Clustern gegen Textchaos

Text-Exploration mit computerlinguistischen Methoden

Nasrin Saef

Nasrin Saef
Universität Köln
Institut für Digital Humanities
saefnasrin@gmail.com

*Vortragsmanuskript (gehalten auf der vfm-Frühjahrstagung des vfm am 9. April 2019)

EINLEITUNG

Noch nie war das Anlegen, Verwerfen, Umsortieren und Kopieren von Dokumenten so einfach wie heute.[1] Ein Computer mit Dateiablage und Textverarbeitungsprogramm reicht, wo lange Zeit Papier, Schreibgeräte, Ordner und Regale von Nöten waren. Dokumente wie Rechnungen, Kontoauszüge oder Steuerbescheide werden in digitaler Form verschickt und abgelegt. Briefe werden durch E-Mails ersetzt oder am PC getippt und ausgedruckt. Kurzum: Die Verwaltung von Schriftgut findet in großem Ausmaß digital statt. Das gilt auch für die öffentliche Verwaltung, die Textverarbeitungsprogramme, E-Mails und abteilungsinterne Dateiablagen zur Durchführung ihrer alltäglichen Geschäfte nutzt.

Für spezialisierte Aufgaben existieren dezidierte Softwareanwendungen, und für die Führung aktenrelevanter Unterlagen werden elektronische Dokumentenmanagementsysteme eingeführt. Doch außerhalb dieser kontrollierten Systeme florieren Ordnersysteme voller Dateien, angelegt von Sachbearbeitern ohne Ausbildung in Schriftgutverwaltung. Solche Ablagen wachsen häufig über Jahren und werden dabei groß, unübersichtlich und individuell. Trotz ihrer ungewöhnlichen Form handelt es sich bei ihnen aber um Verwaltungsschriftgut, welches den deutschen Archivgesetzen unterliegt. Archivare müssen also darüber entscheiden, ob Bestandteile der Ablagen auf Dauer im Archiv aufbewahrt und zugänglich gemacht werden sollen. Dafür ist ein Überblick über die in ihnen enthaltenen Dateien und deren Inhalte nötig – eine große Herausforderung, wenn es um tausende bis zehntausende Dokumente geht und keine offensichtliche Struktur zu erkennen ist. Herkömmliche archivische Methoden verlassen sich für das Schaffen dieses Überblicks auf die Gegenwart von Aktenplänen oder ähnlichen strukturierenden Elementen, die es für die digitalen Ablagen in der Regel nicht gibt – und sind somit nicht ohne Weiteres anwendbar.

In der archivarischen Diskussion wird die Bewertung von Dateiablagen erst seit Kurzem behandelt. Auf dem Deutschen Archivtag 2015 behandelte ein Vortrag Dateiablagen als archivische Quellen und besprach auch explizit ihre Bewertung, beschrieb aber vor allem die dabei auftretenden Schwierigkeiten: Die Ablagen seien häufig groß, unstrukturiert und unübersichtlich und entzögen sich herkömmlichen Bewertungsmethoden.[2] Im Jahr darauf erschien der von der Archivschule Marburg herausgegebene Band „Moderne Aktenkunde“, welcher neben analogem Schriftgut des 20. Jahrhunderts auch beispielsweise E-Mails, Datenbanken und Dateiablagen aktenkundlich untersucht.[3] Darin erfolgte die erste ausführliche Charakterisierung von Dateiablagen durch den Wirtschaftsarchivar Schludi.[4] Ebenfalls 2016 fand der Workshop „Kreative digitale Ablagen und die Archive“ der Staatlichen Archive Bayerns statt.[5] Dieser befasste sich ausdrücklich mit Dateiablagen und rückte Strategien und Tools zu ihrer Handhabung ins Zentrum. Allerdings beschränkten sich die vorgestellten Ansätze weitgehend darauf, mit Hilfe von Formatanalysen und Duplikaterkennung die im Anschluss manuell durchzusehenden Bestände zu reduzieren. In der Fachzeitschrift „Der Archivar“ findet sich der erste Beitrag zu dem Thema im Heft vom Juli 2017.[6] Im gleichen Jahr befasste sich ein Vortrag auf dem Deutschen Archivtag mit der Bewertung großer Dateiablagen.[7] Darin beschrieben Miegel und Rödel die Erfahrungen des Hessischen Hauptstaatsarchivs Wiesbaden mit der Bewertung und Übernahme digitaler Dateiablagen und kamen zu dem Schluss, dass Softwarelösungen benötigt werden, um die inhaltliche Bewertung von Dateiablagen effizient durchzuführen.[8]

Eine Möglichkeit zur IT-gestützten inhaltlichen Auswertung digitaler Volltexte ist die Nutzung von Methoden aus den Feldern Computerlinguistik und Information Retrieval. Ziel der Textanalyse ist es, eine Vorstellung vom Inhalt der Ablage zu vermitteln, Vorschläge zu ihrer Strukturierung zu machen und hilfreiche Informationen wie Personennamen zu finden. So soll die Entscheidung über die Archivwürdigkeit von Dokumenten aus Dateiablagen erleichtert werden.

