Technology
Arts Sciences
TH Köln

Web Scraping als Monitoring-Instrument für Massenmedien im Web

Bachelorarbeit

ausgearbeitet von

Philip Ehnert

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

vorgelegt an der

Technischen Hochschule Köln
Campus Südstadt
Fakultät für Informations- und
Kommunikationswissenschaften

im Studiengang
Angewandte Informationswissenschaft

Erster Prüfer/in:	Prof. Dr. Philipp Schaer Technische Hochschule Köln
Zweiter Prüfer/in:	Dr. Sebastian Stier GESIS – Leibniz-Institut für Sozialwissenschaften in Köln

Köln, im November 2018

Inhaltsverzeichnis

Abstract

Während traditionelle Medieninhaltsanalysen eine etablierte Methode in der empirischen Sozialforschung darstellen, so werden sie doch selten mit analytischen Verfahren zur Verarbeitung großer Dokumentenkollektionen kombiniert (Blei u. a., 2003), die Aufschluss über latente inhaltliche Schwerpunkte einzelner Nachrichtenportale sowie deren relative Themenverteilung liefern können.
Im Vordergrund der Forschungsarbeiten stand daher die technische Realisierung eines automatisierten Verfahrens, das als Instrument zur Beobachtung der massenmedialen Agenda im Web dient. Zu diesem Zweck musste eine eigene Infrastruktur entwickelt werden, welche die Medienbeobachtung verschiedener Kanäle (u.a. "BILD", "Spiegel", "Junge Freiheit") möglich machte. Im Fokus steht hierbei die Entwicklung von drei Kernelementen des Monitoring-Instruments, die für die Archivierung, die Aufbereitung und schließlich die inhaltliche Analyse der Artikel aus den RSS-Kanälen der einzelnen Nachrichtenportale zuständig sind.
Neben dem o.g. Verfahren zur Aufdeckung latenter Themenkomplexe mithilfe des sogenannten LDA Topic Modellings bieten die in strukturierter Form vorliegenden Artikel aus dem Dokumentenkorpus ein breites Spektrum an weiteren Anwendungsmöglichkeiten. So wird das mittels Frequenzanalysen ermittelte Aufkommen von Themen in der massenmedialen Agenda ferner mit Umfragedaten[1][2] kombiniert, die zukünftig Aufschluss über die öffentliche Meinungsbildung zu den in den Medien vorkommenden Themen liefern können.
Insbesondere die Forschungsergebnisse der LDA-Analyse zeigen, dass die relative Themenverteilung der entsprechenden Nachrichtenportale auf einen Blick dargestellt werden kann; somit leisten die erzielten Ergebnisse einen Beitrag, ein tieferes Verständnis von komplexen sozialwissenschaftlichen Zusammenhängen, wie in diesem Fall der Agenda eines massenmedialen Akteurs, zu erlangen und diese mithilfe entsprechender Visualisierungen greifbar zu machen.

[1] https://www.gesis.org/allbus/allbus/
[2] https://www.gesis.org/issp/home/

1 Einleitung

1.1 Forschungsdesign

Im Rahmen meiner Tätigkeit als studentische Hilfskraft bei GESIS – Leibniz-Institut für Sozialwissenschaften habe ich unter der Leitung von Dr. Sebastian Stier bei der Durchführung eines Forschungsprojekts mitgewirkt. Ziel des Projekts ist es, sowohl das Aufkommen als auch die Aufbereitungsweise bestimmter Themen in den traditionellen Medien über einen bisher unbestimmten Zeitraum zu analysieren; anschließend dienen die zuvor erhobenen Daten dazu, Zusammenhänge zwischen der Frequenz des Aufkommens von Themen in den massenmedialen Medien und den parallel erfassten Umfrageergebnissen des ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) und des ISSP (International Social Survey Programme) zu erkennen.
Technisch wird die Medienbeobachtung auf Basis eines Web Scrapings realisiert. Dabei wird zunächst ein repräsentatives Sample aus RSS-Kanälen der traditionellen Medienagenturen ausgewählt, das das gesamte politische Spektrum bestmöglich abbildet. Die Liste der einzelnen RSS-Feeds, die jeweils einen Artikel bzw. Beitrag darstellen, wird daraufhin mithilfe des Web Scrapers in zuvor definierten Zeitintervallen auf neu verfügbare Links überprüft. Wurde der mit dieser per Suchmuster extrahierten URL assoziierte Artikel nicht schon zuvor archiviert, wird dieser automatisiert als HTML-Datei heruntergeladen.
Zum Zweck der weiteren Analyse müssen die bis dato in HTML ausgezeichneten Informationen (bspw. Titel, Textinhalt, Zeitstempel, eingebettete Tweets) der einzelnen Artikel maschinenlesbar gemacht werden. Dabei werden die benötigten Daten mithilfe einer Kombination aus CSS-Selektoren und regulären Ausdrücken aus der HTML-Datei geparsed, in ein homogenes Data Frame überführt und können schließlich in einer MySQL-Datenbank hinterlegt werden. Die nun in strukturierter Form vorliegenden Artikel bilden somit den Korpus für die weitere inhaltliche Analyse.
In einem nächsten Schritt wird die Häufigkeit der in den Artikeln behandelten Themen ermittelt. Dies ermöglicht es, das Aufkommen von Themen in den traditionellen Medien mit den politischen Einstellungen der befragten Personen der sozialwissenschaftlichen Umfragen ALLBUS und ISSP zu vergleichen.
Obwohl es sich bei der allgemeinen Bevölkerungsumfrage ALLBUS um eine Querschnittstudie handelt, deren repräsentative Stichprobe lediglich alle zwei Jahre zufällig ausgewählt wird, erstreckt sich der Zeitraum der Datenerhebung über mehrere Wochen bis Monate. Das ist zum einen mit der schieren Größe der Stichprobe begründet und zum anderen darauf zurückzuführen, dass die Befragung im Rahmen eines ca. einstündigen Interviews durchgeführt wird. Dadurch, dass die Umfrageergebnisse des ALLBUS über eine längere Zeitspanne hinweg erfasst werden, können Frequenzanalysen auch auf dieser ständig aktuellen Datenbasis wiederholt durchgeführt werden. Auf diese Weise ist es darüber hinaus realisierbar, Wechselwirkungen zwischen aufkommen-

den Themen der traditionellen Medien und den Antworten der befragten Personen auch im zeitlichen Verlauf mittels weiterer visueller Aufbereitung der Daten in R zu veranschaulichen.
Neben statistischen Auswertungen wie der o.g. Frequenzanalyse, die das Aufkommen von Termen anhand einer Keywordliste erfasst, oder einer Bestimmung der n-häufigsten Terme (generell sowie gruppiert nach Medienagentur), sind zudem auch weitere komplexere Verfahren für tiefgreifendere Analysen vorgesehen. Die in strukturierter Form vorliegenden Informationen aus dem Dokumentenkorpus bieten dabei ein breites Spektrum an Anwendungsmöglichkeiten, wobei anzumerken ist, dass die inhaltliche Analyse nicht vollständig innerhalb der Bearbeitungsdauer dieser Bachelorarbeit realisierbar ist.
Unter anderem ermöglicht die Auswertung zu erfassen, welche Themen bzw. Terme die unterschiedlichen Medienagenturen Artikeln zuweisen, die denselben im Dokument eingebetteten Tweet zum Thema haben. Dies könnte Aufschluss darüber geben, inwiefern ein bestimmter Tweet von verschiedenen Akteuren gedeutet bzw. selektiv wahrgenommen wird.
Ferner kann mittels Einsatzes eines Algorithmus (Blei u. a., 2003) auf Ebene der Nachrichtenportale ermittelt werden, welchen relativen Anteil die vom Verfahren definierten Themenfelder haben. In diesem Fall kann festgestellt werden, mit welchen Termen die Kernthemen der Umfragen ALLBUS und ISSP von den jeweiligen Akteuren assoziiert werden. Auf diese Weise werden Einblicke in die Art der Aufbereitung von Themen der Medienagenturen gegeben.

1.2 Begründung des Forschungsdesigns

Angesichts der Tatsache, dass es sich bei dieser Forschungsarbeit um ein äußerst vieldimensionales Projekt handelt, ist es notwendig, die interdisziplinären Zusammenhänge näher zu betrachten. So ist es im Hinblick auf die Beweggründe des Projekts unablässig, sowohl auf die politik- als auch sozialwissenschaftlichen Hintergründe einzugehen. Besonders hervorzuheben sind hierbei die Theorien des “Agenda Settings”, das auf das Setzen konkreter Themenschwerpunkte der Öffentlichkeit abzielt, als auch des “Framings” bestimmter Themen in den Medien. Die Einarbeitung in dieses für das Projekt essentielle Grundlagenwissen soll verdeutlichen, warum gerade dieser methodische Ansatz für das Forschungsprojekt gewählt wurde und ferner aufzeigen, welche Motivation sich hinter der Zielsetzung des Projekts verbirgt.
In Vorbereitung auf die technische Umsetzung des Projekts muss in einem nächsten Schritt zunächst fundiert begründet werden, welche Methode zur Extraktion der relevanten Informationen aus den einzelnen Zeitungsartikeln in Frage kommt und zudem für den konkreten Anwendungsfall der Medienbeobachtung sinnvoll ist. So führte die intensive Auseinandersetzung mit dem aktuellsten Stand der Technik in diesem sich rasant entwickelnden Forschungsfeld zu einer breiten Auswahl an unterschiedlichen Lösungsansätzen. Diese galt es unter Berücksichtigung der spezifischen Anforderungen an die Qualität der zu erhebenden Forschungsdaten kritisch zu hinterfragen und dementsprechend die vielversprechendste Methode aus einem breiten Spektrum an unterschiedlichen Optionen auszuwählen.

Allein die Auseinandersetzung mit allen o.g. Aspekten war letztendlich zielführend und leistete damit einen wesentlichen Beitrag zum Erfolg des Forschungsprojekts bei.

1.2.1 Politikwissenschaftliche Begründung

Bereits seit Mitte des letzten Jahrhunderts sind sowohl Politik- als auch Sozialwissenschaftler daran interessiert, ein tieferes Verständnis für die komplexen Dynamiken der öffentlichen Wahrnehmung sowie der politischen Meinungsbildung zu erlangen. Joseph T. Klapper untersuchte schon im Jahr 1960 die Auswirkungen von (Massen-)Medienexposition auf die persönliche Meinungsbildung der Rezipienten und kam bei der Auswertung seiner Studie zu einem Ergebnis, das die moderne Meinungsforschung nachhaltig prägen sollte. Nach Analyse seiner Forschungsdaten kam er zu der Erkenntnis, dass die Meinungen der Rezipienten zu bestimmten in den Massenmedien kommunizierten Themen nur marginal beeinflusst werden. So kam es nur in den seltensten Fällen dazu, dass Rezipienten in Folge der Medienexposition ihre Meinung zu den entsprechenden Themen stark verändert bzw. sogar komplett gewechselt hatten. Er merkte jedoch auch an, dass bereits vorher existierende Meinungen durch die sogenannte persuasive Massenkommunikation verstärkt werden und diese besonders dann ihre Wirkung entfaltet, wenn der Rezipient vor Einflussnahme der Medien noch keine Einstellung zum kommunizierten Thema hatte (Klapper, 1960, S.278).
Cohen fasste die o.g. Erkenntnisse 1963 anschaulich in seinem Werk zusammen:

> *"The press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about."* - Bernhard C. Cohen (Cohen, 1963)

McCombs und Shaw entwickelten 1972 unter Einbezug von Cohens Theorie ein zentrales Konzept, das einen weiteren Meilenstein auf dem Weg zur modernen Medienwirkungsforschung darstellt - die sogenannte Thematisierungs- bzw. Agenda-Setting-Theorie. Ausgangspunkt ist hierbei Cohens grundlegende Schlussfolgerung, dass Massenmedien zwar kaum einen Einfluss darauf haben, welche Einstellung Rezipienten zu einem Thema haben, allerdings eine bedeutsame Einwirkung, über welche Themen sie sich überhaupt Gedanken machen. Um ihre Theorie zu belegen, werteten McCombs und Shaw im Zuge des US-Präsidentschaftswahlkampfs 1968 aus, welche Themenschwerpunkte von den Massenmedien inhaltlich gesetzt wurden, und welchen Einfluss dies auf die Agenda der Öffentlichkeit hat. Auch wenn die Forschungsmethode nicht ohne Kritik blieb, so spricht das Ergebnis - eine hohe positive Korrelation von mehr als 90 Prozent - für sich (McCombs u. Shaw, 1972, S. 176). Seitdem wurde das Thema in diversen Studien und Veröffentlichungen aufgegriffen, wobei bei den meisten lediglich marginale Anpassungen, wie etwa die Beschränkung auf ein bestimmtes Publikum oder ein Medium, vorgenommen wurden. Andere wiederum gingen der Frage der kausalen Richtung nach (Russell Neuman u. a., 2014, S. 194). Dazu beobachteten sie die Agenda der traditionellen Medien und die der Öffentlichkeit parallel über einen längeren Zeitraum. Dies ermöglichte es ihnen, anhand des zeitlich verzögerten Effekts zu bestimmen, welcher Akteur mit welchem Erfolg welches Thema auf die jeweilige Agenda des anderen gesetzt hatte. So führten bspw. Brosius und Kepplinger im Zeitraum des Jahres 1986 eine Längstschnittstudie durch, um die öffentliche

Wahrnehmung wöchentlich anhand von 16 vorher definierten Themen zu erfassen. Die Ergebnisse aus ihrer Umfrage verglichen sie daraufhin mit den parallel erfassten Sendeinhalten der größten deutschen Fernsehsender. Dabei stellten sie fest, dass die Abdeckung von Themen im Fernsehen vor allem bei vier stark vertretenden Themen (z.B. Verteidigungspolitik) zu einer Beeinflussung der öffentlichen Wahrnehmung führte. In entgegengesetzter Richtung wurden immerhin drei Themengebiete (z.B. Rente) ausgemacht, die zu einer vermehrten Ausstrahlung von entsprechenden Inhalten der größten deutschen Fernsehprogramme führten (Brosius u. Kepplinger, 1990).
Aufgrund der stetig wachsenden Verbreitung digitaler Medien in der modernen Informationsgesellschaft wird die Ermittlung von Ursache und Wirkung zunehmend komplexer. Über Social-Media-Plattformen und Blogs ist es heutzutage möglich, seine eigene Meinung zu einem öffentlichen Diskurs in Sekundenschnelle mitzuteilen oder aber eine neue Diskussion auf Knopfdruck aufkeimen zu lassen. Angesichts dieser Tatsache erschwert es auf der einen Seite die Dynamiken der öffentlichen Wahrnehmung im Hinblick auf deren kausale Richtung korrekt zu erfassen und in einem Modell zu veranschaulichen; auf der anderen Seite bietet die immense Anzahl digital vorliegender Daten gleichzeitig auch Chancen, bestehende Theorien der Medienwirkungsforschung anhand von automatisierten Analyseverfahren ("Big Data") auf die Probe zu stellen und diese gegebenenfalls um neue Aspekte zu erweitern (Russell Neuman u. a., 2014, S. 194-195).
Im Hinblick auf deren Einfluss auf die öffentliche Wahrnehmung ordnet Neuman Social-Media-Beiträgen und Blogs eine vergleichsweise untergeordnete, wenn auch stetig wachsende Rolle zu. Dies begründet er insbesondere dadurch, dass Nutzer der Social-Media-Plattformen demographisch (noch) nicht repräsentativ sind und dass Beiträge von Nutzern dieser Kanäle aufgrund der dort vorherrschenden expressiven Gesprächskultur beeinflusst sein könnten. Aufgrund des immensen Ausmaßes an täglich produzierten Beiträgen in den sozialen Medien, misst er ihnen trotz des hohen Anteils an belanglosen Inhalten einen essentiellen Beitrag zur politischen Diskussion in der Öffentlichkeit bei. Gerade im Hinblick auf die aktuelle #MeToo-Debatte wird deutlich, dass Multiplikatoren wie Journalisten, Politiker oder Personen des öffentlichen Lebens aktiv Social-Media und insbesondere Twitter nutzen, um ihrerseits Einfluss auf öffentliche Diskussionen zu nehmen. 2010 zeigte Chadwick mit seiner Studie zur "Bullygate"-Affäre auf, welche weitreichenden Effekte zielgerichtete Kampagnen solcher Multiplikatoren auf die massenmediale Agenda zur Folge haben (Chadwick, 2011). Damals wurde dem amtierenden Premierminister Gordon Brown kurz vor einem entscheidenden Wahlkampf vorgeworfen, er habe Mitarbeiter und Kollegen in seinem Büro psychisch und physisch misshandelt. Chadwick konnte anhand einer umfänglichen Medienbeobachtung festhalten, welche vielfältigen Wechselwirkungen sich daraufhin zwischen der massenmedialen und der Agenda der Öffentlichkeit ergeben haben.
Den traditionellen Medien, speziell den Onlineauftritten der ehemalig vorherrschenden Print- und Massenmedien, wird aber weiterhin signifikante Rolle beim Setzen von Themen in der öffentlichen Agenda zugeschrieben (Russell Neuman u. a., 2014, S. 196) und stützt seine These dabei auf die Auswertungen Hindmans zu diesem Thema, das sich unter anderem mit dem Mediennutzungsverhalten im digitalen Zeitalter auseinandersetzt (Hindman, 2008).

Um ein tieferes Verständnis für die Dynamiken der öffentlichen Wahrnehmung zu erlangen, wertete Neuman 2014 die Agenda sowohl der traditionellen als auch sozialen Medien täglich über einen längeren Zeitraum anhand von 29 vorher definierten Themen aus. Dabei nahm er an, dass die in den sozialen Medien und Blogs vertretenen Einstellungen größtenteils denen der öffentlichen Wahrnehmung entsprechen. Nach Evaluation seiner Ergebnisse kam Neuman zu der Erkenntnis, dass die Beeinflussung der Agenda nicht wie vorher angenommen entweder in einer Richtung von den traditionellen Medien zu den sozialen Medien verläuft und umgekehrt, sondern dass diese in einer ständigen Wechselwirkung zueinander stehen (Russell Neuman u. a., 2014, S. 210).
Im Laufe der Zeit wurde die Thematisierungstheorie ständig weiterentwickelt und schließlich durch das sogenannte Framing, das Neuman als zweite Ebene des Agenda-Settings (Second-Level-Agenda-Setting) betrachtet, ergänzt (Russell Neuman u. a., 2014, S. 196-197). In seiner Definition von 1993 beschreibt Entman das Framing als eine Aufbereitungsweise von Themen, um eine bestimmte selektive Wahrnehmung zu diesem Sachverhalt seitens der Rezipienten zu begünstigen. Rezipienten sollen dabei eine ursprünglich komplexe Information aus einer bestimmten kausalen oder moralischen Perspektive wahrnehmen (Entman, 1993).
In der Folge wurden weitere Studien durchgeführt, die die nachhaltige Wirkung des Framings einzelner Themen auf die öffentliche Wahrnehmung zum Gegenstand ihrer Untersuchung hatten. So konnten Druckman und Nelson 2003 anhand einer Gegenüberstellung im zeitlichen Verlauf beobachten, dass der Effekt eines solchen Framings meist nur von kurzer Dauer ist und damit langfristig betrachtet kaum von der ursprünglichen eigenen Einstellung zu den entsprechenden Themen abwich (Druckman u. Nelson, 2003).

