verbundprojekt postdemokratie
und neoliberalismus

Sebastian Dumm

Topic Modelle

3/5 – Serie „Atomenergiediskurs“

Reihe: ePol Text Mining Verfahren (eTMV)

Was ist eTMV?

Die ePol Text Mining Verfahren (eTMV) stellen in kurzen, in sich abgeschlossenen Einheiten die verschiedenen Analyseverfahren des Leipzig Corpus Miners (LCM) vor (Wiedemann/Niekler 2014). Zur Illustration der Verfahren verwenden alle fünf Module dieser Serie Beispiele aus dem Kontext des Diskurses um die Nutzung der Atomenergie in Deutschland. Ziel der eTMV ist es, für interessierte SozialwissenschaftlerInnen eine best practise der Verwendungsmöglichkeiten und -grenzen von Text Mining anhand des LCM zu veranschaulichen – auch wenn bislang keine Vorerfahrungen mit Text Mining bestehen.

Der Datenbestand des ePol-Projekts besteht aus 3.495.822 deutschsprachigen Zeitungsartikeln, die sich wie folgt zusammensetzen: Frankfurter Allgemeine Zeitung (Stichprobe, 1959–2011, 200.389 Artikel), Süddeutsche Zeitung (6027 Ausgaben, 1992–2011, 1.505.714 Artikel), die tageszeitung (7821 Ausgaben, 1986–2012, 1.391.981 Artikel) und Die Zeit (3841 Ausgaben, 1946–2012, 397.729 Artikel). Für die Funktionalität aller Analysen einschlägig ist Wiedemann, Gregor / Niekler, Andreas (2014): Analyse qualitativer Daten mit dem Leipzig Corpus Miner, Hamburg / Leipzig (=Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus Discussion Paper 5), online unter: http://www.epol-projekt.de/discussion-paper/discussion-paper-5/.

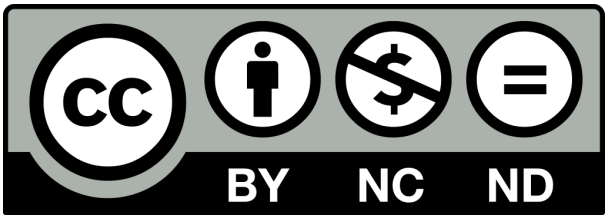
ISSN 2363-6335

Zitationsweise und Lizenz

Dumm, Sebastian (2014): Topic Modelle, Hamburg/Leipzig (=ePol Text Mining Verfahren, Serie „Atomenergiediskurs“, Modul 3/5).

Rückfragen bitte an die unter **www.epol-projekt.de/kontakt** angegebene Email-Adresse.

Alle Teile der Reihe sind unter **www.epol-projekt.de/etmv** online abrufbar. Das gesamte Material der Reihe ist unter **Creative Commons 4.0 international** als **BY-NC-ND** lizenziert.



Inhalt

1. Definition

Topic Modelle identifizieren globale Begriffszusammenhänge auf der Ebene von Dokumentkollektionen. „Die mit diesem Verfahren automatisch berechneten Begriffszusammenhänge, sogenannte Topics, können als latente Sinnkomplexe oder Thematiken interpretiert werden, deren Verteilung über den Korpus insgesamt untersucht werden kann.“ (Wiedemann/Lemke/Niekler 2013, S.109).

2. Erkenntnishorizont

Im Sinne des *Blended-Reading-Ansatzes** (Lemke/Stulpe 2015) ist ***Text Mining**** **ein modularer Prozess**. Dieser setzt sich aus verschiedenen computergestützten Analyseverfahren und von der Forscherin/dem Forscher zu leistenden Interpretationen von Einzeltexten zusammen. *Blended-Reading** nimmt – was die computergestützten Verfahren anbelangt – eine analytisch-prozedurale Gewichtung vor, die von einfachen, strukturierenden hin zu komplexeren, inhaltlichen Verfahren reicht. Topic Modelle gehören, wie die Kookkurrenzanalysen (eTMV 2), zu den **sekundären Operationen des *Text Mining****.

Als Topic Modelle wird eine Klasse von **probabilistischen (wahrscheinlichkeitstheoretischen) Modellen** bezeichnet, die versuchen, thematische Strukturen in Textkorpora zu entdecken. Dabei arbeiten die meisten Modelle unüberwacht, das heißt datengetrieben, ohne die Hinzugabe zusätzlicher Wissensressourcen des Forschers/der Forscherin zum Modellierungsprozess. Topic Modelle basieren auf der Annahme, dass jedes **Dokument eine Mischung aus mehreren Themen** zu unterschiedlichen Anteilen enthält. Diese Themen (engl. Topic) werden als Zusammenhang zwischen verschiedenen Worten angesehen. Dabei sind Anzahl, Umfang und Zusammensetzung der Themen in einem großen Textkorpus nicht direkt beobachtbar, sondern bilden die latenten Variablen im Analyseprozess. Auf die Topics (latente Variablen) wird in einem Rechenverfahren über das gemeinsame Auftreten von Worten (manifeste Variable) im gesamten Textkorpus geschlossen. Diese Topics (latenten Variablen) lassen sich dann als thematische Zusammenhänge interpretieren. Jedem Wort kommt dabei in einem bestimmten Thema eine spezifische aus der Wahrscheinlichkeit seines Auftretens im Thema abgeleitete Gewichtung zu. Die hochgewichteten Worte stellen dann den inhaltlich interpretierbaren thematischen Zusammenhang her. Damit besteht ein Dokument aus einer bestimmten Menge an Topics, wobei die spezifische Wahrscheinlichkeit für jedes Topic angegeben werden kann. Topic Modelle **identifizieren somit die latenten Sinnzusammenhänge** von Worten als Topics über Einzeldokumente hinaus[1].

Zwei Analyseleistungen stehen dabei für einen *Belended-Reading** Prozess im Vordergrund, einerseits die Exploration (1) und anderseits die Filterung (2) eines bestehenden Textkorpus:

[1] Das erste Topic Modell die *Latent Dirichlet Allocation (LDA)** wurde von Blei, Ng und Jordan (2003) vorgestellt.

(1) Topic Modelle geben für jedes Dokument eine Wahrscheinlichkeitsverteilung der in ihm enthaltenen Topics an. Mit dieser Information lassen sich **Dokumente selektieren,** die bestimmte Topics zu einem bestimmten Anteil enthalten. Die Zählung dieser Dokumente, bzw. die Aggregation der in ihnen enthaltenen Wahrscheinlichkeitsmasse für ein Topic lässt sich als Zeitreihe visuell darstellen. Mit Topic Modellen lassen sich latente Sinnzusammenhänge einer Dokumentkollektion sichtbar machen, die als Themen oder Kontexte betrachtet werden können. Zusätzlich können Dokumentkollektionen nach Themen gefiltert und die Präsenz der Themen kann diachron im Korpus als Längsschnitt der Häufigkeiten visualisiert werden. Durch die Nutzung von Topic Modellen ist damit die **Exploration von großen Textkorpora** möglich, da die einzelnen Subthemen innerhalb eines Diskurses automatisch identifiziert werden.