[1] Dieser Beitrag entspricht einer gekürzten und überarbeiteten Fassung meiner 2018 eingereichten Masterarbeit mit dem Titel „Textanalyse mit dem „CollectionExplorer“. Eine Evaluation der Anwendbarkeit computerlinguistischer Methoden zur archivischen Bewertung großer unstrukturierter Textsammlungen“. Die vollständige Arbeit ist zugänglich unter https://kups.ub.uni-koeln.de/9543/.

[2] Vgl. Wendt, Gunnar und Sina Westphal: Eine Herausforderung des Übergangs. In Monika Storm: Transformation ins Digitale. Fulda 2017.

[3] Vgl. Berwinkel, Holger, Robert Kretzschmar und Karsten Uhde: Moderne Aktenkunde. Marburg 2016.

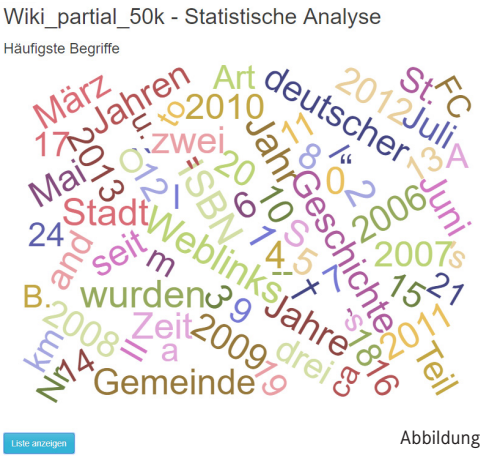


Abbildung 1

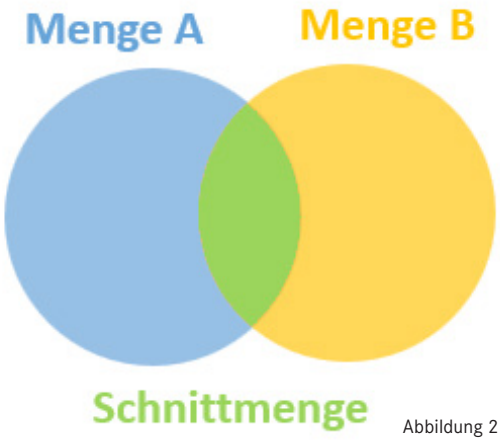


Abbildung 2

Die Anwendbarkeit dieser Methoden beschränkt sich aber nicht auf die Analyse von Textdokumenten, oder auf den Anwendungsfall archivischer Bewertung: Digitalisierte Altbestände, die per OCR erkannt und nie mit ausführlichen Metadaten versehen wurden, können ebenso als Datenquelle genutzt werden wie automatisch erzeugte Transkripte von Audio- und Videodateien. Und die nachfolgend vorgestellten Methoden eignen sich auch als Bausteine eines Recherchetools, für Datenvisualisierungen oder für explorative Zugänge zu noch unbekannten Dateisammlungen.

VERFAHREN ZUR INFORMATIONSGEWINNUNG

Im folgenden Kapitel werden die eingesetzten Verfahren mit ihrem jeweiligen Zweck und dem erwarteten Informationsgewinn kurz vorgestellt. Auf eine detaillierte Beschreibung von Vorverarbeitung und Volltextsuche wird dabei verzichtet, da sie vor allem Grundlagen für die weiteren Analysen sind.

WORTHÄUFIGKEITEN UND N-GRAMME

Das Zählen von Worthäufgkeiten ist ein sehr einfaches Mittel, um einen Eindruck von den in einem Korpus behandelten Themen zu erlangen. Es ist insofern wirkungsvoll, dass die häufigsten Worte einen guten Eindruck der wichtigsten Themen des Bestands vermitteln. Wenn diese aber bereits bekannt sind, werden die häufigsten Worte kaum neue Erkenntnisse generieren. (Vgl. Abbildung 1: Worthäufigkeit)

Ebenfalls analysiert wird die Häufigkeit von N-Grammen. Dabei handelt es sich um Abfolgen von *n* aufeinander folgenden Worten; ein einzelnes Wort ist also ein Unigramm, zwei Worte bilden ein Bigramm, drei ein Trigramm. N-Gramme zeigen typische oder häufige Wortfolgen auf.

Oben wurde die Häufigkeit von Unigrammen diskutiert. Womöglich sind aber auch Bi- und Trigramme für die Ermittlung des Inhalts eines Dokuments oder einer Dokumentensammlung relevant, denn damit finden sich mehrere Worte umspannende Begriffe.[9] Ihre Identifikation wird nach verschiedenen Relevanzkriterien vorgenommen. Das einfachste ist die absolute Vorkommenshäufigkeit. Komplexere Verfahren haben zum Ziel, besonders signifikante N-Gramme zu identifizieren. Signifikant ist ein N-Gramm dann, wenn ein starker Zusammenhang zwischen den Worten im N-Gramm besteht. Dies ist zum Beispiel der Fall, wenn zwei Worte mit sehr hoher Wahrscheinlichkeit gemeinsam auftreten und vergleichsweise selten mit anderen Worten zusammenstehen. Der erwartete Informationsgewinn durch N-Gramme ist ähnlich wie der durch Worthäufigkeiten: Sie werden voraussichtlich die dominanten Themen des Bestands abbilden.