Relevanz für das Projekt

Die o.g. Erkenntnisse dienen uns im Hinblick auf das Projekt als Basis für das im weiteren Verlauf angewandte Forschungsdesign. Die vorangegangenen Forschungen machen deutlich, wie bedeutend das Sammeln von Daten über einen längeren Zeitraum für die anschließende Analyse etwaiger Wechselwirkungen ist. Zu diesem Zweck werden sowohl die traditionellen Medien als auch die Umfrageergebnisse über einen längeren Zeitraum bis zum Abschluss der Feldarbeiten des ALLBUS erfasst. Ferner ist bei der Wahl der Forschungsmethode zu beachten, dass die Agenda der traditionellen Medien bzw. der Öffentlichkeit auch tatsächlich anhand der erfassten Daten und vorher definierten Themen auszumachen ist. Zwar stellt die Wahl von in Onlineauftritten der traditionellen Medien publizierten Artikeln keine bedeutende Abweichung zu vorangegangenen Methoden dar, wohl aber der Vergleich eben dieser Ergebnisse mit den Umfrageergebnissen des ALLBUS. Das ist vor allem damit zu begründen, dass die Ergebnisse zum einen über einen längeren Zeitraum erhoben werden und zudem Einstellungen zu bestimmten Themen explizit abfragt werden. Die während der Umfrage erfassten Themen können somit mit den aktuellen Diskussionen in den traditionellen Medien verknüpft werden.
Diese Verfahren bzw. auch komplexere Verfahren wie das Latent Dirichlet Allocation Topic Modelling lassen zwar keine genaueren Rückschlüsse auf das Framing von

Themen durch die unterschiedlichen Medienagenturen zu, da anhand dieser Methode nicht erfasst werden kann, ob ein Thema bzw. Term entweder positiv oder negativ besetzt ist; das Verfahren kann aber dennoch Hinweise darauf geben, mit welchen Termen das jeweilige Nachrichtenportal bestimmte Themen assoziiert. Zusätzlich lassen sich solche Verfahren auch auf in Artikel eingebettete Tweets anwenden, um besser verstehen zu können, mit welchen Themen die diversen traditionellen Medien ein und denselben Tweet in Verbindung bringen. Zu diesem Zweck werden die IDs der entsprechenden Twitter-Beiträge aus den einzelnen Artikeln extrahiert. In einem darauf folgenden Schritt kann daraufhin ermittelt werden, welche Terme die einzelnen Nachrichtenportale mit den jeweiligen Tweets assoziieren.

1.2.2 Begründung der technischen Methode

Zu Beginn der Forschungsarbeiten musste zunächst erörtert werden, welche Methoden für die Erstellung des Dokumentenkorpus in Betracht gezogen werden können. Dabei war das vorrangige Ziel, folgende Informationen möglichst umfassend aus den einzelnen Nachrichtenbeiträgen zu extrahieren: Den eigentlichen Text, den einleitenden Text bzw. Teaser, die Überschrift, das Datum, welcher Kategorie der Artikel zuzuordnen ist, ob es sich um einen zahlungspflichtigen Beitrag (z.B. BILDplus, Welt+) handelt und ob bzw. welche Tweets in den Artikel eingebettet wurden.
Unter Berücksichtigung dieser Anforderungen stellte sich heraus, dass das Auslesen der Informationen aus den entsprechenden RSS-Kanälen nicht den o.g. Kriterien gerecht wird, da die geforderten Informationen in diesem Format nicht für alle Nachrichtenportale in vollem Umfang vorliegen. Ferner wurde es auf Nachfrage nicht gestattet, Artikel über eine API der jeweiligen Nachrichtenagenturen entgegenzunehmen. Dies hatte zur Folge, dass ein direkter Zugriff auf die meist kostenpflichtigen Datenbanken nicht möglich war. Zur Erfassung aller relevanten Metadaten aus den Artikeln war es daher notwendig, die Informationen den ursprünglichen Websites der jeweiligen Portale zu entnehmen.

State of the Art

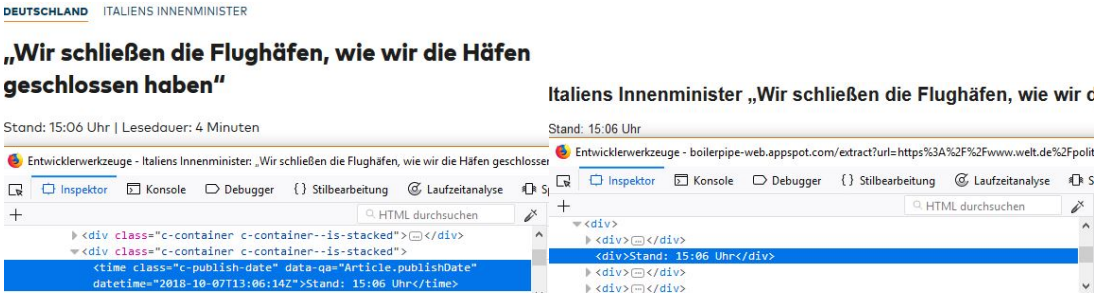
Aus diesem Grund wurde zunächst eine Recherche durchgeführt, die Aufschluss über die aktuellsten Methoden und Entwicklungen in diesem Forschungsfeld gab. Für das Extrahieren von Informationen aus dem Internet (web data extraction) gibt es dabei ein breites Spektrum an Lösungsansätzen, die von einfachen Frameworks zum Scrapen von Inhalten aus Websites, über semi-automatische Web Crawler bis hin zu Gesamtpaketen reichen, die praktisch ohne größere Konfiguration seitens des Nutzers einsetzbar sind.
Nennenswert ist insbesondere das 2017 erschienene Python-Tool “news-please”. Es wurde eigens dafür konzipiert, Nachrichtenartikel für spätere Anwendungszwecke in strukturierter Form abzuspeichern. Hierzu vereint es zwei grundlegende Komponenten.
Einerseits einen Web Crawler, der entweder vorher definierte RSS-Kanäle auf neu erschienene Artikel hin analysiert oder ebenfalls vorher festgelegte URLs der Nachrichtenagenturen direkt crawlt. In diesem Zusammenhang ist außerdem zu erwähnen,

dass der Crawler Websites auch rekursiv durchlaufen kann, demnach auch eigenständig internen Links folgt und dabei gegebenenfalls auch die in der Sitemap enthaltene Information zur Struktur der Website nutzt (Hamborg u. a., 2017).
Andererseits kombiniert news-please aktuelle Tools, um die im HTML-Format vorliegenden Informationen aus den Dokumenten zu extrahieren. Es vereint dabei die bereits zuvor umfangreichen Funktionen der Pakete "newspaper" (Ou-Yang, 2013) sowie "readability" (Baburov, 2010) und erweitert diese mit zusätzlichen Tools wie bspw. einem auf regulären Ausdrücken basierenden Bundle zum Extrahieren von Datumsangaben (Geva, 2016).
Insgesamt stellt news-please damit ein einfach zu bedienendes Komplettpaket dar, das die Stärken aller bisher verfügbaren Extraktoren in sich vereint und zudem das Crawlen der Artikel übernimmt. Bei genauerer Betrachtung der im Forschungsbericht genannten Gesamtperformance ist der Einsatz eines solchen Tools allerdings nicht mit den Anforderungen an die Qualität der für das Projekt benötigten Forschungsdaten vereinbar.
Während 82% des Titels und 76% des Teasers überwiegend vollständig erfasst wurden, so bricht die Performance beim Extrahieren des Datums (70%) und des Textes (62%) erheblich ein Hamborg u. a. (2017). Gerade im Hinblick auf die erst im zeitlichen Verlauf identifizierbaren Dynamiken des Agenda Settings sind solche Evaluationswerte nicht zweckdienlich, da aufgrund der fehlenden Datumsinformationen ein unvollständiges Gesamtbild entsteht. Anzumerken ist außerdem, dass die meisten dort extrahierten Inhalte aus englischsprachigen Quellen (15 von 20) stammen; die Ergebnisse lassen sich folglich nicht auf ein rein deutschsprachiges Sample von Onlineauftritten der Nachrichtenportale übertragen, da die entsprechenden Extraktoren newspaper und readability primär auf englischsprachige Quellen spezialisiert sind und daher auf dieser Dokumentenbasis performanter operieren. Des Weiteren lassen sich mit den vorgefertigten Einstellungen von news-please nicht alle für das Projekt geforderten Metadaten (wie bspw. eingebettete Tweets) erfassen. So werden auf dem Nachrichtenportal der ARD Tweets erst angezeigt, sobald eine Bestätigung der Datenschutzrichtlinie seitens des Nutzers erfolgt ist.
Neben news-please gibt es noch weitere für das Projekt interessante Lösungsansätze zur Generierung von verwertbaren Forschungsdaten aus Nachrichtenbeiträgen, die ebenfalls aus einer Kombination aus Web Crawler und Extraktor bestehen.
Das Tool "Web Scraper" (Wang u. a., 2009) basiert dabei auf einem semi-automatischen Web Crawler zum Extrahieren von Onlineinhalten. Dabei wird der relevante Content zunächst manuell unter Verwendung eines Google Chrome Add-ons vom Nutzer innerhalb des Browsers markiert, sodass der Crawler daraufhin jede Seite dieser Domain automatisch nach den zuvor markierten DOM-Objekten crawlen kann. Ferner ist es mithilfe des Add-ons möglich, eigene Sitemaps der Website zu definieren, die dem Crawler vorgeben, wie er auf der entsprechenden Seite navigieren soll. Anschließend kann der Inhalt über einen cloud-basierten Web Crawler extrahiert und als CSV-Datei ausgegeben werden. Im Prinzip stellt das Tool Web Scraper eine beinahe optimale Lösung dar, da die gewünschten Informationen wie bspw. Überschrift und Text über die zuvor markierten HTML-Tags extrahiert und separat in einer CSV-Datei abgespeichert werden können. Außerdem lässt sich eine auf jedes einzelne Nachrichtenportal individuelle Sitemap erstellen. Ein entscheidender Nachteil dieses Verfahrens ist, dass

sich manche Metadaten wie bspw. das Erstellungsdatum des Dokuments nicht in jedem Fall über die GUI des Browsers markieren lassen, da sie zwar im Quelltext der Seite vorhanden, nicht aber für den Nutzer sichtbar sind. Weiterhin haben erste Versuche mit dem Chrome Add-on im Rahmen des Projekts gezeigt, dass die Markierungen in einigen Fällen zu systemischen Fehlerfassungen führten. Dies ist vor allem darauf zurückzuführen, dass die Struktur eines Artikels desselben Nachrichtenportals so stark voneinander abweichen kann, dass die Markierungen zwar die Struktur eines bestimmten Artikeltyps korrekt erfassen, es aber bei abweichenden Strukturen zu Fehlinterpretationen seitens des Extraktors geführt hat.
Des Weiteren gibt es mehrere Lösungsansätze, die sich mit der Differenzierung von wesentlichen und unwesentlichen Inhalten befassen. Ein Beispiel für eine solche Herangehensweise stellt "Boilerpipe" (Kohlschütter u. a., 2010) dar. Ursprünglich basierend auf der Programmiersprache Java, bietet es heutzutage auch Wrapper zur Nutzung in Python und R (boilerpipeR). Es nutzt unter anderem Konzepte des Machine Learnings, um den eigentlichen Content einer Webseite von überschüssigen Inhalten zu trennen. Die Entscheidungen, ob es sich bei einem DOM-Element tatsächlich um einen relevanten Inhalt einer Webseite handelt, basieren dabei auf sogenannten SVMs (support vector machines) und Entscheidungsbäumen, die die Relevanz auf Grundlage von Faktoren wie Textdichte, linguistischen Eigenschaften des Textinhaltes sowie strukturellen Beschaffenheiten (Nähe zu anderen Elementen etc.) des DOM-Elements einordnen.
Einen vergleichbaren Ansatz bietet das Tool "Web2Text" (Vogels u. a., 2018). Unwesentliche Inhalte werden hier mithilfe eines sogenannten HMM (Hidden-Markov-Model) vom relevanten Inhalt getrennt. Das HMM ist ein probabilistisches Modell, das unter anderem beim maschinellen Lernen angewandt wird. Dabei werden bestimmte Eigenschaften von DOM-Elementen (bspw. Position im Quellcode/zu anderen Elementen, Anzahl Stoppworte, Art des HTML-Tags) mit den Eigenschaften anderer DOM-Objekte innerhalb des Dokuments verglichen und daraufhin die Wahrscheinlichkeit errechnet, ob es sich bei diesem Element um relevanten Inhalt handelt. Zur Verbesserung der Prognose wird das Verfahren mittels eines neuronalen Netzes trainiert.
Beide Ansätze, sowohl "Boilerpipe" als auch "Web2Text", würden die Archivierung der Artikel prinzipiell komfortabler gestalten, da unwesentliche Bestandteile der HTML-Dateien durch Anwendung dieser beiden Verfahren herausgefiltert werden würden. Ein Problem stellt aber auch in diesem Fall das ungewollte Filtern von essentiellen Informationen dar; so könnten Metadaten wie die Datumsangaben von den jeweiligen Algorithmen nicht als relevanter Bestandteil der Artikel erkannt werden. Die Funktionsweise des Tools Boilerpipe lässt sich auf der dazugehörigen Website explorieren[1].
Zur Veranschaulichung des Problems ist in Abbildung 1.1 der Quelltext eines kurz zuvor publizierten WELT-Artikels zu sehen. Neben dem sichtbaren Text "Stand: 15:06" des Elements "time" sind innerhalb des Attributs "datetime" relevante Informationen hinsichtlich des Publikationsdatums enthalten, die auch nach längerer Archivierung des Dokuments (wie bspw. zur Visualisierung des zeitlichen Verlaufs) weiterhin von Nutzen sind.
Lässt man nun den relevanten Inhalt des o.g. Artikels über das Tool Boilerpipe von angeblich irrelevanten Content trennen, so enthält das Resultat lediglich den sicht-

[1] https://boilerpipe-web.appspot.com/

Abbildung 1.1: Datumsinformation im Quelltext eines Artikels vor (links) und nach (rechts) Extraktion der Inhalte durch Boilerpipe



baren Text des Elements. Weiterhin sind die Informationen nunmehr ohne jegliche semantische Auszeichnung vorliegend; sie lassen sich daher nicht mehr ohne weiteres durch CSS bzw. XPath selektieren. Dies hätte zur Folge, dass die ursprünglich enthaltenen relevanten Informationen nicht mehr aus dem nun inhaltsreduzierten Dokument zu entnehmen sind. Ähnlich verhält es sich auch mit den im Artikel thematisierten Tweets; diese werden in einigen Fällen nicht korrekt erkannt, da sie oftmals in einer speziellen standardisierten Form in der HTML-Struktur eingebettet sind.
Ein weiteres, speziell für den Anwendungsfall der Web Data Extraction geschaffenes Tool, stellt die auf XPath basierende Extraktionssprache "OXPath" (Furche u. a., 2013) dar. Die leicht verständliche Syntax bietet dabei auch Personen ohne bereits vorhandene Programmierkenntnisse die Gelegenheit, mit nur wenigen Zeilen Code einen funktionsfähigen Web Crawler zu entwerfen. So war es in einem vorangegangenen Projekt möglich, selbst Personen ohne nennenswerte Vorkenntnisse in der zugrundeliegenden XML-Technologie binnen weniger Stunden einzuarbeiten, sodass diese anschließend eigenständig Metadaten aus digitalen Bibliotheken extrahieren konnten (Neumann u. a., 2017). Des Weiteren bietet OXPath - selbst beim Crawlen großer Kollektionen - eine vergleichsweise hohe Performance hinsichtlich der Belegung des Arbeitsspeichers, da jedes Dokument nacheinander verarbeitet wird. In einem weiteren Projekt konnte somit der bestehende GIRT4-Dokumentenkorpus mit Metadaten aus dem sozialwissenschaftlichen Fachportal Sowiport angereichert werden (Schaer u. Neumann, 2017).
Für das Projekt ist insbesondere die von OXPath simulierte Interaktion mit einzelnen Elementen von Webseiten (wie bspw. das Klicken von Buttons) interessant, da es somit auch für die Extraktion dynamischer Seiteninhalte geeignet ist. Eine solche Vorgehensweise würde unter anderem die Extraktion von mehrseitigen Artikeln erleichtern. Allerdings erfordert die Extraktion von dynamischen Inhalten ein vorheriges Rendering der gesamten Webseite; dies schlägt sich wiederum in einer eher moderaten Bearbeitungsdauer nieder. Ferner könnte die automatisierte Nutzung dynamischer Inhalte auf den Webseiten der Nachrichtenportale IP-Sperrungen zur Folge haben. OXPath bietet zwar auch hier die Gelegenheit, eine Nutzung durch den Menschen mit integrierten Befehlen (Wartezeit zwischen Aktionen etc.) vorzutäuschen; dies würde aber den Prozess des Web Crawlings weiter verlangsamen.
Schließlich bieten Frameworks wie scrapy für Python als auch analog rvest (Wickham,

2016) für die Programmiersprache R die Möglichkeit, mittels bereits vordefinierter Funktionen relevante Inhalte manuell aus den jeweiligen Artikeln zu extrahieren. Die in diesem Fall angewandte Technik des sogenannten Web Scrapings zielt darauf ab, Webseiten zunächst als Dokument abzuspeichern und die gesuchte Information in einem darauf folgenden Schritt aus diesem zu extrahieren. Der wesentliche Unterschied zum Web Crawling besteht also darin, dass das Dokument zunächst in vollem Umfang heruntergeladen wird. Einerseits hat dies den Vorteil, dass sämtliche Informationen auch für spätere Analysezwecke vorliegen; andererseits hat dies aber auch den Nachteil, dass viele weniger relevante Inhalte der Webseite im Zuge des Prozesses mit archiviert werden.
Zwar kostet das Programmieren eines individuell auf den Anwendungsfall zugeschnittenen Web Scrapers viele Ressourcen und ein gewisses Maß an Know-how, allerdings hat diese Form der Extraktion von relevanten Informationen auch das Potenzial, die im Vergleich zu den oben diskutierten Lösungsansätzen besten Evaluationsergebnisse zu erzielen. Angesichts der Tatsache, dass das vorrangige Ziel des Forschungsprojekts die Gewinnung möglichst vollständiger Forschungsdaten für die weiteren inhaltlichen Analysen ist, fiel die Entscheidung schließlich auf die Programmierung eines Web Scrapers.

1.3 Rechtliche und ethische Aspekte

Schließlich muss die Methodik der Forschungsarbeit auch juristisch betrachtet werden, um bereits im Vorfeld zu evaluieren, ob die Realisierung des Projekts mit geltendem Recht konform ist. Diesbezüglich wurde im Vorhinein eine Rechtsrecherche unter Einbezug von juristischen Kommentaren des Datenbankhosts Beck Online durchgeführt, die ausschlaggebende Informationen hinsichtlich der praktischen Umsetzbarkeit eines solchen wissenschaftlichen Projekts zu Tage förderte.

Durchführung der Rechtsrecherche

Durch die fortschreitende Entwicklung moderner Verfahren zur Erhebung- und Analyse großer Datenmengen (Big Data) aus dem Internet steht der Gesetzgeber vor einer besonderen Herausforderung. Auf der einen Seite verspricht der Einsatz solcher Technologien ungeahnte Möglichkeiten für Unternehmen und Forschungsinstitutionen, da mittels Data Mining sowie statistischer Verfahren bisher verborgene Muster bzw. Korrelationen in aggregierten Datenkollektionen aufgedeckt werden können. Auf der anderen Seite stehen die Interessen der privilegierten Institutionen und Rechteinhaber, die der Vervielfältigung, der kostenfreien Verwertung und der öffentlichen Zugänglichmachung ihrer urheberrechtlich geschützten Werke zum Zweck der Datenauswertung kritisch gegenüberstehen. Es gilt einen vertretbaren Kompromiss zu finden, der die Interessen beider Parteien berücksichtigt und zudem die bisher geltende Gesetzgebung an die Anforderungen der modernen Informationsgesellschaft anpasst.
Aus diesem Grund verkündete der Gesetzgeber das seit 01.03.2018 in Kraft getretene UrhWissG ("Urheberrechts-Wissensgesellschafts-Gesetz"). Die Norm setzt dabei weitestgehend die in der europäischen InfoSoc-Richtlinie (2001/29/EG) getroffenen Vereinbarungen zur Vereinheitlichung und Modernisierung des digitalen Urheberrechts

in national geltendes Recht um. Zu diesem Zweck wurde unter anderem ein neuer Unterabschnitt ("Gesetzlich erlaubte Nutzungen für Unterricht, Wissenschaft und Institutionen") in das deutsche UrhG ("Urheberrechtsgesetz") aufgenommen. Dabei ist anzumerken, dass das neue UrhWissG zunächst auf fünf Jahre befristet wurde, um die Auswirkungen des Gesetzes nach einer 4-jährigen Testphase evaluieren zu können.

