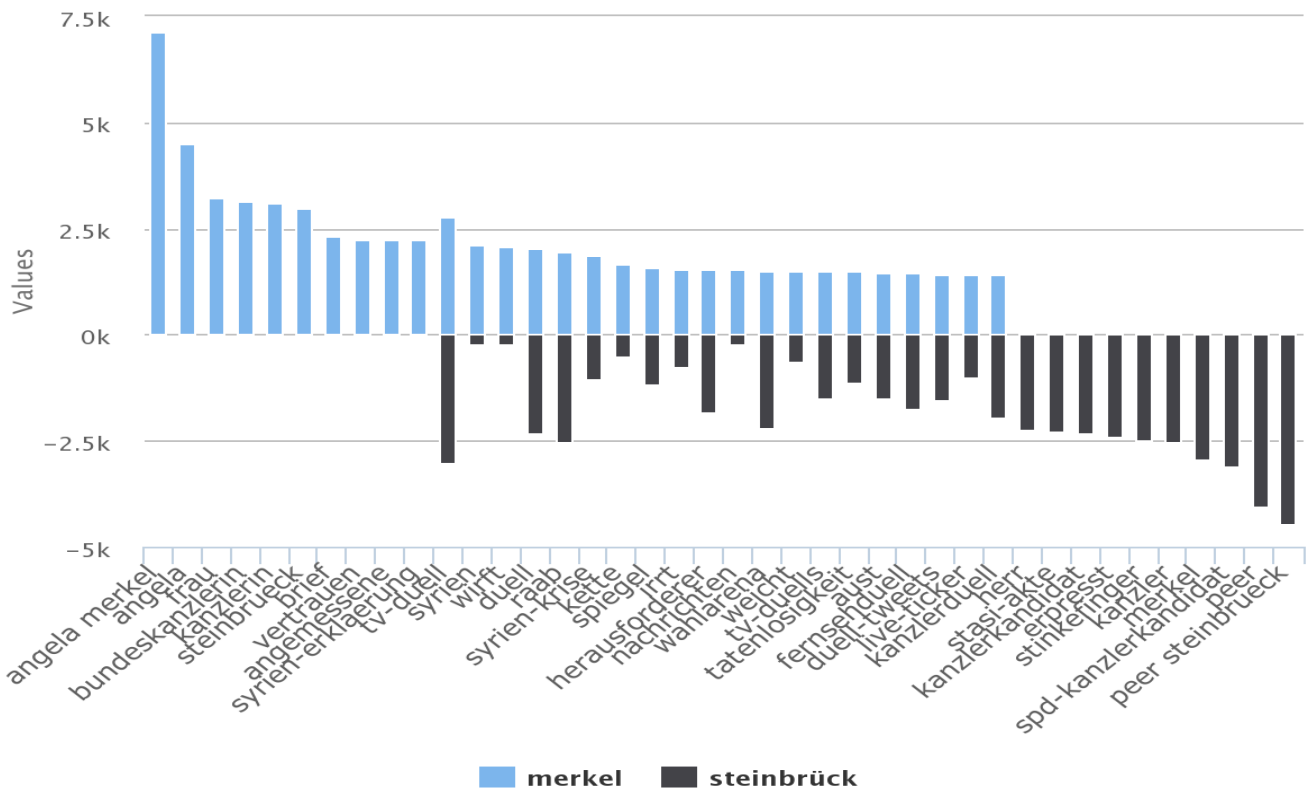
Das Kernstück der kontrastiven Kollokationsanalyse ist die Visualisierung der Wörter in einem Tornadodiagramm. Tornadodiagramme werden häufig genutzt, um zwei Verteilungen zu vergleichen und sind beliebt beim Vergleich von eng verwandten Populationen, wie die Verteilung von Männern und Frauen. Anstatt die Daten in zwei getrennten Histogrammen darzustellen, was es schwierig macht diese zu vergleichen, kombiniert ein Tornadodiagramm beide Histogramme. Dies wird durch das Drehen der Diagramme erreicht, so dass diese horizontal angeordnet sind.

Um die Analyse zu visualisieren, müssen die Knoten und Kanten der Graphrepräsentation in der Datenbank auf die Balken und deren Ausprägung des Tornadodiagramms abgebildet werden. Die Abbildung der unterschiedlichen Strukturen besteht aus drei verschiedenen Datenbankabfragen. Für jeden gegebenen Politiker oder für jede gegebene Partei werden die Knoten bestimmt, welche über eine Relation nur mit dieser Entität verbunden sind. Dafür werden die Knoten durch das ausgewählte Signifikanzmaß in den Eigenschaften der Relation sortiert und von der Datenbank zusammen mit der Relation zurückgegeben. Die Relation wird benötigt, um die Stärke der Assoziation durch das Signifikanzmaß darzustellen. Knoten, die eine Relation mit beiden Entitäten besitzen, werden mit einer getrennten Datenbankabfrage bestimmt. Dies ist performanter als für jede Entität alle Knoten zu bestimmen, die durch eine Relation stark assoziiert sind und die Berechnung serverseitig durchzuführen. Abschließend werden die Ergebnisse auf dem Server zu einem Tornadodiagramm zusammengesetzt und die fertige Analyse an den Client übertragen. Abbildung 6.1a präsentiert dieses erzeugte Tornadodiagramm für die Politiker Angela Merkel und Peer Steinbrück beschränkt auf 40 Begriffe. Neben der kontrastiven Analyse können auch für ein Schlüsselwort stark assoziierte Begriffe ermittelt und visualisiert werden (Abbildung 6.1b).

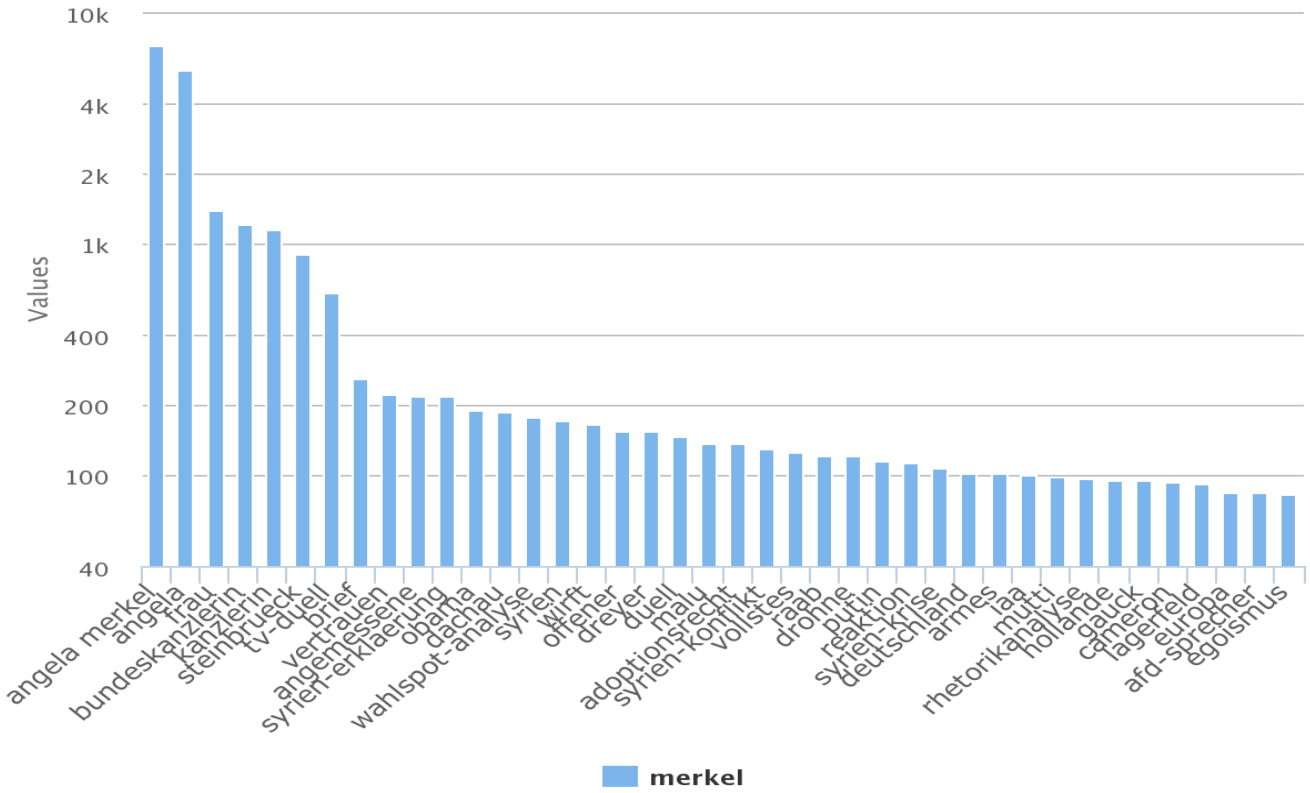
Eine Herausforderung der Visualisierung ist das Skalieren des Wertebereichs, so dass die Diagramme sowohl sehr kleine, als auch sehr große Werte abbilden können. Abbildung 6.2a zeigt die automatische Skalierung der Diagrammbibliothek. Dadurch, dass sehr große und sehr kleine Werte dargestellt werden müssen, sind einige Balken klein skaliert. Eine mögliche Lösung für dieses Problem ist sehr kleine Werte im Diagramm manuell durch einen konstanten Wert anzuheben wie in Abbildung 6.2b dargestellt. Dieser Ansatz löst zwar die Problematik der Skalierung kleiner Werte, hat jedoch einen entscheidenden Nachteil: Werte erscheinen fälschlicherweise gleich groß, da sie die gleiche Ausprägung haben. Diese Problematik lässt sich auch in der automatischen Skalierung durch die Bibliothek beobachten.

Abbildung 6.2c präsentiert eine manuell durchgeführte logarithmische Skalierung. Durch diese Logarithmierung geht jedoch die Unterscheidung großer Werte verloren und der Abwärtstrend ist nicht deutlich erkennbar.

Das Problem lässt sich schließlich mit einer geeigneten statistischen Datentransformation lösen. Die Transformation wird mit Hilfe der inversen Sigmoidfunktion („Logistische“ Funktion) durchgeführt. Diese ist nützlich, um kleine und große Werte in ihrer Darstellung zu trennen und wird im maschinellen Lernen für Klassifizierungsprobleme verwendet [5]. Abbildung 6.1a zeigt das durch die inverse Sigmoidfunktion transformierte Diagramm.

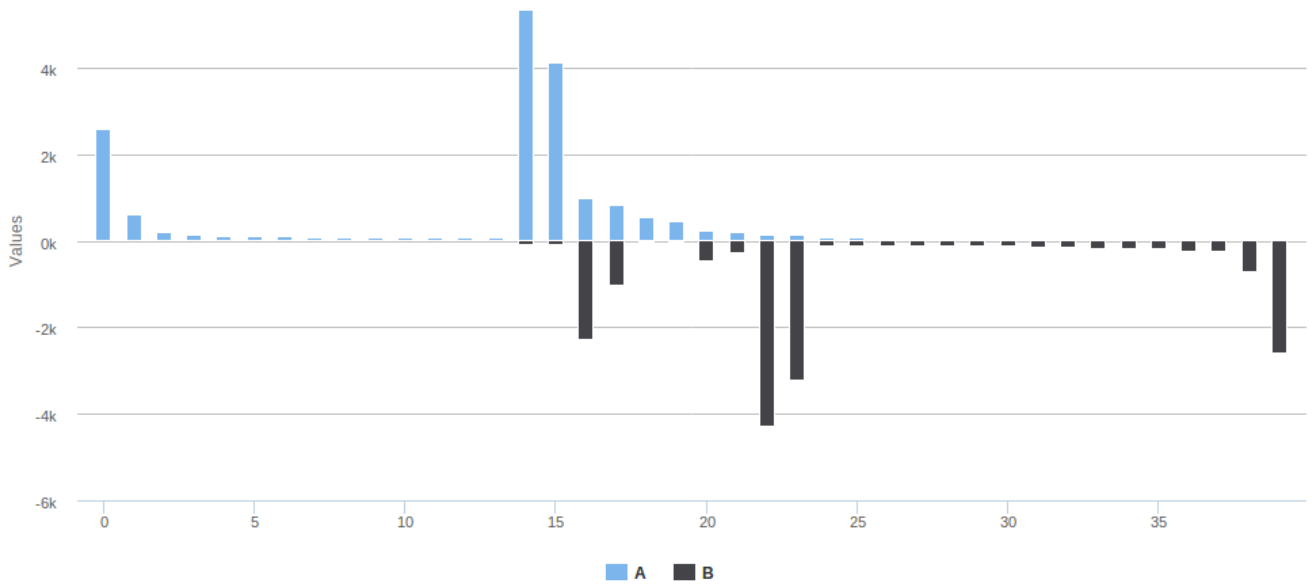


(a) Kontrastive Analyse zwischen den beiden Politikern Angela Merkel und Peer Steinbrück. Angezeigt werden 50 Begriffe, wovon 50 Prozent durch den Schnitt repräsentiert werden. Außerdem werden Usernamen und Hashtags ausgeblendet.