DUPLIKAT- UND VERSIONSERKENNUNG

Ein Problem im Umgang mit Dateiablagen ist die Erkennung von Duplikaten oder Versionen der gleichen Datei. Einmal erkannt können Duplikate ausgeblendet und Versionen des gleichen Dokuments miteinander verknüpft werden, um die finale Fassung eines Dokuments zu identifizieren. (Vgl. Abbildung 2: Jaccard-Index)

Zur Erkennung von Versionen und Duplikaten wird der Jaccard-Index verwendet. Dieser beschreibt die Ähnlichkeit zweier Mengen A und B, indem ihre Schnittmenge durch die Gesamtmenge von A und B geteilt wird. Ein Wert von 1 beschreibt Duplikate, ein Wert von 0 zwei Mengen ohne Schnittmenge. Um die Ähnlichkeit von Dokumenten zu berechnen, können Hashwerte der die in ihnen vorkommenden Bigramme als die zu vergleichenden Mengen verwendet werden.[10] Dieses Verfahren hat gegenüber einem einfachen Stringvergleich den Vorteil, dass auch Texte mit kleinen Textänderungen oder geänderter Reihenfolge einiger Satzbausteine als Versionen voneinander erkannt werden. Ähnliche Technologien werden zum Beispiel in der Plagiatserkennung ein-

[4] Vgl. Schludi, Ulrich: Das Schriftgut der Wirtschaft. In Holger Berwinkel, Robert Kretzschmar, Karsten Uhde: Moderne Aktenkunde. Marburg 2016.

[5] Vgl. Puchta, Michael und Kai Naumann: Kreative digitale Ablagen und die Archive. München 2017.

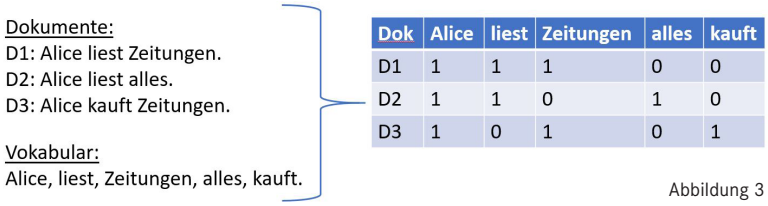
[6] Vgl. Jaeger, Karina und Maria Kobold: Zwischen Datenwust und arbeitsökonomischer Bewertung. In Der Archivar.

[7] Vgl. Miegel, Annekathrin und Eva Rödel: Wege aus dem Daten-Dschungel – Bewertung und Übernahme großer Dateisammlungen. In Klara Deecke, Ewald Grothe: Massenakten - Massendaten. Fulda 2018.

[8] Vgl. ebd., S. 35.

[9] Auf N-Gramme mit n > 3 wird hier verzichtet, da nicht nur die benötigte Rechenleistung mit höheren Werten deutlich steigt, sondern zugleich die Aussagekraft der N-Gramme sinkt. Je länger ein N-Gramm ist, desto seltener werden Wiederholungen davon gefunden werden, was zu immer weniger verwertbaren Ergebnissen führt.

[10] Vgl. Liu, Bing: Web Data Mining. Berlin, Heidelberg 2011, S. 231 f.



Dok	Alice	liest	Zeitungen	alles	kauft
D1	1	1	1	0	0
D2	1	1	0	1	0
D3	1	0	1	0	1

Abbildung 3

gesetzt. Das Verfahren funktioniert vor allem dann gut, wenn es sich um Versionen mit ähnlicher Länge handelt. Bei Versionen stark unterschiedlicher Länge ist die Schnittmenge zu klein, um die Übereinstimmung zu erkennen.[11]

NAMED ENTITIY RECOGNITION

Named Entitiy Recognition (NER) behandelt die Erkennung von Eigennamen. Den erkannten Named Entitys wird eine Kategorie zugewiesen, beispielsweise Person, Ort, Organisation oder Weiteres. Für eine genaue Zählung der Entitäten ist das Verfahren nicht geeignet, da seine Genauigkeit (ohne eigens für den Datenbestand trainierte Modelle) nicht hoch genug ist. Außerdem findet keine Disambiguierung gleichnamiger Named Entitys statt, und die Erkennung zusammen gehöriger Vor- und Nachnamen ist teils schwierig. Die Auflistung der Entitäten ist also nicht präzise, sollte aber einen Eindruck der häufig vorkommenden Namen, Personen und Organisationen vermitteln.

CLUSTERN VON TEXTDOKUMENTEN

Eine Möglichkeit, die in einem Korpus enthaltenen Dokumente thematisch zu gruppieren, ist Clustering. Dabei handelt es sich um ein Verfahren des unüberwachten maschinellen Lernens, welches Daten nach einem gewählten Ähnlichkeitsmaß sortiert.

