verbundprojekt postdemokratie
und neoliberalismus

Schriftenreihe des Verbundprojekts

„Postdemokratie und Neoliberalismus“

Discussion Paper Nr.5

Analyse qualitativer Daten mit dem Leipzig Corpus Miner

Zur Vorbereitung des Workshops „Text Mining in der Politikwissenschaft

Gregor Wiedemann & Andreas Niekler

www.epol-projekt.de

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

UNIVERSITÄT LEIPZIG

HELMUT SCHMIDT
UNIVERSITÄT
Universität der Bundeswehr Hamburg

ISSN 2363-6335

Zitierweise:

Wiedemann, Gregor / Niekler, Andreas (2014):
Analyse qualitativer Daten mit dem Leipzig Corpus Miner

Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus
Discussion Paper
Helmut-Schmidt-Universität Hamburg (UniBw) und Universität Leipzig

Zusammenfassung

Der Leipzig Corpus Miner (LCM) ist eine Webanwendung, die verschiedene Text Mining Verfahren für die Analyse großer Mengen qualitativer Daten bündelt. Durch eine einfach zu bedienende Benutzeroberfläche ermöglicht der LCM Volltextzugriff auf 3,5 Millionen Zeitungstexte, die nach Suchbegriffen und Metadaten zu Subkollektionen gefiltert werden können. Auf dem Gesamtdatenbestand sowie auf den Subkollektionen können verschiedene computergestützte Auswertungsverfahren angewendet und zu Analyseworkflows kombiniert werden. Damit ermöglicht der LCM die empirische Analyse sozialwissenschaftlicher Fragestellungen auf Basis großer Dokumentkollektionen, wobei qualitative und quantitative Analyseschritte miteinander verschränkt werden können. Dieser Artikel gibt einen Überblick über die Analysekapazitäten und mögliche Workflows zur Anwendung des LCM.

Inhaltsverzeichnis

Vorbereitung zum Workshop von Seiten des Teilprojekts Informatik

Der LCM ist eine Serveranwendung, die im Rahmen des Projekts „ePol – Postdemokratie und Neoliberalismus“ in der Abteilung Automatische Sprachverarbeitung an der Universität Leipzig entwickelt wird. Bei der Software handelt es sich um einen Prototypen, der für die Analysezwecke des ePol-Projekts erstellt wurde und auf die Bedürfnisse eines kleinen Nutzerkreises abgestimmt wurde. Folglich sind die Analysekapazitäten für bestimmte Zwecke optimiert. Für den Workshop werden Sie in die Benutzung des LCM eingewiesen, so wie wir sie uns im Rahmen des ePol-Projekts erdacht haben. Darüber hinaus werden Sie aber womöglich Benutzungsweisen anwenden, welche wir nicht vorhergesehen haben. Dadurch sowie durch den erstmaligen gleichzeitigen Einsatz der Software durch Sie als Teilnehmende kann es zu unerwarteten Ereignissen oder Fehlfunktionen kommen. Beim LCM handelt es nicht um ein fehlerfreies, fertiges 'Produkt'. Wir möchten Sie hiermit also bitten, geduldig zu sein, wenn nicht alles reibungslos laufen sollte. Für Fragen zur Software stehen wir Ihnen immer zur Verfügung. Wir hoffen, ein hilfreiches Stück Software für spannenden Analyse entwickelt zu haben.

Andreas Niekler, Gregor Wiedemann

1. Datenbestand

Im Rahmen des Workshops erhalten Sie Zugriff auf lizenzierte Daten von deutschen Tages- bzw. Wochenzeitungen, die Sie für ihre Forschungsfragen auswerten können.

Zeitung	**abgedeckter Zeitraum**	**Anzahl Ausgaben**	**Anzahl Artikel**
Frankfurter Allgemeine Zeitung (FAZ)	1959 – 2011	Stichprobe	200.389
Süddeutsche Zeitung (SZ)	1992 – 2011	6027	1.505.714
die tageszeitung (taz)	1986 – 2012	7821	1.391.981
Die Zeit	1946 – 2012	3841	397.729
Gesamt			3.495.822

Die Zeitungsartikel liegen in einer Datenbank vor, auf die mit Hilfe der Webserveranwendung LCM zugegriffen werden kann. Dort können Sie nach Artikeln suchen, diese im Volltext lesen, Kodierungen im Text vornehmen und verschiedene Text Mining Verfahren auf gefilterten Subkollektionen anwenden.

Für den Zugriff auf den LCM erhalten Sie einen Benutzeraccount zu Beginn des Workshops. Dieser Account dient gleichzeitig dazu, ihre spezifischen Analyseschritte bzw. Ergebnisse abzuspeichern. Auswahlen von Dokumenten werden als Subkollektionen benutzerspezifisch gespeichert. Sie können auch zu mehreren Personen mit einer Subkollektion arbeiten. Das gleiche gilt für Kategoriesysteme für den Fall, dass Sie Kodierungen am Text vornehmen wollen.

2. Analysemöglichkeiten

2.1 Analysen auf dem Gesamtkorpus

2.1.1 Dokumentselektion

Über den LCM erhalten Sie Zugriff auf den Volltext der 3,5 Mio. Zeitungsartikel. Dieser Korpus muss selbstverständlich für die Untersuchung konkreter Forschungsfragen auf möglichst relevante Artikel eingeschränkt werden. Dazu stellt der LCM die Möglichkeit einer komplexen Volltextsuche bereit. Artikel können nach einzelnen Suchbegriffen, Kombinationen von Suchbegriffen und Metadaten (Zeitraum, Zeitung, Ressort, Autor, Begriffsvorkommen in Überschriften oder Artikeltext) gefiltert werden. Für einfache Suchen nach Keywords und Metadaten stehen Formulare auf der Benutzeroberfläche zur Verfügung. Für komplexe Anfragen gibt es eine Anfragesprache (Lucene Query Language) mit der genauere Filterungen vorgenommen werden können.