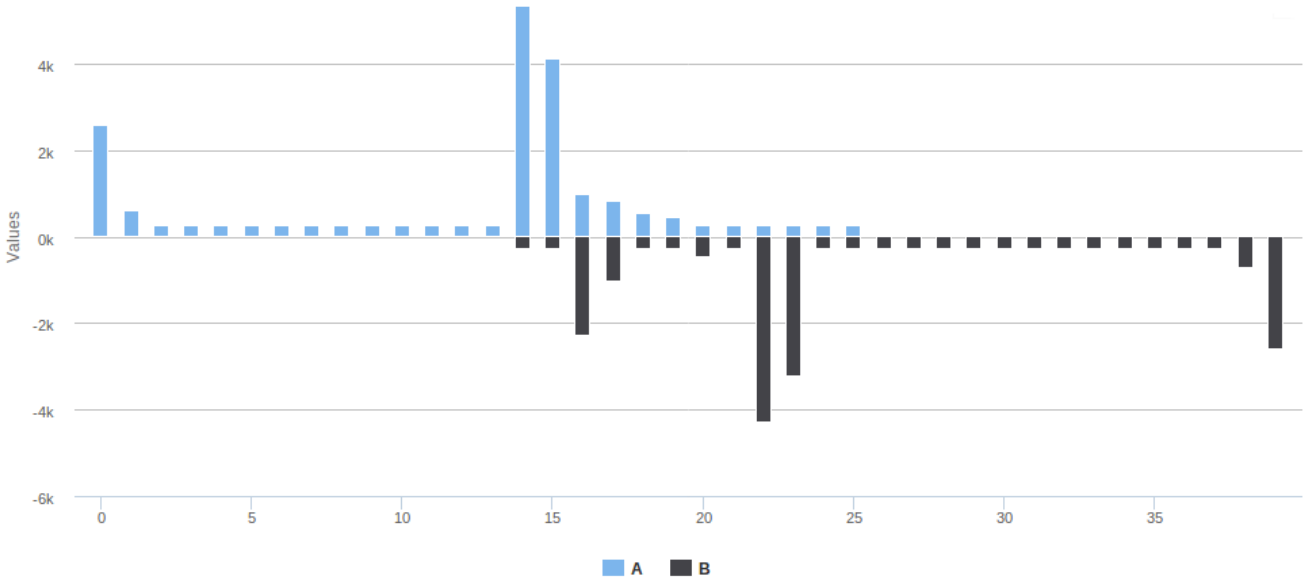


(b) Analyse der stark assoziierten Worte für Angela Merkel

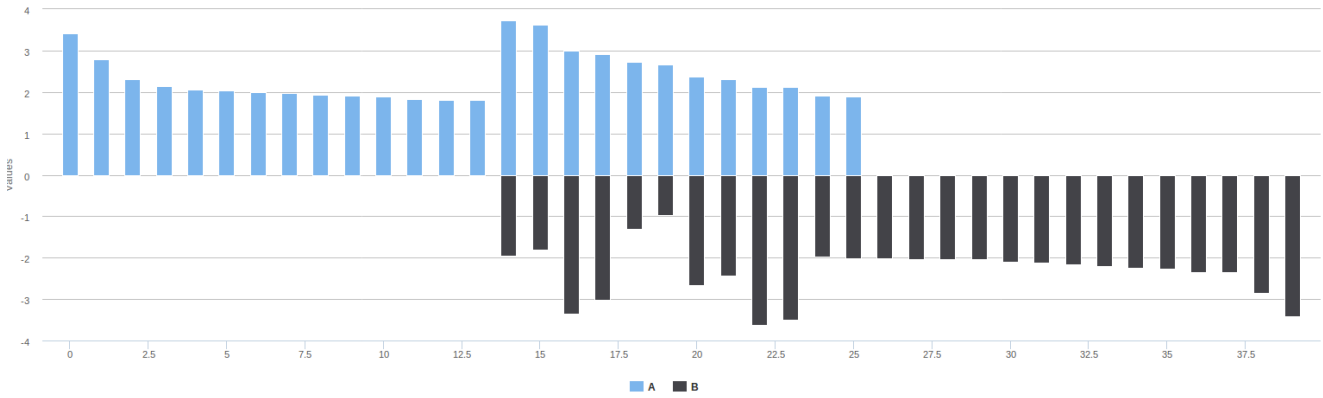
Abbildung 6.1: Kollokationsanalyse für Angela Merkel und Peer Steinbrück.



(a) Standardverhalten der Diagrammbibliothek für die Skalierung der Werte.



(b) Jeder zu kleine Wert wird um 5 Prozent des Maximus angehoben.



(c) Manuelle logarithmische Skalierung der Werte ungleich null.

Abbildung 6.2: Unterschiedliche Skalierungen für die kontrastive Kollokationsanalyse.

6.2 Interaktionsmöglichkeiten

Wie bereits in der Motivation der Arbeit beschrieben, richtet sich die kontrastive Kollokationsanalyse an politisch interessierte Bürger, Journalisten, Schüler und Studenten, sowie Politiker. Ein großes Anliegen ist die Bedienbarkeit der Anwendung ohne spezielles Vorwissen von den Nutzern vorauszusetzen. Deshalb soll die Analyse in einer Standardeinstellung gute Ergebnisse erzielen. Komplexere Einstellmöglichkeiten bleiben den unerfahrenen Benutzern verborgen. Der Fokus liegt auf dem explorativen Erkunden der Ergebnisse und nicht auf dem Konfigurieren der Analyse. Abbildung 6.3a zeigt die Benutzerschnittstelle im Modus für unerfahrene Benutzer, wohingegen Abbildung 6.3b den erweiterten Modus für erfahrene Benutzer darstellt. Dieser Modus kann mit dem Schalter auf der rechten Seite der Konfigurationsparameter eingestellt werden.

Nachfolgend werden die Parameter des erweiterten Modus vorgestellt und die Auswirkung der Einstellung auf die Analyse beschrieben.

Schlüsselwörter

Die beiden Eingabefelder erwarten jeweils einen Politiker oder eine Partei, für welche eine kontrastive Kollokationsanalyse durchgeführt werden soll. Mit Hilfe des Buttons, auf dem ein Minuszeichen abgebildet ist, kann ein Eingabefeld entfernt werden, um eine Analyse mit nur einer Entität durchzuführen. Analog erzeugt der Button mit dem Pluszeichen ein neues Eingabefeld für die kontrastive Analyse. Das Eingabefeld ist außerdem mit einer Autovervollständigungsfunktion ausgestattet, welche Spitzenkandidaten dieser Bundestagswahl unterstützt. Durch betätigen des *Go-Buttons*, wird die Anfrage an den Server übermittelt und anschließend das Ergebnis visualisiert.

Anzahl der Begriffe

Dieses Feld reguliert die Gesamtanzahl der angezeigten Wörter in der Analyse. Die Obergrenze liegt bei 200 Begriffen, da mehr Wörter nicht überschneidungsfrei auf durchschnittlichen Monitoren dargestellt werden können.

Part-of-Speech

Diese Option ermöglicht es, verschiedene Wortklassen von der Analyse auszuschließen. In der Standardeinstellung werden alle Wortklassen betrachtet. Diese Wortklassen repräsentieren neben herkömmlichen Wortarten (z.B. Verben, Nomen und Adjektive) auch Standorte, die in Zusammenhang mit der untersuchten Entität stehen. Außerdem werden twitterspezifische Klassen wie Hashtags und Usernamen unterstützt.

Sonstige Optionen

Zum einen können Tweets nach der Bundestagswahl am 22. September in die Auswertung der Analyse aufgenommen werden. Diese Funktion ist nützlich für Untersuchungen, die speziell auf Auswirkungen vor und nach dem Wahlkampf fokussiert sind. Zum anderen bietet der zweite Schalter die Möglichkeit, die Analyse auf Namensentitäten zu beschränken. Dadurch werden nur Politiker angezeigt, die mit dem gegebenen Politiker oder der gegebenen Partei stark assoziiert sind.

Wortassoziationsmetriken

Hier kann das Wortassoziationsmaß variiert werden. Dieses Maß bestimmt, welche Wörter für die zu untersuchenden Entitäten in die Ergebnisse aufgenommen werden sollen. Es existiert außerdem eine starke Abhängigkeit zu anderen Parametern, wie der Wortfrequenz. Das bedeutet, das Ändern dieses Maßes kann es erforderlich machen, andere Parameter entsprechend anzupassen.

Kantengewicht

Dieser Parameter repräsentiert einen Schwellwert für das minimale Kantengewicht zwischen der gegebenen Entität und dem stark assoziierten Wort. Ein minimales Kantengewicht von drei bedeutet, dass beide Wörter mindestens dreimal als Bigramm im Korpus enthalten sein müssen.

Wortfrequenz A und B

Die Wortfrequenz bestimmt, wie häufig ein Wort im Korpus enthalten sein muss, um in die Ergebnisse aufgenommen zu werden. Dieser Parameter ist wichtig für die Verwendung mit dem PMI Maß und kann genutzt werden, um seltene Bigramme aus dem Ergebnis zu entfernen.

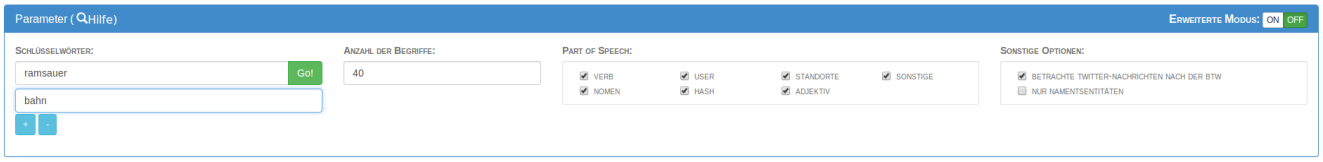
Wortfrequenz AB

Dieser Parameter reguliert explizit die Frequenz der Wörter im Schnitt der Analyse.

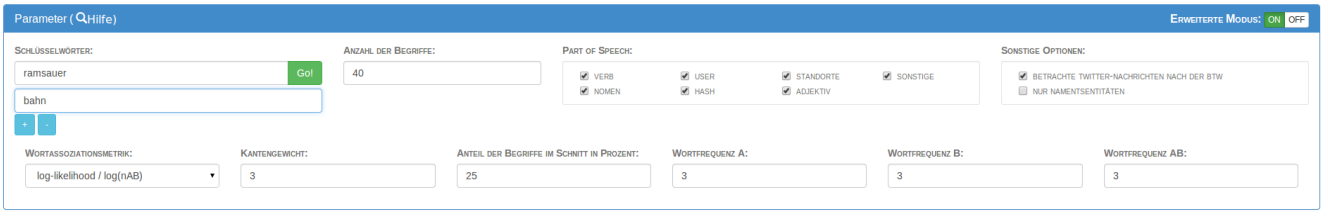
Anteil der Begriffe im Schnitt

Diese Einstellung kontrolliert den Anteil der Gemeinsamkeiten zwischen den betrachteten Entitäten. Der Wert 25 bedeutet, dass mindestens 25 Prozent der Menge an Wörtern im Ergebnis Gemeinsamkeiten repräsentieren müssen. Abbildung 6.4 zeigt eine kontrastive Analyse zwischen Peter Ramsauer und der Deutschen Bahn. Der Parameter für die Gemeinsamkeiten ist auf 50 Prozent eingestellt, wodurch sich das Eingreifen Peter Ramsauers in das Bahnchaos in Mainz in der Analyse widerspiegelt.

Die Benutzer haben die Möglichkeit unter dem Eintrag „Hilfe“ Erklärungen für die einzelnen Parameter einzusehen, um zu verstehen, welche Auswirkung diese haben. Außerdem werden den Nutzern *Tooltips* angezeigt, wenn der Mauszeiger eine kurze Zeit unbewegt über einem entsprechenden Element verweilt. Diese Tooltips sind Hilfetexte, die in kurzer und prägnanter Form den Parameter beschreiben.



(a) Mögliche Parameter für unerfahrene Benutzer, um Analysen zu erstellen.



(b) Parameter für im Umgang mit der Anwendung erfahrene Benutzer.

Abbildung 6.3: Parameter zum Einstellen der kontrastiven Analyse.