UNÜBERWACHTES MASCHINELLES LERNEN MIT DOKUMENTVEKTOREN

Beim unüberwachten maschinellen Lernen wird ohne manuell annotierte Trainingsdaten gearbeitet. Die Bearbeiterin wählt einen Algorithmus und legt dessen Parameter fest, aber die Verteilung der Datensätze auf Gruppen erfolgt ohne ihr Zutun; stattdessen sortiert der Algorithmus die Datensätze automatisch. Bei diesem Vorgehen werden zum Erzielen guter Ergebnisse sehr viele Daten benötigt.

Entscheidend für die Gruppierung ist das Festlegen von Kriterien, anhand derer die Ähnlichkeit ausgemacht wird. Der bloße Dokumenttext taugt dafür nicht, denn er besteht aus einer willkürlichen Abfolge von Zeichen, die keine semantischen Eigenschaften mitbringen. Stattdessen benötigen die Algorithmen eine Zahlenrepräsentation der Texte. Dafür wird zum Beispiel jedem in der Dokumentensammlung vorkommenden Wort ein fortlaufender Schlüssel zugewiesen. Anschließend wird eine Tabelle erstellt, deren Spalten dem Vokabular der Sammlung entsprechen. Für jedes Dokument wird eine eigene Zeile erzeugt und für alle im Dokument vorkommenden Worte ein Wert gesetzt. Das kann eine Eins sein, um das Vorkommen anzuzeigen, oder die Häufigkeit des Worts. (Vgl. Abbildung 3: Eine einfache Dokumentenmatrix)

Die sich so ergebenden Zeilen können auch als Dokumentvektoren betrachtet werden. Vektoren haben den Vorteil, dass sie im Raum angeordnet und über Distanzmaße wie die Euklidische Distanz miteinander verglichen werden können. So kann der Grad der Ähnlichkeit zwischen den Vektoren, und demzufolge auch den Dokumenten, über ihre Distanz im Vektorraum bestimmt werden. Alle in den folgenden Abschnitten vorgestellten Verfahren basieren auf diesem Grundprinzip.

TOPIC MODELLING

Ziel des Topic Modellings ist es, ein Korpus von Dokumenten auf in ihnen vorkommende sprachliche Muster zu untersuchen (sogenannte Topics). Es erfreut sich vor allem unter Literaturwissenschaftlern und Linguisten steigender Beliebtheit, wird aber auch in anderen Feldern vermehrt eingesetzt.[12] So hat beispielsweise die Zeitschrift Signs ihr Archiv mit dem Verfahren aufbereitet.[13] Eine der gängigsten Methoden für Topic Modelling ist Latent Dirichlet Allocation (LDA)[14].

LDA ist eine probabilistische Methode, die latente Muster in den Daten aufdecken soll.[15] Es handelt sich um ein generatives Modell, das zunächst eine Annahme über die Entstehung von Dokumenten trifft: Es existiere eine Menge von Topics, denen verschiedene Wörter mit unterschiedlicher Häufigkeit zugeordnet seien. Jedes Dokument werde je zu einem bestimmten Prozentsatz aus Worten dieser Topics erzeugt. Die Zusammensetzung der Topics könne algorithmisch ermittelt und somit für jedes Dokument bestimmt werden, welches Topic in ihm besonders stark vertreten sei. Die Topics werden über ihre signifikantesten Worte beschrieben, wodurch sie – untypischerweise, für ein Verfahren des unüberwachten maschinellen Lernens – eine Art Label oder Kurzbeschreibung erhalten.

Topic Modelling funktioniert ohne manuelle Annotation, der Bearbeiter muss nur die Zahl der Topics, die ermittelt werden sollen, festlegen. Sie werden über die für sie signifikantesten Worte beschrieben, ihre Interpretation muss der Bearbeiter selbst vornehmen.[16]

[11] Dieses Problem kann mit der Nutzung anderer Methoden umgangen werden. Beispielsweise kann die Schnittmenge durch die Größe der kleineren Menge geteilt werden, um mit unterschiedlichen Dokumentenlängen umzugehen.

[12] Vgl. Fechner, Martin und Andreas Weiß: Einsatz von Topic Modeling in den Geschichtswissenschaften. In Zeitschrift für digitale Geisteswissenschaften, S. 2.

[13] Vgl. Goldstone, Andrew, Susana Galán, C. Laura Lovin et al.: An Interactive Topic Model of Signs.

[14] Vgl. Blei, David M., Andrew Y. Ng und Michael I. Jordan: Latent Dirichlet Allocation. In Journal of Machine Learning Research. Die folgenden Erläuterungen zu LDA nach Blei 2012.

[15] Vgl. ebd., S. 78 f.

[16] Ausführliche Beispiele für die so erzeugten Ergebnisse – auch für die nachfolgend beschriebenen Verfahren – finden sich im Anhang der vollständigen Fassung der Masterarbeit.