Im Volltextindex sind die folgenden Informationen erfasst und durchsuchbar:

1. Überschrift*
2. Unterüberschrift*

3. Dachtitel*
4. Dokumenttext (Absätze)*
5. Ressort
6. Subressort
7. Schlagworte (nur bei manchen ZEIT-Daten)
8. Autor
9. Verlag
10. Nachrichtenagentur
11. Named-Entities
12. Datum
13. Seite

Die mit * gekennzeichneten Datenfelder sind in zwei Varianten volltextindiziert: normal und „raw“:

- *normale Indexierung:* ermöglichst Suche nach Schlüsselbegriffen ohne Berücksichtigung von Groß-Kleinschreibung aber mit Stemming und Entfernung von Stoppworten1
- *„raw“-Indexierung:* ermöglichst Suche nach Schlüsselbegriffen ohne Berücksichtigung von Groß-Kleinschreibung, ohne Stemming und ohne Entfernung von Stoppworten

Im LCM finden sich 3 Such-Möglichkeiten:

- Simple: einfache Suche nach Schlagwortsuche auf den indexierten Textfeldern. Auswahl ob auf "_raw" oder normalen Indexfeldern gesucht wird
- Detailed: zusätzliche Sucheinschränkung nach Metadaten
- Custom: Eingabe eines Suchstrings in Lucene Query Syntax.

Für eine Custom Query kann die mächtige Lucene Query Syntax direkt mit folgenden Feldnamen genutzt werden

1. Title bzw. Title_raw
2. Subtitle bzw. Subtitle_raw
3. Rooftitle bzw. Rooftitle_raw
4. Paragraph bzw. Paragraph_raw
5. Section
6. Subsection
7. Subject
8. Creator
9. Company
10. NewsAgency
11. Entity
12. Date
13. Page
14. PublicationType

Beispiel 1

Eine Suche nach allen Artikeln aus der FAZ oder taz, die "keine Alternative" im Untertitel enthalten

1 Unter Stoppwörtern versteht man die Wortarten, welche in einer Sprache sehr häufig für den Satzaufbau genutzt werden. Sie haben keine semantische Bedeutung für die Textinhalte.

kann folgendermaßen gestellt werden:

- *Subtitle:"keine Alternative" AND (Company:taz OR Company:faz)* --> liefert 2822 Treffer, von denen die viele das Wort "Alternative" oder "Alternativen", aber nicht "keine Alternative" enthalten. Das liegt daran, dass wir hier auf dem Feld Subtitle gesucht haben, bei dem Stoppworte (wie "keine") nicht berücksichtigt werden
- *Subtitle_raw:"keine Alternative" AND (Company:taz OR Company:faz)* --> liefert 166 Treffer, von denen alle die Phrase "keine Alternative" enthalten.
- *Subtitle_raw:"keine Alternativen" AND (Company:taz OR Company:faz)* --> liefert 16 Treffer, von denen alle die Phrase "keine Alternativen" enthalten.
- *(Subtitle_raw:"keine Alternative" OR* Subtitle_raw:"keine Alternativen") AND (Company:taz OR Company:faz) --> liefert 166 + 16 = 182 Treffer

Beispiel 2

Suche nach allen Artikeln von Karen Horn in unserem Korpus:

- *Creator: (Karen Horn)* --> 21 Treffer, nämlich alljene, bei denen genau der Name als Autor-Metadatum in unserer DB ist. Frau Horn schreibt aber auch unter Kürzel "orn"
- *Creator: (Karen Horn) OR Paragraph_raw:orn* --> 169 Treffer, aber auch ein paar Unerwünschte durch schlechtes OCR. Alle die nicht von der FAZ sind könnten wir noch ausschließen
- *(Creator: (Karen Horn) OR Paragraph_raw:orn) AND Company:FAZ* --> 73 Treffer

Beispiel 3

Suche nach Date Ranges:

- Paragraph: Cobain AND Date:[1994-04-05T00:00:00Z TO 1994-04-12T00:00:00Z] -> 8 Treffer aus der Woche in der Kurt Cobain verstarb

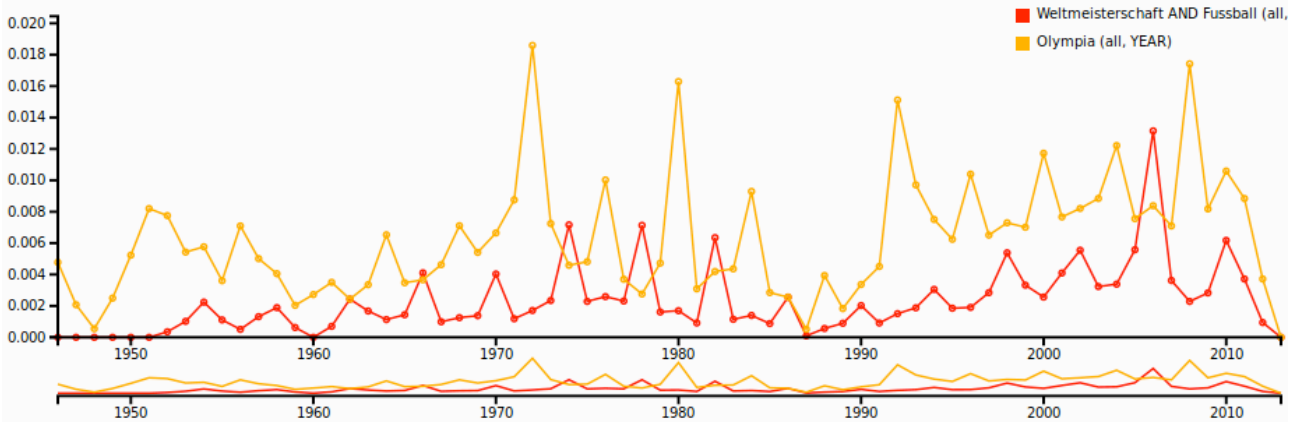
2.1.2 Collections

Sobald sie ihre Suchanfrage passend für ihr Forschungsvorhaben formuliert haben, können Sie aus der Treffermenge eine „Collection“ anlegen, auf der dann weitere Analysen ausgeführt werden können. Collections haben einen selbst zu vergebenden Namen und sind mindestens einer NutzerIn zugeordnet.

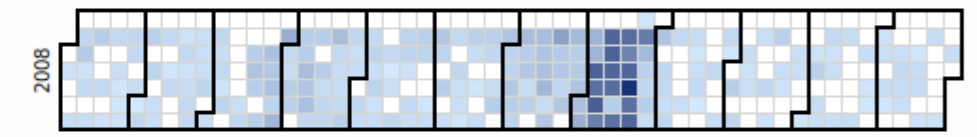
Tipp: Sobald Sie eine Suchanfrage zusammengestellt haben, die Sie als Grundlage für eine Collection nehmen möchten, sollten Sie sich die Suchanfrage separat in ein Textdokument abspeichern.