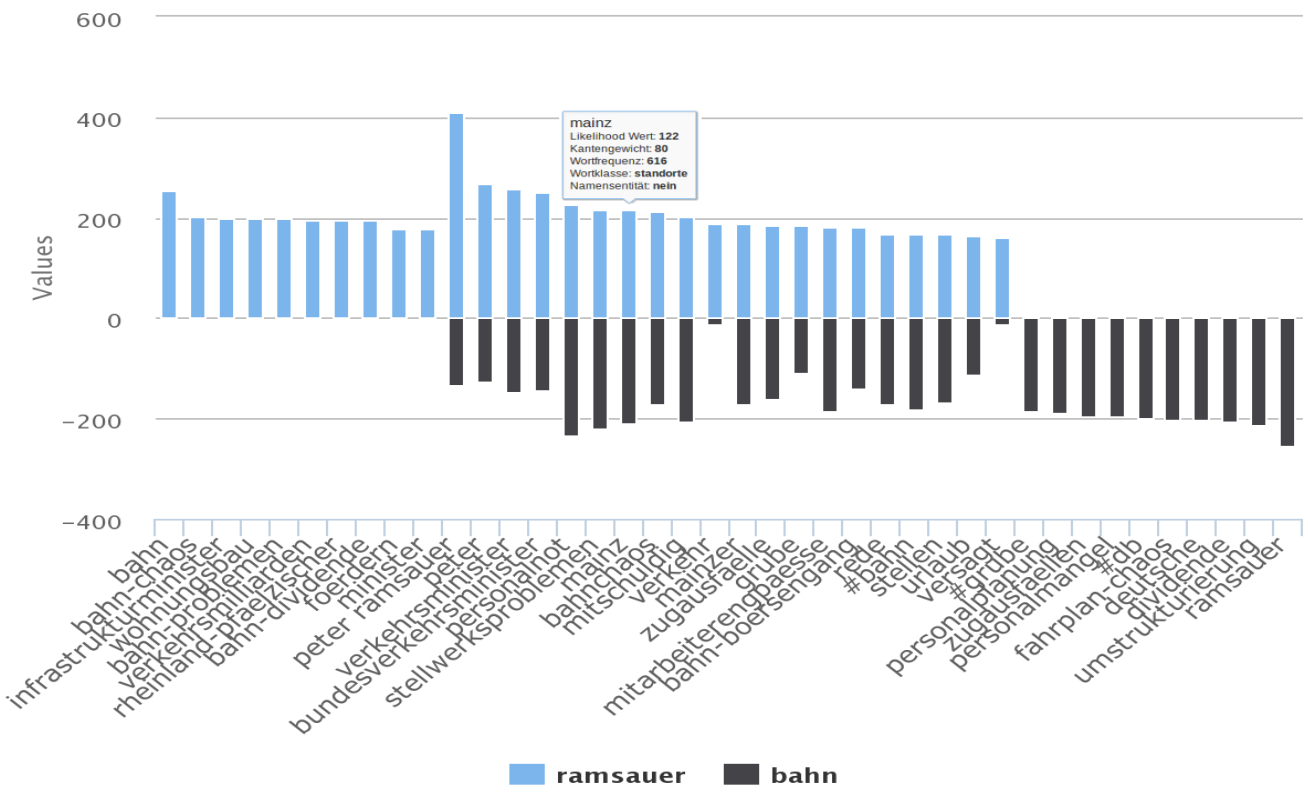


Abbildung 6.4: Kontrastive Analyse zwischen Peter Ramsauer und der Deutschen Bahn. Angezeigt werden 40 Begriffe, wovon 50 Prozent durch den Schnitt repräsentiert werden.

Erkunden der Ergebnisse

Hat ein Benutzer eine Analyse in einer gewünschte Konfiguration erzeugt, kann er die Ergebnisse erkunden. Dazu bietet die Anwendung verschiedene Möglichkeiten.

Popup-Informationen

Verweilt ein Benutzer kurze Zeit unbewegt über einem Balken im Diagramm, werden Informationen zu dem entsprechenden Wort in einem Popup-Fenster angezeigt (Abbildung 6.4). Dieses Popup enthält den Wert des Wortassoziationsmaßes, das Kantengewicht im Korpus, die Wortfrequenz, die Wortklasse und die Tatsache, ob es sich bei dem Wort um eine Namensentität handelt.

Manuelles Anpassen der Ergebnisse und Exportfunktion

Es ist außerdem möglich, bestimmte Wörter aus dem Diagramm zu entfernen. Diese Funktion ist vor allem für Benutzer interessant, die eine Analyse in einer Hausarbeit, einem Zeitungsartikel oder in einem Blogeintrag verwenden möchten. Damit können Wörter aus dem Ergebnis entfernen werden, die beispielsweise Tokenisierungsfehler enthalten. Abbildung 6.5b zeigt diese Möglichkeit. Abschließend lässt sich das Diagramm mit den angewendeten Einstellungen in verschiedene Bildformate exportieren und herunterladen.

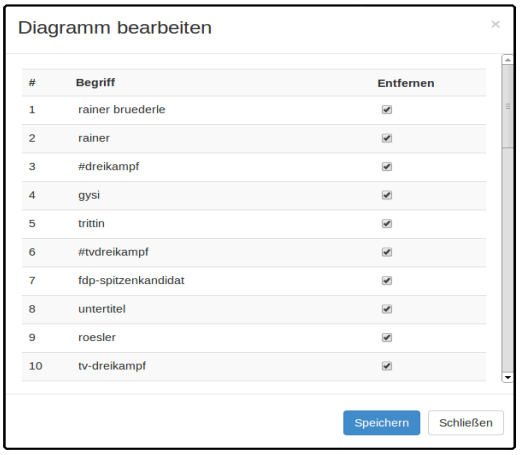
Kontextansicht

Eine weitere Funktion im Umgang mit der Analyse und ein Hilfsmittel zur Interpretation der Daten ist die Möglichkeit Twitter-Nachrichten anzuzeigen, die Grund für die Wortassoziation sind. Durch einen Linksklick auf einen entsprechenden Balken, öffnet sich die Kontextansicht, welche in Abbildung 6.5a betrachtet werden kann.

In dieser Kontextansicht werden alle Twitter-Nachrichten angezeigt, die sowohl den gegebenen Politiker oder die gegebene Partei, als auch das Wort enthalten, welches mit dieser Entität stark assoziiert ist. In der oberen linken Hälfte der Ansicht kann inspiziert werden, wie viele solcher Tweets das Korpus enthält und welcher aktuell betrachtet wird. Mit dem „Nächste Twitter-Nachricht Button“ kann jeweils der nächste Tweet betrachtet werden und der „Schließen Button“ beendet die Kontextansicht.

Mögliche Erweiterungen dieser Ansicht und Visualisierung der Twitter-Nachrichten werden in Kapitel 8 diskutiert.

(a) Die Kontextansicht, welche Twitter-Nachrichten enthält die Grund für die Wortassoziation sind.



(b) Mit Hilfe dieses Editors ist es möglich, ausgewählte Begriffe aus dem Diagramm zu entfernen.

Abbildung 6.5: Möglichkeiten die Ergebnisse zu erkunden und anzupassen.

7 Evaluation

Die Tatsache, dass Twitter als Plattform für politische Deliberation genutzt wird, bedeutet nicht zwangsläufig, dass aussagekräftige Informationen extrahiert werden können oder die Verteilung der Meinung in den Tweets die Meinung der Bevölkerung widerspiegelt. Um die Forschungsfrage „Wie reflektiert Twitter Ereignisse der realen Welt?“ zu beantworten, werden mit vorher gewählten Schlüsselwörtern kontrastive Analysen durchgeführt und untersucht. Die Evaluation gliedert sich in zwei Teile.

Abschnitt 7.1 stellt in einer ersten Untersuchung exemplarisch zwei kontrastive Analysen vor.

Der Abschnitt 7.2 führt eine quantitative Untersuchung durch und präsentiert Ergebnisse, die den Grad der Übereinstimmung der Resultate mit deutschen Tageszeitungen angeben. Diese Ergebnisse werden kritisch analysiert und es werden Probleme mit der Analyse aufgezeigt.

7.1 Betrachtung exemplarischer kontrastiver Analysen mit kritischer Würdigung

Nachfolgend werden zwei kontrastive Analysen präsentiert und die stark assoziierten Wörter in den Kontext der Bundestagswahl 2013 gesetzt. Hiermit zusammenhängend folgt eine Gegenüberstellung der Meinungen der Twitter-Nutzer, repräsentiert durch deren Tweets, mit Zeitungsartikeln. Damit wird begründet, dass die hier vorgestellte Analyse sowohl für das Erforschen von Beziehungen zwischen Politikern geeignet ist, als auch aktuelle Wahlthemen reflektiert werden können.

7.1.1 Angela Merkel und Peer Steinbrück

Diese kontrastive Analyse vergleicht die beiden Kanzlerkandidaten Peer Steinbrück (SPD) und Angela Merkel (CDU). Betrachtet werden 40 Begriffe, wobei 25 Prozent dieser Begriffe den Schnitt repräsentieren. Außerdem werden Usernamen aus den Ergebnissen entfernt und nur Twitter-Nachrichten vor der Bundestagswahl am 22. September betrachtet. Zum Finden der Kollokationen wird das normierte Log-Likelihood Maß verwenden und ein minimales Kantengewicht von drei genutzt.

Peer Steinbrück

Auf den ersten Blick zeigt Abbildung 7.1a einige offensichtliche Fakten. Angela Merkel ist die aktuelle Bundeskanzlerin („bundeskanzlerin“) und Peer Steinbrück der Kanzlerkandidat der SPD („spdkanzlerkanidat“). Das sechststärkste, mit Peer Steinbrück assoziierte Wort „stinkefinger“, ist eher ungewöhnlich für die politische Domäne. „Stinkefinger“ bezieht sich auf ein Bild des SPD-Kanzlerkandidaten der süddeutschen Zeitung, auf dem dieser den Mittelfinger zeigt. Das Bild hat eine Welle der Empörung ausgelöst, wodurch viele Twitter-Nutzer ihren Unmut unter dem Hashtag #Stinkefinger zum Ausdruck brachten.

Die Begriffe „Circus“ und „Halligalli“ referenzieren die Fernsehshow „Circus Halligalli“, bei der Peer Steinbrück während seines Wahlkampfes zu Gast war. Diese Fernsehshow moderiert von Klaas Heufer-Umlauf und Joko Winterscheidt ist eher für ihre Satire, als für ihre politischen Inhalte bekannt. Die Tweets in diesem Zusammenhang wurden überwiegend im Vorfeld der Sendung verfasst und drücken zum einen Erstaunen aus, dass Peer Steinbrück an dieser Sendung teilnimmt. Zum anderen enthalten diese jedoch auch Spott und deuten den Auftritt als Verzweiflungstat.

- „respekt: peer steinbrueck wagt sich, wenige tage vor der wahl, in den circus #HalliGalli“.
- „Wie? Watt? Peer Steinbrück morgen in ‘Circus Halligalli’...? Also das sieht ja schon nach purer Verzweiflung aus...“.

- „Neue Chance auf neue Mittelfingergesten: Peer Steinbrück heute bei ‘Circus HalliGalli’ zu Gast http://t.co/ZoQiizMe5J“.
- „Peer Steinbrück bei Circus HalliGalli war einfach verdammt gut. Lustig, sympathisch und sehr souverän. Schönes Ding. #btw13“.

Die Twitter-Nachrichten zum eigentlichen Auftritt Steinbrücks beschreiben diesen als lustig und souverän. Dies ist in Übereinstimmung mit den Medien, die den Auftritt ebenfalls als souverän und erfolgreich skizzieren, aber als wenig hilfreich angesichts des aussichtslosen Wahlkampfes zum Ende der Wahl hin bewerten (Anhang B.1).

Eine weitere Gruppe von Wörtern („erpressungsversuch“, „erpresst“, „putzfrau“) repräsentiert Twitter-Nachrichten, die ausschließlich von Nachrichtendiensten stammen. Diese thematisieren einen Erpressungsversuch im Wahlkampf. Peer Steinbrück habe angeblich in der Vergangenheit eine illegale Putzhilfe beschäftigt und der Erpresser drohte dies öffentlich zu machen, wenn Steinbrück nicht auf seine Kandidatur verzichte (Anhang B.1). Die Tweets verweisen mit Links auf Zeitungsartikel der entsprechenden Nachrichtendienste und enthalten häufig die oben genannten Schlagworte.

Angela Merkel

Das Thema Syrien-Krise („syrien-erklärung“, „syrien“, „obama“) ist ein sehr präsentes Thema in diesem Wahlkampf. Nach dem Giftgasanschlag auf Damaskus fordert Obama eine Bestrafung in Form eines Militärschlags. Auf dem G-20 Gipfel verweigerte Angela Merkel jedoch ihre Unterschrift auf der Erklärung um eine gemeinsame Strategie in Syrien *(„Merkel hat die Kriegserklärung von Obama nicht unterschrieben, als einzige. Sie hat Bundestagwahl und anscheinend Schiss vor dem Wähler #btw13“)*.

Ein weiteres sehr häufig diskutiertes Thema in dieser Bundestagswahl ist die NSA-Affäre um Edward Snowden. Die Begriffe „offener“, „brief“, „angemessene“ repräsentieren einen Aufruf in Form eines offenen Briefes an Angela Merkel. In diesem Brief forderte die Autorin und Juristin Juli Zeh nach Bekanntwerden der NSA-Affäre zusammen mit anderen Schriftstellern Angela Merkel auf, die volle Wahrheit über den Spähangriff zu sagen. Dieser Brief hat sich zu einem internationalen Aufruf entwickelt, den unter anderem sechs Nobelpreisträger unterzeichnet haben. Die Tweets teilen den Link[22] zu diesem Brief sehr häufig und hoffen auf Unterstützer.