TF-IDF

Die Abkürzung Tf-idf steht für „Termfrequenz – inverse Dokumentfrequenz“ und hat zum Ziel, das für ein Dokument spezifische Vokabular zu ermitteln. Tf-idf legt die Annahme zugrunde, dass Worte, die in allen Dokumenten eines Korpus‘ häufig vorkommen, nicht spezifisch für den Inhalt eines bestimmten Dokuments sind. Kommt ein Wort dagegen im Gesamtkorpus selten vor, aber in einem Dokument sehr häufig, ist es vermutlich für dieses Dokument sehr aussagekräftig.

In vorliegendem Fall wird eine Matrix aus den Identifikatoren aller Dokumente in der einen Dimension und den Td-idf-Repräsentationen der in ihnen vorkommenden Worte in der anderen gebildet. Wenn diese im Vektorraum angeordnet werden, sind Dokumente, die ein ähnliches Vokabular verwenden, nah beieinander. Tf-idf ist daher besonders gut für die Identifikation serieller Quellen geeignet. Auch bei diesem Verfahren müssen die Festlegung der Cluster-Zahl und die Interpretation der Ergebnisse durch den Bearbeiter erfolgen; Labels gibt es hier nicht, der Bearbeiter muss die Themen der Cluster selbst ermitteln.

SEMANTISCHE CLUSTER MIT DOC2VEC

Ein weiterer Versuch, eine semantische Dimension in den Textdaten zu erfassen, wird mit den Verfahren Word2Vec[17] und Doc2Vec[18] unternommen. Word2Vec nutzt ein flaches neuronales Netz, das die Wahrscheinlichkeit maximiert, ein Wort basierend auf seinem Kontext vorherzusagen. Aus diesen Vorhersagen lernt es ein Sprachmodell, das über Wortvektoren abgebildet wird.

Es benötigt große Mengen an Trainingsdaten, um sinnvolle Ergebnisse zu liefern: Der ursprüngliche Entwickler Mikolov und seine Kollegen haben ein Sprachmodell mit einem Google-News-Korpus aus hundert Milliarden Tokens trainiert. Es war unter anderem in der Lage, semantische Beziehungen wie die zwischen einem Land und seiner Hauptstadt abzubilden.[19] Die Herstellung solcher Zusammenhänge anhand von Wortkontexten macht Word2Vec für die vorliegende Problemstellung interessant. Auch wenn die Wortvektoren hauptsächlich ein Nebenprodukt der Erzeugung von Dokumentvektoren sind, können sie beispielsweise die Volltextsuche unterstützen, indem die semantisch ähnlichsten Worte aus dem Vokabular als ähnliche Suchbegriffe vorgeschlagen werden. Es kann außerdem überprüft werden, wie das Suchwort im vorliegenden Korpus verwendet wird.

Mit Doc2Vec werden semantische Dokumentvektoren[20] erzeugt, die zusätzlich zu den Wortvektoren auch den bisherigen Text des aktuell betrachteten Dokuments abbilden. Um sie zu erzeugen, wird für die Vorhersage des nächsten Worts der gesamte bisherige Inhalt des Dokuments genutzt und anschließend der Dokumentvektor mit den neuen Informationen aktualisiert. Der Output von Doc2Vec ist eine Matrix von Dokumentvektoren, die als Input für weitere Anwendungen des maschinellen Lernens wie beispielsweise Clustering genutzt werden kann.[21] Die so erzeugten Cluster unterliegen den gleichen Bedingungen wie bei der Nutzung von Tf-idf.

EVALUATION DES INFORMATIONSGEWINNS DURCH DIE COMPUTERLINGUISTISCHEN VERFAHREN

Nachfolgend sollen alle eingesetzten Verfahren daraufhin evaluiert werden, ob und unter welchen Bedingungen durch sie ein Informationsgewinn erzeugt werden kann. Dafür wurden sie an einem Testbestand aus 50.000 zufällig ausgewählten deutschen Wikipedia-Artikeln sowie an vier sich in ihrer Größe, Zusammensetzung und Sprache unterscheidenden Beständen des Hessischen Hauptstaatsarchivs Wiesbaden getestet.

Die Vorverarbeitung und die Volltextsuche funktionieren problemlos und unterscheiden sich nicht nach Bestandsgröße oder -sprache.

Die Dokumentähnlichkeit über den Jaccard-Index identifiziert Versionen in Stichproben zuverlässig. Problematisch beim Auffinden von Versionen und Duplikaten ist die Performance – bei großen Beständen ist die Verarbeitungsdauer sehr hoch.