2.1.3 Time Series

Treffer zu Suchanfragen können direkt als Zeitreihe visualisiert werden. Dabei lassen sich auch die Ergebnisse mehrerer Anfrage in einer Grafik kombinieren.



Zusätzlich können Treffermengen auch als Heatmaps zu einzelnen Tagen des Suchzeitraums visualisiert werden.



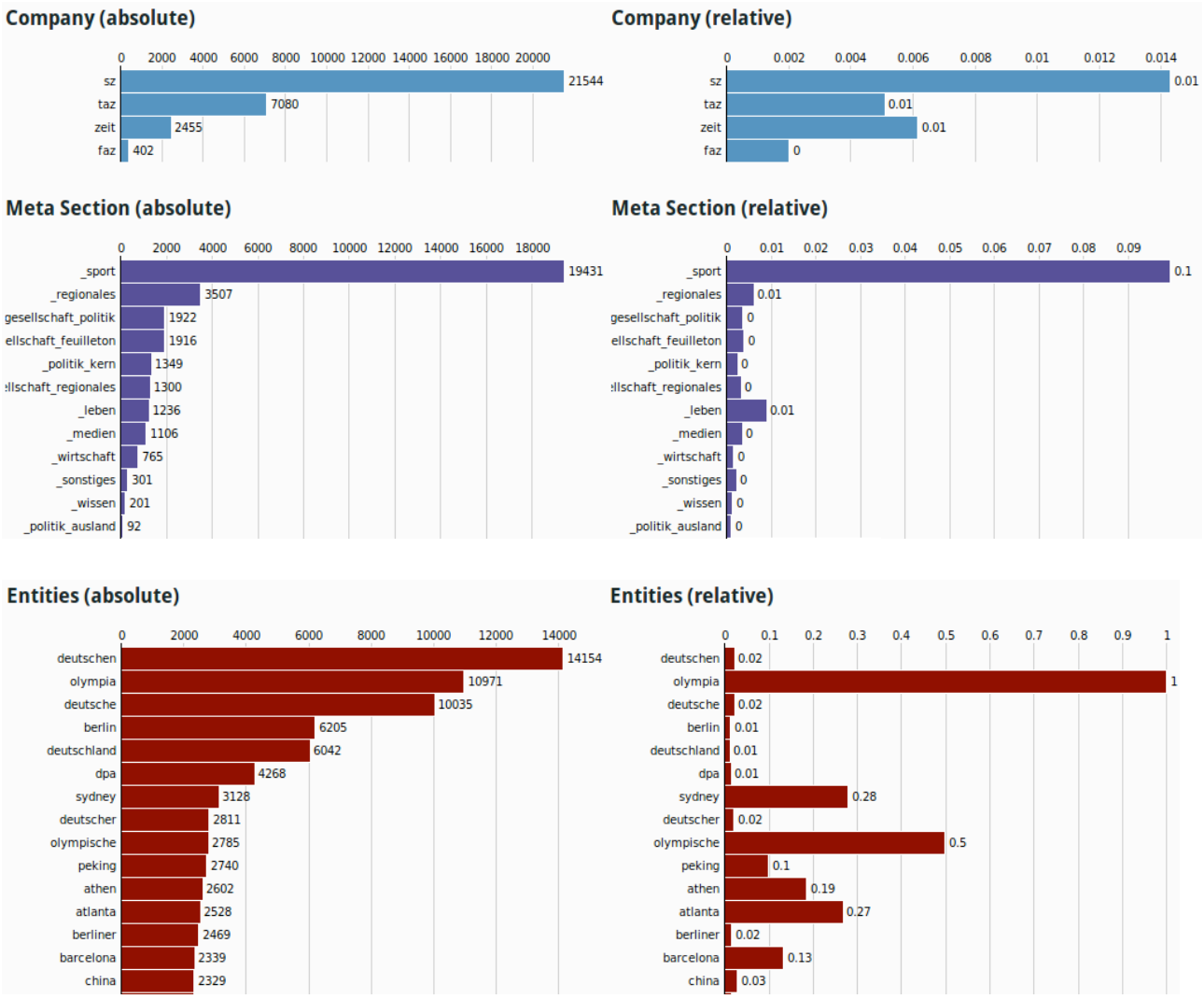
2.1.4 Volltextansicht

Zeitungsartikel können im Volltext gelesen werden. Schlussfolgerungen und Interpretationen auf Basis quantifizierender Auswertungsschritte können und sollten auf diese Weise immer mit einem qualitativen Blick in die Daten validiert werden.

2.1.5 Facets

Bestimmte Metadaten zu Treffern von Suchanfragen können aggregiert dargestellt werden. So ist es möglich zu erfassen, ob eine Suchanfrage in einer bestimmten Publikation, einem bestimmten Ressort oder in Zusammenhang mit einem bestimmten Eigennamen absolut oder relativ häufig gemeinsam miteinander auftritt.

Beispiel: Suchanfrage „Olympia“



2.2 Analysen auf Basis von „Collections“

Die folgenden Analysen können auf nach Suchanfragen gefilterten und anschließend gespeicherten Collections durchgeführt werden.

2.2.1 Frequenzanalyse

Der LCM erlaubt die Berechnung von Begriffsfrequenzen über die Zeit. Ergebnisse lassen sich in absoluten und relativen Häufigkeiten (relativ in Bezug zur Gesamtmenge an Artikeln innerhalb der Collection).