Bedeutung für das Projekt

Da bereits vor Beginn der eigentlichen Projektdurchführung sichergestellt werden musste, dass die wissenschaftliche Auswertung von urheberrechtlich geschützten Werken sowie die Verwendung eines Web Scrapers zur automatisierten Archivierung rechtskonform ist, war eine intensive Auseinandersetzung mit der aktuellen Gesetzesgebung unvermeidbar. Diesbezüglich wurde eine Rechtsrecherche durchgeführt, die den Fokus insbesondere auf zwei der im neuen Unterabschnitt 4 enthaltenen Schrankenbestimmungen §60c ("Wissenschaftliche Forschung") und §60d ("Text und Data Mining") des UrhG legt.

Regelungstechnik

Bereits die Unterteilung der Schrankenbestimmungen in konkrete Anwendungsszenarien, etwa für Unterrichts- und Lehrmedien bzw. wissenschaftliche Forschung, lässt darauf schließen, dass sich die Europäische Kommission bei der Verabschiedung der InfoSoc-Richtlinie (2001/29/EG) nicht am Fair-Use-Beispiel der US-amerikanischen Gesetzgebung orientiert hat. Das Fair-Use-Prinzip zeichnet sich vor allem dadurch aus, dass der Wortlaut des Gesetzestexts absichtlich vage und offen gehalten wird und somit Raum für weitere Interpretationen bietet, um auf etwaige technologische Veränderungen adäquat reagieren zu können. Die europäische Richtlinie sieht hingegen keine allgemeingültige Klausel vor, sondern versucht vielmehr durch die Nennung expliziter Tatbestände größtmögliche Rechtssicherheit für die jeweiligen Anwendungsfelder zu gewährleisten (Schack, 2017, ZUM 2017, 805).

§60c Wissenschaftliche Forschung

Mit Inkrafttreten des UrhWissG wurde die Schrankenbestimmung §60c zugunsten nicht kommerzieller wissenschaftlicher Forschung überarbeitet. In erster Linie soll sie Wissenschaftlern dazu dienen, bis zu 15% eines urheberrechtlich geschützten Werks für ihre Forschungsarbeit vervielfältigen bzw. öffentlich zugänglich machen zu können, ohne vorher das Einverständnis des Rechteinhabers einzuholen (Grübler, 2018, UrhG § 60c Rn. 1). Angesichts der Festlegung auf bis zu 15% des Umfangs eines Werks wird abermals deutlich, dass der Gesetzgeber bestrebt war, die konträren Interessen der Urheber und Forschungsinstitutionen bei der Gesetzgebung zu berücksichtigen und ferner beabsichtigt, möglichst wenig Spielraum für Interpretationen des Gesetzestextes zuzulassen, um somit ein Maximum an Rechtssicherheit zu gewährleisten. Darüber hinaus wird der Personenkreis, der für die Verbreitung und die öffentliche Zugänglichmachung urheberrechtlich geschützter Werkteile in Frage kommt, explizit auf Personen begrenzt, die die Werke für ihre eigene wissenschaftliche Forschung benötigen (Grübler, 2018, UrhG § 60c Rn. 13-14).

Zwar räumt der Gesetzgeber mit der Schrankenbestimmung §60c UrhWissG grundlegende Privilegien für die wissenschaftliche Nutzung von Werken ein, für den konkreten Anwendungsbereich des Forschungsprojekts bietet sie allerdings keine ausreichende Rechtssicherheit. Dies ist vor allem der Tatsache geschuldet, dass der zur späteren inhaltlichen Analyse erforderliche Korpus eine vollständige - wenngleich auch in strukturierter Form - archivierte Kopie der Gesamtheit aller bis zu diesem Zeitpunkt erschlossenen Artikel nachbilden soll. Da §60c jedoch nur zu einer Vervielfältigung von bis zu 15% eines Werkes für Forschungszwecke berechtigt, deckt diese Norm nicht alle rechtlichen Aspekte des Projekts ab.

§60d Text und Data Mining

Im Sinne des öffentlichen Interesses und des immensen Forschungspotenzials, das moderne Text- und Data Mining Verfahren erwarten lassen, hat der Gesetzgeber mit §60d eine völlig neue Wissenschaftsschranke in das UrhG aufgenommen. Diese Norm berechtigt zur systematischen und automatisierten Vervielfältigung einer "Vielzahl von Werken" zur Erstellung des Korpus, sofern diese für nicht kommerzielle wissenschaftliche Forschung verwendet werden (Dreier u. Schulze, 2018, UrhG § 60d Rn. 1). Zudem räumt der Gesetzgeber das Recht ein, Werke ohne vorherige Zustimmung des Rechteinhabers zu vervielfältigen (Dreier u. Schulze, 2018, UrhG § 60d Rn. 12).
Im Hinblick auf das Projekt ist hierbei anzumerken, dass beim Erstellen bzw. Herunterladen der einzelnen Werke darauf geachtet werden muss, den Zugang Dritter zum Ursprungsmaterial nicht zu behindern (Dreier u. Schulze, 2018, UrhG § 60d Rn. 7). Weiterhin ist die Vervielfältigung von Werken nur dann für den Zweck der nichtkommerziellen Forschung schrankenprivilegiert, sofern es sich um nicht durch technische Maßnahmen geschützte Werke handelt und diese für Dritte öffentlich zugänglich sind (Hagemeier, 2018, UrhG § 60d Rn. 13). Ferner erlaubt die Norm die öffentliche Zugänglichmachung des Korpus zur gemeinsamen Forschungsarbeit. Für den Fall, dass die Einhaltung wissenschaftlicher Standards überprüft wird, wie etwa bei einem Peer Review-Verfahren, ist es außerdem zulässig den Korpus Dritten zugänglich zu machen (Dreier u. Schulze, 2018, UrhG § 60d Rn. 8).
Neben den zahlreichen Privilegien, die die Forschungsinstitutionen durch die Schrankenbestimmung §60d genießen, legt ihnen der Gesetzgeber mit §60d Abs. 3 gleichzeitig auch eine Löschungspflicht auf. Unter einer Löschung versteht der Gesetzgeber dabei die vollständige Vernichtung des Korpus sowie aller Vervielfältigungsstücke, die während der Forschungsarbeiten erstellt wurden, nachdem das Forschungsziel erreicht bzw. die Arbeiten daran offiziell für beendet erklärt worden sind (Hagemeier, 2018, UrhG § 60d Rn. 19). Die Tatsache, dass eine Vielzahl an Forschungsprojekten nicht offiziell für beendet erklärt werden, lässt daran zweifeln, inwiefern §60d Abs.3 in der praktischen Umsetzung der Norm greift (Hoeren, 2018, IWRZ 2018, 120). Um eine Umgehung von Abs.3 durch bspw. eine vage Definition der Forschungsziele zu vermeiden und damit zukünftigen Rechtsstreitigkeiten zwischen den Rechteinhabern und den beteiligten Forschungsinstitutionen zu verhindern, schlagen Ahlberg und Götting den Entwurf eines "Löschungskonzepts" vor. Das Konzept soll dabei bereits vor der Initiierung des Forschungsprojekts festlegen, zu welchem Zeitpunkt etwaige technische Maßnahmen zur Einhaltung der Löschungspflicht ergriffen werden müssen, und ent-

sprechend dokumentieren, ob der Bestimmung gemäß §60d Abs.3 Folge geleistet wurde (Hagemeier, 2018, UrhG § 60d Rn. 19). Ausnahmen von der o.g. Löschungspflicht bilden laut Gesetzgeber die in den Normen §§ 60e und 60f genannten privilegierten Institutionen. Diese sind ausdrücklich dazu berechtigt, alle während der Forschungsarbeiten angefertigten Kopien in einem Langzeitarchiv zu speichern, vorausgesetzt, die Archivierung dient einem nicht kommerziellen Zweck (Hagemeier, 2018, UrhG § 60d Rn. 21).
Archiven, die im öffentlichen Interesse tätig sind, wird mit §60f Abs.2 zudem das Recht zugesprochen, zum Zwecke der langfristigen Archivierung der Werke Vervielfältigungen selbst bzw. diese gegebenenfalls durch Dritte erstellen zu lassen. Diese Sonderstellung der Archive wird damit begründet, dass die Archivierung eines Werkes in elektronischer Form gleichzeitig stets mit einer Vervielfältigung des Werks verbunden ist, und Archiven als einzige Institution das Recht zugesprochen wird, Werke in elektronischer Form zu archivieren (Dreier u. Schulze, 2018, UrhG § 60f Rn. 7).

Fazit der Rechtsrecherche

Nach Abschluss der Rechtsrecherche ist festzuhalten, dass die Durchführung des Forschungsprojekts, insbesondere durch Inkrafttreten der Schrankenbestimmung §60d (Text und Data Mining), rechtskonform ist. Dabei wird die Erhebung der Werke mittels RSS-Scrapings und deren anschließende Vervielfältigung in einer MySQL-Datenbank zur Erstellung des erforderlichen Korpus durch §60d Abs 1.1 zulässig. Bei der Erhebung der Artikel ist jedoch darauf zu achten, dass keine Schutzmaßnahmen umgangen werden und Dritten der Zugang zum Ursprungsmaterial durch unverhältnismäßige Inanspruchnahme der Bandbreite des RSS-Servers nicht verwehrt wird.
§60d Abs. 1.2 berechtigt außerdem dazu, die in der MySQL-Datenbank abgespeicherten strukturierten Daten zum Zwecke der gemeinsamen Forschung anderen Personen (öffentlich) zugänglich zu machen. Zur Einhaltung der Löschungspflicht sollte nach §60d Abs. 3 und den Empfehlungen der Autoren des Beck'schen Onlinekommentars ein Konzept entworfen werden, das die Forschungsziele klar definiert und terminiert, um auch dieser Bestimmung gerecht zu werden.
Um die Einhaltung wissenschaftlicher Standards bei der Durchführung des Forschungsprojekts auch zukünftig nachvollziehen zu können, wäre es außerdem sinnvoll, vor Einhaltung der Löschungspflicht das Korpus langfristig intern im DAS (Datenarchiv für Sozialwissenschaften) zu archivieren. In Hinblick auf mögliche Folgeprojekte ist zudem darauf hinzuweisen, dass bei der Verwendung personenbezogener Daten den Anweisungen der "Leitlinie zur Entwicklung eines Löschkonzepts mit Ableitung von Löschfristen für personenbezogene Daten" (DIN 669398) Folge geleistet werden sollte (Hagemeier, 2018, UrhG § 60d Rn. 19).

2 Erstellung des Dokumentenkorpus

Nach intensiver Auseinandersetzung mit den politikwissenschaftlichen Hintergründen, der grundlegenden technischen Methodik und der Ergründung der aktuellen Rechtslage kann in einem nächsten Schritt mit der konkreten Umsetzung des Projekts begonnen werden.
Aus informationswissenschaftlicher Sicht ist dabei vor allem die technische Umsetzung von großem Interesse. Das Spektrum an Anwendungsmöglichkeiten reicht dabei von der Entwicklung eines funktionsbasierten Web Scrapers in der statistischen Programmiersprache R über den Gebrauch von CSS-Selektoren sowie regulärer Ausdrücke zur Extraktion von relevanten Informationen aus den archivierten HTML-Dokumenten bis hin zur Visualisierung der Evaluationsergebnisse des massenmedialen Monitorings in R. Zur Erstellung des für die weiteren inhaltlichen Analysen essentiellen Dokumentenkorpus wurde darüber hinaus ein relationales Datenbankmodell erstellt, in dem die extrahierten Daten in strukturierter Form hinterlegt werden konnten. Ferner waren grundlegende Kenntnisse der deskriptiven Statistik und des Information Retrieval notwendig, um die Analyse durchführen zu können.

2.1 Vorbereitende Maßnahmen

Bevor mit der Programmierung der Software begonnen werden konnte, musste zunächst reflektiert werden, wie sich die Methode des Web Scrapings auf den konkreten Fall der Medienbeobachtung übertragen lässt. Weiterhin musste eine eigene Infrastruktur entwickelt werden, die den Hard- und Softwareanforderungen des Projekts entspricht.

RSS-Kanäle als Basis des Monitorings

Das Fundament jedes weiteren Vorgehens bildet dabei die Auswahl der RSS-Kanäle. Von zentraler Bedeutung ist zum einen, dass das Sample an RSS-Kanälen das gesamte politische Spektrum umfasst, also neben der politischen Mitte auch Onlineauftritte von Medien mit linker bzw. rechter Tendenz in der Berichterstattung abdeckt. Zum anderen muss durch die Wahl der RSS-Kanäle auch gewährleistet sein, dass die zentralen Themengebiete (Politik, Wirtschaft, Sport) für die einzelnen Nachrichtenportale zu möglichst gleichen Teilen erfasst werden, sofern diese überhaupt über mehr als einen RSS-Kanal verfügen. Zu diesem Zweck wurden Ressourcen der Abteilung DBG (Dauerbeobachtung der Gesellschaft) in Mannheim in Anspruch genommen. Die dort ausgearbeitete Liste umfasst alle relevanten RSS-Kanäle von insgesamt 17 Online-Nachrichtenportalen. Für das Monitoring der medialen Agenda sind vor allem die traditionellen Massenmedien von besonderer Bedeutung; diese setzen sich zusammen aus den 12 Kanälen von ARD, BILD, FAZ, FOCUS, Freitag, Junge Freiheit, Spiegel Online, Stern, SZ, Tagesspiegel, Welt und der Zeit. Zur Erfassung aller im Zeitraum

der Medienbeobachtung publizierten Artikel der o.g. Portale ist es außerdem erforderlich, die Länge der einzelnen RSS-Listen und die Frequenz zu erfassen, in der die jeweiligen Portale ihre Beiträge veröffentlichen. Diese Informationen sind insbesondere deshalb relevant, um Aussagen darüber zu machen, in welchen zeitlichen Abständen der Web Scraper die einzelnen Kanäle künftig auf neu zu erfassende Beiträge prüfen muss. Veröffentlicht bspw. die BILD innerhalb von einer Stunde 20 Artikel auf ihrem RSS-Kanal, der auf 10 Listenplätze beschränkt ist, so würde ein Web Scraper, der diesen RSS-Kanal lediglich jede Stunde scannt, nur eine Teilmenge der 20 publizierten Artikel korrekt erfassen können. Entsprechende Beobachtungen zur Vermeidung dieses Problems ergaben, dass im kritischsten Fall ca. 30 neue Feeds in einem RSS-Kanal mit 20 Listenplätzen in einem Zeitraum von 30 Minuten veröffentlicht wurden. Um einen gewissen Sicherheitspuffer zu wahren, wurde daher beschlossen, dass der Scraper zukünftig alle 10 Minuten eingesetzt werden muss.

Einrichtung der Hard- und Softwareumgebung

In einem weiteren Schritt musste sichergestellt werden, dass eine für das Vorhaben des Projekts geeignete Hard- und Softwareinfrastruktur bereitgestellt wird. Erschwerend kam in diesem Fall hinzu, dass der lokale Archivierungsserver mittelfristig aufgrund von länger andauernden Wartungsarbeiten nicht zur Verfügung stand und daher alternative Lösungswege gefordert waren.
Die Wahl fiel diesbezüglich auf das Anmieten eines AWS (Amazon Web Services) Servers. Gerade im Hinblick auf die erforderliche Skalierbarkeit bot dieser im Vergleich zu Konkurrenzprodukten die besseren Optionen. So war es möglich, die Festplattenkapazität dynamisch an die jeweiligen Anforderungen anzupassen, was gerade im Zusammenhang mit der für einen bisher unbestimmten Zeitraum geplanten Archivierung der Artikel ein nicht zu unterschätzender Vorteil ist. Des Weiteren bietet Amazon ein hohes Maß an Verfügbarkeit (99.9%[1]), was insbesondere für die Erfassung aller im Zeitraum publizierten Artikel essentiell ist. Betrachtet man die für das Projekt erforderlichen Hardwareressourcen, so ist die Performance des Remote Servers hingegen nur zweitrangig zu bewerten. In Anbetracht dieser Anforderungen fiel der Entschluss auf einen 2-Kern Server mit lediglich 4 GB Arbeitsspeicher, einer skalierbaren Festplatte (90-275GB) und dem Betriebssystem Ubuntu 16.04. Die Konfiguration erfolgte über die Remote Software WinSCP, die einerseits eine grafische Oberfläche zur Verwaltung des Dateisystems bot und andererseits einen PuTTY-Client zum direkten Zugriff auf den Server über ein Shell-Terminal ermöglichte.
Um abschließend die genauen Anforderungen an die Softwareumgebung zu planen, musste zunächst ergründet werden, welche Programmierumgebung für die künftige Entwicklung des Web Scrapers in Betracht gezogen werden kann. Bei einem Vergleich der beiden Programmiersprachen “R” und “Python” stellte sich heraus, dass es für den speziellen Anwendungsfall des Web Scrapings und der Datenanalyse einige Parallelen gibt. So basiert das auf Web Scraping ausgelegte R-Paket “rvest” größtenteils auf einer Python-Bibliothek namens “BeautifulSoup” (Richardson, 2015); entsprechend marginal sind die Unterschiede im Hinblick auf Syntax und Funktionsumfang. Wiederum ist das Python-Paket zur Datenanalyse “pandas” (McKinney, 2015) inspiriert von nativen