Die Methoden zur statistischen Auswertung eines Bestands – Worthäufigkeiten, N-Gramme und NER – funktionieren unabhängig von der Bestandsgröße, sie sind allerdings von der Eingabesprache abhängig. Auch wenn die Sprache stimmt und die Verarbeitung funktioniert, ist der Erkenntnisgewinn durch Worthäufigkeiten und N-Gramme eingeschränkt. Denn in der Regel liegen genug Informationen über einen Bestand vor, um die wichtigsten Themen im Vorhinein einschätzen zu können – und zu darüber hinausgehenden Erkenntnissen führen die Worthäufigkeiten meist nicht.[22] Der geringe Erkenntnisgewinn durch N-Gramme erklärt sich aus der Art ihrer Ermittlung. Die häufigsten N-Gramme verhalten sich wie die häufigsten Worte. Die nach Signifikanz ermittelten N-Gramme finden Wortfolgen, die sehr viel wahrscheinlicher gemeinsam auftreten als mit anderen Worten. Dieses Vorgehen bevorteilt Fachbegriffe, fremdsprachige und unübliche Worte, die nicht Teil des normalen Sprachgebrauchs sind, sondern (fast) nur gemeinsam auftreten. Diese Methoden haben in allen Testfällen bestehende Annahmen über den Inhalt des Bestands gestützt, ohne signifikante neue Informationen beizutragen. Die Auflistung von Named Entitys dagegen scheint hilfreich, um die prominentesten Akteure im Bestand zu identifizieren.

[17] Vgl. Mikolov, Tomas und Quoc Le: Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning.

[18] Vgl. Mikolov, Tomas, Kai Chen, Greg Corrado et al.: Efficient Estimation of Word Representations in Vector Space.

[19] Vgl. ebd., S. 5–7.

[20] Mikolov spricht in seinem Paper von „paragraph vectors“. Grundsätzlich kann ein paragraph eine Einheit von wenigen Sätzen bis hin zu einem ganzen Dokument umfassen (vgl. Mikolov, Tomas und Quoc Le: Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, S. 1).

[21] Vgl. Mikolov und Le, S. 3.

[22] Auf Dokumentebene sind sie interessanter – denn während der Inhalt der gesamten Sammlung ungefähr bekannt ist, ist das nicht zwingend für die einzelnen darin enthaltenen Dokumente der Fall. Dort kann eine Word Cloud mit den häufigsten Begriffen schnell Aufschluss über die grobe inhaltliche Einordnung des Texts geben.

Die semantische Auswertung – Doc2Vec-Cluster, Topic Models und Tf-idf-Cluster – benötigt für die Erzeugung sinnvoller Ergebnisse eine Mindestmenge an Dokumenten. Ab circa 5.000 Dokumenten hat dies in den Tests für Topic Modelling und Tf-idf-Cluster gut funktioniert, die Bestände werden in (größtenteils) sinnvolle Gruppen eingeteilt. Dabei fällt auf, dass Topic Models die Dokumente eher nach thematisch-inhaltlichen Gesichtspunkten gruppieren, Tf-idf-Cluster dagegen serielle Dokumente verlässlich gemeinsam einordnen. Alles in allem sind die Parallelen zwischen den durch die beiden Methoden gebildeten Gruppen aber groß.

Die Tests mit Doc2Vec waren dagegen nur teilweise erfolgreich. Verweise auf ähnliche Dokumente werden (ebenfalls ab etwa 5.000 Dokumenten) zuverlässig gesetzt, meist besteht zwischen zwei so verknüpften Dokumenten ein inhaltlicher Zusammenhang. Die Analyse des Vokabulars liefert ebenfalls passende Wortähnlichkeiten. Clustern mit Doc2Vec hat dagegen keine sinnvollen Gruppierungen erzeugt. Möglicherweise lag dies an zu geringen Textmengen in drei der vier Testbestände. Aber auch Tests mit dem größten Bestand, der rund 150.000 Dokumente umfasst, sind gescheitert. Allerdings nicht (nur) an wenig sinnvollen Gruppierungen, sondern auch an der Menge der Dokumente. Grundsätzlich gilt: je mehr Daten, desto besser. Die Cluster müssen allerdings von einem Menschen daraufhin durchgesehen werden, ob zwischen den Dokumenten ein thematischer Zusammenhang besteht. Dies ist bereits mit wenigen tausend Dokumenten pro Cluster nicht immer einfach. Bei dutzenden Clustern mit jeweils zehntausenden Dokumenten kann ein Mensch kaum eine Übersicht über den Inhalt jedes Clusters erlangen. So große Bestände sollten vor dem Clustern durch beispielsweise den Ausschluss von Duplikaten, nicht-führenden Versionen und wenig interessanten Unterordnern reduziert werden, vielleicht wäre auch ein Umstieg auf überwachtes maschinelles Lernen sinnvoll.

Zusammenfassend kann festgestellt werden, dass die Methoden ab circa 5.000 Dokumenten, und sofern die Texte zum größten Teil auf Deutsch verfasst sind, sinnvoll eingesetzt werden können. Die statistischen Methoden können zwar auch für kleinere Bestände verwendet werden, geben aber insgesamt weniger Aufschluss über den Bestand als die semantischen Methoden.

Bei sehr großen Beständen mit über 100.000 Dokumenten sind die Clustering-Verfahren, die bei den mittelgroßen Beständen die meisten Erkenntnisse über den Inhalt der Ablage hervorgebracht haben, nicht mehr praktikabel. Um sie zu bearbeiten, müssen mindestens Anpassungen an den Bestand und eine Reduktion der Gesamtmenge vorgenommen werden.