Beispiel: Collection „Mindestlohn“ in den 1990er Jahren, Relative Frequenz der Begriffe „Arbeitszeit“ und „Arbeitsbedingungen“

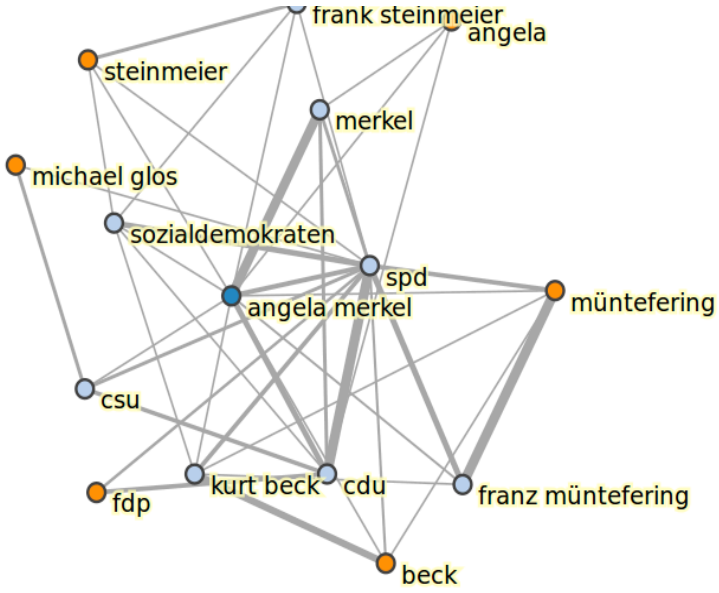
2.2.2 Kookkurrenzanalyse

Statistisch signifikant überzufällig häufig miteinander auftretende Begriffe innerhalb von Sätzen, Absätzen oder Dokumenten lassen sich als Graphen darstellen.

- gerichteter Graph: linke / rechte Nachbarn eines Terms innerhalb des Texts
- ungerichteter Graph: gemeinsames Vorkommen innerhalb einer Kontexteinheit (Satz, Absatz, Dokument) zweier Begriffe
- Matrix: gemeinsames Vorkommen innerhalb einer Kontexteinheit

Für die Berechnung der Kookkurrenzen ist eine Filterung nach Eigennamen möglich. So ist es beispielsweise möglich gemeinsam innerhalb eines Dokuments auftretende Eigennamen (Named Entities) zu messen.

Beispiel: Collection „Mindestlohn“, gemeinsames Auftreten von Named Entities zu „Angela Merkel“ in Dokumenten



2.2.3 Topic Modelle

Topic Modelle bilden globale Begriffszusammenhänge auf Ebene einer Dokumentkollektion mit Hilfe latenter Variablen, den sogenannten Topics, ab (Blei 2012). Die mit diesem Verfahren automatisch berechneten Begriffszusammenhänge, können als latente Sinnkomplexe oder Thematiken interpretiert werden, deren Verteilung über den Korpus insgesamt untersucht werden kann. Diese Verteilungen wiederum lassen sich zur Themenidentifikation anwenden, mit deren Hilfe auf Artikelebenen bestimmte Themenkategorien in der Untermenge von Artikeln aus dem Politik-Ressort des Zeitungskorpus separiert werden können. Der LCM erlaubt die Berechnung von Topic Models mit Hilfe einer Online-LDA Implementierung, mit der schnell sehr große Kollektionen analysiert werden können. Als zweites steht ein Hierarchical Pitman-Yor Process zur Verfügung, der ggf. exaktere Topics aber mit wesentlich längerer Berechnungszeit liefert.

Berechnete Topics aus einem Topic Modell auf Basis einer Collection können:

- zur Identifikation von Themen innerhalb einer Kollektion herangezogen werden (Topic-Wahrscheinlichkeiten in der Kollektion und der die Topics konstitiuerenden Begriffslisten)
- absolut oder relativ über die Zeit visualisiert werden
- zur Verfeinerung der Collection genutzt werden (z.B. Filterung aller Dokumente die mit dem Thema Musik in Zusammenhang stehen aus einer Collection die anhand des Suchbegriffes „Funk“ erzeugt wurde)

2.2.4 Term Extraktion

Topic Modellen extrahieren globale Begriffszusammenhänge über latente Variablen. Die Begriffe, die Topics repräsentieren können auch genutzt werden, um Listen wichtiger Begriffe einer Collection generell zu beschreiben. Der LCM bietet die Möglichkeit solche Listen von Schlüsselbegriffen auf Basis von Topic-Modellen zu berechnen und sich Beispiele für das Auftreten solche Begriffe in der Collection anzeigen zu lassen.

2.2.5 KWIC

Begriffe aus der Liste der extrahierten Terme lassen sich als „Keyword in Context“ anzeigen. Dabei werden aus der aktuellen Collection Beispiel-Abschnitte extrahiert, die den ausgewählten Begriff enthalten und so eine qualitative Überprüfung der Verwendungsweise eines Begriffes ermöglichen.

2.2.6 Manuelle Annotation

Methodologien der qualitativen Datenanalyse arbeiten in der Regel mit Kategorien, die entweder induktiv aus dem empirischen Material durch gründliches Lesen und interpretieren gewonnen werden oder deduktiv auf Basis existierender Theorien angeleitet und operationalisiert werden (z.B. mit Hilfe von Begriffslisten, sogenannten Diktionären). Der LCM unterstützt solche Arten der Analyse mit folgenden Funktionen:

- Definition hierarchischer Kategoriesysteme
- Kodierung von Dokumenten oder Textstellen mit Kategorien des Kategoriesystems (vgl. zu Code-and-Retrieve-Software a wie etwa MAXQDA oder Atlas.ti)
- Messung von Intercoder-Übereinstimmungen bei der Vergabe von Codes durch

unterschiedliche Benutzer

Beispiel Collection „Mindestlohn“: Kodierte Sätze, die eine Zustimmung bzw. Ablehnung zur Forderung nach einem Mindestlohn zum Ausdruck bringen

	Positiv	09.04.2014	Olaf Scholz hingegen warb in einem Brief an die Koalitionsfraktionen für den Mindestlohn	7799-788
	Positiv	09.04.2014	Nur so lasse sich verhindern, dass »der Staat als dauerhafter Lohnzahler in Anspruch genommen wird«, weil die Menschen zusätzlich zum Lohn Unterstützung brauchten	8038-820
	Negativ	09.04.2014	Die Rechtsunsicherheit empfinden viele bereits als schlimmer als den Mindestlohn selbst	9190-927
	Positiv	09.04.2014	Arbeitgeber wollen den Mindestlohn	33-67
	Positiv	09.04.2014	Akteure für Mindestlöhne in Deutschland starkgemacht haben	191-257
	Positiv	09.04.2014	dass Mindestlöhne die Marktwirtschaft »in ihren Grundfesten« beschädigten und zu Jobverlusten »erschütternden Ausmaßes« führen müssten	1393-152
	Positiv	09.04.2014	Mindestlöhne seien Ausdruck der Sozialpartnerschaft	1657-170
	Positiv	09.04.2014	Man benötige den Gesetzgeber, fügt Hundt hinzu	1732-177
	Positiv	09.04.2014	Die fürchten sich überdies, weil der deutsche Arbeitsmarkt im Mai 2011 für osteuropäische Arbeitnehmer geöffnet wird	2149-226
	Positiv	09.04.2014	Es sei denn, es gibt einen Mindestlohn	2443-248