- „Unterstützt bitte den offenen Brief von Juli Zeh an Angela Merkel - http://t.co/cUdGb7L4kq“.
- „Habe unterschrieben: Offener Brief an Bundeskanzlerin Angela Merkel: Angemessene Reaktion auf die NSA-Affäre https://t.co/Ide63ZR3B5“.

Gemeinsamkeiten

Die Mehrheit der Wörter stehen in Verbindung mit dem TV-Duell, an dem sowohl Angela Merkel, als auch Peer Steinbrück teilnahmen und das unter dem Hashtag #tvduell diskutiert wurde. Zusätzlich zu diesem Hashtag befindet sich auch der Moderator Stefan Raab, ein deutscher Komödiant und Entertainer im Schnitt der Analyse. Das Wort „Kette“ steht ebenfalls mit dem TV-Duell in Zusammenhang, ist jedoch stärker mit Merkel, als mit Steinbrück assoziiert. Der Begriff steht für eine Halskette in den Deutschlandfarben, die Angela Merkel beim TV-Duell getragen hat. Das Accessoire erregte viel Aufmerksamkeit auf Twitter und hat mittlerweile sogar seinen eigenen Twitter-Account[23]. Die mit dieser Halskette assoziierten Tweets sind überwiegend positiv für Merkel und enthalten ironische Kommentare bezüglich ihres Konkurrenten Peer Steinbrück. Über den Ausgang und Verlauf des Duells finden sich keine Wörter im Schnitt der Analyse.

- „Die Merkel hat ’ne Schwarz-Rot-Goldenen Kette um, Steinbrück einen einfachen seriösen Anzug...Sagt viel über die Kandidaten! #tvduell #btw13“.
- „Steinbrück chancenlos, allein Merkels Kette ist die Vorentscheidung! #tvduell :D“.

[22] `http://change.org/nsa`
[23] https://twitter.com/schlandkette

- „Steinbrück sieht die Merkel in etwa so an, als würde er jeden Moment ihre Kette stehlen wollen. #tvduell“.

Die Syrien-Krise ist eine weitere Gemeinsamkeit der beiden Politiker und lässt sich zeitlich in die Endphase des Wahlkampfes einordnen. Beide Politiker müssen sich zum Giftgasanschlag in Damaskus äußern und wollen nicht als Kriegstreiber gelten, da die überwiegende Mehrheit der Bevölkerung nach einer Forsa-Umfrage einen Militärschlag ablehnt. Der Begriff „Syrien-Krise“ ist stärker mit Merkel als mit Steinbrück assoziiert und die Ausprägung könnte repräsentieren, dass Merkel aktiver in dieses Thema involviert ist. Diese Interpretation lässt sich durch Medienberichte bestätigen. Nach Aussagen des Magazin Spiegel (Anhang B.1) hält Merkel eine internationale Reaktion auf den Giftgasanschlag für unabdingbar. Entsprechende Konsequenzen für die Bundesregierung lässt sie jedoch offen. Steinbrück hingegen äußert sich im Hamburger Abendblatt zurückhaltend zu diesem Thema (Anhang B.1).

Abbildung 7.1b zeigt eine auf Adjektive beschränkte kontrastive Analyse zwischen Angela Merkel und Peer Steinbrück. Die Adjektive im Schnitt der Analyse zeigen ein Meinungsbild der Twitter-Nutzer über den Verlauf des TV-Duells. Das Wort „souveräner“ ist viel stärker mit Steinbrück assoziiert, als mit Merkel. Die Tweets der Kollokation „Steinbrück – souveräner“ beschreiben Steinbrück überwiegend souveräner als Merkel im Umgang mit Zwischenfragen und bewerten dessen klare Aussagen als positiv.

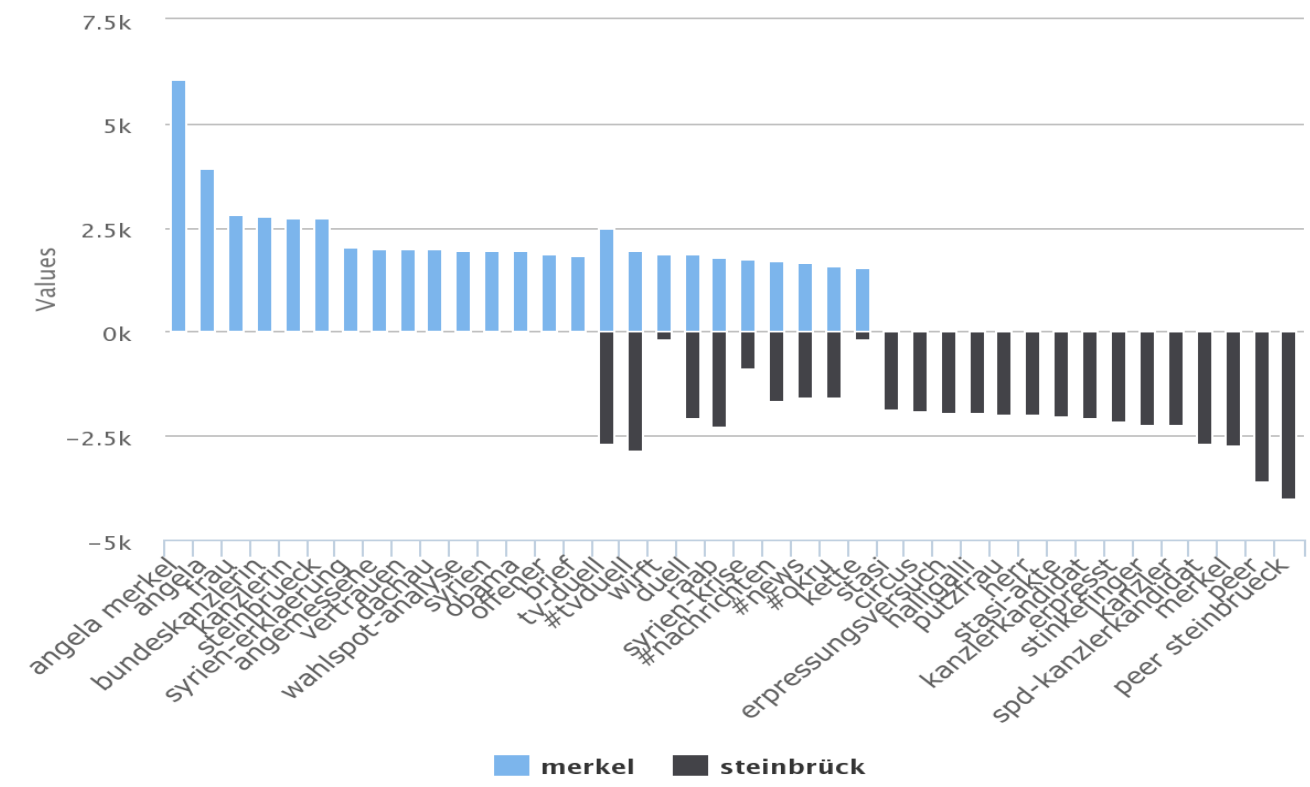
- „Steinbrück hat seine Phrasen erheblich souveräner vorgetragen als Merkel. #tvduell“.
- „Steinbrück souveräner als Merkel - und auch sachlicher. #tvduell“.

Durch diese Vergleiche wird Merkel jedoch auch automatisch mit dem Wort „souverän“ assoziiert. Ohne den zugrundeliegenden Kontext könnte dies zu Fehlinterpretationen führen. Die Ausprägung der Wörter innerhalb der Analyse ist demnach nur ein Indikator und der Kontext unterstützt den Benutzer eine genaue Aussage treffen zu können.

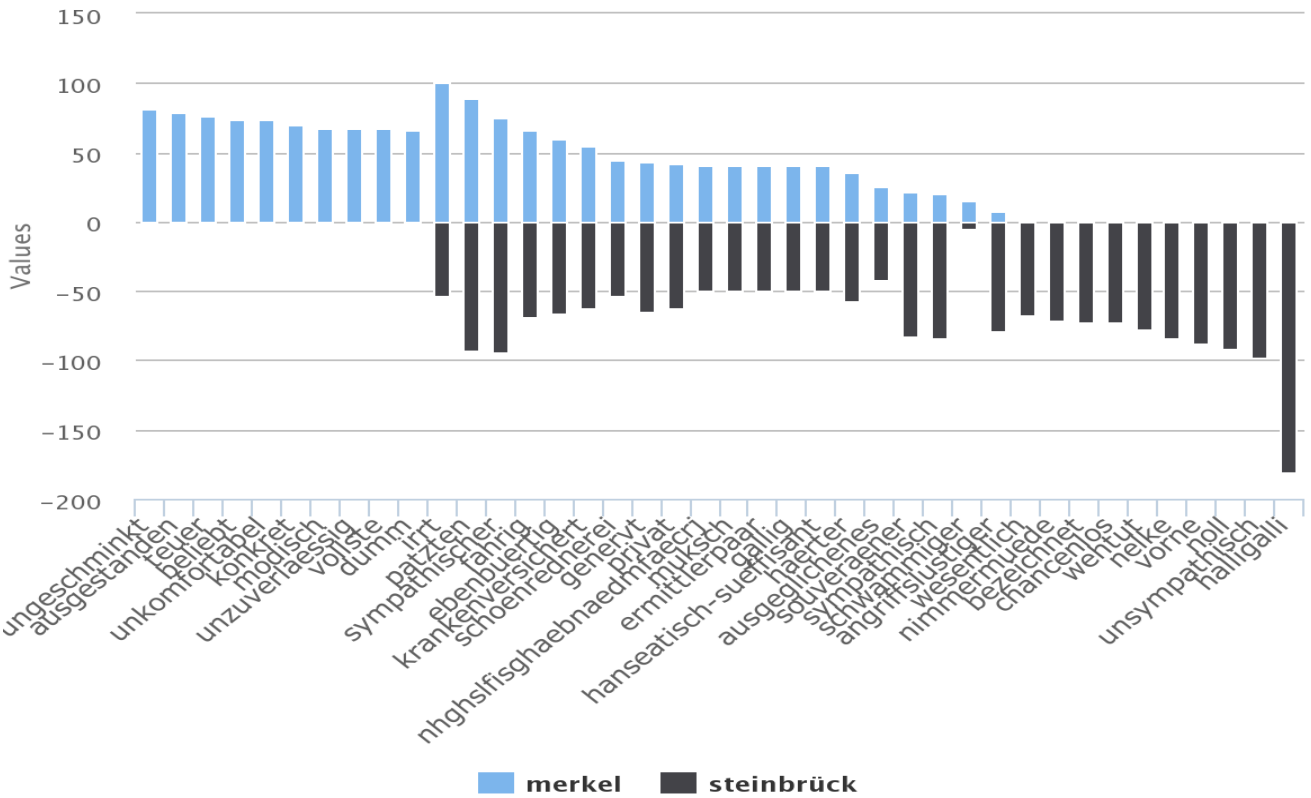
Der umgekehrte Fall lässt sich bei dem Adjektiv „genervt“ beobachten. Das Wort ist stärker mit Steinbrück assoziiert und lässt darauf deuten, dass Steinbrück genervt wirkt. Der Kontext beschreibt Steinbrück jedoch als ruhig und sachlich und Merkel als genervt und unsympathisch. Es wird deutlich, dass bei der gemeinsamen Betrachtung der Politiker der zugrundeliegende Kontext eine entschiedene Rolle spielt.

Steinbrück wird außerdem sowohl als sympathisch, als auch unsympathisch charakterisiert, wobei das Adjektiv unsympathisch stärker mit ihm assoziiert ist. Die Tweets reflektieren das kontroverse Meinungsbild der Twitter-Nutzer bezüglich des Kanzlerkandidaten Peer Steinbrück.

Basierend auf der Analyse lässt sich Steinbrück als souverän und aggressiv („angriffslustiger“) beschreiben und Merkel eher als ausweichend und genervt. Diese Einschätzung ist in Übereinstimmung mit den Berichten der Medien zu diesem TV-Duell. Die beschreiben das Aufeinandertreffen als gepflegte Langeweile und Herunterbeten von Parteiprogrammen. Angela Merkel wich unangenehmen Fragen aus und umarmte den Gegner. Steinbrück hingegen wollte aggressiv angreifen. Die Versuche prallten jedoch chancenlos an Merkel ab (Anhang B.1).



(a) Kontrastive Kollokationsanalyse zwischen Angela Merkel und Peer Steinbrück. Angezeigt werden 40 stark assoziierte Begriffe, wovon 25 Prozent den Schnitt repräsentieren.



(b) Eine auf Adjektive beschränkte kontrastive Kollokationsanalyse zwischen Angela Merkel und Peer Steinbrück. Angezeigt werden 40 stark assoziierte Begriffe, wovon 50 Prozent den Schnitt repräsentieren.

Abbildung 7.1: Kontrastive Kollokationsanalyse zwischen Angela Merkel und Peer Steinbrück

7.1.2 Rainer Brüderle und Gregor Gysi

Abbildung 7.2 zeigt die kontrastive Analyse zwischen zwei anderen bekannten Politikern: Gregor Gysi (Die Linke) und Rainer Brüderle (FDP). Als Parameter werden die gleichen Einstellungen wie für die Betrachtung der beiden Kanzlerkandidaten verwendet und die Menge der Wörter im Schnitt auf 30 Prozent erhöht.

