Mönche, Schnee und Algorithmen – Eine Anwendung von Topic Modeling auf die Wetterbeobachtungen des Einsiedler Paters Joseph Dietrich (1645-1704)

Lukas Heinzmann

Im vorliegenden Beitrag wird exemplarisch die Eignung computergestützter Textanalyseverfahren für die Untersuchung frühneuzeitlicher Texte diskutiert. Zu diesem Zweck wurden Pater Joseph Dietrichs (1645-1704) Wetterbeobachtungen mit Hilfe von Topic Modeling analysiert. In einem ersten Schritt erfolgte die Konzeption eines geeigneten Workflows. Eine Besonderheit der gewählten Vorgehensweise bestand darin, dass die Anwendung des Algorithmus mit unterschiedlichen Arten der Segmentierung vorgenommen wurde. Dadurch konnten in einem zweiten Schritt sowohl inhaltliche Muster als auch orthografische und stilistische Veränderungen über längere Zeiträume herausgearbeitet werden. Die Ergebnisse bilden eine wesentliche Grundlage für die Entschlüsselung der Schreib- und Arbeitsweise des Autors sowie für das Verständnis seiner Naturwahrnehmung.

Cette contribution se penche sur la pertinence des méthodes d'analyse de texte assistées par ordinateur en prenant pour exemple des textes du début de l'époque moderne. Dans ce but, les observations météorologiques du Père Joseph Dietrich (1645-1704) ont été analysées à l'aide de la modélisation thématique (Topic Modeling). La première étape a consisté à concevoir un flux de travail approprié. Une particularité de la procédure choisie a été d'appliquer l'algorithme avec différents types de segmentations. Cela a permis, dans un deuxième temps, de mettre en évidence aussi bien des modèles de contenu que des changements orthographiques et stylistiques sur le long terme. Les résultats constituent une base essentielle pour le décryptage de la méthode d'écriture et de travail de l'auteur ainsi que pour la compréhension de sa perception de la nature.

This paper examines the relevance of computer-assisted text analysis methods, using texts from the early modern period as an example. To this end, the meteorological observations of Father Joseph Dietrich (1645-1704) were analysed using Topic

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.16

Modeling. The first step was to design an appropriate workflow. A particular feature of the procedure chosen was to apply the algorithm with different types of segmentations. This enabled us to identify both content patterns as well as orthographic and stylistic changes over longer periods of time. The results provide an essential basis for deciphering the author's method of writing and working, and for understanding his perception of nature.

1 Einleitung

Für Geisteswissenschaftlerinnen und Geisteswissenschaftler ist es selbstverständlich, sich bei der Suche nach Fachliteratur und Forschungsmaterial auf die Resultate von Services zu verlassen, die unter anderem auf Grundlage textbasierter algorithmischer Berechnungen zustande kommen. Mit Ausnahme von wenigen Spezialgebieten kommen für die weiteren Arbeitsschritte in der Regel jedoch überwiegend erprobte Methoden der jeweiligen Disziplin zur Anwendung. Insbesondere Textanalyseverfahren bergen jedoch ein grosses Potenzial für innovative Ansätze und werden im Hinblick auf die gewaltige Menge an verfügbaren Textmaterials, die im Kontext rein numerischer Produktion und systematischer Retrodigitalisierung exponentiell zunimmt, in Zukunft erheblich an Bedeutung gewinnen. Eine zielführende Adaption bedingt jedoch, dass sich die einzelnen Disziplinen intensiver mit den Potenzialen, Einsatzmöglichkeiten und Grenzen digitaler Methoden auseinandersetzen, ohne sich von der eigenen epistemologischen Basis zu trennen. Der Einsatz digitaler Methoden führt nämlich nicht zu einer Veränderung der grundlegenden Fragen einer Disziplin, sondern eröffnet neue Perspektiven und Zugangswege zur Beantwortung derselben.[1]