[1] https://aws.amazon.com/compute/sla/

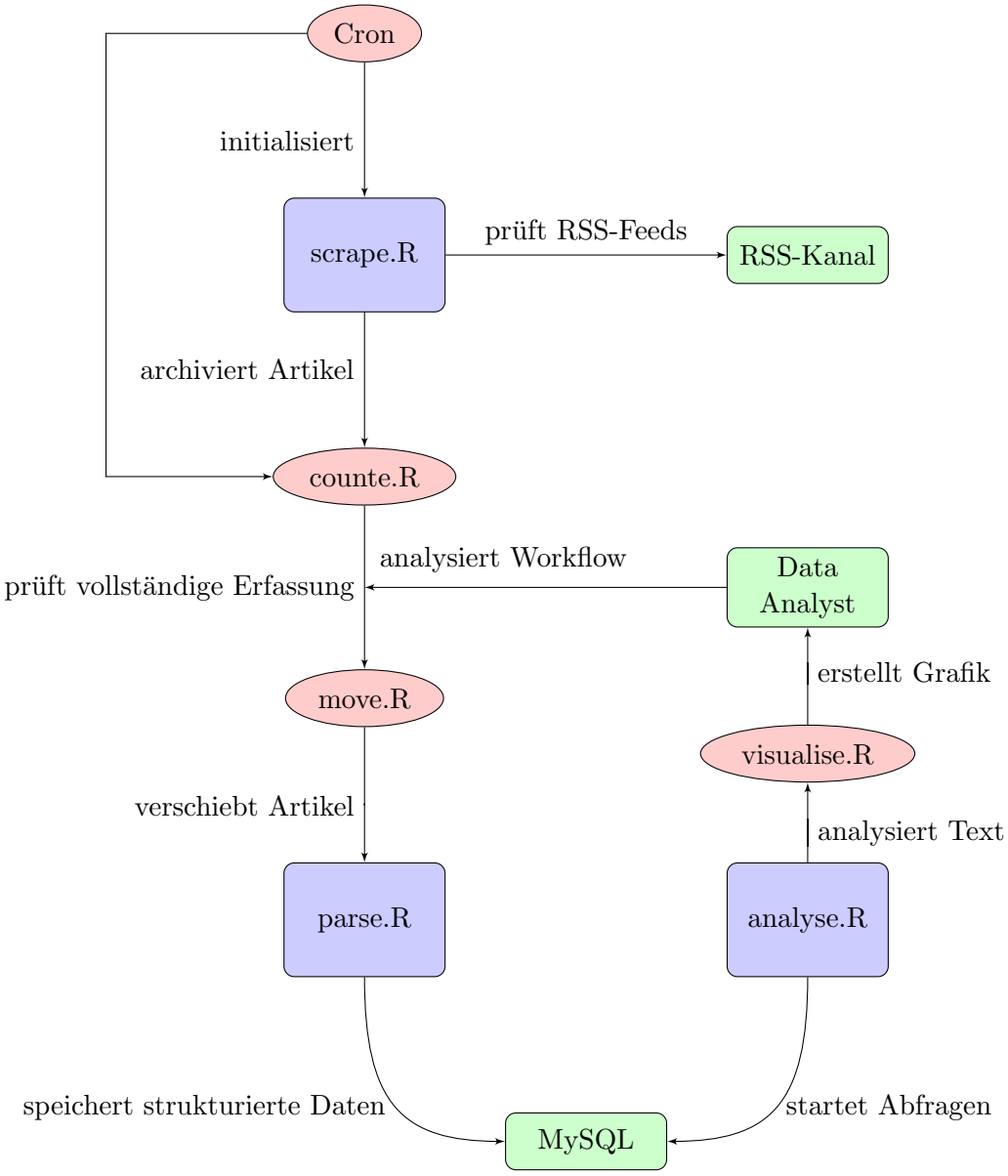
R-Funktionen. Vor allem hinsichtlich der inhaltlichen Textanalyse bietet R mit "quanteda" (Benoit u. Nulty, 2016) ein Paket mit enormem Funktionsumfang zur Erstellung und anschließender Auswertung eines Dokumentenkorpus. Ferner ist es bezüglich der Performanz für die Analyse großer Datenbestände ausgelegt. Da bereits ein gewisses Know-how bezüglich der Programmierung in R intern vorhanden war und zudem Teile eines Web Scrapers eines vorangegangenen Projekts vorlagen, fiel die Wahl schließlich auf die Programmiersprache R.
Das Know-how fällt dabei weitestgehend zurück auf die im Werk von Simon Munzert (Munzert u. a., 2014) beschriebenen Prozesse zur Extraktion von Informationen mittels Web Scraping und bildet damit die Grundlage bei der Programmierung und Weiterentwicklung meiner eigenen Skripte. Die Auseinandersetzung mit den dort angewandten Methoden führten außerdem zu der Entscheidung, die Archivierung der Artikel mittels einer SQL-Datenbank zu realisieren; die dort bereits nach erfolgtem Parsen der Dokumente in strukturierter Form hinterlegten Daten werden demnach als Basis für den Aufbau eines Dokumentenkorpus genutzt.
Zusammenfassend wurden folgende Softwarekomponenten auf den jeweiligen Systemen installiert:

	Software:	**Anwendungszweck:**
Testumgebung:	RStudio	Entwicklung und Test von Skripten
	XAMPP	Entwicklung des relationalen DBMS
	WinSCP	Remote Accesss
Laufzeitumgebung:	R-Base & Packages	Ausführen der R-Skripte
	LAMPP	Bereitstellung der MySQL-Datenbank
	Cron	zeitbasierte Ausführung der Skripte

2.2 Entwicklung des Monitoring-Instruments

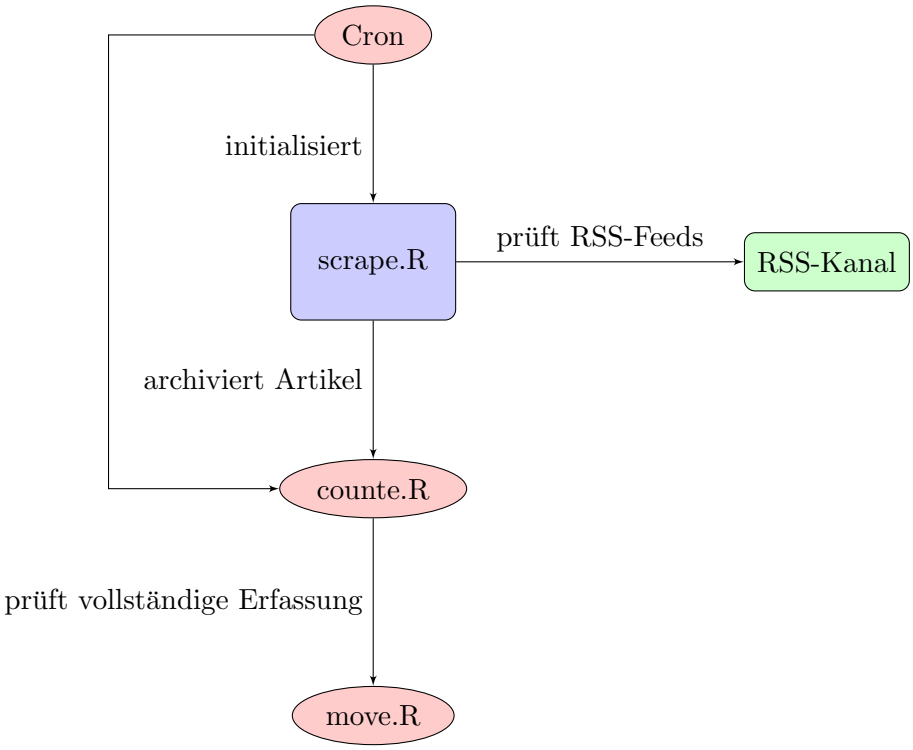
Nach Abschluss der obligatorischen Vorbereitungen konnte schließlich mit der Programmierung der eigentlichen Skripte begonnen werden. Technisch betrachtet soll dabei der gesamte Prozess, der von der Archivierung der HTML-Dokumente über das Parsen der einzelnen Artikel bis hin zur inhaltlichen Analyse reicht, anhand von drei separaten funktionsbasierten Skripten abgedeckt werden.
Des Weiteren werden optional einsetzbare Skripte geringeren Umfangs vorgestellt, die insbesondere der Qualitätssicherung dienen und ferner bei eventuell auftretenden Komplikationen Optionen zum frühzeitigen Eingreifen bieten sollen. Abbildung 2.1 gibt dazu eine Übersicht über den gesamten Prozess des Media Monitorings und soll gleichzeitig Aufschluss über die zugrunde liegende Softwarearchitektur geben.

Abbildung 2.1: Gesamtübersicht über die zugrundeliegende Softwarearchitektur



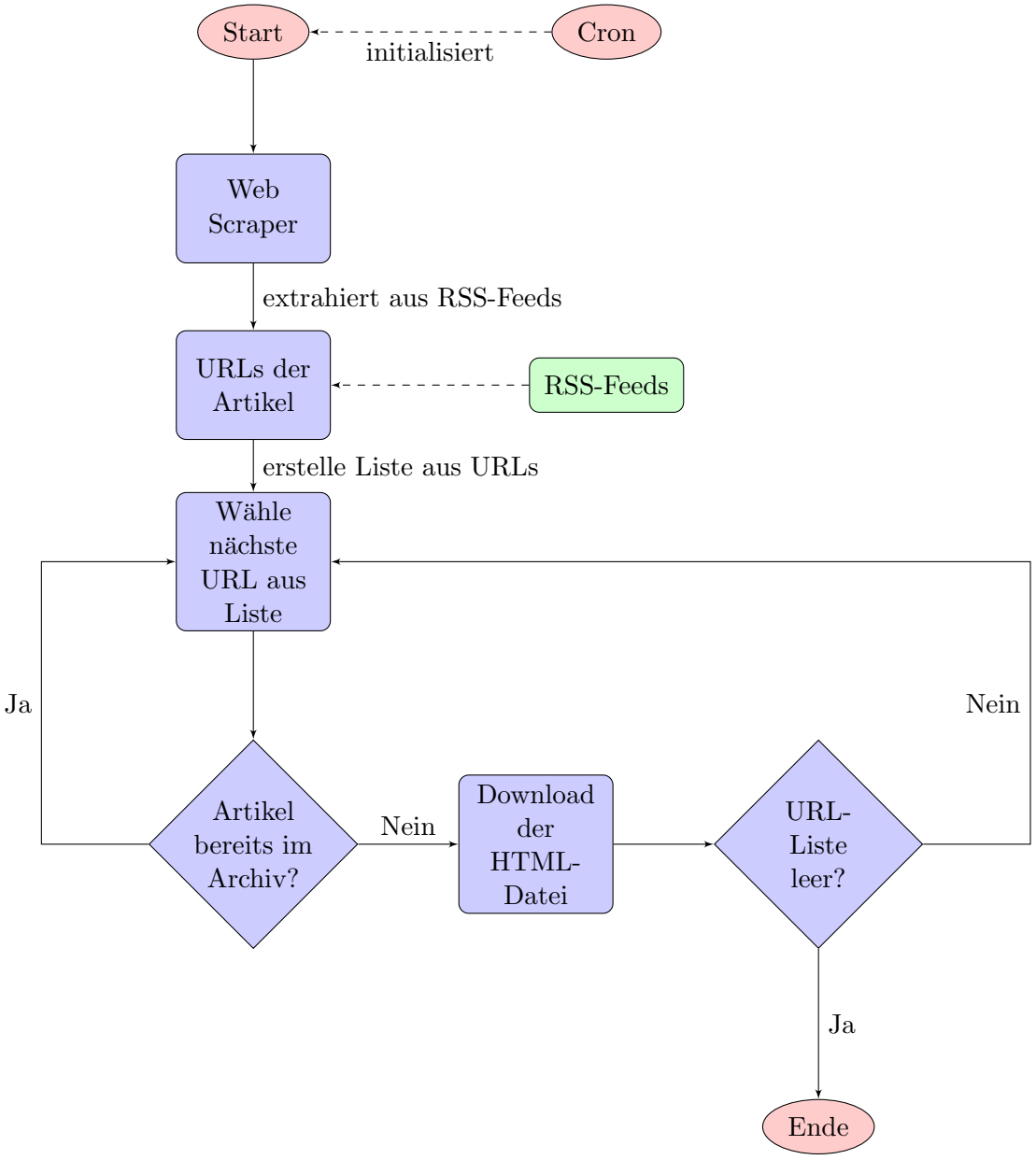
Die Visualisierung des in Abbildung 2.1 vorgestellten Prozessablaufs soll verdeutlichen, dass die zentralen Funktionen vor allem von den drei Kernelementen (scrape.R, parse.R und analyse.R) übernommen werden. Die genauere Funktionsweise der einzelnen Skripte wird im Folgenden detailliert beschrieben.

2.2.1 Entwicklung des Web Scrapers



Der "scrape.R" stellt das erste der drei zentralen Elemente des Monitoring-Instruments dar. Die Hauptfunktion dieses Skripts ist insbesondere die Generierung von Rohdaten, die im Zuge der weiteren Aufbereitung schließlich die Basis für spätere inhaltliche Analysen bilden.
Das Programm nutzt dabei Methoden des Web Scrapings, um URLs von neu erschienenen Artikeln aus den entsprechenden RSS-Kanälen zu extrahieren und lädt diese, sofern sie nicht bereits zu einem früheren Zeitpunkt archiviert wurden, in einem nächsten Schritt herunter (s. Abbildung 2.2). Als Basis dient dabei die von der Abteilung DBG erstellte Zusammenfassung aller für das Forschungsziel relevanten Feeds. Um auch im Nachhinein rückverfolgen zu können, aus welchem RSS-Kanal jeder einzelne Tweet stammt, wird außerdem eine Kopie des aktuellen RSS-Kanals abgespeichert. Aufgrund der hohen Fluktuation der Feeds innerhalb der RSS-Kanäle ist es darüber hinaus erforderlich, den Prozess in zeitlichen Abständen von 10 Minuten zu wiederholen, um eine vollständige Erfassung aller in diesem Zeitraum publizierten Artikel zu gewährleisten. Zu diesem Zweck wurde ein Crontab erstellt, der den Start des Scrapers in den dafür vorgesehenen zeitlichen Intervallen automatisch initialisiert.

Abbildung 2.2: Prozessaufbau Web Scraping

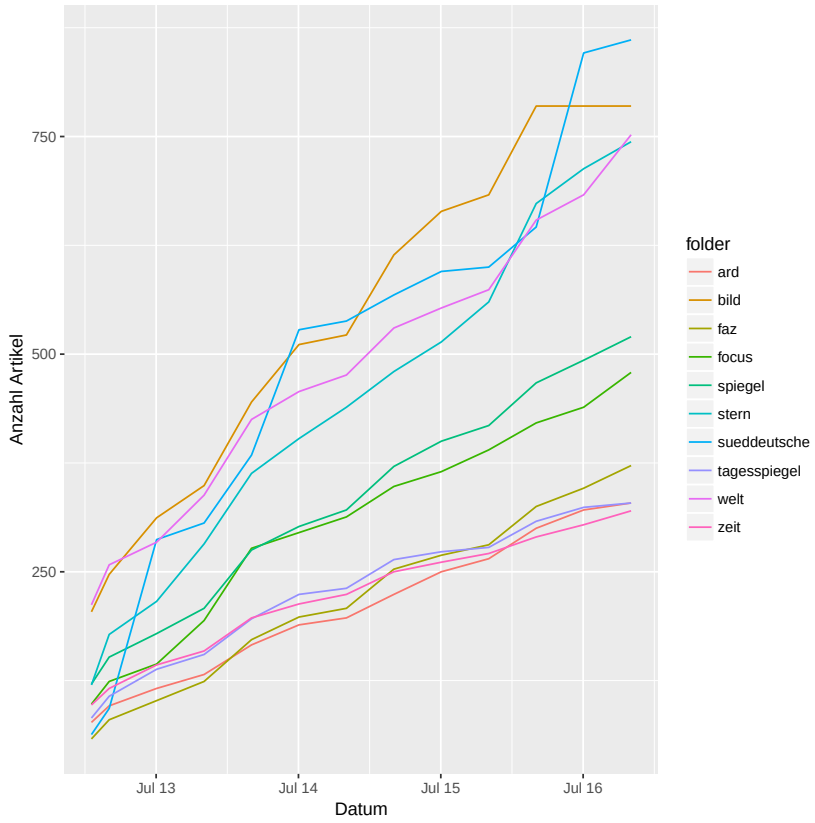


Funktionen der Skripte counte.R und move.R

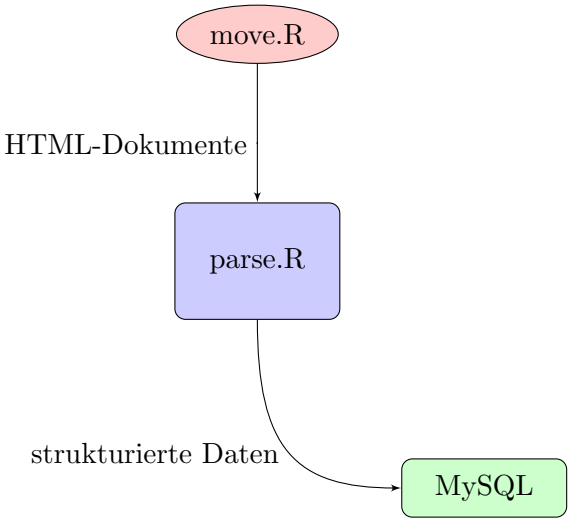
Um mögliche Komplikationen wie bspw. den Ausfall eines RSS-Kanals möglichst frühzeitig zu erkennen, werden die anhand der URLs heruntergeladenen Artikel zunächst in einer separaten Ordnerstruktur zwischengespeichert. Dabei werden sowohl die im HTML-Format vorliegenden Artikel als auch die im XML-Format vorliegenden RSS-Kanäle - getrennt nach den einzelnen Nachrichtenportalen - aufbewahrt. Dies ermöglicht dem “counte.R”, die bis zu den jeweiligen Zeitpunkten heruntergeladenen Artikel zu zählen. Zu diesem Zweck wurde ebenfalls ein weiterer Crontab eingerichtet, der den

Prozess des counte.Rs in Zeitintervallen von 8 Stunden automatisch startet. Im Anschluss wird eine Grafik erstellt, die die Anzahl der bis dahin erfassten Artikel im zeitlichen Verlauf darstellt. Durch die in zeitlichen Abständen von ca. einer Woche durchgeführten Checkups ist es dem Nutzer somit möglich, Ausfälle oder Umzüge einzelner RSS-Kanäle auf einem Blick zu erkennen und entsprechende Gegenmaßnahmen zeitnah einzuleiten. Das in Abbildung 2.3 dargestellte fiktive Beispiel soll veranschaulichen, welche Schlüsse der Betrachter aus der vom counte.R erstellten Grafik ziehen kann. Die Grafik zeigt, dass im Zeitraum vom 15.07 16:00 Uhr - 16.07 8:00 Uhr keine zusätzlichen Artikel aus dem RSS-Kanal des Nachrichtenportals BILD heruntergeladen werden konnten. Es ist daher naheliegend, dass der RSS-Kanal der BILD-Zeitung nicht mehr über die vorherige URL erreichbar ist. Dies ermöglicht es dem Nutzer die veraltete URL möglichst zeitnah durch die nun aktuellere Version in der dafür vorgesehenen Liste der RSS-Kanäle im scrape.R zu ersetzen. Nach erfolgter Analyse durch den Nutzer, können die bisher temporär zwischengespeicherten Artikel schließlich für die weitere Aufbereitung archiviert werden. Um mögliche Fehlerquellen beim manuellen Verschieben der Dateien zu vermeiden, wurde der Prozess mithilfe der Funktion “move.R” automatisiert.