(2) Die durch das Topic Modell berechneten Themenanteile können auch dazu genutzt werden, bestimmte **Themen aus dem Korpus gezielt auszuschließen**. Dies ist vor allem dann sinnvoll, wenn durch Doppeldeutigkeiten eine Themenvermischung entstanden ist. Des Weiteren können Topic Modelle in einem bestehenden Korpus dazu genutzt werden, **spezifische Subdiskurse zu extrahieren** und so *Subkorpora** für die weitere Analyse zu erstellen.

Topic Modelle ermöglichen somit die inhaltliche Erschließung und Bereinigung von Korpora für weitere Analyseschritte.

3. Vorbereitung der Analyse

Bei der Vorbereitung der Analyse eines Korpus mittels eines Topic Modelles ist das angestrebte Erkenntnisinteresse im Rahmen des Oben beschriebenen Erkenntnishorizonts von zentraler Bedeutung. Soll das Topic Modell (1) zur **Exploration** eines bestehenden Korpus genutzt werden, so steht die **Beschreibung des Korpus** und der im Korpus vorhandenen Subthemen im Vordergrund der Analyse. In diesem Fall ist eine auf adäquaten Suchbegriffen aufgebaute Dokumentenkollektion Ausgangspunkt der Untersuchung. Des Weiteren kann ein Topic Modell auch als Zwischenschritt im Rahmen des *Blended-Reading-Ansatzes** verwendet werden, um so (2) eine **Bereinigung eines Korpus** oder die **Extraktion von Subthemen** aus einem bestehenden Korpus zu ermöglichen und damit den Untersuchungskorpus zu verfeinern.

Wie bereits in eTMV 1 beschrieben, ist der Atomenergiediskurs durch die Suche des Wortes **Atomenergie** nicht vollständig zu beschreiben. Durch die Nutzung von *Diktionären** konnte gezeigt werden, dass sich der Atomenergiediskurs über den Begriff der Atomenergie nur unzureichend beschreiben lässt. Durch die **Kookkurrenzanalysen** (eTMV 2) konnten signifikante Wortzusammenhänge im erweiterten Textkorpus identifiziert werden. Die Frage, welche Subthemen im Diskurs enthalten sind, blieb jedoch unbeantwortet. Hierfür eignet sich das Verfahren der **Topic Modelle**. Diese stehen im Leipziger Corpus Miner (LCM) in zwei verschiedenen Varianten zur Verfügung: einerseits als *Online-LDA* * Implementierung und andererseits als *Hierarchical Pitman-Yor Prozess**.

Durch diese Verfahren lassen sich die latenten Sinnzusammenhänge im LCM in unterschiedlichen **Zeitintervallen** darstellen (Tage, Wochen, Monate, Jahre). Die Darstellung der **Dokumentenanzahl**, welche mit einer bestimmten (einstellbaren) **Auftretenswahrscheinlichkeit** in einem Topic enthalten sind, ist ebenso möglich wie der relative Anteil des ermittelten Subthemas in Bezug auf den analysierten Korpus. Bei der Berechnung der Topics können minimale und maximale Schwellenwerte des Wortvorkommens angegeben werden, die den

Einbezug von Worten in die Analyse begrenzen. Dieses Vorgehen ermöglicht den Ausschluss von *Stoppwörtern** und Spezialvokabular[2].

[2] Als Untersuchungseinheit lassen sich Uni-, Bi- und Trigramme einstellen. Es ist aber in verschiedenen Abstufungen auch die Kombination von verschiedenen *N-grammen** möglich, d.h. es können Ein- oder Mehrworteinheiten oder eine Kombination von verschiedenen Mehrworteinheiten zur Analyse genutzt werden.

4. Topic Modelle im Kontext des „Atomenergiediskurses“

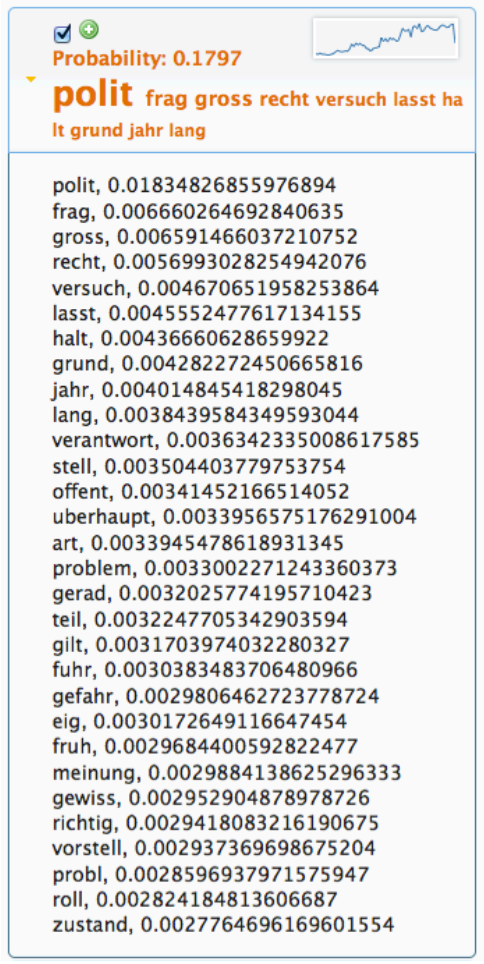
Durch Topic Modelle werden die in einem Korpus existierenden latenten Sinnzusammenhänge automatisch als einzelne Topics dargestellt. Diese repräsentieren Subthemen in einem thematisch eingegrenzten Korpus. Dies ist als **explorative Methode** besonders dann sinnvoll, wenn ein Korpus dahingehend untersucht werden soll, in welche **Subthemen** sich ein Diskurs aufspaltet. Auf diese Weise wird ein Überblick über die in einem thematisch zusammengestellten Korpus, zum Beispiel dem Atomenergiediskurs, existierenden Subthemen hergestellt. Diese Herangehensweise ermöglicht eine **Beschreibung des Untersuchungskorpus** hinsichtlich der unterschiedlichen in einem Thema vorkommenden Diskursstränge.

Die Berechnung eines Topic Modells auf Grundlage des bekannten Atomenergiediskurs-Korpus[3] mittels der *Online-LDA** Implementierung ergibt eine Einteilung des Korpus in 25 Subthemen[4]. Der LCM gibt die Subthemen einer Kollektion mit der Wahrscheinlichkeit ihres Auftretens im Korpus als Zahl aus. Ebenso werden die Worte benannt, die am häufigsten im Topic verwendet werden. Die Auftretenswahrscheinlichkeit der einzelnen zum Topic gehörenden Worte ist per Drop-down Menü einzusehen (Abb. 4.1), wobei die Topics in absteigender Wahrscheinlichkeit ihres Auftretens im Untersuchungskorpus dargestellt werden.

[3] Die nachfolgende Analyse basiert auf dem bereits aus den ersten beiden Einheiten bekannten Korpus zum Atomenergiediskurs, der aus der Suchanfrage „(atomenergie OR kernenergie OR nuklear* OR kernkraft) AND deutsch*“ entstanden ist und aus 17.331 Dokumenten besteht.