Die vorliegende Arbeit kann als anwendungsorientiertes Beispiel dienen, wie digitale Methoden mit klassisch hermeneutischen Ansätzen in Beziehung gesetzt werden können. Der Hintergrund bildet die Auseinandersetzung des Verfassers mit dem rund 12'000-seitigen Einsiedler Kloster-Tagebuch von Pater Joseph Dietrich (1645-1704) im Rahmen eines laufenden Dissertationsprojekts, bei welchem der Fokus auf textgenetischen und klimageschichtlichen Fragen liegt. Aufgrund verschiedener individueller Faktoren und externer Einflüsse ist das Tagebuch von Veränderungen und Brüchen gekennzeichnet, deren Dekon-

1 Vgl. Graham/Milligan/Weingart, Exploring Big Data, S. 31-33.

struktion insbesondere hinsichtlich stilistischer und orthografischer Eigenheiten mit Close Reading[2] allein schwierig ist.

Aufgrund der beschriebenen Unzulänglichkeiten stellte sich die Frage, ob ein komplementärer Zugang mit Hilfe einer computergestützten Textanalyse gewinnbringend sein könnte. Diese übergeordnete Frage wird in der vorliegenden Arbeit am Beispiel des Ansatzes Topic Modeling vertieft behandelt. Als Grundlage wird nicht das gesamte Tagebuch, sondern nur die Natur- und Wetterbeobachtungen, welche zum Zweck einer klimageschichtlichen Analyse extrahiert und transkribiert wurden, verwendet.[3] Die Beschränkung auf ein inhaltlich und formal weitgehend homogenes Korpus soll eine differenziertere Betrachtungsweise ermöglichen.

1.1 Kontext und Sample

1.1.1 Autor und Werk

Als Sohn des Schultheissen Johann Peter Dietrich (1611-1681) in Rapperswil geboren, legte Joseph Dietrich im Jahr 1662 im Kloster Einsiedeln Profess ab. Ende 1669 erlangte er mit seiner Priesterweihe schliesslich den Status eines Vollmitglieds. Bis zu seinem Tod bekleidete Dietrich zahlreiche klosterinterne Ämter, wobei der Schwerpunkt auf der wirtschaftlichen Verwaltung und der juristischen Vertretung des Klosters lag. Daneben fungierte er zeitweise auch als Bibliothekar, Archivar und Direktor der Stiftsdruckerei. Im Rahmen seiner Tätigkeiten wurde er ab 1688 insgesamt achtmal versetzt und verbrachte rund zehn Jahre als Statthalter in den klösterlichen Aussenstationen in Freudenfels (TG) und Päfffikon (SZ) sowie als Beichtvater im Kloster Fahr (AG), wo er infolge eines dreiwöchigen Fiebers im Alter von 59 Jahren starb.[4] Im Folgenden sind die Aufenthaltszeiträume pro Standort aufgeführt:

- 21.01.1662 – 26.11.1688: Kloster Einsiedeln
- 26.11.1688 – 07.12.1690: Schloss Freudenfels
- 07.12.1690 – 28.07.1692: Kloster Einsiedeln

2 Der Terminus „Close Reading“ bezeichnet das Lesen von Texten und bildet den Gegenpart zum Begriff „Distant Reading“, welcher sich auf Erschliessungsmethoden von Texten ohne Lektüre derselben bezieht. Vgl. Viehhauser, Mittelalterliche Texte, S. 24-25.

3 Das gesamte Tagebuch wird im Rahmen eines digitalen Editionsprojekts aufbereitet. Zum digitalen Editionsprojekt vgl. http://www.dietrich-edition.unibe.ch, Stand: 30.06.2023.