Für beide Politiker ist das am stärksten assoziierte Wort ihr eigener Vorname, gefolgt von dem Vornamen des jeweils anderen. Das ist ein Indikator, dass die Benutzer auf Twitter in einer bestimmten Verbindung über beide Politiker reden. Die Tweets bezüglich der Vornamen referenzieren überwiegend einen sogenannten Dreikampf. Dieser Dreikampf ist ein Fernsehduell zwischen den Spitzenkandidaten der Freien Demokratischen Partei, Bündnis 90/Die Grünen und der Linken und wurde kurz nach dem Fernsehduell zwischen den beiden Bundeskanzlerkandidaten Angela Merkel und ihrem Rivalen Peer Steinbrück ausgestrahlt. Die oben genannten Tweets bringen Vorfreude auf eine spannende Konfrontation, nach dem eher ruhigen TV-Duell am Tag zuvor zum Ausdruck.

- „Das #tvduell ist der bisherige Tiefpunkt des Wahlkampfes. Morgen auf ARD wird es spannender mit Gysi, Trittin und Brüderle.“.
- „ich schau mir morgen den TV-Dreikampf zwischen Rainer Brüderle, Gregor Gysi und Jürgen Trittin an. Wird sicher spannender als das #tvduell“.
- „So, und morgen dann das #tvduell zwischen Gysi, Brüderle und Trittin. Das wird dann hoffentlich etwas spannender...“.

Alle Wörter im Schnitt der Analyse beschreiben das Aufeinandertreffen der beiden Politiker in diesem Dreikampf. Der dritte Teilnehmer des Duells, Jürgen Trittin von den Grünen, findet sich auch in den Gemeinsamkeiten wieder und ist vergleichbar stark mit den beiden Politikern assoziiert, wie deren jeweilige Vornamen.

Der Dreikampf wurde in Twitter unter dem Hashtag #Dreikampf diskutiert und befindet sich auch im Schnitt der Analyse. Ein zentrales Thema der Diskussion war, ob ein gesetzlicher Mindestlohn notwendig ist. Diese Debatte wird durch das Wort „mindestlohn“ im Schnitt dargestellt und repräsentiert Tweets, die das Thema Mindestlohn kontrovers diskutieren.

- „Brüderle hat recht was das Thema Mindestlohn angeht! #fdp #zdf“.
- „Brüderle will keinen Mindestlohn. Sollte mal #WISO schauen. Da war ne Frau, die 3 Jobs braucht, um über die Runden zu kommen. #TVDreikampf“.
- „Sehr gut. @JTrittin erklärt dem Kollegen Brüderle mal die Sinnlosigkeit des Widerstands gegen Mindestlohn #3kampf“.

Brüderle spricht sich gegen einen gesetzlichen Mindestlohn aus: *„Die Friseure hätten gezeigt, dass sie eine Regelung ohne Staat finden könnten“*. Gysi kontert: *„#Gysi: Lohn für Friseure in Thüringen bei 3,50 Euro. "Vergessen Sie es, Herr Brüderle". #Dreikampf“* (Anhang B.2).

Die verbleibenden Wörter im Schnitt der Analyse können mit dieser Debatte in Verbindung gebracht werden. Zusammenfassend identifizieren die Benutzer auf Twitter Rainer Brüderle überwiegend als Schuldigen.

- „Der Brüderle kann so lange reden wie er will, die Mehrheit der Deutschen will weiter einen gesetzlichen Mindestlohn #dreikampf“.
- „Ohje. Der Brüderle meint, Mindestlohn sei nur was für Berufseinsteiger. #fail #dreikampf“.

- „Brüderle nennt Forderung nach Mindestlohn "Gejammer". Aha. Aber er scheint relativ nüchtern zu sein. #dreikampf“.

Die Presse hingegen analysiert Gysis Argument kritisch (Anhang B.2). Ab dem 1. August 2013 wird für Friseure in Thüringen und allen ostdeutschen Bundesländern erstmals ein Mindestlohn von 6,50 Euro eingeführt. Eine weitere Erhöhung auf 8,50 Euro zum 1. August 2015 wird folgen und es gilt dann in Ost- und Westdeutschland ein einheitlicher Mindestlohn im Friseurhandwerk. Gregor Gysi ignoriert diese Tatsache gänzlich in seiner Argumentation, so dass Rainer Brüderle für die Twitter-Nutzer als Verlierer aus diesem Duell heraustritt.

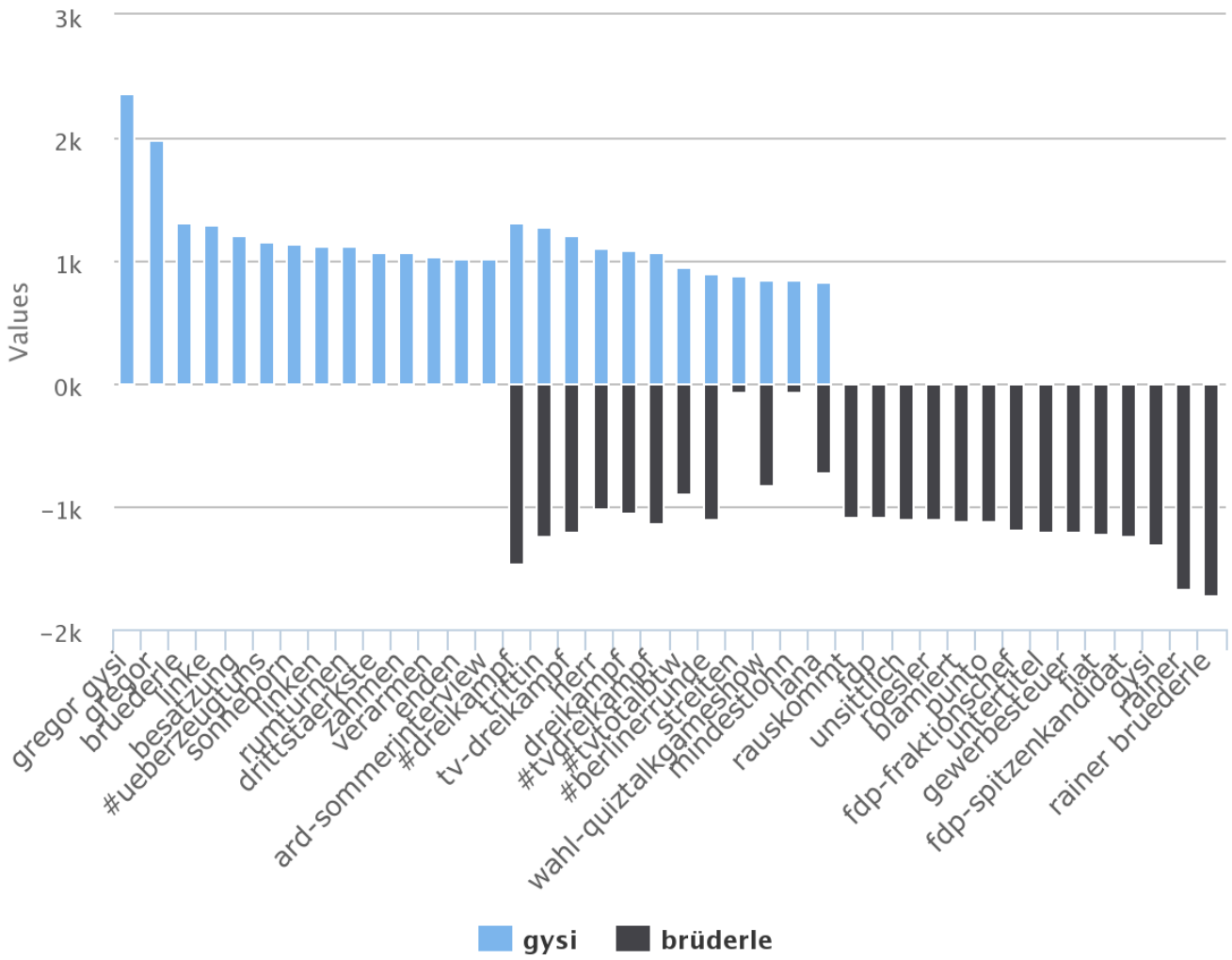


Abbildung 7.2: Kontrastive Analyse für die Politiker Rainer Brüderle und Gregor Gysi. Betrachtet werden 40 Begriffe, wovon 30 Prozent durch den Schnitt repräsentiert werden.

7.2 Ergebnisse und Diskussion der quantitativen Untersuchung

Die Entscheidung, ob eine Kollokation eine valide oder invalide Kombination von Wörtern ist, stellt eine schwierige Aufgabe dar. Werden Kollokationen als Mehrwortausdruck betrachtet, lässt sich diese Entscheidung mit Hilfe von speziellen Kollokationswörterbüchern, wie dem *Oxford Collocations Dictionary (OCD)* [1] treffen. Das OCD wurde von Lexikographen erstellt und enthält alle Wörter, die häufig in Kombination mit einem Stichwort verwendet werden: Nomen, Verben, Adjektive, Adverbien und Präpositionen, sowie häufig genutzte Phrasen. Das Wörterbuch umfasst damit über 150.000 Kollokationen und 9.000 solcher Stichwörter.

Kollokationen wie „Steinbrück – Stinkefinger“ sind jedoch in keinem Kollokationswörterbuch enthalten. Dadurch, dass Kollokationen nicht als Mehrwortausdruck betrachtet werden, sondern eine quantitative Untersuchung von stark assoziierten Wortpaaren durchgeführt wird, kann diese Möglichkeit der Evaluation damit nicht verwendet werden.

Deshalb wird die Relevanz der Ergebnisse mit Hilfe von Artikeln der deutschen Tageszeitung und Online-Newsdiensten evaluiert. Um Unterschiede und Gemeinsamkeiten aufzuzeigen, wird außerdem der dem Wort zugrundeliegende Kontext betrachtet. Dieser wird durch die Tweets dargestellt, die Grund für die Assoziation sind. Dadurch können die Begriffe besser mit den Zeitungsartikeln verglichen werden und es besteht die Möglichkeit auch Verbindungen herzustellen, die nicht direkt aus dem Wort abgeleitet werden können.

Um die Ergebnisse mit der Tagespresse zu vergleichen, wird die Plattform „Wörter des Tages“[24] [14] von der Universität Leipzig verwendet. Diese Plattform zeigt Begriffe, die besonders relevant für einen ausgewählten Tag sind in Hinblick auf unterschiedliche Tageszeitungen. Ob ein Begriff relevant ist, wird bestimmt durch seine Häufigkeit an diesem Tag, verglichen mit seiner durchschnittlichen Häufigkeit über einen längeren Zeitraum hinweg. Für jeden Begriff werden exemplarisch Auszüge verschiedener Tageszeitungen präsentiert, die diesen enthalten. Zusätzlich wird ein Häufigkeitsvergleich angegeben, der die Häufigkeit des Wortes über einen festen Zeitraum beschreibt. Außerdem kann ein Wortkollokationsgraph betrachtet werden. Dieser enthält Wörter, die für den ausgewählten Begriff an diesem Tag besonders stark assoziiert sind. Die Begriffe sind unterteilt in die Kategorien: „Sportler“, „Sport“, „Politiker“, „Organisationen“, „Ereignis“, „Schlagwort“, „Ort“, „Personen aus Kunst, Kultur und Wissenschaft“, sowie einer Kategorie für sonstige Personen des öffentlichen Lebens.

Zur Durchführung der Evaluation werden alle Politiker der Plattform „Wörter des Tages“ als Schlüsselwörter für die Analyse verwendet. Die Tabelle C.1 im Anhang beinhaltet die identifizierten Begriffe, die mit dem jeweiligen Politiker in den Tageszeitungen vom 8. August bis 16. Oktober stark assoziiert wurden. Es liegen keine Daten für den Zeitraum vom 24. September bis zum 9. Oktober vor. Es folgt ein Vergleich der gesammelten Begriffe mit den Resultaten der entsprechenden Analyse.

Für diesen Vergleich wird zwischen exakter Übereinstimmung der Begriffe und einer Übereinstimmung durch Zunahme des Wortkontextes unterschieden. Eine exakte Übereinstimmung liegt vor, wenn zwei Wörter gleich sind oder diese die gleiche Bedeutung haben. Dies ist beispielsweise der Fall bei den Wörtern „Hilfspaket“ und „Rettungspaket“. Die beiden Wörter beschreiben das gleiche Thema und sind in diesem Zusammenhang als Synonym zu verstehen. Eine Übereinstimmung durch den Kontext liegt vor, wenn das Wort mit Hilfe der zugrundeliegenden Tweets thematisch zugeordnet werden kann.