[4] Bei der Berechnung der Topics wurden alle Worte die in weniger als 0.05 % und in mehr als 99 % aller Dokumente vorkommen entfernt. Als untersuchte Worte wurden, wenn nicht anders beschrieben, *gestemmte** Unigramme verwendet.

Abb. 4.1 - Topic aus dem Atomdiskurs mit der höchsten Auftretenswahrscheinlichkeit



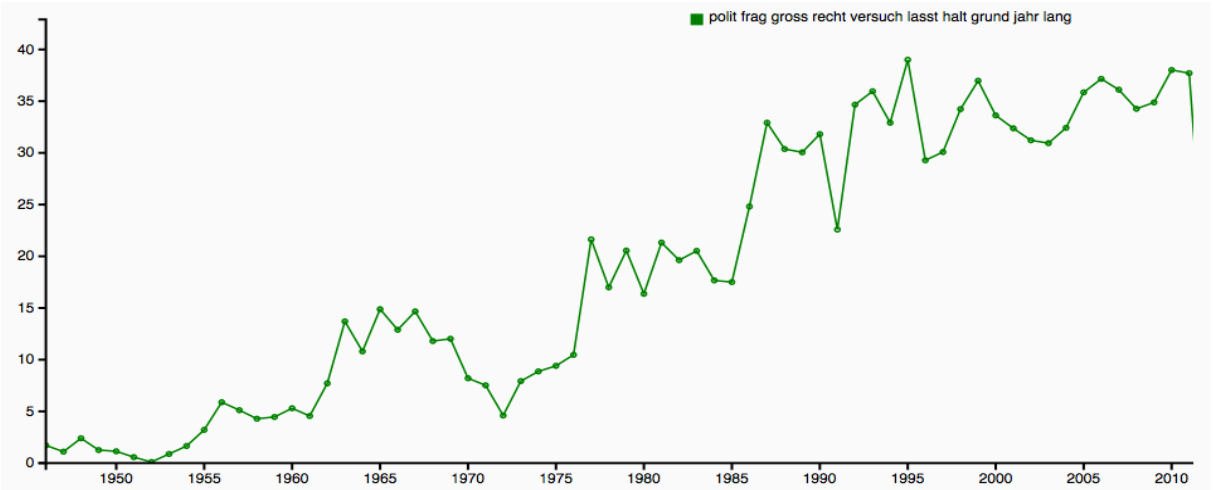
Wie aus dieser Abbildung ersichtlich wird, ist das Topic mit der größten Auftretenswahrscheinlichkeit[5] ein sehr allgemeiner Themenbereich. Da die Topics nur aus der Angabe der Auftretenswahrscheinlichkeit einzelner Worte bestehen, ist es Aufgabe der Forscherin / des Forschers , die **inhaltliche Dimension der Topics** thematisch zu kennzeichnen bzw. herauszuarbeiten und dem Topic einen am Inhalt orientierten Namen zu geben. Diese interpretatorische Aufgabe des Forschers / der Forscherin gibt dann der statistisch errechneten Auftretenswahrscheinlichkeit erst eine inhaltliche Sinndimension und ermöglicht damit die weitere Analyse der einzelnen Subthemen. Für das Topic in Abbildung 4.1 wäre ein theoretisch hergeleiteter Name beispielsweise „Allgemeine politische Zuschreibung“. Erleichtert wird dieser

[5] Die Auftretenswahrscheinlichkeit eines Topics in einem Korpus wird in Abbildung 1 durch den Wert: „Probability“ ausgegeben und bezeichnet den Anteil den das Topic am Untersuchungskorpus einnimmt.

Prozess der inhaltlichen Bestimmung eines Subthemas durch die zu einem Topic zugehörigen Dokumente, welche auch in einem Close Reading Prozess einzeln gesichtet werden können.[6] Dass es sich bei dem Topic mit der **höchsten Auftretenswahrscheinlichkeit** um ein sehr allgemeines Topic handelt, ist nicht weiter verwunderlich, denn grundlegend ist bei der Analyse von Topic Modellen zu beachten, dass ein Text sich nicht nur mit einem Thema beschäftigt. Ein Text, in unserem Fall ein Zeitungsartikel, ist zwar zu einem bestimmten Thema geschrieben worden, aber zur Erklärung dieses Themas sind weitere Sinnbezüge zu Nachbarthemen notwendig. Für das in Abbildung 4.1 dargestellte Topic ist dies sehr gut verständlich, da zur politischen Auseinandersetzung mit dem Atomenergiediskurs auch der politische Prozess der BRD im Allgemeinen mitgebetrachtet werden muss und so in sehr vielen der Dokumente auffindbar ist.

Dieses unspezifische Topic „Allgemeine politische Zuschreibung“ (Abb. 4.2) verläuft annähernd parallel zu der absoluten Berichterstattung, wie diese sich auf Grundlage des Korpus Atomenergie darstellt (Abb. 1.3).

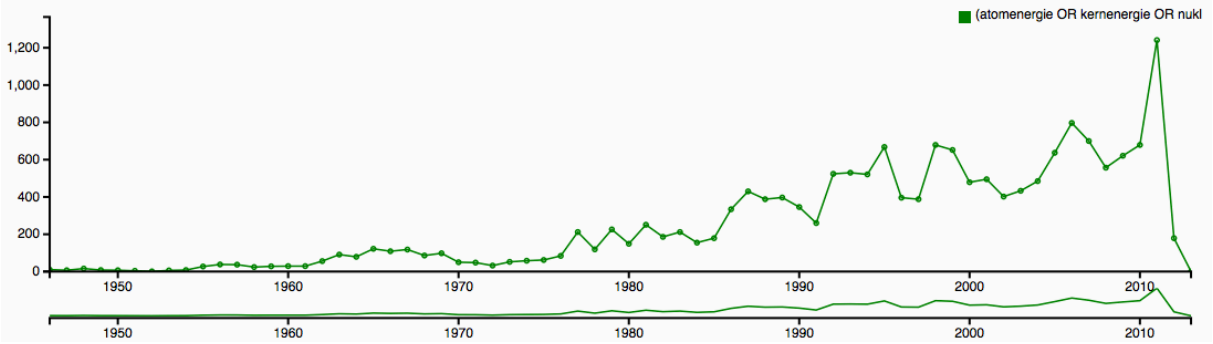
Abb. 4.2 – Dokumente, die das Topic „Allgemeine politische Zuschreibung“ enthalten[7]



[6] Für einen kurzen Überblick im Close Reading Prozess ist es hilfreich nur die Dokumente auszuwählen, die das Topic mit einer hohen Wahrscheinlichkeit beinhalten. Im Falle des Topics „Politik Allgemein“, dass mit einer Wahrscheinlichkeit von 0.1797% in den Dokumenten enthalten ist, sollte dann ein prozentualer Wert oberhalb der durchschnittlichen Auftretenswahrscheinlichkeit gewählt werden, um hinreichend aussagekräftige Artikel zu erhalten.