2.2.7 Automatische Klassifikation

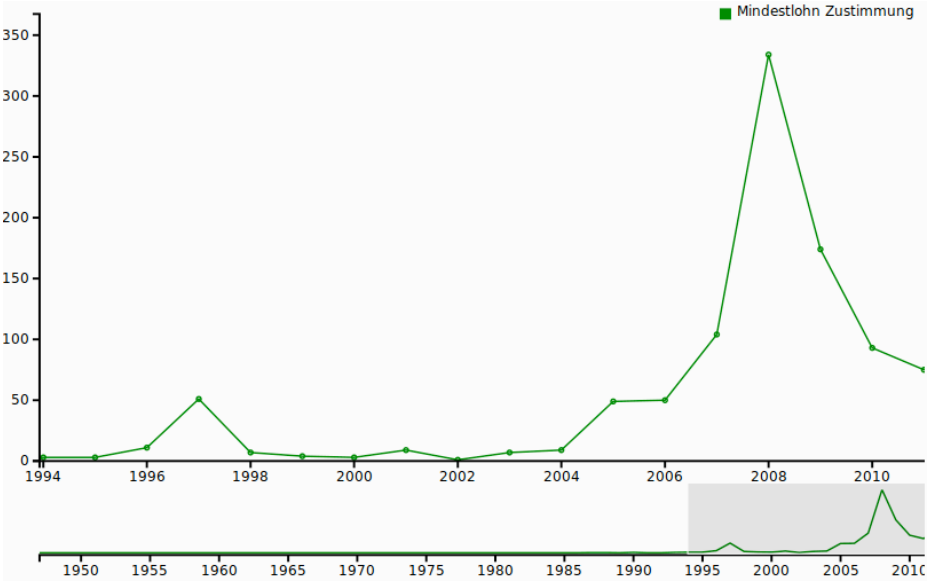
Kategorien, für die sich eine hohe Intercoder-Reliabilität erzielen lässt, die also hinreichend eindeutig definiert sind, so dass verschiedene Kodierer (größtenteils) die gleichen Kategorien für Dokumente oder Textstellen vergeben, eignen sich auch für automatische Klassifikationsverfahren. Überwachte maschinelle Lernverfahren zur Klassifikation lernen anhand gegebener Trainingsbeispiele die Einteilung in Zugehörigkeiten einer Datenentität zu einer definierten Klasse. Übersetzt für die Arbeit mit Texten bietet der LCM die Möglichkeit, Textstellen die einer Kategorie X zugeordnet wurden, als Trainingsbeispiele für die Klasse „Zugehörigkeit zu X“ anzunehmen. Textstellen in annotierten Dokumenten, die nicht mit X kodiert sind, werden zu Trainingsbeispielen mit der Annahme „nicht Zugehörig zu X“. Aus genügend vielen Trainingsbeispielen für hinreichend diskriminative Kategorien lernt der LCM Textmerkmale (Wortvorkommen und deren Kombinationen), die auf das Vorhandensein einer bestimmten Klasse schließen lassen. Somit ist es möglich, zu bereits kodierten Textstellen neue Textstellen in Dokumenten einer Kollektion zu identifizieren, die höchstwahrscheinlich zur Kategorie passen. Auf diese Weise wird es möglich, sehr große Dokumentmengen zu kodieren. Zur Bestimmung der Qualität der automatischen Klassifikation kommen Evaluierungsmaße auf den Trainingsdaten zum Einsatz (Niekler/Dumm 2014).

Zu iterativen Verbesserung der automatischen Klassifikation können die automatisch ermittelten Zuordnungen bestätigt bzw. korrigiert werden (*„Active Learning“*). So wird das automatische Klassifikationsergebnis schrittweise verbessert. Auf diesem Wege können wenige Hand-kodierte Textabschnitte (n < 100) mit relativ wenig Aufwand auf größere Mengen von Beispieltextabschnitten erweitert werden. Sobald die Evaluationsmaße für die automatische Klassifikation befriedigende Ergebnisse liefern, kann mit der Anwendung des Klassifikators auf eine „Collection“ eine zuverlässige Aussage über die Verteilung von Kategorien, z.B. im Zeitverlauf getroffen werden.

Beispiel:

Häufigkeit der Kategorie „Mindestlohn Zustimmung“ aus automatisch klassifizierten Sätzen wie

„DGB-Chef Michael Sommer hat einen gesetzlichen Mindestlohn von 7,50 Euro gefordert.“ oder „Die Vorwürfe zum Post-Mindestlohn seien völlig ungerechtfertigt“



3 Workflows und Hinweise zur Umsetzung

Die oben vorgestellten Text Mining Verfahren, die im LCM zur Verfügung gestellt werden, können als Einzelverfahren zur Unterstützung von Inhaltsanalysen genutzt werden. Durch die Kombination von Verfahren lassen sich für die Analyse aber erheblich mehr Erkenntnisse aus den produzierten Ergebnissen ableiten. Die durch Text Mining Verfahren produzierten Ergebnisse sind mitunter schwer als Einzelergebnisse für sich stehend zu interpretieren. Erkenntnisse können in diesem Fällen über den Vergleich von Einzelergebnissen auf unterschiedlichen Grundgesamtheiten von Daten gewonnen werden.