Betrachtet werden für jeden Politiker jeweils 50 und 100 Begriffe mit einem minimalen Kantengewicht von drei. Des weiteren werden Hashtags und Usernamen aus den Ergebnissen ausgeblendet und Tweets vor und nach der Bundestagswahl betrachtet. Die Betrachtung von Hashtags und Usernamen ist nicht notwendig, da diese nicht in Zeitungen verwendet werden und damit keine für den Vergleich relevanten Begriffe darstellen. Die Evaluation wird mit dem normierten Log-Likelihood Maß durchgeführt.

[24] http://wortschatz.uni-leipzig.de/

	Log-Likelihood			
	50 betrachtete Begriffe		100 betrachtete Begriffe	
Politiker	**Korrekt**	**Korrekt mit Kontext**	**Korrekt**	**Korrekt mit Kontext**
Horst Seehofer	68,4%	89,5%	-	-
Rainer Brüderle	34,5%	52%	45%	55%
Sigmar Gabriel	50%	75%	-	-
Peter Altmaier	66,6%	-	83,3%	-
Frank-Walter Steinmeier	45,5%	73,3%	-	-
Katrin Göring-Eckardt	80%	-	-	-
Christine Lieberknecht	100%	-	-	-
Dirk Niebel	66,6%	83,3%	-	100%
Hermann Gröhe	75%	-	-	-
Katja Kipping	66,6%	-	-	-
Sabine Leutheusser-Schnarrenberger	35,7%	64,3%	-	-
Christian Ude	62,5%	75%	87,5%	-
Norbert Lammert	75%	-	-	-
Ronald Pofalla	66,6%	-	83,3%	-
Daniel Bahr	83,3%	-	-	-
Peter Ramsauer	58,3%	75%	66,6%	83,3%
Armin Laschet	66,6%	83,3%	-	-
Jürgen Trittin	61,5%	69,2%	77%	-
Hans-Peter Friedrich	83,3%	-	-	-
Ilse Aigner	83,3%	-	-	-
Wolfgang Schäuble	59%	100%	-	-
Joachim Gauck	100%	-	-	-
Thomas de Maizière	100%	-	-	-
Peer Steinbrück	70%	80%	-	-

Tabelle 7.1: Evaluationsergebnisse der Kollokationsanalyse für die ersten 50 und 100 stark assoziierten Begriffe. Zum Finden der Kollokationen wurde das Log-Likelihood Maß verwendet.

Tabelle 7.1 zeigt die von der Plattform „Wörter des Tages“ identifizierten Politiker über den oben genannten Zeitraum mit den Ergebnissen der Evaluation. Die „Korrekt“ Spalte beschreibt die exakte Übereinstimmung der Wörter, wohingegen die „Korrekt mit Kontext“ Spalte sowohl Wörter durch exakte Übereinstimmung enthält, als auch Wörter die erst durch den Kontext als übereinstimmend identifiziert werden konnten.

Durchschnittlich steigt die Übereinstimmung durch Zunahme des Kontextes unter den ersten 50 Begriffen um zehn Prozent. Unter den ersten 100 Begriffen ist dieser Anteil mit 1,8 Prozent deutlich kleiner. Das Wachstum von 3,4 Prozent der Übereinstimmungen von 50 zu 100 Begriffen ist ein Indiz, dass eine Betrachtung von weiteren Begriffen keine größere Übereinstimmung liefern wird.

Insgesamt erreicht die Analyse unter Berücksichtigung der eingestellten Parameter und untersuchten Politiker eine Übereinstimmung mit Kontext von 82,5 Prozent für die ersten 100 untersuchten Begriffe und 79,1 Prozent für die ersten 50 Begriffe.

Die kleinste exakte Übereinstimmung unter den ersten 50 stark assoziierten Begriffen liegt für die Politiker Sabine Leutheusser-Schnarrenberger (35,7 Prozent) und Rainer Brüderle (34,5 Prozent) vor. Es folgt eine kritische Betrachtung dieser quantitativen Ergebnisse unter Berücksichtigung des zugrundeliegenden Mediums Twitter.

Für die Politikerin Sabine Leutheusser-Schnarrenberger kann für drei thematische Gruppen (Anhang C.1) keine exakte Übereinstimmung gefunden werden:

Forderung nach einem Pornofilter im Internet

Dieses Thema beschäftigt sich mit der Forderung der CSU nach einem Pornofilter im Internet. Die Bundesjustizministerin Leutheusser-Schnarrenberger lehnt diese Forderung ab und sagte der Saarbrückener Zeitung in der Mittwochsausgabe am 7.8.2013: *„Wer Pornofilter beim Internetprovider fordert, sollte sich die Debatten der vergangenen Jahre ausdrucken lassen - und vielleicht das Grundgesetz“*. Die Plattform „Wörter des Tages“ präsentiert für den 8. August 2013 Ausschnitte aus Zeitungen, die dieses Zitat vom 7. August referenzieren. Das Korpus enthält das Wort „Pornofilter“, sowie Wörter die damit in Zusammenhang stehen jedoch nicht. Die Online-Suche von Topsy zeigt, dass Twitter-Nachrichten vom 7. August dieses Thema behandeln. Da das Korpus nur Tweets nach diesem Tag enthält, ist diese Thematik auch nicht Teil des Korpus. Eine Schlussfolgerung aus dieser Gegebenheit ist, dass das Medium Twitter Themen schneller adressiert als die Printmedien. Eine aktuelle Studie der Universität Edinburgh [35] hat dies untersucht und ist zu dem Schluss gekommen, dass Twitter nur bei bestimmten Ereignissen das schnellste Medium ist. Neben Sport waren dies auch Meldungen über politische Ereignisse.

Boykott der olympischen Winterspiele

Ähnlich verhält es sich mit dem Aufruf der Ministerin zum Boykott der olympischen Winterspiele in Sotschi am 4. August. Zeitungsausschnitte der Plattform „Wörter des Tages“ thematisieren diese Aussage erst am 8. August und stammen alle von einer Online-Zeitung mit russischer Domäne. Tweets zu diesem Thema können in Topsy am 4. August gefunden werden, weshalb das Korpus keine Informationen zu diesem Ereignis enthält.

NSA-Affäre

Die Wörter zum Thema NSA-Affäre finden keine exakte Übereinstimmung, können aber aus der Analyse mit dem Wort „Datenschutz“ in Verbindung gebracht werden. Das Wort repräsentiert Tweets, die mit Links auf externe Nachrichtenseiten verweisen. Diese Seiten behandeln die Forderung der Ministerin, dass für die EU-Ebene der deutsche Datenschutzmaßstab gelten sollte. Diese Debatte steht in direktem Zusammenhang mit den Wörtern zur NSA-Affäre. Dadurch gelten die Wörter erst durch den Kontext als übereinstimmend. Viele Informationen in Twitter werden durch Links verbreitet und können deshalb erst durch den Kontext identifiziert werden. In diesem konkreten Fall, sind viele der Wörter einem Thema zugehörig das in einer kleinen exakten Übereinstimmung resultiert.

Diese exemplarischen Auszüge und Erläuterungen zeigen, dass die quantitative Evaluation nur bedingt aussagekräftig ist und die Ergebnisse für die vorgestellten Politiker positiver sind, als diese auf den ersten Blick scheinen mögen. Nach der dadurch gewonnenen Einschätzung lässt sich feststellen, dass die Ergebnisse in hoher Übereinstimmung mit der Tageszeitung sind und die kontrastive Kollokationsanalyse politische Ereignisse reflektiert.

Diese Art der Analyse ist für Journalisten, Wähler und Politiker interessant, um die heiß diskutierten Themen der Bundestagswahl 2013 zu erfahren. Politiker können einsehen, wie die bei Twitter aktiven Bürger einen bestimmten öffentlichen Wahlkampfauftritt wahrnehmen und bewerten. Interessierte Bürger haben außerdem die Möglichkeit, politische Vorgänge transparenter zu erfahren und zu verstehen. Die Analyse ist damit ein wichtiges Instrument für den Wahlkampf und lässt sich als Spiegel der öffentlichen Meinung beschreiben. Ferner kann die Analyse auf eine andere Domäne übertragen werden und beispielsweise als Meinungsbarometer für die Fußball-Weltmeisterschaft 2014 dienen. Die vorgestellte Methode ist außerdem nicht auf Twitter-Daten beschränkt, sondern lässt sich auch auf andere Social-Media Daten anwenden. Dazu müssen nur die Vorverarbeitungsschritte dem jeweiligen Medium angepasst werden.

Der Umfang dieser Analyse ist jedoch auf die Anzahl der aktiven Twitter-Nutzer beschränkt. Offizielle Zahlen von Twitter zur Anzahl dieser Nutzer in Deutschland liegen nicht vor. Es existieren verschiedene Studien zu diesem Thema, die mit Hilfe von Befragungen und Schätzungen die Zahl der aktiven Nutzer bestimmen.

Aufgrund der unterschiedlichen Durchführung der Untersuchung unterscheiden sich die Studien deutlich in ihren Ergebnissen. Zum einen muss definiert werden, ob Twitter-Nutzer in Deutschland gemeint sind oder Nutzer, die deutschsprachige Tweets verfassen. Zum anderen muss aber auch diskutiert werden, ob inaktive Nutzer in die Studie aufgenommen werden sollen. Aufgrund dieser unterschiedlichen Betrachtungen und Ergebnisse lässt sich die Frage „Wieviele Twitter-Nutzer gibt es in Deutschland?“ nicht eindeutig beantworten.

Außerdem reflektiert die Analyse nicht alle Bevölkerungsschichten gleichermaßen. Aus einer Studie[25] der *Bitkom Research* vom 29. Juli 2013 geht hervor, das 68 Prozent der Internetnutzer zwischen fünfzig und vierundsechzig Jahren bei mindestens einem Social-Media Dienst registriert sind. Von den dreißig- bis vierzigjährigen Internetnutzern sind 76 Prozent und von den vierzehn- bis neunundzwanzigjährigen Nutzern 90 Prozent bei einem Netzwerk wie Facebook oder Twitter registriert. Befragt wurden 1016 Internetnutzer ab vierzehn Jahren.

Möglichkeiten, die vorgestellte Analyse zu verbessern, werden in Kapitel 8 diskutiert.

[25] `http://bitkom.org/de/themen/36444_76863.aspx`

8 Ausblicke und Verbesserungen

Verschiedene Aspekte des Softwaresystems können adressiert werden, um die Ergebnisse zu verbessern und die Benutzerfreundlichkeit der Analyse zu erhöhen. Basierend auf dem aktuellen Entwicklungsstand des Systems und den Evaluationsergebnissen können die folgenden Aspekte für zukünftige Arbeiten erweitert oder angepasst werden.

Präprozessor

Die Evaluation hat gezeigt, dass viele der wichtigen Informationen nicht direkt im Tweet zu finden sind. Es werden häufig Webseiten getwittert, die das eigentliche Thema referenzieren. Der Inhalt des Tweets besitzt in diesen Fällen keinen Informationsgehalt für die Analyse. Da diese Informationen jedoch hoch relevant sind, stellt das Aufnehmen dieser Informationen in das Korpus eine wichtige Aufgabe dar.

Wie bereits in Abschnitt 3.1 thematisiert haben Duplikate und Quasi-Duplikate eine enorme Auswirkung auf die Ergebnisse der Analyse. Exakte Duplikate werden im Präprozessorschritt entfernt, Quasi-Duplikate können mit diesem Verfahren jedoch nicht erkannt werden. Es sind diese Duplikate, die problematisch für die Analyse sind und Ergebnisse verfälschen. Zukünftige Arbeiten sollen sich mit diesem Thema befassen und sehr ähnliche Tweets entfernen. Das können beispielsweise Tweets sein, die sich nur in einer Zahl oder in der Art der Anführungszeichen unterscheiden.

Eine weitere Verbesserung umfasst das Ersetzen des NER Taggers im Präprozessor. Sehr bekannt ist der *Standord Names Entity Recognizer* [17] von der *Natural Language Processing Group*[26] der Universität Stanford. Dieser Tagger ist ein *Conditional Random Field (CRF)* Klassifizierer, dem ein wahrscheinlichkeitsbasiertes Modell zugrunde liegt. Ein deutsches Sprachmodell wird von Faruqui und Padó [16] zu Verfügung gestellt.

Signifikanzanalysesystem

Damerau [11] präsentiert eine Möglichkeit, welche die relative Frequenz von Wörtern zwischen zwei Korpora betrachtet, um domänenspezifische Kollokationen zu extrahieren. Mit Hilfe dieser relativen Wortfrequenzen lassen sich Kollokationen extrahieren, die charakteristisch für einen Korpus sind, wenn dieser mit einen anderen verglichen wird. Dieses Verfahren kann verwendet werden, um nur solche Kollokationen zu extrahieren, die aus der politischen Domäne stammen. Ein Problem im Zusammenhang mit diesem Verfahren ist jedoch, dass Wörter mit geringer Frequenz im Vergleichskorpus ein hoher Wert zugeordnet wird. Das Implementieren und Evaluieren dieses Ansatzes verbleibt als Erweiterung.