[7] Wobei sich die Angaben der Skala links auf die Wahrscheinlichkeit des Auftretens in allen Artikeln über den Beobachtungszeitraum eine Prozentangabe ist. Dieses Topic umfasst damit 7.540 des 17.331 Dokumente umfassenden Korpus.

Abb. 4.3 - Dokumente im Korpus Atomenergiediskurs



Anders stellt sich der Verlauf von anderen Subthemen im Korpus dar welche einem **spezifischeren Thema** zugeordnet werden können. So ist die bereits in den vorangegangenen Einheiten zu unterscheidende Debatte zwischen der Wortverwendung von Kernenergie und Atomenergie abzulesen. Das Topic: „Bundesregierung und Kernenergie“ (Abb. 4.4) hat beispielsweise eine Auftretenswahrscheinlichkeit von 0.0556 im Atomenergiekorpus.

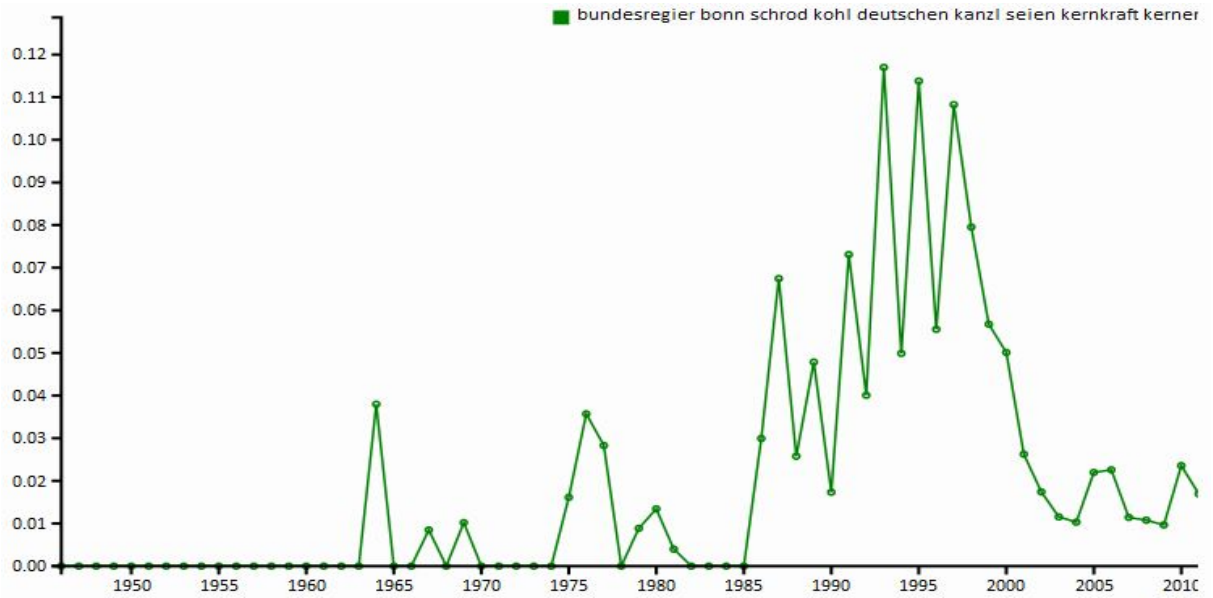
Abb. 4.4 - Topic „Bundesregierung und Kernenergie“

Der Verlauf dieses Subthemas „Bundesregierung und Kernenergie“ im Atomenergiediskurs (Abb. 4.5) ist ein anderer als der des Gesamtdiskureses über die Atomenergie (Abb. 4.3). Wie bereits in eTMV 1 beschrieben, steigt die Relevanz dieses Subthemas mit dem Reaktorunfall in Tschernobyl, fällt dann aber um das Jahr 2000 mit dem beschlossenen Ausstieg aus der Atomenergie durch die Rot-Grüne Bundesregierung wieder ab.

Abb. 4.5 – Dokumente, die das Topic „Bundesregierung und Kernkraft“ enthalten

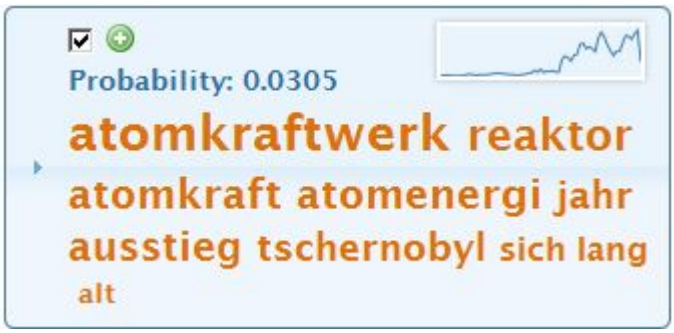
Auch der Blick auf die Verteilung des Topics „Bundesregierung und Kernkraft“ relativ zum gesamten Atomenergiediskurs bestätigt den Rückgang des Themas im Korpus. Abbildung 4.6 zeigt den relativen Anteil an Dokumenten die das Topic im Verhältnis zum gesamten Korpus enthalten. Dabei wird der Anteil des Topics am Korpus für Dokumente dargestellt, die das Topic „Bundesregierung und Kernkraft“ mit einer Mindestauftretenswahrscheinlichkeit von 20 Prozent enthalten.

Abb. 4.6 – Dokumente, die das Topic „Bundesregierung und Kernkraft“ zu mindestens 20 % enthalten



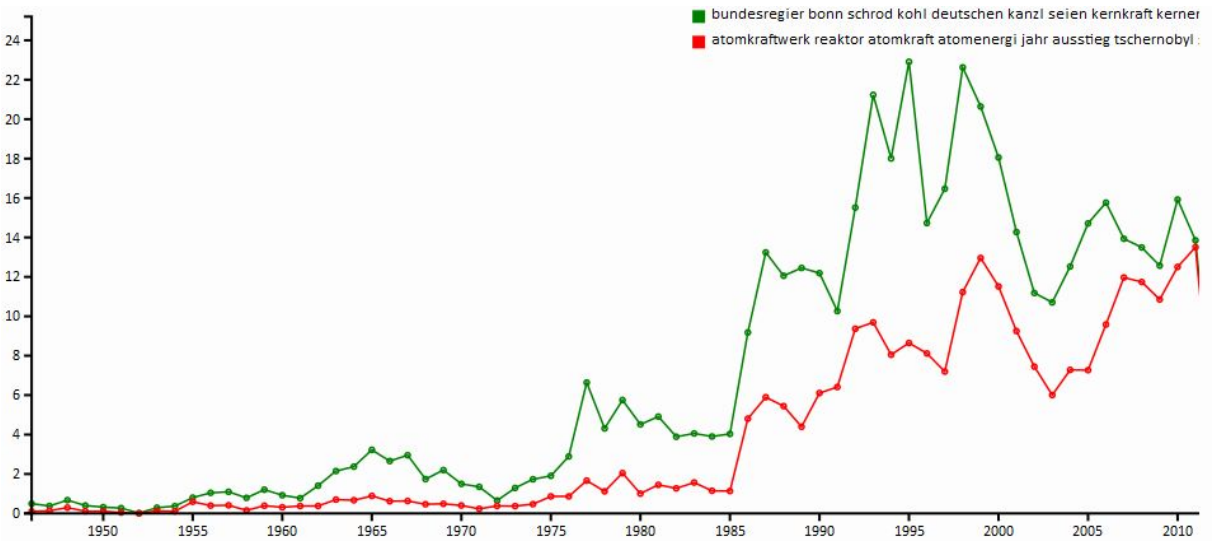
Parallel zu diesem Subthema „Bundesregierung und Kernenergie“ entwickelt sich aber mit dem Reaktorunfall in Tschernobyl ein weiteres Topic, das sich dezidiert mit den Gefahren der Atomenergie und dem Ausstieg aus dieser Form der Energiegewinnung beschäftigt. Im Atomenergiediskurs hat das Topic „Atomkraftwerke und Ausstieg“ die etwas geringere Auftretenswahrscheinlichkeit von 0.0305 (Abb. 4.7).