PERSPEKTIVEN UND POTENTIALE

Dieser Beitrag konzentriert sich auf das Problem der Bewertung von Dateiablagen, für die vorgestellten Methoden sind aber auch andere Einsatzgebiete denkbar. Die Erschließung von Archivalien erfolgt auf Grundlage von Metadaten und dem Überfliegen des Volltexts eines Dokuments. Eine zusammenfassende Ansicht mit den häufigsten Wörtern und den signifikantesten Begriffen nach Tf-idf wäre für die schnelle inhaltliche Einordnung eines Dokuments hilfreich. Named Entitys würden sofort Informationen darüber liefern, ob und welche Personen im Dokument genannt werden und kombiniert mit Suchfunktionen eine schnelle Prüfung von Schutzfristen oder Verschlagwortung mit Personennamen ermöglichen. Die Verknüpfung von Duplikaten und Versionen kann helfen, Redundanzen zu vermeiden.

Für Archivnutzer wäre eine Anwendung im Stil des CollectionExplorers als Ergänzung zu herkömmlichen Datenbanken denkbar. Zum einen bietet sie mit Volltextsuche, der Suche nach Named Entitys und dem Verlinken ähnlicher Dokumente komfortable Funktionalitäten, die die Recherche vereinfachen würden. Weiterhin könnte sie als Ersatz dienen, wenn für Archivmaterial noch keine Erschließungsdaten vorhanden sind.

An vielen Stellen hätte die Anwendung außerdem noch verbessert oder ausgebaut werden können. Es wurden nur originär digitale Textdokumente in die Verarbeitung einbezogen. Möglich wäre auch die Nutzung von über Speech Recognition transkribierten Tondokumenten, und mittels OCR erkannten Scans. Dies würde allerdings einen Schritt zur Qualitätssicherung oder zumindest eine automatische Bewertung der Erkennungsqualität benötigen. Weitere noch einzubeziehende Quellen wären (in großem Maß in den Testbeständen vorkommende) Bild- und Videodateien, die beispielsweise automatisch beschrieben oder nach Ähnlichkeit der Pixel oder der abgebildeten Objekte gruppiert werden könnten. Die Metadaten der verarbeiteten Dateien wurden weitgehend ignoriert, da sie zu inkonsistent waren; ein Versuch, die Dateien auf Basis von Informationen wie ihrem Autor oder Entstehungszeitraum zu vernetzen, könnte aber unternommen werden.

FAZIT

Es wurde untersucht, ob computerlinguistische Methoden dafür geeignet sind, sich einen Überblick übergroße, unstrukturierte Dokumentensammlungen zu verschaffen. Diese Fragestellung wurde vorrangig im Kontext archivischer Bewertung betrachtet, da immer mehr Archive mit solchen Sammlungen konfrontiert werden, ohne bisher eine Methodik für den Umgang mit ihnen entwickelt zu haben.

Der für die Bewertung nötige Überblick über die Inhalte einer solchen Ablage soll durch automatische Auswertung mit Hilfe computerlinguistischer Methoden geschaffen werden. Dafür wird eine Reihe von Methoden herangezogen. Eine Volltextsuche macht die Texte der Dokumente durchsuchbar und fungiert als Bindeglied zwischen den Funktionalitäten, um beispielsweise erkannte N-Gramme oder Entitäten in Dokumenten zu finden. Über die Schnittmenge von Bigrammen werden Dokumente identifiziert, die möglicherweise Versionen voneinander sind. Worthäufigkeiten und N-Gramme werden herangezogen, um die Sammlung dominierende Themen zu erkennen. Mit Hilfe von Named Entity Recognition werden Personen, Orte und Organisationen in den Dokumenten gefunden. Topic Modelling, Tf-idf und Doc2Vec werden eingesetzt, um die Dokumente nach inhaltlich-textlichen Aspekten zu gruppieren. Außerdem können über Doc2Vec das Vokabular der Sammlung untersucht und einander ähnelnde Dokumente gefunden werden.

Der Einsatz einiger der vorgestellten Methoden wäre auch für weitere Szenarien im archivischen Arbeitsprozess denkbar. So könnte beispielweise die Erschließung durch Analyse des zu bearbeitenden Dokuments unterstützt oder den Archivbenutzern das Browsing noch nicht erschlossener Unterlagen ermöglicht werden. Außerdem können auch weitere Quellenarten, wie Transkripte audiovisueller Quellen, mit den beschriebenen Methoden verarbeitet werden.