Visuelles System

Eine wichtige Untersuchung im Zusammenhang mit der Studie zur Bundestagswahl ist die Veränderung von Meinungen bzw. Wortassoziationen über die Zeit. Damit kann ergründet werden, welche Auswirkung ein politisches Ereignis in Twitter hat und auch, ob bestimmte Wahlkampfauftritte eine Veränderung im Verlauf der Wahl bewirken. Der aktuelle Entwicklungsstand unterstützt bereits die Möglichkeit Wortassoziationen vor und nach der Bundestagswahl zu betrachten. Es wird angestrebt, dies auf beliebige Zeitpunkte zu erweitern, so dass unterschiedliche Fixpunkte betrachtet werden können.

Die Kontextansicht visualisiert im aktuellen Entwicklungsstand alle einzigartigen Tweets, die Grund für die Wortassoziation sind. Eine mögliche Erweiterung ist das Anzeigen der Anzahl von Retweets (Duplikate) für jeden Tweet, um zu verdeutlichen welche Relevanz der jeweilige Tweet bzw. die dar-

[26] `http://nlp.stanford.edu/`

in enthaltene Aussage hat. Außerdem kann aus allen verfügbaren Tweets eine Menge von Tweets ausgewählt werden die präsentiert werden sollen. Ein Auswahlkriterium könnte beispielsweise die Häufigkeit der Retweets sein.

Sonstiges

Wie in Kapitel 3 bereits aufgeführt, können dem Kollokationsgraphen zur Laufzeit keine Daten hinzugefügt werden. Damit neue Twitter-Daten dem System während des Betriebs hinzugefügt werden können, müssen die Frequenzen der Wörter explizit in einer Datenstruktur, die einen schnellen Zugriff ermöglicht, gespeichert werden. Diese Frequenzen werden benötigt, um die neue Signifikanz für bestehende Kollokationen unter Berücksichtigung der neuen Daten zu berechnen. Mit dieser zusätzlichen Datenstruktur, soll die Auslastung der Graphdatenbank während des Betriebs minimiert werden. Die Datenstruktur lässt sich als Kollokationsmatrix $K_{m,n}$ modellieren und wird zum Start der Webanwendung aus der relationalen Datenbank in den Speicher geladen. Die Matrix repräsentiert eine Zuordnung von Wortpaaren zu Wortfrequenzen.

$$\mathrm{K}_{m,n} = \begin{pmatrix} 0 & & & \\ f_{2,1} & 0 & & \\ \vdots & \vdots & \ddots & \\ f_{m,1} & f_{m,2} & \cdots & 0 \end{pmatrix}$$

Abbildung 8.1: Kollokationsmatrix $K_{m,n}$.

Das Hinzufügen neuer Daten ist nachfolgend schematisch skizziert:

(1) Bestimmen von satzbasierten Kollokationen und deren Frequenzen für die neuen Daten.

(2) Die Kollokationsmatrix aktualisieren, falls der Wortpaareintrag vorhanden ist. Ansonsten muss ein neuer Eintrag für das Wortpaar in der Matrix erzeugt werden.

(3) Die neue Signifikanz für die zu aktualisierenden Kollokationen unter Verwendung der Kollokationsmatrix $K_{m,n}$ berechnen.

(4) Für diese Kollokationen Updatequeries generieren und zu einer Transaktion zusammenfassen.

(5) Die Transaktion wird abschließend in der Nacht auf der Graphdatenbank ausgeführt, so dass die neuen Daten den Benutzern am nächsten Morgen zu Verfügung stehen. Alternativ kann die Transaktion auch zu einem anderem beliebigen Zeitpunkt ausgeführt werden.

9 Konklusion

Das Ziel dieser Arbeit war es zu zeigen, in welchem Maße Twitter Ereignisse der realen Welt reflektieren kann. Diese Untersuchung basiert auf einer Studie zur deutschen Bundestagswahl 2013 in Twitter und umfasst über zehn Millionen Twitter-Nachrichten, die nach festgelegten Suchbegriffen ausgewählt wurden. Um die Forschungsfragen zu beantworten, wurde die kontrastive Kollokationsanalyse vorgestellt, die es ermöglicht, zwei Politiker oder Parteien anhand ihrer stark assoziierten Wörter zu vergleichen. Damit können Gemeinsamkeiten und Unterschiede der gegebenen Entitäten betrachtet und ferner analysiert werden, welche Wörter typischerweise mit der gegebenen Entität assoziiert sind. Diese Analyse basiert auf überzufällig häufigen Wortpaaren, sogenannten Kollokationen, welche in einer Vorverarbeitung aus den Twitter-Daten extrahiert und in einem Kollokationsgraphen gespeichert wurden. Zum Finden von aussagekräftigen Kollokationen sind verschiedene Wortassoziationsmaße vorgestellt und deren Vor- und Nachteile diskutiert worden.

Mit der entwickelten Webanwendung lassen sich kontrastive Analysen durchführen. Der Fokus dieser Anwendung liegt auf der einfachen Bedienbarkeit und soll auch für nicht erfahrene Benutzer leicht verständlich sein. Die kontrastive Kollokationsanalyse wird mit einem Tornadodiagramm dargestellt, welches für den Vergleich von Verteilungen geeignet ist und benutzergenerierte Inhalte intuitiv visualisiert, so dass eine kontrastive Untersuchung von zwei Politikern oder Parteien möglich ist. Außerdem bietet die Webanwendung dem Benutzer verschiedene Möglichkeiten an, die Ergebnisse der Analyse zu erforschen, wodurch dieser einen transparenteren Einblick in politische Vorgänge erhält.

Die Relevanz der Ergebnisse wurde mit Hilfe von Artikeln verschiedener deutscher Tageszeitungen evaluiert. Die Ergebnisse zeigen, dass der Schnitt der kontrastiven Kollokationsanalyse eine sinnvolle Reflektion politischer Ereignisse ist und die meisten relevanten Themen aus Zeitungsartikeln bezüglich dieser Analyse in Twitter abgebildet werden.

Projektseite

Die Projektseite dieser Arbeit ist unter der Adresse `http://maggie.lt.informatik.tu-darmstadt.de/thesis/bachelor/BTWitter/` erreichbar. Dort ist der Quellcode veröffentlicht und eine Demo der Webanwendung installiert.

A Suchbegriffe für die Datenerfassung

Tabelle A.1 zeigt die Suchbegriffe, welche in der Vorbereitung zur Datenerfassung identifiziert wurden. Die Suchbegriffe sind sortiert nach Kategorien und jeweils in Themen unterteilt.

Kategorie	Thema	Hashtag(#-Prefix)
Parteinamen	CDU	cdu
		union
	CSU	csu
	SPD	spd
	FDP	fdp
	Die Linke	linke
		linken
		linkspartei
	Die Grünen	grüne
		grünen
Spitzenkandidaten	Angela Merkel	merkel
		angie
		angelamerkel
		angela_merkel
	Peer Steinbrück	steinbrück
		peer_steinbrück
		peersteinbrück
	Rainer Brüderle	brüderle
	Gregor Gysi	gysi
	Nicole Gohlke	
	Jan van Aken	
	Caren Lay	
	Klaus Ernst	ernst
	Dietmar Bartsch	bartsch
	Sarah Wagenknecht	wagenknecht
	Diana Golze	
	Jürgen Trittin	trittin
	Katrin Göring-Eckardt	göring
Sonstige Politiker	Horst Seehofer	seehofer
	Philipp Rösler	rösler
	Claudia Roth	roth

	Sigmar Gabriel	gabriel
	Katja Kipping	kipping
	Wolfgang Schäuble	schäuble
	Christian Lindner	lindner
	Peter Altmaier	altmaier
	Frank-Walter Steinmeier	steinmeier
	Ronald Pofalla	pofalla
	Guido Westerwelle	westerwelle
	Thomas de Maizière	maiziere
	Ursula von der Leyen	vonderleyen
	Peter Ramsauer	ramsauer
	Ilse Aigner	aigner
	Sabine Leutheusser-Schnarrenberger	leutheusser
	Dirk Niebel	niebel
	Daniel Bahr	bahr
	Hans-Peter Friedrich	friedrich
	Christine Lieberknecht	lieberknecht
	Volker Bouffier	bouffier
Wahl 2013		btw2013
		btw13
		bundestagswahl
		wahl2013
		btw
		wahl
		bundestagswahl2013
		bundestagswahl13
		gehtwählen
Fernseh-Wahlkampf	Kanzler TV-Duell	tvduell
	Wie geht's Deutschland	wiebitte
		wiegehts
	Steinbrück Twitter Hall Meeting	fragpeer
	TV-Dreikampf	dreikampf
	TV-Elefantenrunde	elefantenrunde
Wahlthemen	Energiepolitik	energiepolitik
		energiewende
		energiewirtschaft
		atomenergie
	Finanzpolitik	solidaritätszuschlag

	soli
	steuerpolitik
	steuerreform
	steuern
Arbeitspolitik	arbeitslosigkeit
	hartz4
	hartz
Lohnpolitik	lohnpolitik
	mindestlohn

Tabelle A.1: Suchbegriffe für Twitter und Topsy.

B Zeitungsartikel zur exemplarischen Betrachtung der Analysen

B.1 Kontrastive Analyse: Angela Merkel und Steinbrück

Steinbrück - Circus Halligalli

[...] Die Antwort liegt auf der Hand: Eine knappe Woche vor der Wahl setzt Peer Steinbrück jetzt alles auf eine Karte. Wenn ein Stinkefinger im Magazin der „Süddeutschen Zeitung“ für Furore sorgt, muss wohl auch die lustig-infantile Schreisendung „Circus Halligalli“ auf ProSieben in Ordnung sein. Also wankte der Kandidat zu Rockmusik die Showtreppe herunter, das Licht der Scheinwerfer im Gesicht, vor eine Tribüne kreischender Jugendlicher. [..] Geht das? Wahlkampf machen neben zwei Moderatoren, die immer nur auf den nächsten Lacher aus sind? Beäugt von Schülern und Studenten, die sich auf Twitter schon vor der Sendung belustigten über dieses abstruse Zusammentreffen? So viel sei schon hier gesagt: Es ging erstaunlich gut. [...] Steinbrück meisterte das alles recht gut. Es gehört wohl zum Jobprofil eines Vortragsreisenden, solche Show-Angebote anzunehmen und dann auch gut zu funktionieren. [...] Steinbrück reagierte mit den üblichen Formeln: der Ehrgeiz, das Land nicht nur verwalten, sondern auch gestalten zu wollen. Er wolle keinen Kreisverkehr wie Merkel, sondern ab durch die Mitte, nach vorne. [...] Bei aller Chancenlosigkeit im Hinblick auf die Wahl, einen Erfolg hat Steinbrück inzwischen sicher erzielt: den Beweis seiner Behauptung, er sei „nicht so langweilig“ wie Merkel. Die Bundeskanzlerin nämlich hatten Winterscheidt und Heufer-Umlauf auch für einen Auftritt angefragt - und eine Absage bekommen.

Quelle: Spiegel Online, Dienstag, 17. September 2013 – 08:53 Uhr

Steinbrück - Erpressung

[...] Zuvor war bekannt geworden, dass ein Erpresser versucht hatte, Steinbrück zum Rückzug von seiner Kanzlerkandidatur zu zwingen. Einem Bericht der Bild-Zeitung zufolge warf der mutmaßliche Täter dem Ehepaar Steinbrück vor, vor etwa 14 Jahren eine Putzfrau illegal beschäftigt zu haben. Er drohte der Zeitung zufolge, dies öffentlich zu machen, wenn Steinbrück nicht auf seine Kandidatur verzichte.

Quelle: Süddeutsche Online, Sonntag, 8. September 2013 - 12:50 Uhr

Gemeinsamkeiten - Syrien-Konflikt

[...] Die SPD pocht daher auf eine politische Lösung. Eine Reaktion des Westens auf die mutmaßlichen Chemiewaffeneinsätze müsse „genau abgewogen werden“, mahnte Steinbrück im „Hamburger Abendblatt“. Bevor „leichtfüßig einer militärischen Logik gefolgt“ werde, müssten sich die Anstrengungen darauf richten, eine gemeinsame Position des Uno-Sicherheitsrats „zu scharfen Sanktionsmaßnahmen“ gegen Syrien zu finden. [...] In den vergangenen Tagen verschärfte sie allerdings den Ton, hält eine internationale Reaktion auf den Giftgaseinsatz in Syrien jetzt für „unabdingbar“. Merkel sprach am Mittwochabend am Telefon mit dem britischen Premierminister David Cameron. Beide seien sich einig gewesen: „Dieser Giftgasangriff ist eine Zäsur in dem schon lange andauernden internen Konflikt. Das syrische Regime darf nicht hoffen, diese Art der völkerrechtswidrigen Kriegführung ungestraft fortsetzen zu können“, erklärte Regierungssprecher Steffen Seibert. [...] Welche Konsequenzen für die Bundesregierung nach dem Giftgasangriff in Frage kommen, lässt sie offen.