Abbildung 2.3: Anzahl der im Zeitraum vom 12.07-16.07.18 erfassten Artikel



2.2.2 Entwicklung des Parsers



Der "parse.R" bildet das zweite der drei Kernelemente des Monitoring-Instruments. Hauptaufgabe dieses Skriptes ist die Aufbereitung der im vorangegangenen Prozess gewonnenen Rohdaten, um diese in strukturierter Form für die weitere Verarbeitung in einer relationalen Datenbank zu hinterlegen.
Hierzu ist es zunächst erforderlich, die relevanten Informationen aus dem Quelltext der im HTML-Format vorliegenden Artikel zu extrahieren. Vorgesehen ist dabei die möglichst einheitliche Erfassung folgender Informationen:

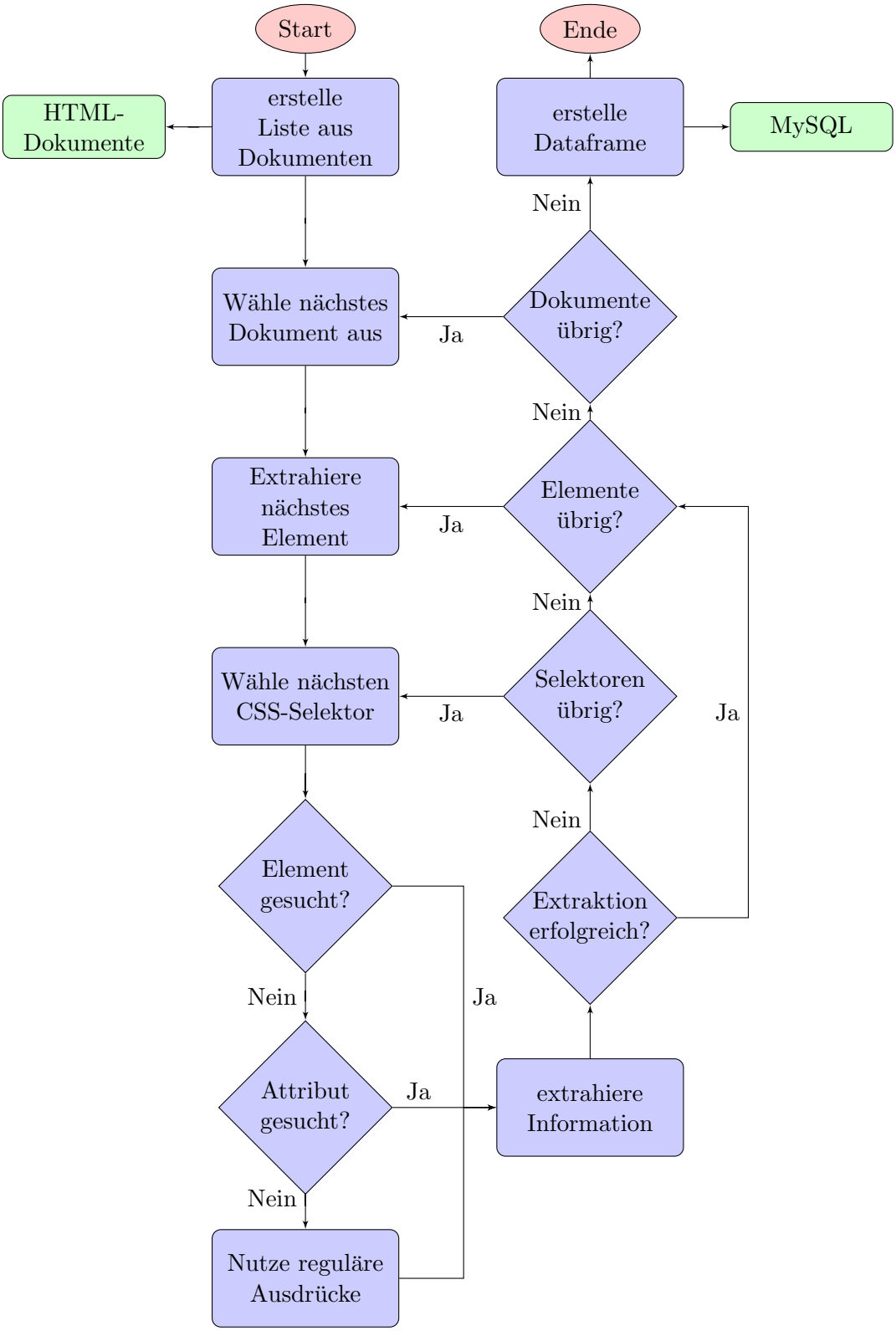
- *Schlagzeile:* Die Überschrift des Artikels
- *Dachzeile:* Eine Ergänzung zur Schlagzeile, die meist aus wenigen prägnanten Stichworten besteht
- *Lead:* Ein einleitender Text, der die wichtigsten Aussagen kurz zusammenfasst
- *Text:* Der eigentliche Text des Artikels
- *Datum:* Auskunft über den genauen Zeitpunkt der Publikation des Artikels
- *Ressort:* Welchem Ressort (Wirtschaft, Politik etc.) ist der Artikel zuzuordnen?
- *Paywall:* Handelt es sich um einen kostenpflichtigen Artikel?
- *Tweets:* IDs von eventuell im Artikel eingebetteten Tweets

Zu diesem Zweck werden entweder einzelne oder wahlweise auch gleich mehrere Knotenpunkte des DOM (Document Object Model), die die gewünschte Information beinhalten, mithilfe eines Selektors angesprochen. Dabei kann die Selektion der DOM-Objekte entweder über CSS- oder XPath-Selektoren erfolgen. Beide Varianten unterscheiden sich lediglich im Detail voneinander und liefern in der Kombination mit dem R-Paket

"rvest" eine vergleichbare Performance bei der Extraktion von einzelnen Elementen.
Gerade die einheitliche Erfassung aller o.g. Informationen stellte angesichts der starken strukturellen Unterschiede der Webseiten eine Herausforderung dar. So enthält bspw. der RSS-Kanal der ARD neben Beiträgen der Tagesschau auch Artikel der zugehörigen regionalen Rundfunkanstalten (z.B. BR, NDR, SWR), die hinsichtlich des Seitenaufbaus z.T. beachtliche strukturelle Unterschiede aufweisen. Das führt insbesondere dazu, dass die für die Extraktion benötigten CSS-Selektoren individuell an die einzelnen Seitenformate angepasst werden müssen. Der Parser muss folglich in der Lage sein, mit den ihm zur Verfügung gestellten Selektoren mehrere verschiedene Seitenformate in einem Vorgang gleichzeitig bedienen zu können. Um ein solch breites Spektrum an unterschiedlichen Seitenformaten zu parsen und dabei auf dynamische Veränderungen der strukturellen Beschaffenheit zeitnah reagieren zu können, war es erforderlich, unterstützende Software einzusetzen. Maßgeblich war in diesem Fall das Chrome Add-on SelectorGadget; durch die Markierung entsprechender Elemente im Browser gibt das Tool die erforderlichen CSS-Selektoren aus und gestaltet dadurch die Anpassung an die entsprechenden Seitenformate wesentlich effizienter.
Weiterhin muss der Parser neben der Extraktion von Informationen aus Elementen oder Attributen auch den Einsatz regulärer Ausdrücke beherrschen, um die ge–wünschten Metadaten gezielt im Quelltext ausfindig zu machen. Gerade im Fall der FAZ ist ein solches auf regulären Ausdrücken basiertes Vorgehen notwendig, um an die im Quellcode eingebetteten IDs der jeweiligen Tweets zu gelangen.
Die genaue Funktionsweise des Parsers wird im Folgenden detailliert in Abbildung 2.4 veranschaulicht, zudem steht das Skript im entsprechenden GitHub-Repository zur freien Verfügung[2].

[2]https://github.com/gesiscss/media-monitoring-mass-media-scraper

Abbildung 2.4: Prozessaufbau Parser



2.3 Statistische Betrachtung und Evaluation des Dokumentenkorpus

Vor Durchführung der inhaltlichen Analysen musste zunächst evaluiert werden, inwiefern die nun für die weitere Verarbeitung aufbereiteten Forschungsdaten den Anforderungen an das Projekt entsprechen. Zu diesem Zweck wurde vorab eine Zusammenfassung (s. Tab. 2.1) erstellt, die einen ersten Überblick über die Art und Beschaffenheit des Dokumentenkorpus ermöglichen soll. Dabei ist anzumerken, dass in der Berechnung des prozentualen Anteils kostenpflichtiger Artikel (Paywall%) lediglich diejenigen Nachrichtenportale mit einbezogen werden, die entsprechend kostenpflichtige Inhalte über ihre RSS-Kanäle publizieren.

In einem weiteren Schritt wird anschließend die prozentuale Erfassung der einzelnen

Portal	**Artikel**	**Wachstum/Woche**	**Tweets(%)**	**Paywall(%)**
ARD	11.540	405	3.7	0
Bild	40.859	1.434	5.9	12.2
FAZ	21.610	758	5.3	20.3
Focus	24.284	852	2	0
Freitag	1.622	57	0	0
Jungefreiheit	1.558	55	10.3	0
Spiegel	24.717	867	9.9	3.6
Stern	35.558	1.248	6	0
SZ	38.858	1.363	0.9	4.7
Tagesspiegel	17.033	598	0.3	0
Welt	31.468	1.104	1	9.2
Zeit	13.242	467	3.8	5.6
Gesamt:	262349	9205	4	9.6

Tabelle 2.1: Betrachtung des Dokumentenkorpus

Elemente dargestellt (s. Tab. 2.2). Die Aussagekraft dieser prozentualen Berechnung ist allerdings äußerst kritisch zu bewerten. Zwar könnte man auf Basis der Tabelle davon ausgehen, dass bspw. 89% des Haupttextes korrekt erfasst wurden; der tatsächliche Wert wird aber u.U. stark von diesem Messwert abweichen. Das ist vor allem damit zu begründen, dass die Messung einerseits lediglich erfasst, dass Informationen für das entsprechende Element extrahiert wurden, nicht aber, ob es sich tatsächlich um den korrekten Inhalt handelt. Andererseits wird das Ergebnis dadurch verfälscht, dass Artikel in der Berechnung berücksichtigt werden, die selbst in der entsprechenden Referenz über keinen Haupttext verfügen. Das ist insbesondere dann der Fall, wenn es sich bei dem jeweiligen Artikel um ein Video o.ä. Inhalt handelt.

Gleichwohl kann die Ermittlung der prozentualen Erfassung Anhaltspunkte bieten, ob bspw. einzelne Formate eines spezifischen Nachrichtenportals systematisch nicht durch die zuvor definierten CSS-Selektoren erfasst wurden.

Um letztendlich eine Aussage darüber treffen zu können, ob auch der korrekte Inhalt extrahiert worden ist, wurde im Folgenden eine detaillierte Evaluation durchgeführt.

Portal	Dachzeile	Kopfzeile	Lead	Text	Ressort	Datum
ARD	75.9	99	74.2	78	91.9	99.8
Bild	99.4	99.9	93.4	91.8	99.5	100
FAZ	99.8	99.8	98.2	97.9	98.4	99.9
Focus	54.6	99.9	96.8	97.7	99.8	99.9
Freitag	99.9	99.9	99.9	97.8	99.9	99.9
Junge Freiheit	99.8	100	98.6	99.9	99.9	100
Spiegel	57.7	74.5	98	80.5	99.8	99.9
Stern	70.3	99.9	99.8	87.4	83.5	100
SZ	99.9	99.9	83.7	99.8	99.9	99.9
Tagesspiegel	99.9	100	99.9	99	99.6	99.9
Welt	99.4	99.7	99.4	66.8	88.4	99.9
Zeit	91.3	99.2	94.4	86.7	95.7	100
Gesamt:	86	97.5	94.4	89	95.5	99.9

Tabelle 2.2: Prozentuale Erfassung der einzelnen Elemente

Evaluation

Für die nachfolgende Evaluation wurde zusätzlich zu den beiden klassischen Evaluationswerten Precision und Recall ein weiterer Messwert mit in die Analyse des Dokumentkorpus einbezogen. Die sogenannte "Slot Error Rate" (Makhoul u. a., 1999) kombiniert dabei die Aussagekraft von Precision und Recall in einem einzigen Messwert. Während die Precision Ersetzungen und Ergänzungen einzelner Elemente korrekt messen kann, eignet sich der Messwert des Recalls dazu, Ersetzungen und Teilerfassungen von Elementen zu bestimmen. Die SER berücksichtigt dagegen alle der drei o.g. potenziellen Fehlerquellen und bietet zusätzlich die Möglichkeit, die einzelnen Fehler entsprechend der gegebenen Umstände zu gewichten. Wird bspw. anstelle einer Überschrift eine Datumsinformation aus einem Artikel für den jeweiligen Slot extrahiert, so ist dieser Fehler - relativ betrachtet - schwerwiegender zu beurteilen als eine einfache Nichterfassung dieses Slots. Dies ist vor allem damit zu begründen, dass solche falsch positiven Ergebnisse zu einer fehlerhaften Datenbasis und somit im weiteren Verlauf zu Fehlinterpretationen führen können. Aus diesem Grund wird ein solcher Ersetzungsfehler (Fehlertyp s) entsprechend stärker gewichtet (x1.5) als eine Nichterfassung (Fehlertyp d, x1). Wird das gesuchte Element zwar vollständig korrekt erfasst, jedoch fälschlicherweise zusätzlicher Inhalt extrahiert, so handelt es sich um einen Ergänzungsfehler (Fehlertyp i). Zwar wurde das gesuchte Element korrekt erfasst, da aber die zusätzlich extrahierten Informationen auch in diesem Fall zu einer fehlerhaften Datenbasis führen, wird dieser Fehlertyp mit einem Multiplikator von 1.25 gewichtet.
Um den Messwert an die spezifischen Anforderungen der Informationsextraktion aus Nachrichtenartikeln anzupassen, wurde die SER um ein weiteres Fehlermerkmal ergänzt. Der Fehlertyp "Partial" (Fehlertyp p) stellt dabei die Summe aller teilweise erfassten Elemente dar. Wird bspw. ein Absatz innerhalb des Haupttextes eines Artikels nicht erfasst, so ist dies als weniger gravierend einzuordnen als die Nichterfassung des gesamten Haupttextes. Folglich fällt auch die Gewichtung geringer aus (x0.75).
Um nachvollziehen zu können, wie die einzelnen Messwerte berechnet worden sind,

werden die entsprechenden Variablen (angelehnt an das Verfahren von Makhoul u. a. (1999)) im Folgenden zunächst definiert:

- n = Summe aller Slots in der Referenz
- m = Summe aller Slots in der Stichprobe
- c = Summe aller korrekt erfassten Slots in der Stichprobe
- s = Summe aller fehlerfassten Slots in der Stichprobe (Substitutions, Gewichtung x1.5)
- i = Summe aller fälschlicherweise mit weiteren Inhalten ergänzten Slots (Insertion, Gewichtung: x1.25)
- d = Summe aller nicht erfassten Slots in der Stichprobe (Deletions, Gewichtung: x1)
- p = Summe aller teilweise erfassten Slots in der Stichprobe (Partial, Gewichtung: x0.75)

Die Evaluationsmesswerte Precision (P), Recall (R), Slot Error Rate (SER) und die gewichtete Slot Error Rate (SER_g) wurden anhand der folgenden Formeln berechnet:

$$P = \frac{C}{M}$$

$$R = \frac{C}{N}$$

$$SER = \frac{s + i + d + p}{n}$$

$$SER_g = \frac{(s * 1.5) + (i * 1.25) + (d * 1) + (p * 0.75)}{n}$$

Die Evaluation (s. Tabelle 2.3) wurde auf Basis einer Stichprobe von 100 per Zufall ausgewählten Artikeln aus dem Dokumentenkorpus durchgeführt. Als Referenz dienten die bis dato (24.10.18) auf den entsprechenden Internetportalen der Nachrichtenagenturen auffindbaren Artikel, wobei 12 der 100 Artikel nicht mehr öffentlich zugänglich waren. Weiterhin handelte es sich bei 6 der verbleibenden 88 Artikel um RSS-Feeds ohne dazugehörigen Text (Videos, kostenpflichtige Inhalte etc.); diese wurden bei der weiteren Evaluation ebenfalls nicht berücksichtigt. Daraus ergibt sich eine Stichprobenmenge

Element	m	n	c	$P(\%)$	$R(\%)$	$(s/i/d/p)$	$SER(\%)$	$SER_g(\%)$
Schlagzeile	80	82	80	100	97.6	(0/0/2/0)	2.4	2.4
Dachzeile	71	75	71	100	95	(0/0/4/0)	5	5
Lead	77	80	77	100	96.3	(0/0/3/0)	3.7	3.7
Text	73	74	70	95.9	94.6	(0/1/1/2)	5.4	5.1
Ressort	78	82	77	98.7	93.9	(0/0/4/1)	6.1	5.8
Datum	82	82	82	100	100	(0/0/0/0)	0	0
Paywall	5	5	4	80	80	(1/0/0/0)	20	30
Tweets	8	10	7	87.5	70	(1/0/2/0)	30	35

Tabelle 2.3: Evaluation des Dokumentenkorpus anhand der SER (Slot Error Rate)

von insgesamt 82 Artikeln[3]. Die Evaluationswerte wurden dabei für jedes einzelne Element (z.B. Überschrift, Text) ermittelt, um ein möglichst detailliertes Ergebnis zu erzielen. Um weiterhin Aussagen über die statistische Signifikanz treffen zu können, wurde ferner ein Zweistichproben t-Test bei abhängigen Stichproben durchgeführt. Zu diesem Zweck wurden zunächst jeweils die traditionelle SER sowie die gewichtete SER für jeden einzelnen der insgesamt 82 Artikel ausgerechnet. Im Anschluss daran konnten die Ergebnisse mithilfe des t-Tests analysiert werden, um schließlich den p-Wert zur Beurteilung der statistischen Signifikanz zu ermitteln (s. Tabelle 2.4).

	$SER(\%)$	$SER_g(\%)$
Mittelwert	0.0304878	0.02820122
Varianz	0.00446025	0.00406563
Beobachtungen	82	82
Pearson Korrelation	0.82615963	
Hypothetische Differenz der Mittelwerte	0	
Freiheitsgrade (df)	81	
t-Statistik	0.5364709	
P(T<=t) einseitig	0.29655176	
Kritischer t-Wert bei einseitigem t-Test	1.66388391	
P(T<=t) zweiseitig	0.59310352	
Kritischer t-Wert bei zweiseitigem t-Test	1.98968632	

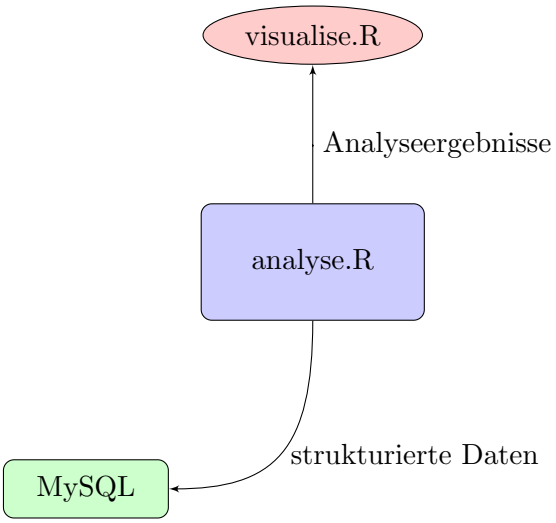
Tabelle 2.4: Zweistichproben t-Test bei abhängigen Stichproben zur Ermittlung der statistischen Signifikanz

[3]In insgesamt 4 der 82 Artikel konnte lediglich eine aktualisierte Version des gleichen Artikels als Referenz ausfindig gemacht werden; da allerdings offensichtlich nur marginale Anpassungen an einzelnen Elementen vorgenommen wurden, konnten die ansonsten korrekt erfassten Slots mit in die Evaluation aufgenommen werden.

Fazit der Evaluation

Die Evaluationsergebnisse deuten darauf hin, dass eine reine Betrachtung der klassischen Messwerte - Precision und Recall - meist nicht sonderlich zweckdienlich ist. Eine Übereinstimmung der Ergebnisse ($Recall + SER_g = 100$) liegt dabei nur in den Fällen vor, in denen lediglich Nichterfassungen (Fehlertyp d) aufgetreten sind. Betrachtet man dagegen bspw. die Messwerte für die Erfassung des Textes sowie der eingebetteten Tweets, so lässt sich feststellen, dass der Wert für die SER weniger gravierend bzw. schwerwiegender ausfällt, als es der Messwert des Recalls erwarten lassen würde. Diese Differenz ist vor allem auf die unterschiedlichen Gewichtungen der einzelnen Fehlertypen zurückzuführen. Mithilfe einer solchen an den Anwendungsfall angepassten Gewichtung kann sich die Aussagekraft der gewichteten SER im Vergleich zu den traditionellen Evaluationsmesswerten erhöhen. Die in Tabelle 2.4 dargelegten Ergebnisse deuten allerdings darauf hin, dass die Abweichungen zwischen der traditionellen SER und der gewichteten SER bei einer solchen Stichprobenanzahl statistisch nicht signifikant sind. Der kritische p-Wert für die zweiseitige Abweichung liegt dabei mit einem Wert von 0.59 deutlich über der vorher definierten Irrtumswahrscheinlichkeit von 5%.
Insgesamt weisen die Evaluationsergebnisse darauf hin, dass die einzelnen Elemente weitestgehend korrekt von den Selektoren bzw. den regulären Ausdrücken des Parsers extrahiert wurden. Auffallend sind allerdings insbesondere die Ergebnisse der Datumsangaben sowie der Tweets und der Erfassung der Paywall. Während die Datumsangaben vollständig korrekt erfasst wurden, scheint die Extraktion der eingebetteten Tweets und der Paywall-Information bisweilen fehleranfällig zu sein. Hierbei ist jedoch anzumerken, dass es sich um eine relativ kleine Stichprobe handelt und somit einzelne Fehler z.T. gravierende Auswirkungen auf das Messergebnis haben können.