Abb. 4.7 - Topic „Atomkraftwerke und Ausstieg“



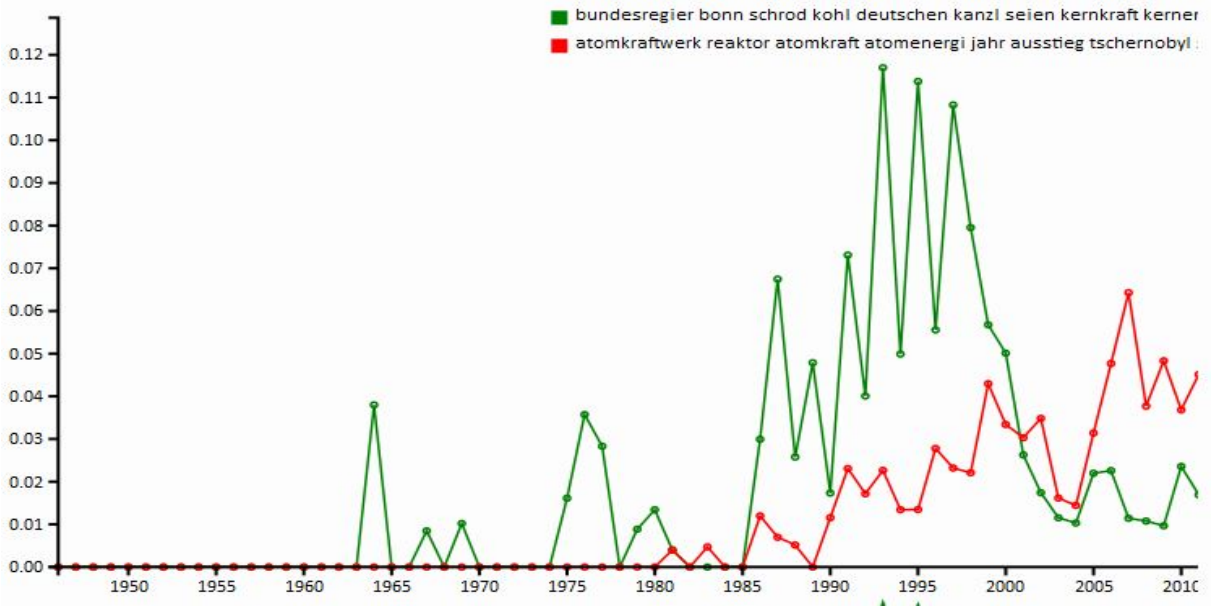
In Abbildung 4.8 werden die Graphen der Auftretenswahrscheinlichkeit der beiden Subthemen zusammen dargestellt. Hier zeigt sich, dass das Topic „Bundesregierung und Kernenergie“ (grün) häufiger im Atomenergiediskurs vorhanden ist als das Topic „Atomkraftwerke und Ausstieg“ (rot). Während das Topic „Bundesregierung und Kernenergie“ (grün) zum Ende des Beobachtungszeitraumes an Relevanz verliert, steigt jedoch die Bedeutung des Topics „Atomkraftwerke und Ausstieg“ (rot) tendenziell an (Abb. 4.8).

Abb. 4.8 – Dokumente, die das Topic „Atomkraftwerke und Ausstieg“ und „Bundesregierung und Kernkraft“ enthalten



Der Vergleich der beiden Topics relativ zum Gesamtkorpus und mit einer Mindestauftretenswahrscheinlichkeit von 20 Prozent im Dokument bestätigt diesen Trend (Abb. 4.9). In der relativen Ansicht offenbart sich sogar, dass das Topic „Atomkraftwerke und Ausstieg“ (rot) ab dem Jahr 2001 einen höheren relativen Anteil im Verhältnis zum Gesamtkorpus aufweist als das Topic „Bundesregierung und Kernkraft“ (grün).

Abb. 4.9 - Dokumente, die das Topic „Atomkraftwerke und Ausstieg“ und „Bundesregierung und Kernkraft“ zu mindestens 20 % enthalten

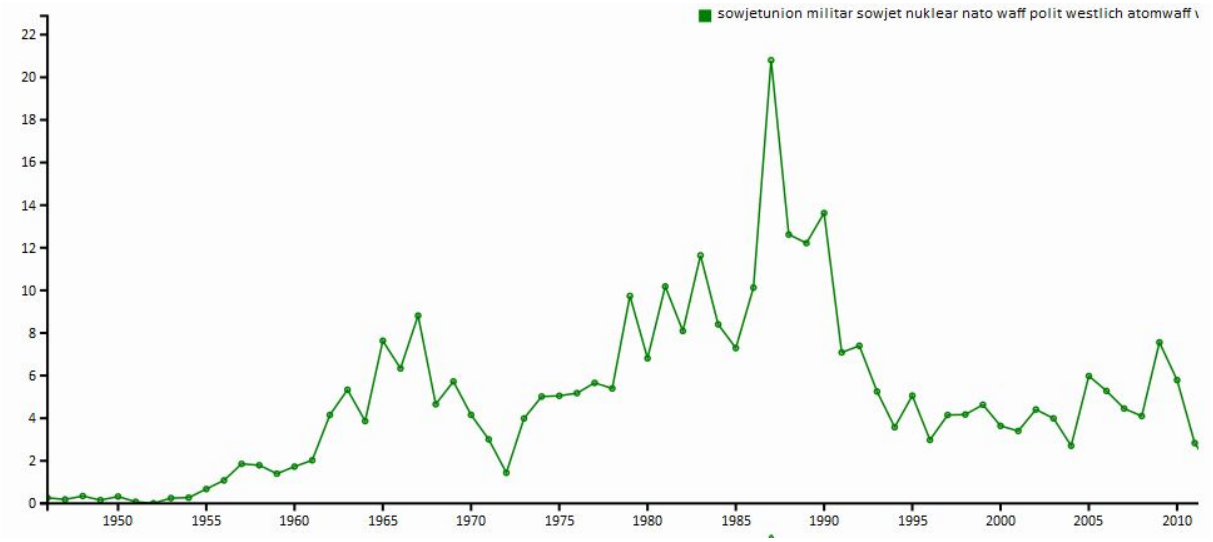


Die Topic Modelle ermöglichen also eine in Abschnitt 2 zum Erkenntnishorizont (1) angesprochene **explorative Untersuchung** der im Korpus enthaltenen Themenbereiche. Durch diese Analyseart kann der Forscher / die Forscherin relativ leicht einen Überblick über die thematische Bandbreite seines Korpus erhalten und so zu einer **Bewertung der Korpuszusammensetzung** für die Beantwortung seiner Forschungsfrage gelangen.