Quelle: Spiegel Online, Donnerstag, 29. August 2013 – 08:25 Uhr

Gemeinsamkeiten - TV-Duell

[...] Stattdessen: gepflegte Langeweile, Herunterbeten von Parteiprogrammen. Keine Leidenschaft, nirgends. Ein Null zu Null zwischen den Kontrahenten ist das Ergebnis. Das war's. Dieses TV-Duell war kein Beispiel für lebendige Demokratie, sondern eine Enttäuschung. [...] Merkel merkelte, umarmte den Gegner, wich unangenehmen Fragen aus. Ihr Motto ist das immer gleiche: weiter so. [...] Dagegen der Herausforderer Peer Steinbrück: Er wollte angreifen. Doch jeder Versuch prallte an Merkel ab.

Quelle: Spiegel Online, Sonntag, 1. September 2013 – 22:45 Uhr

B.2 Kontrastive Analyse: Rainer Brüderle und Gregor Gysi

Mindestlohn - TV-Duell

[...] Stattdessen wollen sie beweisen, dass sie wissen, wie wenig manche verdienen: Die Bäckersfrau in Weimar (Trittin: 3,40 Euro), Friseure in Thüringen (Gsyi: 3,50 Euro). Brüderle hakt ein, dass Friseure sich den Mindestlohn selbst erstritten hätten. Äber Herr Brüderle", zischt Gysi.

Quelle: Süddeutsche Online, Dienstag, 3. September 2013 - 10:34 Uhr

Mindestlohn im Friseurhandwerk in Thüringen

[...] Ab Anfang August gilt für Friseure in Thüringen erstmals ein Mindestlohn von 6,50 Euro. [...] Zum 1. August 2014 steigt der Mindestlohn in den ostdeutschen Bundesländern 7,50 Euro in der Stunde. Mit einer weiteren Erhöhung auf 8,50 Euro zum 1. August 2015 gilt dann erstmals in Ost- und Westdeutschland ein einheitlicher Mindestlohn für die Friseurbranche.

Quelle: Thüringer Allgemeine, Sonntag, 20. Juli 2013 - 05:30 Uhr

C Stark assoziierte Begriffe der Politiker von der Plattform „Wörter des Tages“

Tabelle C.1 zeigt die stark assoziierten Begriffe der Plattform „Wörter des Tages“ für die in der Evaluation untersuchten Politiker. Politiker wie die Bundeskanzlerin Angela Merkel sind sehr präsent in deutschen Tageszeitungen und werden von der Plattform gefiltert.

Politiker	**Stark assoziierte Begriffe**
Horst Seehofer	Horst, Steinbrück, Ministerpräsident, CSU, Pkw-Maut, Schwesterpartei, Koalitionsvertrag, ADAC, Christian Ude, Merkel, Ausländer, CSU-Chef, WDR, ausweisen, WDR-Journalisten, Bayern, raus, Journalisten
Rainer Brüderle	Rainer, FDP-Spitzenkandidat, Solidaritätszuschlag, Wolfgang Schäuble, Lindner, Soli, Datenschutz, Guttenberg, Schavan, Rot-Rot-Grün, Solidarpakt, Energiewende, Sturz, Blamage, Bahn-Mitarbeiter, Personalmangel, Bahn-Vorstand, Trittin, Seehofer, Gysi, Rösler, FDP, Mindestlohn, Schlagabtausch, Spitzenkandidaten, Gregor, TV-Dreikampf, Gewerbesteuer, Post
Sigmar Gabriel	Sigmar, Angela Merkel, Spähaffäre, Peer Steinbrück, Steuererhöhungen, SPD-Chef, Koalition, Ude
Peter Altmaier	Peter, Bundesumweltminister, Umweltminister, CDU, Brüderle, Energiewende, Habeck, Strompreisbremse
Frank-Walter Steinmeier	Stromsteuer, Frank-Walter, Drohnen-Affäre, Thomas de Maizière, SPD-Franktionschef, Parlamentarischen Kontrollgremium, Kontrollgremium, Geheimdienste, NSA-Spähaffäre, Mitteldeutschen Zeitungen, Geheimdienstkoordinator
Katrin Göring-Eckardt	Katrin, Linkspartei, Geheimdiensten, Grünenspitzenkandidatin, Trittin, Spitzenkandidatin, Andreae, Franktionschefin, Fraktionsspitze, Hofreiter
Christine Lieberknecht	Christine, Ministerpräsidentin, CDU, Immunität, Thüringens, Staatsanwaltschaft
Dirk Niebel	Dirk, Entwicklungsminister, Uganda, Menschenrechtsorganisation, Interventionen, FDP-Kandidat
Hermann Gröhe	Hermann, Linkspartei, CDU-Generalsekretär, Sondierungsgespräch
Katja Kipping	Katja, Doppelspitze, Riexinger
Sabine Leutheusser-Schnarrenberger	Sabine, EU-Ebene, Justizministerin, Spähaffäre, Maßstab, Boykott, Winterspiele, CDU, FDP, Pornofilter, Sotschi, Friedrich, Innenmister, Datenschutzstandards
Christian Ude	Christian, Länderfinanzausgleich, Horst Seehofer, Spitzenkandidaten, SPD-Spitzenkandidat, Drohbrief, Hermann, SPD, Klage
Norbert Lammert	Doktorarbeit, Bundespräsident, Norbert, Wahlumfragen
Ronald Pofalla	Kontrollgremium, PKG, CDU, Spähaffäre, Nachrichtendiensten, Oppermann, Geheimdienst, NSA-Affäre, NSA, BND, Prism, Ronald, Kanzleramtchef
Daniel Bahr	Bundesgesundheitsminister, Organspende, Krankenversicherung, FDP, Versicherten, Privatversicherungen
Peter Ramsauer	Grube, Bahnchef, BER, Trittin, Leutheusser-Schnarrenberger, Mainz, Bahn, Hauptbahnhof, Zugausfälle, Verkehrsmister, Peter, Bundesverkehrsminister
Armin Laschet	CDU-Vorsitzende, Koalitionsvertrag, Rebellen, Friedrich, CDU-Vizevorsitzender, Armin

Jürgen Trittin	Jürgen, Kantinen, Gregor, Rainer Brüderle, TV-Dreikampf, Spitzenkandidaten, Mindestlohn, Schlagabtausch, Steuererhöhungen, Göring-Eckardt, Pädophilie-Debatte, Missbrauch, Grünen
Hans-Peter Friedrich	Hans-Peter, Bundesinnenminister, Leutheusser-Schnarrenberger, Militärschlages, CSU, EU-Flüchtlingspolitik
Ilse Aigner	Vergleichsportal, Verbraucherschutzministerin, CSU, Roth, Söder, Ilse
Wolfgang Schäuble	Griechenland, Hilfspaket
Joachim Gauck	Joachim, Bundespräsident
Thomas de Maizière	Thomas, Verteidigungsmister
Peer Steinbrück	TV-Duell, Angela Merkel, SPD-Herausforderer, SPD-Kanzlerkandidat, Fernsehduell, Schlagabtausch, Mittelfinger, Kanzlerkandidat, Osnabrück

Tabelle C.1: Stark assoziierte Begriffe der Plattform „Wörter des Tages“ für die in der Evaluation untersuchten Politiker.

Abkürzungsverzeichnis

POS Part-of-Speech

NER Named Entity Recognition

CAS Common Analysis System

JVM Java Virtual Machine

NP Nominalphrasen

PMI Pointwise Mutal Information

CRF Conditional Random Field

OCD Oxford Collocations Dictionary

JSON JavaScript Object Notation

LLC Leipzig Corpus Collection

UIMA Unstructured Information Management applications

lol laughing out loud

UTC Universal Time Coordinated

API Application Programming Interface

GPS Global Positioning System

XML Extensible Markup Language

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] *Oxford Collocations Dictionary for Students of English*. Oxford University Press, USA, 2002.

[2] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, und Aitor Soroa. A Study on Similarity and Relatedness Using Distributional and Wordnet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[3] Sabine Bartsch. *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Narr, 2004.

[4] Chris Biemann. Unsupervised Part-of-Speech Tagging Employing Efficient Graph Clustering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, COLING ACL '06, pages 7–12, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

[6] Norbert Blenn, Kassandra Charalampidou, und Christian Doerr. Context-Sensitive Sentiment Classification of Short Colloquial Text. In *Proceedings of the 11th International IFIP TC 6 Conference on Networking - Volume Part I*, IFIP'12, pages 97–108. Springer, 2012.

[7] Stefan Bordag. Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation. In *EACL*. The Association for Computer Linguistics, 2006.

[8] Dipak L. Chaudhari, Om P. Damani, und Srivatsan Laxman. Lexical Co-occurrence, Statistical Significance, and Word Association. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1058–1068, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] Kenneth W. Church und Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. In *Proc. ACL-1989*, pages 76–83, Vancouver, Canada, 1989.

[10] Bethany A. Conway, Kate Kenski, und Di Wang. Twitter Use by Presidential Primary Candidates During the 2012 Campaign. In *American Behavioral Scientist*, 2013.

[11] Fred Damerau. Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. *Inf. Process. Manage.*, 29(4):433–448, 1993.

[12] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3): 297–302, 1945.

[13] Ted E. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[14] Unni C. Eiken, Anja T. Liseth, Hans F. Witschel, Matthias Richter, und Chris Biemann. Ord i Dag: Mining Norwegian Daily Newswire. In *Advances in Natural Language Processing*, volume 4139 of *Lecture Notes in Computer Science*, pages 512–523. Springer, 2006.

[15] Stefan Evert und Brigitte Krenn. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proc. EACL-2001*, pages 188–195, Toulouse, France, 2001.

[16] Manaal Faruqui und Sebastian Padó. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany, 2010.

[17] Jenny R. Finkel, Trond Grenager, und Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[18] John R. Firth. Modes of Meaning. In *Papers in Linguistics 1934-1951*, pages 190–215. Oxford University Press, 1957.

[19] Winthrop N. Francis und H. Kucera. Brown Corpus Manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[20] Evgeniy Gabrilovich und Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[21] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, und Noah A. Smith. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[22] Thilo Götz und Oliver Suhre. Design and Implementation of the UIMA Common Analysis System. *IBM Syst. J.*, 43(3):476–489, 2004.

[23] Courtenay Honeycutt und Susan C. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42). Los Alamitos, CA.*, pages 1–10, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[24] John S. Justeson und Slava M. Katz. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, 1(1):9–27, 1995.

[25] Lars Kaczmirek, Philipp Mayr, Ravikiran Vatrapu, Arnim Bleier, Manuela Blumenberg, Tobias Gummer, Abid Hussain, Katharina Kinder-Kurlanda, Kaveh Manshaei, Mark Thamm, Katrin Weller, Alexander Wenz, und Christof Wolf. Social Media Monitoring of the Campaigns for the 2013 German Bundestag Elections on Facebook and Twitter. *CoRR*, abs/1312.4476, 2013.

[26] Anders O. Larsson und Hallvard Moe. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media and Society*, 14(5):729–747, 2012.

[27] Bing Liu und Lei Zhang. A Survey of Opinion Mining and Sentiment Analysis. In *Mining Text Data*, pages 415–463. Springer US, 2012.

[28] Christopher D. Manning und Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[29] M. Meckel und K. Stanoevska-Slabeva. Auch Zwitschern muss man üben: Wie politiker im deutschen Bundestagswahlkampf 'twitterten'. *Neue Zürcher Zeitung*, 2009.

[30] David Milne und Ian H. Witten. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of AAAI 2008*, 2008.

[31] David Nadeau und Satoshi Sekine. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

[32] Melanie Neunerdt, Michael Reyer, und Rudolf Mathar. A POS Tagger for Social Media Texts trained on Web Comments. *Polibits*, 48:61–68, 2013.

[33] Jerzy Neyman und Egon S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337.

[34] Daisuke Okanohara und Jun ichi Tsujii. Text Categorization with All Substring Features. In *SDM*, pages 838–846. SIAM, 2009.

[35] Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, und Luke Shrimpton. Can Twitter Replace Newswire for Breaking News? In *ICWSM*. The AAAI Press, 2013.

[36] Addison Phillips und Mark Davis. BCP 47 – Tags for Identifying Languages, 2006.

[37] Uwe Quasthoff, Matthias Richter, und Chris Biemann. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC*, pages 1799–1802, Genoa, 2006.

[38] Reinhard Rapp. The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[39] Ines Rehbein. Fine-grained POS Tagging of German Tweets. In *Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*, volume 8105, pages 162–175. Springer, 2013.

[40] Anne Schiller, Simone Teufel, Christine Stöckert, und Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, Universität Stuttgart, Universität Tübingen, Stuttgart, Germany, 1999.

[41] Helmut Schmid. Improvements In Part-of-Speech Tagging With an Application To German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, 1995.

[42] Uladzimir Sidarenka, Tatjana Scheffler, und Manfred Stede. Rule-Based Normalization of German Twitter Messages. In *Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*. Springer, 2013.

[43] John Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, 1991.

[44] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, und Isabell Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proc. AAAI-2010*, pages 178–185, Atlanta, GA, USA, 2010.

[45] Mathukumalli Vidyasagar. The Complete Realization Problem for Hidden Markov Models: A Aurvey and Some New Results. *MCSS*, 23(1-3):1–65, 2011.

[46] Hao Wang, Dogan Can, Abe Kazemzadeh, Francois Bar, und Shrikanth S. Narayanan. A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of ACL*, 2012.