3.1 Aufteilung von Datengrundlagen

Collections aus Zeitungsartikeln als Repräsentanten einer Grundgesamtheit in Bezug auf eine bestimmte Fragestellung können am einfachsten nach Meta-Daten aufgeteilt werden. So lässt sich eine Collection, die anhand bestimmter Suchbegriffe identifiziert wurde, beispielsweise in verschiedene *Zeitabschnitte* oder nach *Zeitung* oder *Zeitungsressort* oder Kombinationen aus all diesen Metadaten aufteilen. Der Vergleich von Ergebnissen aus Kookkurrenzanalysen zu bestimmten Schlüsselbegriffen kann so zum Beispiel zur Erfassung von Bedeutungswandel über die Zeit beitragen (vgl. Lemke 2014 zum Begriff „Soziale Marktwirtschaft“).

3.2 Kombination von Einzelverfahren

Die zusammenhängende Betrachtung von Ergebnissen aus den Einzelverfahren ermöglicht interessante Analyseprozesse. So lassen sich zum Beispiel aus der Beobachtung von Frequenzen über den Zeitverlauf Zeitabschnitte erkennen, die auf eine gleichartige Verwendungsweise von Begriffen hindeuten und Zeitpunkte, an denen Veränderungen stattfinden. Diese Zeitabschnitte

können zur Aufteilung einer Collection in Sub-Collections genutzt werden, auf denen dann weitere vergleichende Analysen wie Topic Modell- oder Kookkurrenzberechnungen angewendet werden können. Topic Modelle selbst eignen sich über die Beschreibung von thematischen Zusammenhängen hinaus zur Filterung von Collections. So können spezifische semantische Zusammenhänge selektiert werden, mit denen die Aussagekraft von Frequenzanalysen deutlich erhöht werden kann. Im Rahmen des ePol-Projekts haben wir beispielsweise das Vorkommen von sogenannter „Alternativlosigkeitsrhetorik“ in den Zeitungsartikeln gemessen. Dabei handelt es sich im Wesentlichen um ein Dictionary, dass Begriffe wie „alternativlos“, „keine Alternative“, „unvermeidbar“, „unumgänglich“ etc. enthält. Eine Messung der Begriffe auf dem Gesamtkorpus kann kaum aussagekräftig sein, da nichts über die Verwendungskontexte der Begriffe bekannt ist. Selektiert man aber Artikel aus dem Ressort „Politik“ und filtert diese über ein Topic Model zu einer Dokumentmenge, die Dokumente mit einem Themenanteil-Anteil „EU, Europa, Europäische Integration“ von mindestens 20% enthalten, so werden Zählungen von Begriffen oder Konzepten (Begriffslisten) durch ihre Kontextualisierung deutlich aussagekräftiger.

3.3 Collection-Größe und Berechnungszeiten

Der LCM erlaubt das Ausführen der o.g. Analyseschritte und das Einstellen von bestimmten Parametern über die Webanwendung. Einzelne Analyseschritte können dabei sehr lange Berechnungszeiten in Anspruch nehmen. Eine Kollektion mit ca. 80.000 Dokumenten benötigt zur Zeit für die Berechnung einer Frequenzextraktion ca. 11 Stunden, für ein Topic Modell ca. 12 Std. Dies sollten Sie für ihre Analyseplanung im Hinterkopf behalten. Für die Berechnungen wird außerdem sehr viel Arbeitsspeicher verbraucht, s.d. nur eine begrenzte Anzahl an Berechnungen parallel auf dem ePol-Server laufen kann. Das bedeutet, dass die von Ihnen beauftragten Analysen vom Server unter Umständen einige Zeit in einer Warteschleife hängen. Denken Sie also daran, wenn Sie Analysetasks während des Workshops starten, dass Sie den anderen TeilnehmerInnen nicht durch ein Nacheinanderauslösen sehr vieler aufwändiger Tasks die Nutzungsmöglichkeiten zu sehr einschränken.

Für den Workshop werden wir Kollektionen auf maximal 20.000 Dokumente zu beschränken. Schränken Sie bei größeren Suchergebnismengen ihre Suche nach Zeitung, Zeitraum etc. ein und berechnen Sie Ergebnisse auf kleineren Collections.

3.4 Parameter

Für Text Mining Anwendungen werden Textdaten in numerische Daten umgewandelt. In der Regel werden dafür Wortvorkommen in Dokumenten (bzw. Absätzen, Sätzen) gezählt. Die Zählungen über dem gesamten Vokabular einer Dokumentkollektion (also alle enthalten verschiedenen Wortformen, sog. *Types*) werden pro Dokument in einem Vektor gespeichert. Alle Dokument-Vektoren einer Kollektion bilden eine Term-Dokument-Matrix (TDM), auf der dann verschiedene textstatistische Auswertungen gemacht werden können. Solch eine TDM bildet die Bag-of-Words-Hypothese ab, bei der lediglich das Vorhandensein von Wortvorkommen Berücksichtigung findet, nicht jedoch die spezifische Reihenfolge. Dies ist eine krasse Vereinfachungsannahme zur Modellierung von Sprachbedeutung, die sich jedoch für viele Anwendungen als nützlich herausstellt.

Von besonderer Bedeutung für die Analysen sind die Vorverabreitungsschritte, wie diese TDM

erzeugt wird. Zum Beispiel sollten Wortvorkommen unter einem gewissen Schwellwert ignoriert werden. Vorkommen der Wortvorkommen „Haus“, „Häuser“, „Hauses“ sollten ggf. zu „Haus“ unifiziert werden. Für bestimmte Analysen sollen gar nur Nomen oder Eigennamen berücksichtigt werden. Wenig Bedeutung tragende Worte wie „der“, „die“, „das“, sogenannte Stoppworte, sollten für die meisten Anwendungen generell unberücksichtigt bleiben.