Topic Modelle ermöglichen es aber auch, wie oben im Erkenntnishorizont (2) beschrieben, einzelne **Subthemen aus einem bestehenden Korpus zu selektieren** um anschließend Dokumente, die dieses Thema zu einem bestimmten Anteil enthalten aus der Kollektion zu entfernen oder in einen eigenständigen Korpus zu überführen. Beide Funktionen erlauben eine **Verfeinerung des Untersuchungskorpus** für weitere Analyseschritte. Der erste Fall der Bereinigung eines Korpus ist vor allem dann einzusetzen, wenn aufgrund einer zu weitgefassten Suchanfrage oder aufgrund von Doppeldeutigkeiten im Korpus unerwünschte Bestandteile die weitere Analyse verfälschen würden. Der zweite Fall der Extraktion und Überführung eines Subthemas in einen eigenständigen Korpus ist dann besonders sinnvoll, wenn eine vertiefende Analyse eines Subdiskurses durchgeführt werden soll. So kann zum Beispiel der bereits in eTMV 1 identifizierte Diskurs *zu erneuerbaren Energien* innerhalb des Atomenergiediskurses extrahiert und genauer untersucht werden.

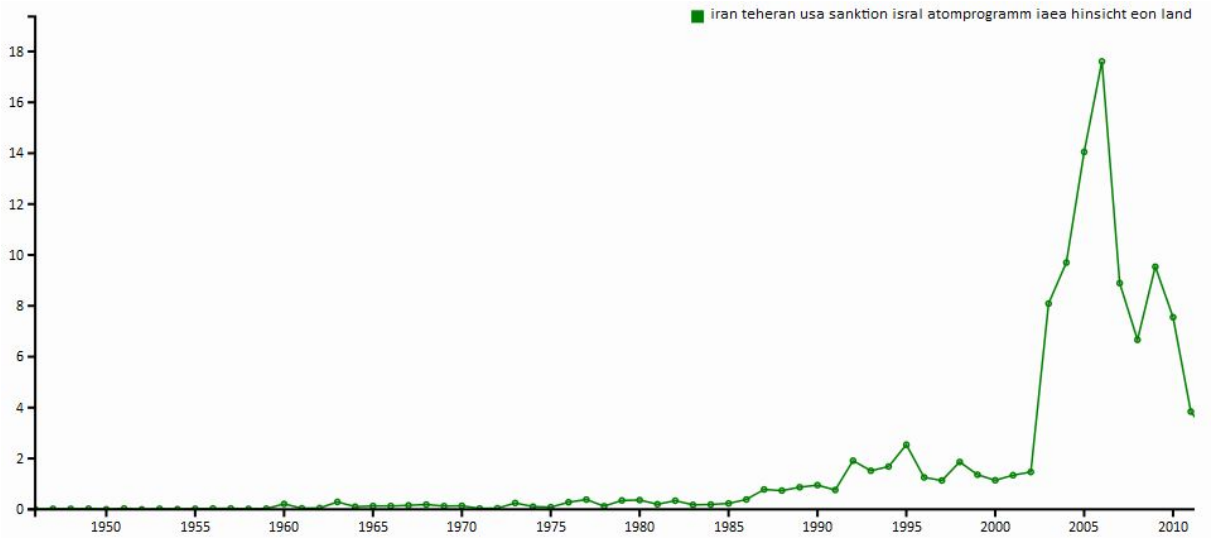
Der erste Fall der **Bereinigung von Textkorpora** kann als Zwischenschritt in einem *Blended-Reading** Prozess eingesetzt werden, um den Zuschnitt des Untersuchungskorpus zu verbessern. So zeigt sich, dass im Korpus zum Atomenergiediskurs, bedingt durch die Suchanfrage (vgl. eTMV 1) Themen Eingang gefunden haben, die hinsichtlich der Energieversorgung durch Atomenergie nur peripher von Interesse sind. Beispielsweise ist das Thema der nuklearen Auseinandersetzung zwischen der NATO und der Sowjetunion (Abb. 4.10) im Atomenergiediskurs vorhanden, obschon es für die Frage nach dem Umgang mit einer Form der Energiegewinnung irrelevant ist.

Abb. 4.10 - Dokumente, die das Topic „Sowjetunion und NATO“ enthalten[8]



Auch Themen, die nicht dem deutschen Atomenergiediskurs zugrechnet werden können, jedoch nicht durch den Suchanfragenzusatz **„ … AND deutsch*“** herausgefiltert werden konnten, sind im Korpus eindeutig identifizierbar. So unter anderem die internationale Diskussion um das Atomprogramm und die damit mögliche atomare Bewaffnung des Irans (Abb. 4.11)[9].

Abb. 4.11 - Dokumente, die das Topic „Atomprogramm Iran“ enthalten

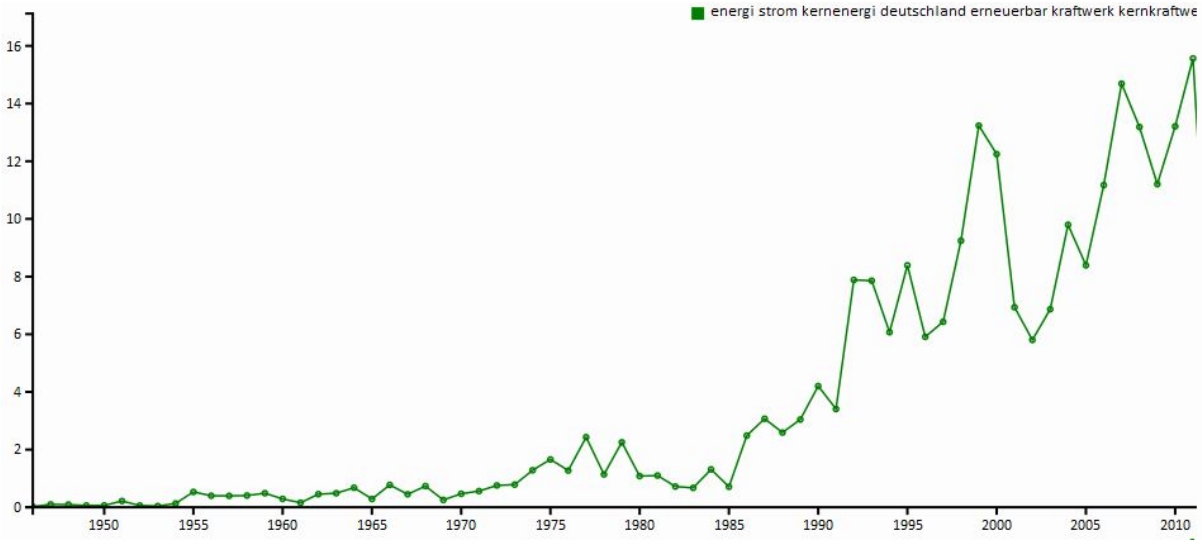


[8] Das „Topic Sowjetunion und NATO“ hat eine Auftretenswahrscheinlichkeit von 0.0437 % im Untersuchungskorpus.
[9] Das Topic „Atomprogramm Iran“ hat eine Auftretenswahrscheinlichkeit von 0.0128 % im Untersuchungskorpus.