3 Durchführung der inhaltlichen Analyse

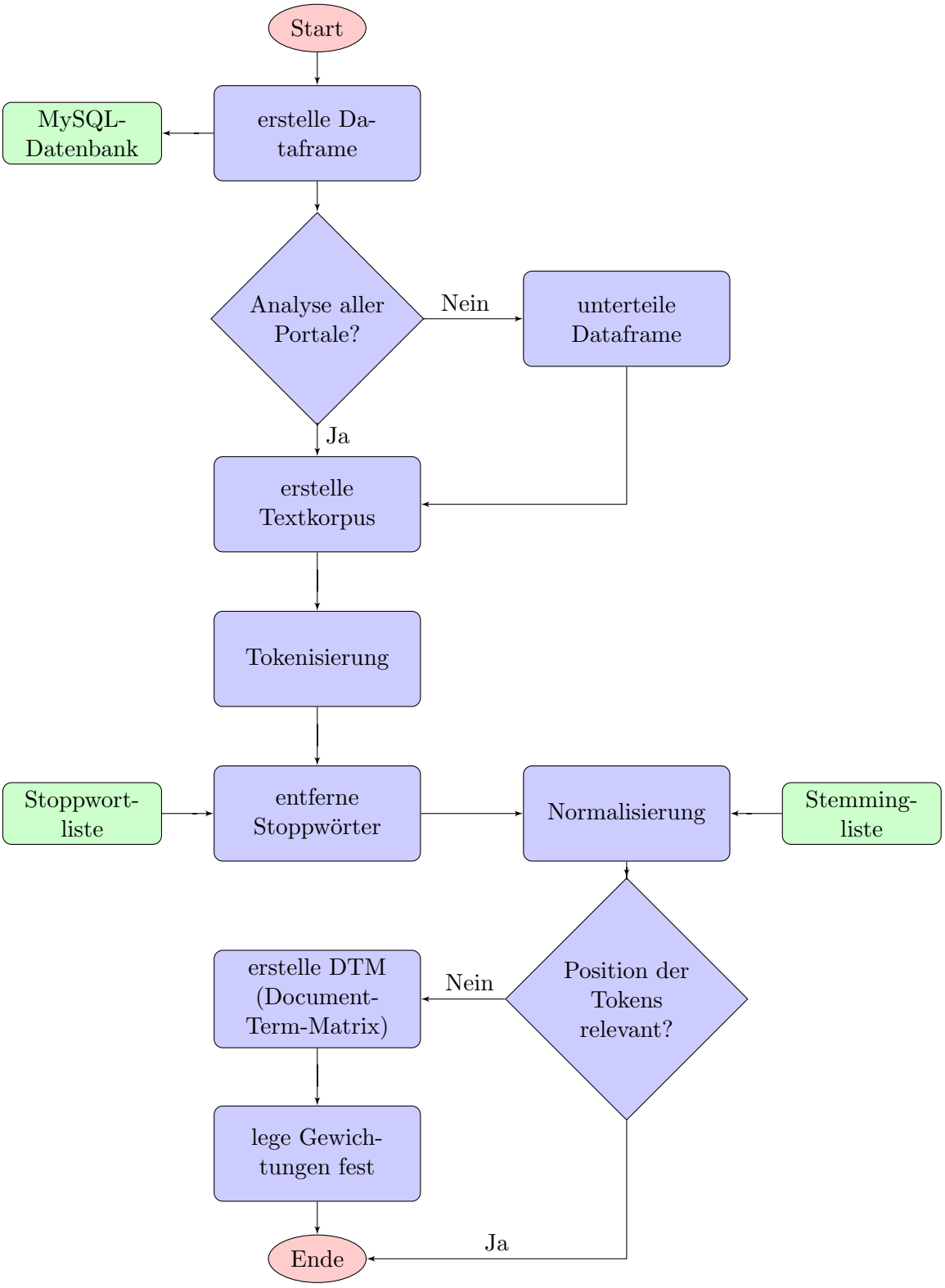


Im Anschluss an die zuvor durchgeführte Evaluation des Dokumentenkorpus können die nun in strukturierter Form archivierten Artikel aus der MySQL-Datenbank in einem nächsten Schritt für die weitere inhaltliche Analyse aufbereitet werden. Diese Aufgabe wird unter anderem vom letzten der drei Kernelemente des Monitoring-Instruments, dem sogenannten "analyse.R", übernommen. Vorrangiges Ziel dieses Verarbeitungsprozesses ist es, den bis dato in natürlicher Sprache vorliegenden Haupttext der Artikel zum Zwecke der weiteren automatisierten Analyse in ein Format zu transformieren, das den unterschiedlichen Anforderungen der jeweiligen Analyseverfahren entspricht. Die einzelnen Teilprozesse der in Abbildung 3.1 veranschaulichten Vorverarbeitungspipeline des Analysers werden im Folgenden kurz vorgestellt. Der Prozessablauf ist dabei angelehnt an das Verfahren von Benoit u. a. (2018).

3.1 Die Vorverarbeitungspipeline

In einem ersten Schritt wird mittels einer SQL-Abfrage auf den Server der gesamte Inhalt des Dokumentenkorpus ausgegeben und anschließend in einem entsprechenden Dataframe abgespeichert. Dieses Dataframe bietet die Basis für das weitere Vorgehen und kann auf Beiträge eines bzw. mehrerer Nachrichtenportale reduziert werden, sofern nicht die Gesamtheit aller Artikel analysiert werden soll.
Zur Vorbereitung der bevorstehenden Aufbereitungsprozesse wird das nun angepasste Dataframe zunächst in ein Korpus-Objekt umgewandelt. In diesem Stadium liegt

Abbildung 3.1: Vorverarbeitungspipeline des Analysers



der Haupttext innerhalb des Korpus-Objekts jedoch noch als reine Zeichenkette vor; folglich wäre die Durchführung einer Textanalyse, insbesondere im Hinblick auf die enorme Menge an zu verarbeitenden Daten, zu diesem Zeitpunkt noch nicht realisierbar. Zielt das Analyseverfahren darauf ab, einzelne Variablen im zeitlichen Verlauf zu visualisieren, so ist es fortan möglich, Artikel anhand einzelner Dokumentenvariablen (wie bspw. der Datumsinformation) zu gruppieren.
Im Zuge der Tokenisierung wird der bis dato als Zeichenkette vorliegende Text auf Wortebene segmentiert. Die einzelnen daraus resultierenden Tokens werden im Anschluss in einer Vektorliste hinterlegt und bilden das Token-Objekt. Dabei ist anzumerken, dass die Position der einzelnen Tokens im Dokument weiterhin zurückverfolgt werden kann. Dies ist vor allem von besonderer Bedeutung für Analyseverfahren, die bspw. Kollokationen oder N-Gramme einzelner Tokens ermitteln oder etwa die natürliche Sprache ("Natural Language Processing") analysieren sollen (Welbers u. a. (2017, S. 250)).
In einem weiteren Schritt werden die nun in Token segmentierten Texte zunächst anhand einer Liste von Stoppwörtern und anschließend von Interpunktionen befreit (Porter (2001)). Insbesondere für spätere Frequenzanalysen ist das Ausschließen dieser Terme relevant, da Stoppwörter i.d.R. in natürlichsprachigem Texten die am häufigsten vorkommende Wortgattung darstellen und gleichzeitig keine besondere inhaltliche Relevanz aufweisen. Ferner ist es für die erfolgreiche Durchführung einer Frequenzanalyse sinnvoll, die nun indexierten Terme auf ihren Wortstamm zu reduzieren. Durch den Einsatz eines Stemmming-Algorithmus (Porter (2001)) ist es in der Folge möglich, alle Varianten eines Wortes im Zuge der Frequenzanalyse zu erfassen.
Ist die Position der Tokens für die nachfolgende Analyse essentiell, so ist die Vorverarbeitung nach Durchführung der Normalisierung abgeschlossen. Das ist vor allem darauf zurückzuführen, dass der bis dato als "string-of-words" vorliegende Haupttext der Artikel im nachfolgenden Verarbeitungsschritt in ein sogenanntes "bag-of-words"-Modell umgewandelt wird, das keine Rückschlüsse auf die Position der jeweiligen Tokens zulässt. Ist die Position der Tokens dagegen für die Anwendung des anschließenden Analyseverfahrens nicht relevant, wird der zuvor segmentierte Text in eine für die weitere Verarbeitung wesentlich effizientere Dokument-Term-Matrix umgewandelt.
Schließlich wird das Vokabular der DTM mittels Anwendung weiterer Filterungs- bzw. Gewichtungsverfahren dahingehend reduziert, dass Terme mit geringem Informationsgehalt nach Möglichkeit ausgeschlossen werden. Gerade für besonders rechenintensive Verfahren wie das LDA Topic Modelling führt dies einerseits zu einer deutlichen Steigerung der Performance, andererseits erhöht dies zusätzlich auch die Genauigkeit solcher Verfahren, da für den jeweiligen Analysezweck weniger relevante Terme im Vorhinein ausgeschlossen wurden Welbers u. a. (2017, S. 253). Neben einer manuell angelegten Filterliste können zudem Gewichtungsverfahren angewandt werden, um das untersuchte Vokabular weiter einzugrenzen. Unter Anwendung des tf-idf Verfahrens (term frequency - inverse document frequency) ist es bspw. möglich, in besonders vielen Dokumenten vorkommende Terme durch Zuteilung einer geringeren Gewichtung vom weiteren Verarbeitungsprozess auszuschließen.
Abschließend ist anzumerken, dass die vorgestellte Vorverarbeitungspipeline in Art und Reihenfolge der einzelnen Teilprozesse variiert werden kann bzw. muss, um sie individuell an den jeweiligen Anwendungsfall anzupassen.

3.2 Vorstellung der Analyseverfahren

Nach Durchführung aller vorbereitenden Prozesse ist es nun möglich, auf Basis des im vorangegangenen Verarbeitungsschritt aufbereiteten Inhalts der Artikel Analyseverfahren anzuwenden. Die im Dokumentenkorpus archivierten Artikel bieten dabei ein breites Spektrum an potenziell durchführbaren Analysen, von denen im Folgenden eine Auswahl von drei Verfahren vorgestellt wird, die sich sowohl in ihrer Vorgehensweise, als auch in der Art der Visualisierung der Ergebnisse stark unterscheiden.

3.2.1 Kollokationsanalyse

Gegenstand der ersten der drei durchgeführten Analyseverfahren war die Bestimmung sogenannter Kollokationen eines zuvor definierten Terms. Dabei versteht man in der Sprachwissenschaft unter einer Kollokation das häufige gemeinsame Aufkommen von zwei oder mehreren Wörtern (Lemnitzer u. Zinsmeister (2006, S. 30)). Im folgenden Beispiel einer solchen Kollokationsanalyse wurde nach den im direkten Umfeld befindlichen Termen gesucht, die am häufigsten gemeinsam mit dem Wort “Flüchtling” vorkommen. Um eventuell auftretende ideologische Kontraste zwischen den einzelnen Akteuren sichtbar zu machen, wurden die am häufigsten mit dem Begriff “Flüchtling” vorkommenden Terme des linksliberalen Nachrichtenportals “Der Freitag”[1], der öffentlich-rechtlichen Institution “ARD” sowie des sich selbst als konservativ bezeichnenden Verlags der Wochenzeitung “Junge Freiheit”[2] miteinander verglichen.

Vorgehensweise

Vor Durchführung der Vorverarbeitung wird der Dokumentenkorpus zunächst auf alle Artikel der Nachrichtenportale “Junge Freiheit”, “Der Freitag” und “ARD” reduziert. Wie bereits in Kapitel 3.1 erwähnt, muss der Vorverarbeitungsprozess für jeden Anwendungsfall individuell angepasst werden. So ist es im Falle der Kollokationsanalyse für das weitere Vorgehen essentiell, dass die Positionen der einzelnen Tokens im Text auch in Folge der Vorverarbeitung erhalten bleiben. Aus diesem Grund wird auf die Erstellung einer DTM zugunsten eines weiterhin im “string-of-words”-Format vorliegenden Inhalts verzichtet. Zudem wurden die einzelnen Token zunächst durch Stemming auf ihren Wortstamm reduziert und anschließend mittels tf-idf gewichtet.
Nach Abschluss der Vorverarbeitung wird der Inhalt der Artikel auf diejenigen Token reduziert, die in einem maximalen Abstand von 10 Token zu dem gesuchten Term “Flüchtling” auftreten. In einem darauffolgenden Schritt werden die einzelnen Dokumente nach dem jeweiligen Nachrichtenportal gruppiert. Nun kann das Aufkommen der entsprechenden Token anhand einer einfachen Frequenzanalyse zunächst erfasst und anschließend in Form einer Wordcloud visualisiert werden. Die folgende Grafik ist das Resultat einer solchen Kollokationsanalyse und zeigt die 80 am häufigsten, im direkten Umfeld (max. 10 Wörter) zum Wort “Flüchtling” liegenden Token der drei untersuchten Nachrichtenportale.

[1] https://www.freitag.de/ueber
[2] https://jungefreiheit.de/informationen/ueber-den-verlag/

Abbildung 3.2: In Form einer kombinierten Wordcloud visualisiertes Ergebnis der Kollokationsanalyse

Interpretation

Zunächst ist anzumerken, dass das als Wordcloud vorliegende Ergebnis der Kollokationsanalyse keine konkreten Rückschlüsse auf ein bestimmtes Framing des Begriffs “Flüchtling” zulässt. Dies ist vor allem damit zu begründen, dass lediglich die Frequenz nahestehender Terme zum Wort “Flüchtling” betrachtet wird; ob die im Kontext zum Wort “Flüchtling” verwendeten Formulierungen darauf abzielen, lediglich bestimmte Interpretationen des Lesers zu diesem Thema zuzulassen, kann auf Basis der erzielten Ergebnisse nicht dargestellt werden. Allerdings zeigt das Ergebnis der Analyse deutlich, dass die am häufigsten im Zusammenhang mit dem Begriff “Flüchtling” verwendeten Token des linksliberalen Nachrichtenportals “Freitag” größtenteils positiver Natur sind, die Token der Wochenzeitung “Junge Freiheit” größtenteils negativ geprägt sind und die resultierenden Token der “ARD” darauf hindeuten, dass eine vergleichsweise neutrale Berichterstattung zugrunde liegt.

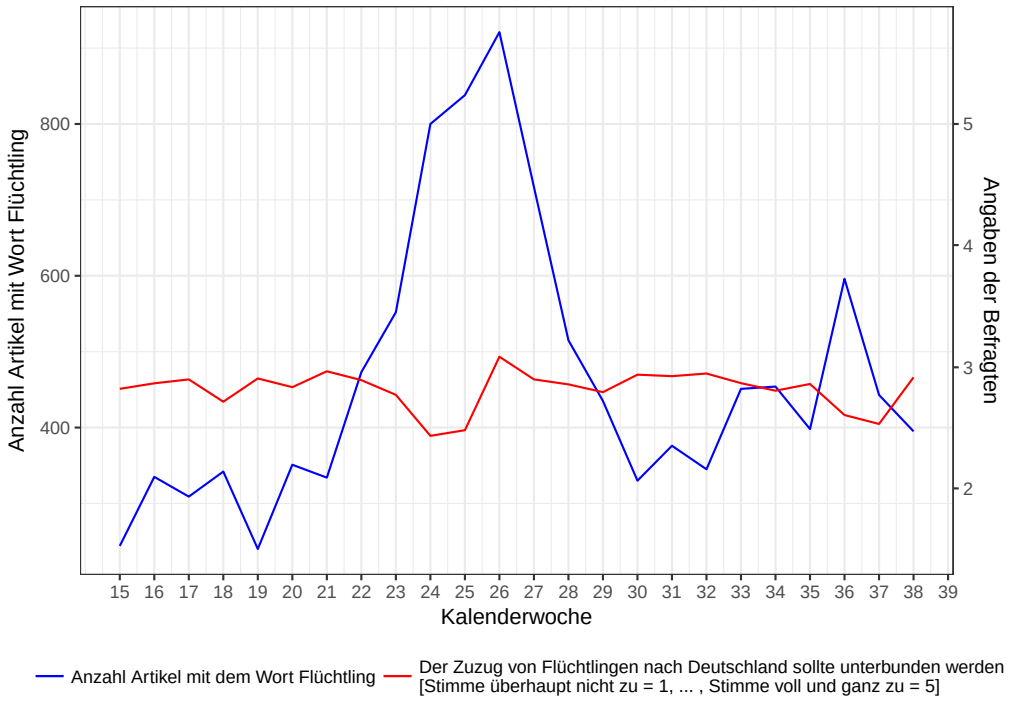
3.2.2 Analyse der ALLBUS-Umfrage

Anlässlich der anfänglich gestellten Forschungsfrage nach möglichen Zusammenhängen zwischen dem Aufkommen bestimmter Themen in den Massenmedien und den Umfrageergebnissen des ALLBUS wurden verschiedene quantitative Verfahren angewandt, um eventuell auftretende Korrelationen sichtbar zu machen.

Vorgehensweise

Vor Durchführung der nachfolgenden Analyseverfahren werden zunächst nur die relevanten Dokumente des Textkorpus für die weitere Verarbeitung selektiert. Dabei werden ausschließlich Artikel ausgewählt, die im Erhebungszeitraum des ALLBUS (09.04.2018 - 23.09.2018) publiziert wurden. In einem weiteren Schritt werden die Dokumente wie bereits in Vorbereitung auf die o.g. Kollokationsanalyse tokenisiert, gestemmt und von Stoppwörtern befreit. Darüber hinaus wird der segmentierte Text in eine DTM umgewandelt und schließlich nach Tagesdatum gruppiert. Anschließend wird die Dokumentenfrequenz pro Tag des Wortstamms des gesuchten Terms ermittelt (z.B. Flüchtling). In einem letzten Schritt wird die Dokumentenfrequenz der einzelnen Tage auf Wochenebene (Kalenderwochen) aggregiert. Die daraus resultierenden wöchentlichen Dokumentenfrequenzen werden daraufhin mit den Angaben der Befragten der ALLBUS-Umfrage verglichen. Analog zur Vorbereitung der Texte des Dokumentenkorpus werden auch diese zunächst nach Datum sortiert und entsprechend mittels Bestimmung des arithmetischen Mittelwerts auf Wochenebene aggregiert.

Abbildung 3.3: Gegenüberstellung von Artikeln mit dem Term “Flüchtling” und den Angaben der Befragten des ALLBUS zum Thema Zuwanderung



Einstellung gegenüber Zuwanderung

In einer ersten Analyse wurden mögliche Zusammenhänge zwischen dem Aufkommen des Terms “Flüchtling” in den Massenmedien und den Umfrageergebnissen des ALLBUS bezüglich migrationspolitischen Fragen untersucht. Zu diesem Zweck wurde

die Dokumentenfrequenz pro Woche des Wortes "Flüchtling" mit dem wöchentlichen Durchschnittswert der Antworten der Befragten zu folgender Fragestellung verglichen: "Der Zuzug von Flüchtlingen sollte unterbunden werden. Die Antwortmöglichkeiten reichten von 1 ("Stimme überhaupt nicht zu") bis 5 ("Stimme voll und ganz zu").
Wie dem in Abbildung 3.3 dargestellten Verhältnis zwischen dem Aufkommen des Wortes "Flüchtling" und den Umfrageergebnissen zu entnehmen ist, gibt es keinen sichtbaren sowie messbaren Zusammenhang der beiden Variablen. Lediglich während des Zeitraums von KW 24 - KW 27 ist eine marginale Abhängigkeit beider Variablen zu beobachten.