Die Vorverarbeitung im LCM kann durch Benutzer über Parametereinstellungen im Task Scheduler gesteuert werden. Folgende Parameter können standardmäßig eingestellt werden:

- „Replace token with multi-word-units and entities“: Vorberechnete Mehrworteinheiten wie „Soziale Marktwirtschaftlich“ oder „Ludwig Erhardt“ werden als ein Wort statt zwei Worte behandelt
- „Transform entities to canonical form“: Verschiedene Varianten von Personennamen wie „Barack Obama“ und „Barack Hussein Obama“ werden auf eine Form „Barack Obama“ unifiziert
- Baseform: ein Wort wird auf sein Lemma unifiziert (ging → gehen, geht → gehen)
- Stemming: ein Wort wird auf seinen Stamm unifiziert (ging → ging, gehen → geh, geht → geh)
- „Remove Stopwords“: lässt eine Liste mit ca. 1000 dt. Stoppworten für die Analyse unberücksichtigt
- „Transform to lowercase“: Unifiziert Großbuchstaben auf Kleinbuchstaben
- N-Gram:
    - 1 = Unigram, zählt in Dokumenten Einzel-Tokens für die Term-Dokument-Matrix
    - 2 = Bigram, zählt zusätzlich zu den Unigrammen Vorkommen zweier aufeinanderfolgender Tokens
    - 3 = Trigram, zählt zusätzlich zu den Unigrammen Vorkommen dreier aufeinanderfolgender Tokens
    - Die Benutzung von N-Grams > 1 kann für die Klassifikation von Textabschnitten nützlich sein, um die strikte „Bag-of-Words“ Hypothese aufzuweichen, indem kurze Kontexteinheiten durch aufeinanderfolgender Worte berücksichtigt werden
- Pruning: Für die Bedeutungskonstitution von Aussagen sind Worte die besonders häufig bzw. besonders selten vorkommen in der Regel weniger relevant als Worte, die zwischen diesen Extremen liegen. Zur Verringerung der zu bearbeitenden Datenmengen und zur Verbesserung von Ergebnissen sollten deshalb bestimmte besonders häufige/seltene Worte für die Erstellung der TDM nicht berücksichtigt werden:
    - relatives Pruning (min x/max y): Unberücksichtigt bleiben alle Worte, die in weniger als x Prozent, in mehr als y Prozent aller Dokumente einer Collection vorkommen
    - absolutes Pruning (min x/max y): Unberücksichtigt bleiben alle Worte, die seltener als x, häufiger als y mal in einer Collection vorkommen
- Append_POS: Part-of-Speech Tagging ist eine Text Mining Aufgabe, bei der zu jedem Wort

ein Wortartenlabel zugeordnet wird (z.B. Nomen, gebeugtes Verb, Adjektiv, …). Beim Einsatz von Stemming wird bspw. „Europa“ und „europäisch“ zur „europa“ unifiziert. Das Anhängen des POS-Labels kann helfen etwa für eine Frequenz-Zählung „europa_NE“ als Eigennamen und „europa_ADJ“ als Adjektiv zu unterscheiden

- „POS-Types“: ermöglicht eine generelle Filterung der TDM nach POS-Types. So können z.B. für bestimmte Analysen nur Nomen oder Verben Berücksichtigung finden
- Token Minimum Length: Lässt alle Worte unberücksichtigt, die kürzer als n Zeichen sind

Wir haben uns bemüht, für die meisten Anwendungen sinnvolle Default-Parameter-Werte einzustellen. Weitere Erläuterungen dazu erfolgen während des Workshops.

3.5 Task spezifische Parameter

- **Kookkurrenzanalyse**
  - Minimum Cooccurrence Frequency: Gespeichert werden nur Wort-Paare, die mindestens n mal gemeinsam miteinander in einer Kontexteinheit auftreten
  - Context Unit: Die Zählung gemeinsamer Wortvorkommen bezieht sich auf ungerichtete Kookkurrenzen innerhalb eines Satzes, Absatzes oder Dokuments
  - Neighborhood horizon: die Zählung bezieht sich auf gerichtete Kookkurrenzen (linke, rechte Nachbarn) im Abstand von n tokens
- **Topic Models**
  - „use paragraph as document“: Themenmischungen innerhalb eines Absatzes (anstelle des ganzen Dokuments) inferieren. Diese Einstellung ist nur bei sehr langen Dokumenten sinnvolle (etwa Buchkapitel)
  - Minimum length of document: Dokumente, die kürzer als n Zeichen sind, werden bei der Modellberechnnung nicht berücksichtigt
- **Klassifikation**
  - Context unit: Kontexteinheit, der ein Code aus ihrem Kategoriesystem per Klassifikation zugeordnet werden soll. Wenn Sie im Prozess der manuellen Annotation Textabschnitte über mehrere Sätze kodiert haben, ist Absatz hier wahrscheinlich die beste Wahl. Ist ihre Kodierung vorwiegend auf Einzelsätzen bzw. Satzfragmenten erfolgt, dann wählen Sie hier auch Sätze aus.
  - Project / Category: Auswahl des Projekts / der Kategorie des Kategoriensystems aus der manuellen Annotation zu der eine automatische Klassifikation durchgeführt werden soll.
  - Mode:
    - Evaluate: Führt eine 10-Fold-Crossvalidierung auf den Trainingsdaten (manuell vergebene Codes und ggf. bestätigte/korrigierte Ergebnisse aus der automatischen Klassifikation) aus und gibt die Ergebnisse im Log des Task Schedulers aus (Dumm/Niekler 2014)

- Evaluate Optimize: Führt eine 10-Fold-Crossvalidierung auf den Trainingsdaten mit verschiedenen Parameter-Werten (Positive Score, C-Wert) aus und gibt die beste Konfiguration für die aktuelle Trainingsdatensituation im Log des Task Scheduler aus; ACHTUNG: Das Ausprobieren der verschiedenen Wertekombinationen kann sehr viel Zeit in Anspruch nehmen
- Classify: Führt eine Klassifikation mit den angegebenen Parametern auf der ausgewählten Kategorie aus. Es werden neue Beispiele für den Active Learning Prozess generiert. Zudem werden Frequenzzählungen der vergebenen Kategorie für Zeitreihendarstellungen auf der gesamten Collection produziert.