Diese Beispiele zeigen wie wichtig eine **detaillierte Untersuchung der Korpuszusammensetzung** ist. Durch die Auswahl der relevanten Topics und einer damit einhergehenden Bereinigung und Neuzusammenstellung der relevanten Themenaspekte des Atomenergiediskus lassen sich Irritationen, welche aufgrund inadäquater Diskurselemente auftreten, vermeiden. Damit ist die **Bereinigung des Untersuchungskorpus** ein wichtiger Schritt im *Blended-Reading-Ansatz**, da der Korpus auf diese Weise für einen detaillierten Close Reading Prozess zugeschnitten werden kann.

Der zweite Fall wie ein Topic Modell als Zwischenschritt in den *Blended-Reading-Ansatz** eingesetzt werden kann, besteht in der **Extraktion eines Subthemas** und der damit einhergehende Erstellung eines eigenständigen *Subkorpus**. Ein Beispiel für den Zuschnitt eines *Subkorpus** ist das bereits aus den vorangegangen Einheiten bekannte Subthema der erneuerbaren Energien im Atomenergiediskurs. Dieses Subthema ist auch im Topic Modell als eigenständiges Diskurselement aufzufinden. Dabei findet eine Vermischung der Themen Kernenergie, Atomenergie bzw. Atomausstieg und erneuerbare Energien statt (Abb. 4.12). Dieses Thema gewinnt ab 1986 an Relevanz im Atomenergiediskurs und hat in den beiden Zeiträumen um den Atomausstieg der Rot-grünen Bundesregierung und dem von Kanzlerin Merkel vollzogenen „Zickzack-Kurs“ mit abschließendem Ausstieg zwei relevante Spitzen.

Abb. 4.12 - Dokumente, die das Topic „erneuerbaren Energien im Atomenergiediskurs“



Soll dieses Subthema einer differenzierten Untersuchung unterzogen werden, besteht die Möglichkeit, die Artikel dieses Topics in einen eigenständigen Korpus zu überführen. Dafür

muss eine adäquate Auftretenswahrscheinlichkeit des Topics in den zu extrahierenden Artikeln gefunden werden. Bei der Bestimmung der Auftretenswahrscheinlichkeit eines Topics pro Artikel gilt es, zwischen einer ausreichenden Anzahl von Artikeln und einer zu undifferenzierten Zurechnung der Artikel für eine weitere Analyse abzuwägen [10]. Wie bereits oben beschrieben, besteht ein Artikel nicht nur aus einem Thema. Da das Thema Atomenergie und erneuerbare Energien eine Auftretenswahrscheinlichkeit von 0.0297 % aufweist, sollte der Schwellenwert für die einzelnen Artikel nicht zu hoch angesetzt werden. Des Weiteren sollte der Untersuchungszeitraum für die Analyse eingeschränkt werden. Artikel die sich vor 1986 mit dem Thema Atomausstieg und erneuerbare Energie auseinandersetzten, scheinen nicht dem für uns relevanten Atomenergiediskurs zu entstammen. Bei einem Schwellenwert der Auftretenswahrscheinlichkeit von 20% und einem Untersuchungszeitraum ab 1986 erhalten wir einen neuen *Subkorpus** der aus 502 Artikeln besteht. Diese Artikel können dann in einen eigenständigen *Subkorpus** überführt werden und erneut mit einem Topic Modell berechnet werden. Bei einer anschließenden inhaltlichen Sichtung der Topics dieses neuen *Subkorpus** kann verifiziert werden, dass diese den Themenbereich Atomenergieausstieg und erneuerbare Energie repräsentiert (Abb. 4.13).[11]

[10] Wird der Schwellwert des Anteils des Topics „erneuerbare Energien und Atomenergie“ beispielsweise auf 50% gesetzt (d.h. das Topic macht 50% der einzelnen Artikel aus), werden nur zwei Artikel aus dem Jahr 2011 diesem Topic zugerechnet. Wird der Schwellenwert des Anteils des Topics „erneuerbare Energien und Atomenergie“ beispielsweise auf 5% gesetzt (d.h. das Topic macht 5% der einzelnen Artikel aus), werden diesem Topic 4216 Artikel zugerechnet. Ein standardisierter Schwellenwert kann hier nicht angegeben werden, da dieser nur im Analyseprozess durch Einbezug von Close Reading Schritten ermittelt werden kann.

[11] Diese 502 Artikel stellen nicht den Diskurs um erneuerbare Energie im Gesamten dar, sondern nun die Diskurselemente die explizit in Zusammenhang mit dem Atomausstieg diskutiert wurden.

Abb. 4.13 – Topics im *Subkorpus** zum Thema Atomausstieg und erneuerbare Energie

Die fünf wahrscheinlichsten Topics dieses neu gebildeten Korpus zeigen, dass die Auswahl durch die Extraktion mittels eines Topic Modells den Themenbereich Atomenergieausstieg und erneuerbare Energie repräsentiert. Mit 502 Artikeln ist dieser *Subkorpus** dann auch einer weiteren Analyse in einem Close Reading Prozess zugänglich.

Abschließend existiert durch das Analyseverfahren der Topic Modelle auch noch eine weitere Möglichkeit die Analyseergebnisse zu nutzen. Die in den Topics enthaltenen Worte kön-

nen auch mittels einer **Term Extraktion** aus dem Datensatz exportiert werden und so als Grundlage für weitere Suchen, beispielsweise unter Anwendung der bereits in eTMV 1 vorgestellten Suchverfahren, genutzt werden. Im vorliegenden Fall können dann zum Beispiel die Worte aus dem erneuerbaren Energiediskurs im Atomdiskurs als Ausgangspunkt für weitere Suchanfragen dienen (Abb. 4.14).