3.2.3 LDA Topic Modelling

Neben den bisher durchgeführten deskriptiven Analysen wird im Folgenden ein Verfahren angewandt, das auf einem "unsupervised machine learning"-Algorithmus basiert. Ziel des sogenannten LDA Topic Modellings ist die Visualisierung von latenten Themenfeldern ("Topics"), die mithilfe probabilistischer Verfahren auf Grundlage der in den Artikeln verwendeten Wörtern aufgedeckt werden. Im weiteren Verlauf ist es zudem möglich, die relative Verteilung der Themenfelder für die einzelnen Nachrichtenportale zu bestimmen. Ein anschließender Vergleich der relativen Anteile der entsprechenden Topics zwischen den Nachrichtenportalen wird darüber hinaus Hinweise geben, welche Themenschwerpunkte von den jeweiligen Portalen gesetzt werden und inwiefern sie sich hinsichtlich ihrer Themenverteilung unterscheiden.

Sampling und Vorverarbeitung

Um gewährleisten zu können, dass jedes einzelne Nachrichtenportal in gleicher Anzahl vertreten ist und gleichzeitig eine möglichst hohe Performanz sicherzustellen, wurde zunächst ein repräsentatives Sample der Artikel aus dem Dokumentenkorpus ausgewählt; dieses besteht aus insgesamt 12000 Dokumenten, wobei jeweils 1000 Artikel einem Nachrichtenportal zugeordnet sind und zudem zufällig aus der Kollektion des entsprechenden Portals selektiert wurden. Aufgrund der Tatsache, dass 1265 der 12000 Artikel (10.5%) über keinen Text verfügten, wurden diese für die weitere Bearbeitung nicht berücksichtigt. Daraus ergibt sich ein Dokumentenkorpus bestehend aus insgesamt 10735 verschiedenen Dokumenten.
Im Zuge der Vorverarbeitung wird das Sample auf die nachfolgenden Analyseprozesse präpariert. Dabei wird der Haupttext jedes Dokuments tokenisiert, anhand einer Stemming-Liste auf den Wortstamm reduziert und schließlich in eine für die weitere Verarbeitung effizientere DTM transformiert. Um die Anzahl an Token zugunsten einer höheren Performanz weiter zu reduzieren und entsprechend weniger relevante Terme zu filtern, wurde die minimale Termfrequenz auf 5 gesetzt. Dies ermöglichte eine Reduktion des Vokabulars von 175.238 auf 36.867 Token.

Topic Modelling

Anschließend kann auf Grundlage der nun im "bag-of-words"-Format vorliegenden Dokumententexte das auf probabilistischen Methoden basierende Analyseverfahren LDA Topic Modelling angewandt werden. Dabei wird eine zuvor festgelegte Anzahl (k=100)

an latenten Themenfeldern ("Topics") innerhalb des Dokumentenkorpus modelliert, die wiederum aus Wörtern gebildet werden, die auf Grundlage probabilistischer Vorhersagen am ehesten dem jeweiligen Topic zugeordnet werden können (Blei u. a. (2003)).
Zu Beginn nimmt das Model an, dass die Wörter innerhalb eines Dokuments thematisch verwandt sind. In einem darauf folgenden Schritt wird unter Berücksichtigung der zuvor definierten Anzahl an Topics (k) die Wahrscheinlichkeit jedes Tokens errechnet, einem bestimmten Topic zugeordnet zu werden. Dabei kann ein Token mehreren Themenfeldern zu unterschiedlichen prozentualen Wahrscheinlichkeiten angehören. Im Umkehrschluss kann in einem weiteren Schritt bestimmt werden, wie hoch die Wahrscheinlichkeit eines Dokumentes ist, einem Themenfeld zugeordnet zu werden. Auch in diesem Fall ist es möglich, dass ein Dokument zu unterschiedlichen prozentualen Wahrscheinlichkeiten mehreren Themenfeldern zugewiesen werden kann.
Aufgrund der Tatsache, dass sowohl die zuvor errechneten Wahrscheinlichkeiten als auch die Zugehörigkeit eines Artikels zu einem bestimmten Nachrichtenportal auf Dokumentenebene bekannt sind, können diese Wahrscheinlichkeitswerte mithilfe der Ermittlung des entsprechenden arithmetischen Mittelwerts auf Ebene des Nachrichtenportals aggregiert werden. Hieraus ergibt sich für jedes einzelne Nachrichtenportal ein Vektor, der die prozentualen Wahrscheinlichkeiten aller Topics beinhaltet. Zusätzlich werden die fünf Wörter, die die höchste Wahrscheinlichkeit besitzen dem jeweiligen Themenfeld zugeordnet zu werden, als Label des entsprechenden Topics genutzt.

Interpretation des LDA Topic Modellings mit k = 100 Topics

Die in Abbildung 3.5 visualisierten Endergebnisse der ersten LDA-Analyse zeigen, dass die Themenschwerpunkte abhängig vom jeweiligen Nachrichtenportal z.T. stark variieren. Ferner lassen die vom LDA-Algorithmus zu einem Themenfeld zusammengefassten Wörter darauf schließen, dass die Anzahl der zuvor definierten Topics (k=100) zu konsistenten Wortgruppen geführt hat. Zur besseren Übersichtlichkeit und um den thematischen Kontrast untereinander detaillierter darstellen zu können, bezieht sich die nachfolgende Auswertung der Ergebnisse auf die in Abbildung 3.4 getroffene Auswahl der Nachrichtenportale "Bild", "Focus" und "Junge Freiheit".
Betrachtet man das Nachrichtenportal "Focus", so ist erkennbar, dass ein klarer Schwerpunkt auf Wirtschafts- und Finanzthemen gelegt wird. Dies impliziert das gehäufte Aufkommen von themenverwandten Begriffen ("Aktie", "Dollar, "Kurs" etc.) innerhalb der fünf Topics mit dem größten relativen Anteil. Einen gänzlich anderen Fokus setzt dagegen das Portal "Junge Freiheit". Im Falle des rechtskonservativen Portals stehen vor allem innen- und migrationspolitische Themen (Schiff/Italien, Chemnitz/Demonstrationen, Merkel/Seehofer etc.) im Vordergrund, deren deutlich überrepräsentierter Anteil auf eine ideologisierte Verwendung der Wörter innerhalb der entsprechenden Themenfelder hindeutet. Hinsichtlich der relativen Themenverteilung der Bild-Zeitung ist anzumerken, dass das Gesamtbild hauptsächlich durch typische Themengebiete des Sensationsjournalismus (Mann, verletzt, Frau, Beamte, Polizei etc.) als auch des Sports geprägt ist. Wobei die Themen mit klarem Bezug zum Sport überwiegend der Kategorie Fußball zuzuordnen sind (Trainer, FC, Minute, Spiel etc.).

Abbildung 3.4: Ergebnisse des LDA Topic Modellings (k=100) für die Nachrichtenportale “Bild”, “Focus” und “Junge Freiheit”

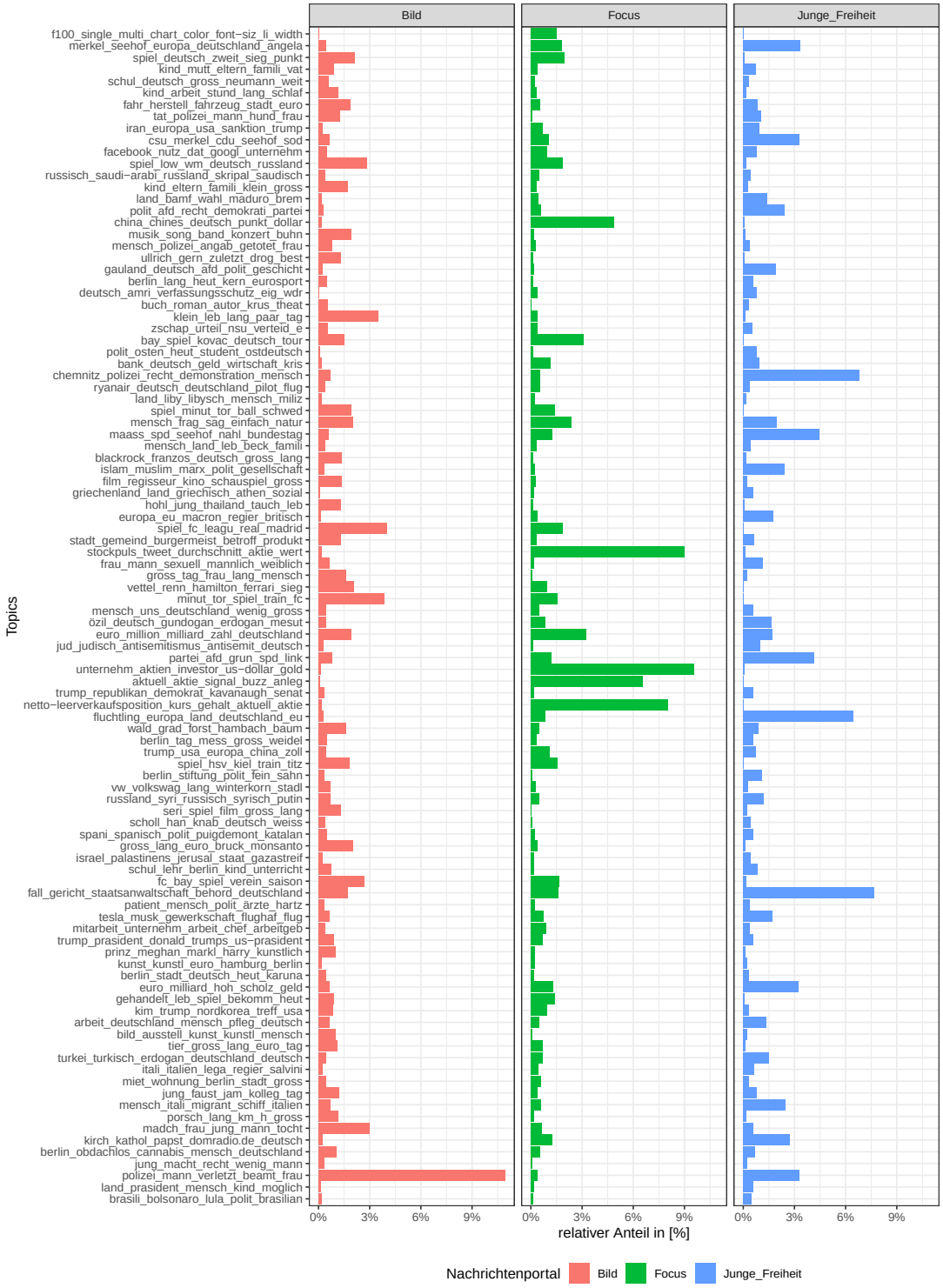


Abbildung 3.5: Ergebnisse des LDA Topic Modellings (k=100)

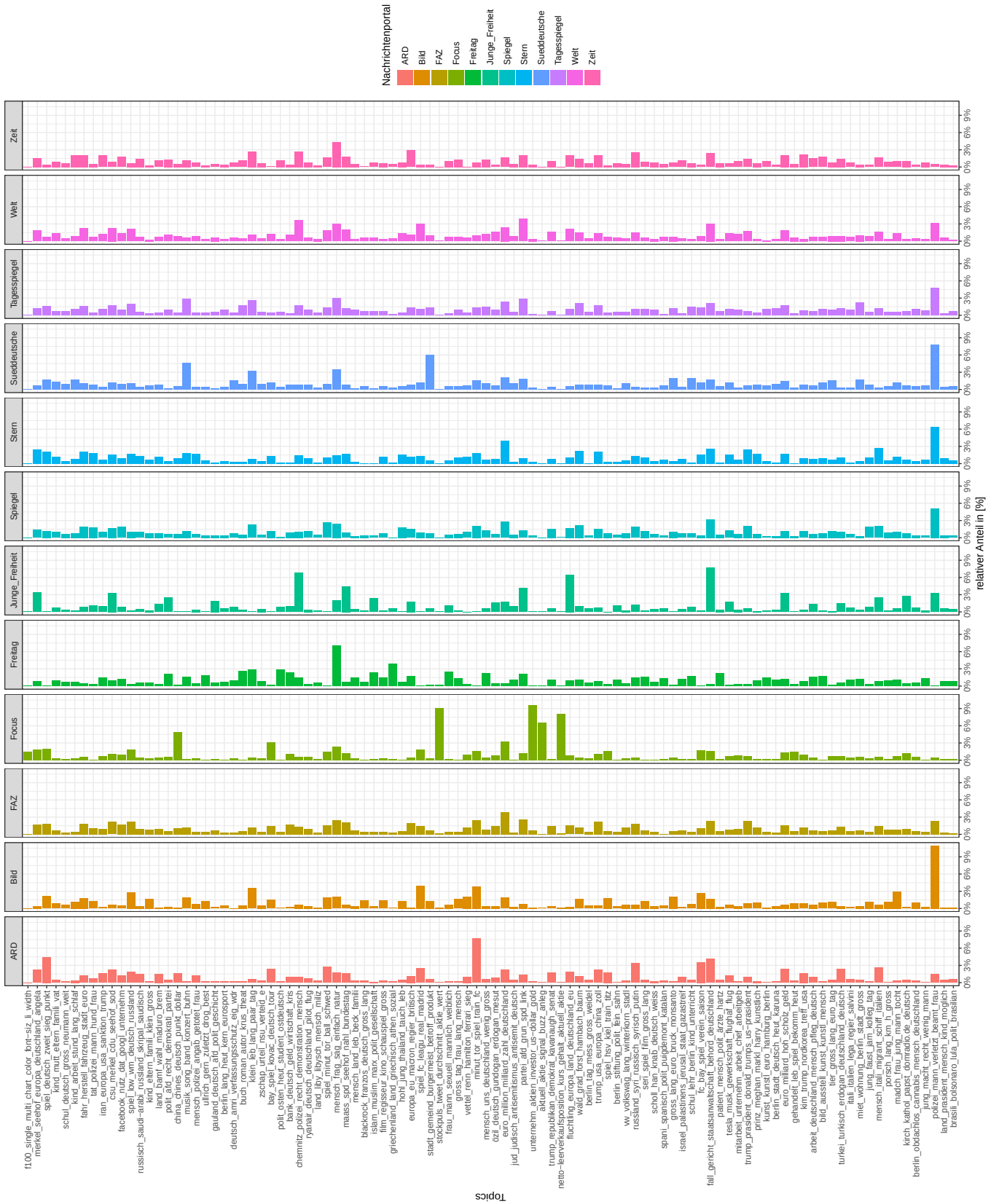
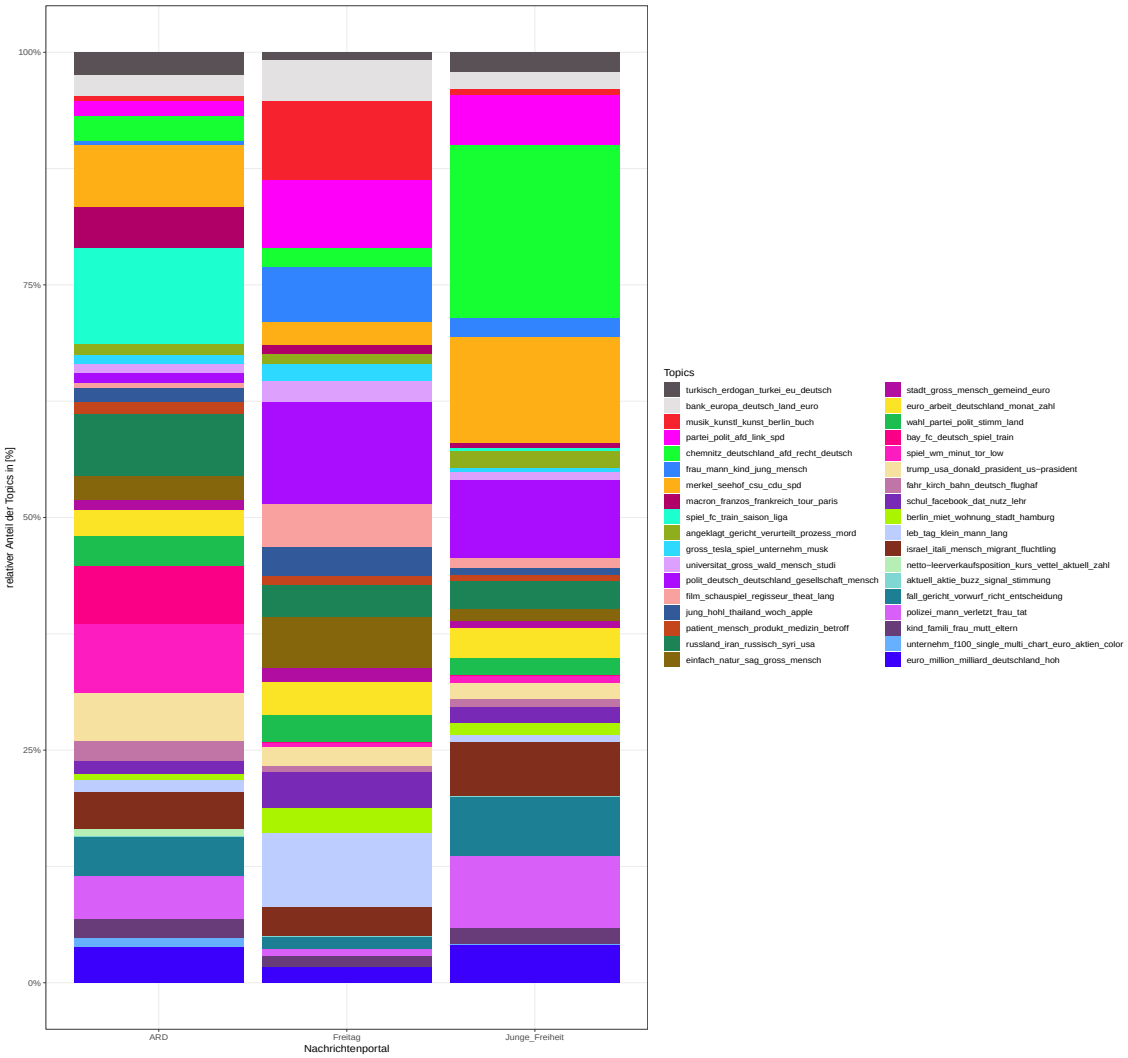


Abbildung 3.6: relative Topicverteilung des LDA Topic Modellings (k=36) für die Nachrichtenportale “ARD”, “Freitag” und “Junge Freiheit”



Interpretation des LDA Topic Modellings mit k = 36 Topics

Im Anschluss an das vorangegangene Verfahren wurde eine weitere Analyse mit einer reduzierten Anzahl an zuvor festgelegten Topics (k = 36) durchgeführt. Entgegen den Erwartungen blieb nicht nur die Konsistenz der einzelnen Wortgruppen weitestgehend erhalten, durch die Reduktion der Topics konnten sogar z.T. redundante Themenfelder eliminiert werden. Ferner konnten die Ergebnisse aufgrund der nun reduzierten Anzahl der Topics in einer angepassten Darstellungsform (s. Abbildung 3.6) visualisiert werden, die den direkten Vergleich der Themenverteilung zwischen den jeweiligen Nachrichtenportalen deutlich erleichtert.
Vergleicht man die relative Themenverteilung des Nachrichtenportals “Junge Freiheit” mit dem vorherigen Analyseergebnis, so kann beobachtet werden, dass bspw.

die beiden Themenfelder “Merkel-Seehofer-Europa-Deutschland-Angela” sowie “CSU-Merkel-CDU-Seehofer-Söder” aufgrund der verringerten Anzahl an Topics zusammengefasst wurden. Das Beispiel soll verdeutlichen, dass eine Reduktion der definierten Anzahl an Topics auch mit einem Informationsverlust einher gehen kann, da der inhaltliche Schwerpunkt der jeweiligen Themenfelder z.T. variieren kann. Dabei ist anzumerken, dass eine Erhöhung der zuvor festgelegten Anzahl an Topics zu einem breiteren Spektrum an detaillierteren Themenfeldern mit einer zunehmend höheren Chance auf Redundanzen führt. Dagegen zieht eine Reduktion der zuvor definierten Anzahl an Themenfeldern zugunsten einer verbesserten Visualisierbarkeit möglicherweise unerwünschte Informationsverluste nach sich. Zur Durchführung einer LDA-Analyse ist es also von essentieller Bedeutung, das richtige Verhältnis (bzw. die Anzahl an Topics) auf den individuellen Fall anzupassen. Insgesamt bleiben solche – in Bezug auf ihren Informationsgehalt – verlustbehafteten Kombinationen im Falle des zugrundeliegenden Analyseverfahrens allerdings die Ausnahme. Vergleicht man bspw. den Inhalt der Themenkomplexe und die jeweilige relative Themenverteilung für das Nachrichtenportal “Junge Freiheit” beider LDA-Resultate (Abbildungen 3.4 und 3.6) so ist festzustellen, dass sowohl die Relationen als auch die inhaltliche Bedeutung der entsprechenden Themenfelder beinahe deckungsgleich sind.
Betrachtet man die relativen Anteile der Nachrichtenportale “ARD”, “Freitag” und “Junge Freiheit”, so ist eine Identifizierung der Themenschwerpunkte der jeweiligen Portale ohne Weiteres möglich. So ist deutlich erkennbar, dass das linksliberale Portal “Der Freitag” einen klaren Fokus auf kultur- und gesellschaftsbezogene Themen (Musik-Kunst-Gesellschaft etc.) legt. Gleichzeitig ist auffällig, dass Themengebiete, die dem Sensationsjournalismus zugeordnet werden können beinahe vollständig entfallen.
Die Themenverteilung der öffentlich-rechtlichen “ÀRD” ist zunächst unauffällig, da dominante Themenfelder auf den ersten Blick nicht ersichtlich sind. Bei näherer Betrachtung ist allerdings der Themenkomplex des (internationalen) Sports (Spiel-WM-Löw) gegenüber den restlichen Nachrichtenportalen überrepräsentiert. Eine mögliche Erklärung für diese Beobachtung könnten die von den öffentlich-rechtlichen Portalen erworbenen Übertragungslizenzen der diesjährigen Fußball-Weltmeisterschaft darstellen.