- Save to Database: Speichert die Ergebnisse in der Datenbank. Wählen Sie diese Option nur, wenn die Evaluationsergebnisse anzeigen, dass Sie einige richtige Ergebnisse aus dem Prozess der automatischen Klassifikation erwarten können. Andernfalls probieren Sie / optimieren Sie verschiedene der zwei folgenden Parameter, annotieren Sie manuell mehr Trainingsbeispiele oder überarbeiten Sie ihr Kategoriesystem
- Positive score threshold: Der eingesetzte SVM-Klassifikator gibt für die Wahrscheinlichkeit dass das ausgewählte Kategorielabel an die Kontexteinheit vergeben wird einen Wert zwischen 0 und 1 aus. Ein Schwellwert von 0.5 bedeutet für die Vorhersage, dass die Zuordnung des Labels wahrscheinlicher ist, als dass es nicht vergeben werden sollte. Für geringe Trainingsmengen am Anfang eines Klassifikationsprozesses macht es jedoch ggf. Sinn, auch bei kleineren Wahrscheinlichkeiten ein Label zu vergeben (Erhöhung des *Recall*), um genug Beispiele für den „Active Learning“ Prozess zu generieren (Evaluation und ggf. Berichtigung der automatisch gefundenen Textsstellen). Für abschließende Klassifikationen zu finalen Auswertung einer Kategorie können Werte > 0.5 sinnvolle sein, um möglichst nur die bestpassendsten Textstellen zu berücksichtigen (Erhöhung der *Precision*)[2]
- SVM C value: Ein Optimierungsparameter des SVM-Klassifikators, der entscheidet, wie gut der Klassifikator auf die Trainingsdaten passend trainiert wird. Daumenmaß: wenig Trainingsdaten → kleiner C-Wert; viele Trainingsdaten → großer C-Wert

4. Vorbereitung auf den Workshop / Operationalisierung

Dieser Paper soll Ihnen einen Vorgeschmack auf die Möglichkeiten der Text Mining gestützten Inhaltsanalyse mit dem LCM geben. Für die optimale Nutzung der Zeit während des Workshops würde wir Sie bitten, sich schon einmal ein paar Gedanken zur Operationalisierung ihrer Forschungsfrage zu machen. In (Dumm/Niekler 2014) wird eine Strategie für die Operationalisierung von Mixed-Method-Ansätzen vorgeschlagen. Demnach sollte die Operationalisierung folgende Schritte beachten, um eine methodische Nachvollziehbarkeit und Qualitätssicherung zu implementieren.

2 Siehe (Dumm/Niekler 2014)

1. Die gesamte Forschungsfrage muss in geeignete Teilaufgaben zerlegt werden. Dies beinhaltet die Analyse des Forschungsprozesses und Identifikation von Teilergebnissen, die zur Beantwortung der Forschungsfrage beitragen.
2. In jeder der Teilaufgaben muss die Forschungsmethode identifiziert und festgelegt werden. Es muss definiert werden, ob eine Teilaufgabe quantitativer oder qualitativer Natur ist und ob die Erkenntnisse deduktiv oder induktiv erzeugt werden.
3. Durch die genaue Einteilung der Teilaufgaben ist es möglich, die aus den Methodenkatalogen der Disziplinen gültigen Mess-, Bewertungs-, Güte- oder Evaluierungsverfahren zu identifizieren und einzusetzen.

Ihre Vorbereitung auf den Workshop sollte die folgenden Schritte umfassen:

1. Dokumentselektion: Welche Suchbegriffe / Kombination von Suchbegriffen (Einschluss, Ausschluss, Nähe von Begriffen,) und Meta-Daten (Daten, Zeitschrift, Ressort, …) können geeignet sein, ihre Grundgesamtheit zu bestimmen?
2. Welche Verfahren des LCM wollen Sie einsetzen? Arbeiten Sie eher explorativ / induktiv? Dann sind Topic Modelle, Frequenz- und Kookkurrenzanalysen primär interessant für Sie. Haben Sie ein theoretisches Modell, dass Sie in ein Kategoriensystem übersetzen und deduktiv testen wollen? Dann sind manuelle Kodierung und Klassifikation interessante Anwendungen. Evtl. möchten Sie beides kombinieren?
3. Dokumentieren Sie ihre Vorgehensweise und ihre Eindrücke während der Arbeit mit dem LCM. Wir betrachten das ganze auch als NutzerInnenexperiment, das uns hilft, die Software zu verbessern. Vor allem aber möchten wir gerne die methodische und methodologische Reflexion über den Einsatz von Text Mining Verfahren in den Sozialwissenschaften voran bringen.

Weiterführende Literatur zur Vorbereitung

- Dumm, Sebastian / Niekler, Andreas (2014): Methoden und Gütekriterien. Computergestützte Diskurs- und Inhaltsanalyse zwischen Sozialwissenschaft und Automatischer Sprachverarbeitung, Discussion Paper 4, ePol-Schriftenreihe.
- Niekler, Andreas / Wiedemann, Gregor / Heyer, Gerhard (2014): Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis, http://hal.archives-ouvertes.fr/hal-01005878.
- Lemke, Matthias (2014): Das Verschwinden der Sozialen Marktwirtschaft. Analyse der Verwendungskonjunkturen eines politisch-kulturellen Kernbegriffs der Bundesrepublik Deutschland mit Text Mining Verfahren, im Review aber per Mail an die Workshop-TN.
- Grimmer, Justin; Stewart, Brandon (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In: Political Analysis, S. 1–31.
- Blei, David M. (2012): Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. Communications of the ACM, 55(4), 77-84.

Schriftenreihe des Verbundprojekts
„Postdemokratie und Neoliberalismus“

Auswahl der Diskussionspapiere

(Stand Juli 2014)

Nr. 1 Neoliberalismus und Postdemokratie: Bausteine einer kritischen Gesellschaftstheorie.
Gary S. Schaal & Claudia Ritzi

Nr. 2 Die Ökonomisierung des Politischen.
Entdifferenzierungen in kollektiven Entscheidungsprozessen
Matthias Lemke

Nr. 3 Argumentmarker. Definition, Generierung und Anwendung im Rahmen eines semi-automatischen Dokument-Retrieval-Verfahrens
Sebastian Dumm & Matthias Lemke

Nr. 4 Methoden und Gütekriterien. Computergestützte Diskurs- und Inhaltsanalysen zwischen Sozialwissenschaft und Automatischer Sprachverarbeitung
Sebastian Dumm & Andreas Niekler

Nr. 5 Analyse qualitativer Daten mit dem Leipzig Corpus Miner. Zur Vorbereitung des Workshops „Text Mining in der Politikwissenschaft
Gregor Wiedemann & Andreas Niekler

Die Arbeitspapiere können bestellt werden/The discussion papers can be ordered:
Helmut-Schmidt-Universität Hamburg
Forschungsprojekt Postdemokratie und Neoliberalismus
Professur für Politische Theorie und Ideengeschichte
z. H. Susanne Kirst
Holstenhofweg 85
22043 Hamburg