Kann nun mit computerlinguistischen Mitteln ein Überblick über große unsortierte Dateiablagen gewonnen werden? Der Eindruck aus den Testanalysen legt nahe, dass dies zu einem gewissen Grad funktioniert. Von einer ohne Weiteres nutzbaren automatischen Sortierung kann nicht gesprochen werden, dafür sind die Ergebnisse zu unzuverlässig. Aber ein menschlicher Bearbeiter kann durch die oben genannten Methoden, vor allem durch die Topic Models und die Tf-idf-Cluster, einen Eindruck von der inhaltlichen Struktur der Dateisammlung bekommen und über NER prominente Personen, Orte und Organisationen identifizieren. Dafür müssen bestimmte Voraussetzungen erfüllt sein: Die Bestandsgröße muss innerhalb eines bestimmten Fensters liegen und eine Anpassung an seine (Haupt-)Sprache muss vorgenommen werden. Die Ergebnisse davon sind nur eine Annäherung an den Inhalt. Aber eine Annäherung ist besser als der Status Quo: unüberschaubare Ordnersysteme ohne jegliche Dokumentation. Eine Anwendung im Stile des CollectionExplorers würde auch mit mehr Entwicklungszeit dem Archivar nicht die Strukturierung des Bestands abnehmen oder eine verlässliche Zusammenfassung aller Inhalte liefern. Sie sollte aber dazu taugen, den Dschungel unbekannter Dokumente ein wenig zu lichten und eine Grundlage für ihre weitere Sortierung und Einordnung zu schaffen.

LITERATURVERZEICHNIS

Berwinkel, Holger, Robert Kretzschmar und Karsten Uhde (Hrsg.): Moderne Aktenkunde [Veröffentlichungen der Archivschule Marburg 64]. Marburg 2016.

Blei, David M.: Probabilistic topic models. In *Communications of the ACM* 55 (2012) H. 4, S. 77–84. DOI: 10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng und Michael I. Jordan: Latent Dirichlet Allocation. In *Journal of Machine Learning Research (*2003) H. 3, S. 993–1022. URL: *www.jmlr.org/papers/volume3/blei03a/blei-03a.pdf*.

Deecke, Klara und Ewald Grothe (Hrsg.): Massenakten - Massendaten. Rationalisierung und Automatisierung im Archiv [Tagungsdokumentationen zum Deutschen Archivtag 22]. Fulda 2018.

Fechner, Martin und Andreas Weiß: Einsatz von Topic Modeling in den Geschichtswissenschaften. Wissensbestände des 19. Jahrhunderts. In *Zeitschrift für digitale Geisteswissenschaften (*2017), ohne Paginierung. DOI: 10.17175/2017_005.

Goldstone, Andrew, Susana Galán, C. Laura Lovin, Andrew Mazzaschi und Lindsey Whitmore: An Interactive Topic Model of Signs. Edited by Andrew Goldstone 2014 [Signs at 40]. URL: *signsat40.signsjournal.org/topic-model (*20.04.2019).

Jaeger, Karina und Maria Kobold: Zwischen Datenwust und arbeitsökonomischer Bewertung. Ein Werkstattbericht zum Umgang mit unstrukturierten Dateisammlungen am Beispiel des Bestandes der Odenwaldschule. In *Der Archivar* 70 (2017) H. 3, S. 307–311.

Liu, Bing: Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data [Data-Centric Systems and Applications]. Berlin, Heidelberg 2011. URL: *dx.doi.org/10.1007/978-3-642-19460-3*.

Miegel, Annekathrin und Eva Rödel: Wege aus dem Daten-Dschungel – Bewertung und Übernahme großer Dateisammlungen. In Klara Deecke, Ewald Grothe (Hrsg.): *Massenakten - Massendaten.* Rationalisierung und Automatisierung im Archiv [Tagungsdokumentationen zum Deutschen Archivtag 22]. Fulda 2018, S. 27–36.

Mikolov, Tomas, Kai Chen, Greg Corrado und Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. o. O. 2013. URL: *arxiv.org/pdf/1301.3781v3*.

Mikolov, Tomas und Quoc Le: Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning* 32 (2014) H. 2, S. 1188–1196.

Puchta, Michael und Kai Naumann (Hrsg.): Kreative digitale Ablagen und die Archive. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23.11.2016 in der Generaldirektion der Staatlichen Archive Bayerns [Sonderveröffentlichungen der Staatlichen Archive Bayerns 13]. München 2017.

Saef, Nasrin: Textanalyse mit dem „CollectionExplorer“. Eine Evaluation der Anwendbarkeit computerlinguistischer Methoden zur archivischen Bewertung großer unstrukturierter Textsammlungen. Köln 2018.

Schludi, Ulrich: Das Schriftgut der Wirtschaft. In Holger Berwinkel, Robert Kretzschmar, Karsten Uhde (Hrsg.): *Moderne Aktenkunde* [Veröffentlichungen der Archivschule Marburg 64]. Marburg 2016, S. 93–108.

Storm, Monika (Hrsg.): Transformation ins Digitale. 85. Deutscher Archivtag in Karlsruhe [Tagungsdokumentationen zum Deutschen Archivtag 20]. Fulda 2017.

Wendt, Gunnar und Sina Westphal: Eine Herausforderung des Übergangs. Fileablagen als Quellen der digitalen Überlieferungsbildung. In Monika Storm (Hrsg.): *Transformation ins Digitale.* 85. Deutscher Archivtag in Karlsruhe [Tagungsdokumentationen zum Deutschen Archivtag 20]. Fulda 2017, S. 105–114.