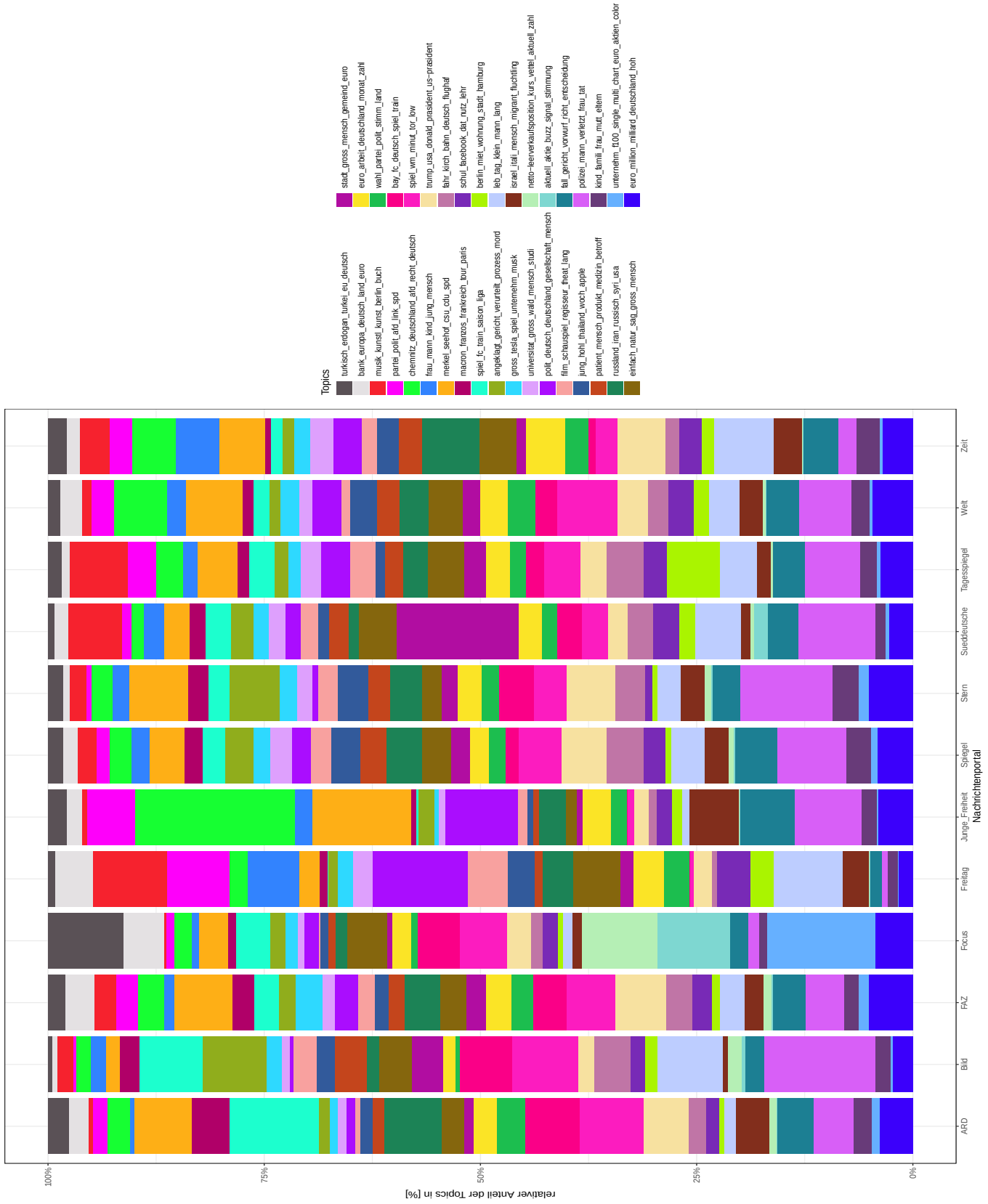


Abbildung 3.7: Ergebnisse des LDA Topic Modellings (k=36)

4 Diskussion

Insgesamt zeigen die Ergebnisse des Forschungsprojekts, dass moderne Verfahren zur Analyse großer Datenmengen (Big Data) durchaus dazu beitragen können, komplexe sozial- und politikwissenschaftliche Fragen zu beantworten. Die Ermittlung der relativen Themenverteilung innerhalb der einzelnen Nachrichtenportale trägt dazu bei, ein genaueres Verständnis über die Agenda der jeweiligen Portale zu erlangen und ist somit hinsichtlich des anfänglich definierten Ziels, sowohl das Aufkommen als auch die Aufbereitungsweise bestimmter Themen in den Massenmedien zu erfassen, als Erfolg zu bewerten.

Die Grundlage erfolgreicher Analysen

Für die erfolgreiche Durchführung der Analyseverfahren war vor allem die Qualität der Rohdaten von besonderer Bedeutung. Insbesondere die vollständige und korrekte Extraktion aller Elemente innerhalb der entsprechenden Artikel hat dazu beigetragen, einen Dokumentkorpus zu erstellen, der den anfangs definierten Anforderungen gerecht wird. Im Zuge der Evaluation stellte sich diesbezüglich heraus, dass sich die Programmierung des eigens für diesen Zweck erstellten Parsers ausgezahlt hatte. Sowohl die einfachen prozentualen Erfassungsraten als auch die anschließende SER-Evaluation machten deutlich, dass der manuell erstellte Parser generischen state-of-the-art Tools in vielerlei Hinsicht überlegen ist. Während bspw. das Tool "news-please" (Hamborg u. a. (2017)) 82% der Titel, 70% der Datumsinformationen und lediglich 62% des Haupttexts überwiegend vollständig erfassen kann, erzielte der für dieses Projekt eingesetzte Parser auch abzüglich weiterer, durch die SER identifizierten Fehlerquellen, weitaus höhere Evaluationswerte (Titel = 97.6%, Datumsinformationen = 100%, Haupttext = 94.6%)[1]. An dieser Stelle ist allerdings anzumerken, dass der Einsatz eines generischen Tools, wie bspw. "news-please", je nach zukünftigen Anwendungsfall trotzdem zu erwägen ist. Insbesondere im Falle von umfassenderen Projekten erfordern die nötigen individuellen Anpassungen des Parser an die einzelnen Seitenformate der Nachrichtenportale einen erheblichen Wartungs- und Ressourcenaufwand.

Die Suche nach Zusammenhängen

Ein weiterer Bestandteil der anfangs definierten Forschungsfrage war das Aufdecken von eventuell auftretenden Korrelationen zwischen dem Aufkommen von Themen in der massenmedialen Agenda und den Umfrageergebnissen des ALLBUS zu dem entsprechenden Themengebiet. Dabei stellte sich die Durchführung der zu diesem Zweck

[1] Die Angaben beziehen sich auf die während der Evaluation ermittelten Messwerte für die ungewichtete SER

erforderlichen Analysen zunächst als problematisch heraus. Zum einen trafen die Ergebnisse der ALLBUS-Umfrage erst gegen Ende der Bearbeitungsdauer dieser Bachelorarbeit ein; zum anderen variierte die Anzahl der durchgeführten Interviews pro Tag z.T. stark, sodass zunächst eine Aggregation auf Wochenebene erforderlich war, um eine relativ stabile Datenbasis zu gewährleisten.
Trotz der o.g. Umstände konnten dennoch erfolgreich Analysen durchgeführt werden, die jedoch keine signifikante Korrelation der beiden Forschungsvariablen nachweisen konnten. Die Umfrageergebnisse des ALLBUS bieten dabei allerdings ein sehr breites Spektrum an verschiedenen Fragestellungen, die eventuell unerwartete Zusammenhänge offenbaren könnten und somit auch weiterhin Gegenstand nachfolgender Forschungsvorhaben sein werden.

LDA Topic Modelling

Gewöhnlich werden die in der massenmedialen Agenda aufkommenden Themen auf Basis reiner Frequenzmessungen des gesuchten Begriffs oder mithilfe von mit Synonymen angereicherten Wortlisten identifiziert (Russell Neuman u. a. (2014)). Dieses Vorgehen hat sich in der Praxis bewährt und bietet zudem in Kombination mit entsprechenden Datumsinformationen die Grundlage, um bspw. Granger-Kausalitäten zwischen der massenmedialen und öffentlichen Agenda zu ermitteln.
Die Ermittlung latenter Themenfelder durch das LDA Topic Modelling stellt hingegen eine alternative bzw. ergänzende Option dar, um die Agenda bestimmter Akteure auf Basis von natürlichsprachigem Text zu untersuchen. Im Gegensatz zu rein frequenzbasierten Verfahren ist es mithilfe der Ergebnisse der LDA-Analysen möglich, gesuchte Begriffe auch im Kontext der mit den einzelnen Topics assoziierten Wörter zu betrachten und darüber hinaus unterschiedliche Verwendungen des gleichen Wortes bei verschiedenen Nachrichtenportalen mithilfe der vorgestellten Visualisierungen auf einen Blick zu erfassen.

Ausblick

Die in Abbildung 3.4 bzw. 3.6 visualisierten Ergebnisse der LDA-Analysen zeigen die vielfältigen Einsatz- und Darstellungsmöglichkeiten, die dieses Verfahren zur Beobachtung bestimmter Themenverteilungen bietet. So könnte bspw. in Abbildung 3.4 ein weiterer Linienverlauf implementiert werden, der den durchschnittlichen relativen Anteil dieses Topics aller Nachrichtenportale darstellt. Negative als auch positive Abweichungen zu diesem Mittelwert wären somit auf einem Blick ersichtlich und könnten die Auswertung insbesondere bei weniger stark vertretenen Themenfeldern erleichtern. Einen weiteren Mehrwert könnte man zudem durch die Darstellung von zeitlichen Verläufen generieren. Ein interessantes Anwendungsszenario wäre bspw. die Darstellung der relativen Themenverteilung bei Bundestagswahlen im zeitlichen Verlauf.
Abschließend muss allerdings darauf hingewiesen werden, dass die Ermittlung von Themenfeldern mithilfe von technischen Methoden, wie in diesem Fall dem LDA Topic Modelling, lediglich eine Annäherung an die tatsächliche Themenverteilung in der Realität darstellen kann.

Literaturverzeichnis

[Baburov 2010] Baburov, Y: *python-readability. Retrieved from https://github.com/buriy/python-readability.* 2010

[Benoit u. Nulty 2016] Benoit, Kenneth ; Nulty, Paul: quanteda: Quantitative analysis of textual data. In: *R package version 0.9* 8 (2016)

[Benoit u. a. 2018] Benoit, Kenneth ; Watanabe, Kohei ; Wang, Haiyan ; Nulty, Paul ; Obeng, Adam ; Müller, Stefan ; Matsuo, Akitaka: Quanteda: An R Package for the Quantitative Analysis of Textual Data. In: *Journal of Open Source Software* 3 (2018), Nr. 30, S. 774

[Blei u. a. 2003] Blei, David M. ; Ng, Andrew Y. ; Jordan, Michael I.: Latent dirichlet allocation. In: *Journal of machine Learning research* 3 (2003), Nr. Jan, S. 993–1022

[Brosius u. Kepplinger 1990] Brosius, Hans B. ; Kepplinger, Hans M.: The Agenda-Setting Function of Television News: Static and Dynamic Views. In: *Communication Research* 17 (1990), Nr. 2, S. 183–211

[Chadwick 2011] Chadwick, Andrew: The political information cycle in a hybrid news system: The British prime minister and the "bullygate" affair. In: *The International Journal of Press/Politics* 16 (2011), Nr. 1, S. 3–29

[Cohen 1963] Cohen, Bernard C.: The press and foreign policy. In: *Princeton, NJ: Princeton University Press. Cohen, J.(1988) Statistical Power Analysis for the Behavioral Sciences* (1963)

[Dreier u. Schulze 2018] Dreier, Dr. T. ; Schulze, Dr. G.: *Dreier/Schulze Urheberrechtsgesetz, 6. Aufl.* 2018

[Druckman u. Nelson 2003] Druckman, James N. ; Nelson, Kjersten R.: Framing and deliberation: How citizens' conversations limit elite influence. In: *American Journal of Political Science* 47 (2003), Nr. 4, S. 729–745

[Entman 1993] Entman, Robert M.: Framing: Toward clarification of a fractured paradigm. In: *Journal of communication* 43 (1993), Nr. 4, S. 51–58

[Furche u. a. 2013] Furche, Tim ; Gottlob, Georg ; Grasso, Giovanni ; Schallhart, Christian ; Sellers, Andrew: OXPath: A language for scalable data extraction, automation, and crawling on the deep web. In: *The VLDB Journal—The International Journal on Very Large Data Bases* 22 (2013), Nr. 1, S. 47–72

[Geva 2016] Geva, R: *article-date-extractor. Retrieved from https://github.com/Webhos/article-date-extractor.* 2016

[Grübler 2018] Grübler, Dr. U.: *BeckOK UrhR/Grübler, 20. Ed.* 2018

[Hagemeier 2018] Hagemeier, Stefanie: *BeckOK UrhR/Hagemeier, 21. Ed.* 2018

[Hamborg u. a. 2017] Hamborg, Felix ; Meuschke, Norman ; Breitinger, Corinna ; Gipp, Bela: news-please: A Generic News Crawler and Extractor. In: *Proceedings of the 15th International Symposium of Information Science*, 2017

[Hindman 2008] Hindman, Matthew: *The myth of digital democracy.* Princeton University Press, 2008

[Hoeren 2018] Hoeren, Prof. Dr. T.: Das Urheberrechts-Wissensgesellschafts-Gesetz. In: *IWRZ 2018*. 120 : Arbeitsgemeinschaft Internationales Wirtschaftsrecht im Deutschen Anwaltverein, 2018

[Klapper 1960] Klapper, Joseph T.: *The Effects of Mass Communication.* Free Press, 1960

[Kohlschütter u. a. 2010] Kohlschütter, Christian ; Fankhauser, Peter ; Nejdl, Wolfgang: Boilerplate detection using shallow text features. In: *Proceedings of the third ACM international conference on Web search and data mining* ACM, 2010, S. 441–450

[Lemnitzer u. Zinsmeister 2006] Lemnitzer, Lothar ; Zinsmeister, Heike: *Korpuslinguistik: Eine Einführung.* Gunter Narr Verlag, 2006

[Makhoul u. a. 1999] Makhoul, John ; Kubala, Francis ; Schwartz, Richard ; Weischedel, Ralph u. a.: Performance measures for information extraction. In: *Proceedings of DARPA broadcast news workshop* Herndon, VA, 1999, S. 249–252

[McCombs u. Shaw 1972] McCombs, Maxwell E. ; Shaw, Donald L.: The agenda-setting function of mass media. In: *Public opinion quarterly* 36 (1972), Nr. 2, S. 176–187

[McKinney 2015] McKinney, Wes: pandas: a Python data analysis library. In: *see http://pandas. pydata. org/. Google Scholar* (2015)

[Munzert u. a. 2014] Munzert, Simon ; Rubba, Christian ; Meissner, Peter ; Nyhuis, Dominic: *Automated data collection with R: A practical guide to web scraping and text mining.* John Wiley & Sons, 2014

[Neumann u. a. 2017] Neumann, Mandy ; Steinberg, Jan ; Schaer, Philipp: Web-Scraping for Non-Programmers: Introducing OXPath for Digital Library Metadata Harvesting. In: *Code4Lib Journal* (2017), Nr. 38

[Ou-Yang 2013] Ou-Yang, L: *Newspaper: Article scraping. Retrieved from http://newspaper.readthedocs.io/en/latest/.* 2013

[Porter 2001] Porter, Martin F.: *Snowball: A language for stemming algorithms.* 2001

[Richardson 2015] Richardson, Leonard: *Beautiful Soup Documentation.* 2015

[Russell Neuman u. a. 2014] Russell Neuman, W ; Guggenheim, Lauren ; Mo Jang, S ; Bae, Soo Y.: The dynamics of public attention: Agenda-setting theory meets big data. In: *Journal of Communication* 64 (2014), Nr. 2, S. 193–214

[Schack 2017] Schack, Haimo: Das neue UrhWissG – Schranken für Unterricht, Wissenschaft und Institutionen. In: *ZUM 2017*. 805 : Prof. Dr. Albrecht Hesse, Prof. Roland Bornemann, Dr. Tilo Gerlach, Prof. Dr. Michael Grünberger, LL.M., Dr. Harald Heker, Prof. Dr. Nadine Klass, LL.M., Prof. Dr. Johannes Kreile, Dr. Urban Pappi, Dr. Robert Staats, 2017

[Schaer u. Neumann 2017] Schaer, Philipp ; Neumann, Mandy: Enriching Existing Test Collections with OXPath. In: *International Conference of the Cross-Language Evaluation Forum for European Languages* Springer, 2017, S. 152–158

[Vogels u. a. 2018] Vogels, Thijs ; Ganea, Octavian-Eugen ; Eickhoff, Carsten: Web2Text: Deep Structured Boilerplate Removal. In: *European Conference on Information Retrieval* Springer, 2018, S. 167–179

[Wang u. a. 2009] Wang, Junfeng ; He, Xiaofei ; Wang, Can ; Pei, Jian ; Bu, Jiajun ; Chen, Chun ; Guan, Ziyu ; Lu, Gang: News article extraction with template-independent wrapper. In: *Proceedings of the 18th international conference on World wide web* ACM, 2009, S. 1085–1086

[Welbers u. a. 2017] Welbers, Kasper ; Van Atteveldt, Wouter ; Benoit, Kenneth: Text analysis in R. In: *Communication Methods and Measures* 11 (2017), Nr. 4, S. 245–265

[Wickham 2016] Wickham, Hadley: Package 'rvest'. In: *URL: https://cran.r-project.org/web/packages/rvest/rvest.pdf* (2016)

Eidesstattliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten Anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben.

Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Köln, 24. November 2018

Philip Ehnert