Abb. 4.14 - Term Extraktion aus dem *Subkorpus** Atomenergie und erneuerbare Energie

Term Extraction Result

Seite 1 von 2 25 Zeige 1 – 25 von 46

Wordform	LDA Stochastic Inference Weigl
prozent	1.334762512620062
energi	1.2449782942237837
strom	1.0996888814287151
jahr	0.691724879311445
erneuerbar	0.6143926607521817
deutschland	0.5371933885662934
klimaschutz	0.3466549580463082
ausstieg	0.3404408131924237
energietrag	0.30494704852014526
kernenergi	0.27707225268373953
kohl	0.23770768262311226
kraftwerk	0.2122926387966655
kilowattstund	0.1977514936956092
energiew	0.1690243964079996
studi	0.15987980148815933
atomenergi	0.15761674023436884
gesellschaft	0.15323733398340816
atomausstieg	0.1432734486512897
erdgas	0.14064710209825274
kohlekraftwerk	0.13937945379503255
energieversorg	0.1340990161388211
erzeugt	0.126731825995444
milliard	0.12016002539894438
megawatt	0.11855575039206097
lang	0.11350022786693822

Download Seite 1 von 2 25 Zeige 1 – 25 von 46

Der Vorteil dieser Anwendung von Topic Modellen und Term Extraktion besteht darin, dass nicht eine möglicherweise unvollständige Wortliste als Operationalisierung eines Sachverhaltes verwendet wird. Durch die Verbindung der beiden Verfahren kann eine umfangreiche

Liste, die auf tatsächlich in diesem Diskurs verwendeten Worten beruht, zur Suche herangezogen werden.

5. Glossar

Begriff	Erläuterung
Blended-Reading-Ansatz	Der Blended-Reading-Ansatz (Lemke/Stulpe 2015) verweist mit Blick auf die von Franco Moretti etablierte Unterscheidung von Close und Distant Reading auf die Notwendigkeit, beide Ebenen im Forschungsprozess integriert zu behandeln. Eine bestimmte Verfahrensfolge im Text Mining (von der Verwendungskonjunktur zum Verwendungskontext) ist durch den jederzeit möglichen Rückgriff auf die der Sprachstatistik zugrunde liegenden Einzeltexte zu begleiten.
Diktionär	Ein Diktionär ist eine beliebig lange, mindestens aber zwei Worte umfassende Liste von Worten, die ein abstraktes oder komplexes Konzept inhaltlich bestimmt und für eine Suchanfrage erschließt. Abstrakte oder komplexe Konzepte, wie etwa Ökonomisierung oder Atomenergiediskurs, die nicht aus sich selbst heraus den Inhalt repräsentieren, für den sie stehen, bedürfen konkretisierender Begriffe, die im Rahmen von Suchanfragen diejenigen Texte auffinden, die das Konzept am besten abbilden.
Hierarchical Pitman-Yor Prozess	Der HPY ersetzt die Dirichlet-Verteilung eines LDA-Prozesses, so dass die optimale Anzahl an Topics anhand der Daten abgeschätzt werden kann. Für Modelle die mit dem HPY berechnet werden muss also kein K als Prozessparameter vorab definiert werden.
LDA	Latent Dirichlet Allocation (LDA) ist das erste Topic Modell, dass von Blei; Ng & Jordan vorgestellt wurde (2003). Der Grundgedanke besteht in der probabilistischen Modellierung eines generativen Prozesses, bei dem Dokumente als eine Auswahl von K Themen aus einer Zufallsverteilung angesehen werden. Einzelne Themen wiederum bestehen aus einer Zufallsverteilung aus Wörtern aus dem gesamten Vokabular einer Dokument-Kollektion. Die Parameter dieser Verteilungen können in einem aufwändigen Rechenverfahren aus den empirischen beobachtbaren Wort-Verteilungen in den Dokumenten eines Korpus inferiert werden. Die A-posteriori-Wahrscheinlichkeit eines Modells gibt dann Auskunft darüber, welches der K Topics zu welchen Anteilen in einem Dokument vorkommt sowie welche Begriffe des Vokabulars mit welcher Wahrscheinlichkeit in einem Topic auftreten.

N-gramm	Im Rahmen der Zerlegung von Texten in einzelne Fragmente, etwa Worte, bezeichnen N-Gramme die Anzahl der betrachteten Wortgruppen. Ein Uni- oder auch Monogramm ist dabei die Bezeichnung für eine Einworteinheit. Eine Zweiworteinheit wird als Bigramm, eine Dreiworteinheit dementsprechend als Trigramm bezeichnet.
Online-LDA	Online-LDA ist eine spezifische Implementierung des LDA-Modells, bei dem die Modell-Parameter in einem Datenstream dokumentweise angepasst werden. Auf diese Weise lässt sich die Berechnung eines Topic Modells erheblich beschleunigen, so dass sehr große Collections analysiert werden können
Subkorpus	Ein Subkorpus ist eine Teilmenge von Artikeln aus dem Gesamtkorpus, die auf Basis einer Artikelsuche (im Modus Simple, Detailed oder Custom) generiert wurde und Gegenstand weiterer Analyseverfahren ist.
Stemming	Beim Stemming werden im Prozess der Verarbeitung die Worte auf ihren Wortstamm reduziert. (Bsp. ging → ging, gehen → geh, geht →geh)
Stoppwort	Stoppworte sind wenig Bedeutung tragende Worte wie „der“, „die“, „das“ die bei den meisten Text Mining Anwendungen generell nicht berücksichtigt.
Text Mining	Nach Gerhard Heyer et al. (2006: 3) bezeichnet der Begriff „[...] computergestützte Verfahren für die semantische Analyse von Texten [...], welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten, unterstützen.“

6. Verwendete Literatur

Blei, David. M. / Ng, Andrew. Y./ Jordan, Michael. I. (2003): Latent dirichlet allocation. In: The Journal of Machine Learning Research, Jg. 3, S. 993–1022.

Heyer, Gerhard et. al (2006): Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. Bochum.

Lemke, Matthias / Stulpe, Alexander (2015): Text und soziale Wirklichkeit. Theoretische Grundlagen und empirische Anwendung durch Text Mining Verfahren am Beispiel des Bigrams ‚soziale Marktwirtschaft', erscheint in Zeitschrift für Germanistische Linguistik, Themenheft „Automatisierte Textanalyse“, hg. von Noah Bubehofer und Joachim Scharloth.

Wiedemann, Gregor/Lemke, Matthias/Niekler, Andreas(2013): Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentation in der Bundesrepublik Deutschland 1949-1011. Ein Werkstattbericht. In Zeitschrift für Politische Theorie 4 (1), 99-115.

7. Weiterführende Literatur

Blei, David M. (2012): Probabilistic Topic Models. Surveying a suite of algorithms that offer a solution to managing large document archives. In: Communications of the ACM 55, 77–84.

Wiedemann, Gregor / Niekler, Andreas (2014): Analyse qualitativer Daten mit dem Leipzig Corpus Miner, Hamburg / Leipzig (=Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus Discussion Paper 5).

Wiedemann, Gregor / Niekler, Andreas (2014): Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries. In: Terminology and Knowledge Engineering 2014, Berlin.

eTMV-Serie „Atomenergiediskurs“

1 – Frequenzanalyse und Diktionäransatz
2 – Kookkurrenzanalyse
3 – Topic Modelle
4 – Annotation und Klassifikation
5 – Sentimentanalyse

www.epol-projekt.de/etmv