4 Vgl. Henggeler, Professbuch Einsiedeln, S. 325-328.

- 28.07.1692 – 25.08.1693: Schloss Pfäffikon
- 25.08.1693 – 30.10.1694: Schloss Freudenfels
- 30.10.1694 – 03.06.1695: Kloster Einsiedeln
- 03.06.1695 – 29.11.1698: Schloss Freudenfels
- 29.11.1698 – 17.06.1701: Kloster Einsiedeln
- 17.06.1701 – 05.04.1704: Kloster Fahr

Im Klosterarchiv Einsiedeln ist ein grösstenteils von Dietrichs Hand stammendes Tagebuch überliefert. Dieses wurde von einem Mitkonventualen am 9. Juli 1670 begonnen und rund ein Jahr später von Dietrich, der es bis zum 19. März 1704 fortführte, übernommen. Der überwiegende Teil des Tagebuchs ist in deutscher Sprache verfasst, wobei es sprachgeschichtlich am Übergang zum Frühneuhochdeutschen zu verorten ist und dialektale Einschläge aufweist. Vor allem im Zusammenhang mit dem klösterlichen Ritus kommen häufig lateinische Begriffe und Phrasen vor. Das Tagebuch umfasst insgesamt 12'232 beschriebene Seiten in 18 Bänden, wobei der Umfang und der zeitliche Bezugsrahmen der einzelnen Bände stark variiert. Dietrich führte das Tagebuch auch während seiner Aufenthalte in den Aussenstationen fort.

Die Entstehung des Tagebuchs verlief nicht linear, sondern weist aufgrund seiner häufigen Ortswechsel und diversen externen Faktoren viele formale und inhaltliche Unregelmässigkeiten oder Besonderheiten auf. Im Weiteren zeigen sich über den Gesamtzeitraum auch Veränderungen bei der Schreibpraxis. In diesem Zusammenhang ist insbesondere der Übergang von einer unregelmässigen zu einer täglichen Tagebuchführung im Jahr 1693 erwähnenswert. Für die vorliegende Arbeit sind weniger die Hintergründe als vielmehr die Konsequenzen der vielschichtigen Textgenese des Werks von Bedeutung, zumal sie bei der Wahl und Aufbereitung der Datengrundlage sowie bei der Interpretation der Ergebnisse berücksichtigt werden müssen.

1.1.2 Datensample

Die Grundlage der vorliegenden Arbeit bilden die im Einsiedler Kloster-Tagebuch enthaltenen Wetterbeobachtungen. Es handelt sich hierbei nicht um Messungen, sondern um narrative Beschreibungen der gefühlten Temperatur, der Intensität und Dauer von Niederschlägen, Windstärke- und Richtung, Himmelsbedeckung usw. Das Wetter bildet nicht das ausschliessliche Thema des Tagebuchs, sondern erscheint in vielfältigem Bezug zur Denk- und Lebenswelt des Autors und seiner Zeitgenossen, so unter anderem unter Hintergrund landwirt-

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.16

schaftlicher oder kultureller Praktiken im Kloster. Im Hinblick auf eine klimageschichtliche Auswertung und eine spätere Übertragung in die Datenbank Euro-Climhist[5] wurden die Wetterbeobachtungen von Pater Joseph Dietrich, die entweder direkt oder indirekt mit der Witterung und ihren Folgen (z.B. Naturkatastrophen) in Beziehung standen, händisch transkribiert und gemäss den Vorgaben der genannten Datenbank strukturiert. Die Beschreibungen wurden dazu jeweils mit einem Datum, Ort und weiteren Metainformationen abgebildet. Für das Datensample in der vorliegenden Arbeit, welches 5'178 tägliche Wetterbeobachtungen umfasst, wurden nur zeitnahe Beobachtungen ab 1676[6] von Dietrichs Hand berücksichtigt.

1.2 Topic Modeling

Topic Modeling ist vereinfacht ausgedrückt eine computergestützte Methode, mit Hilfe derer zusammenhängende Informationen in grossen Datenmengen (Texte, biologische Daten (DNS), Bilder, Musiknoten usw.) sichtbar gemacht werden können.[7] In Bezug auf Texte und Textsorten lassen sich generell zwei übergeordnete Verwendungszecke ausmachen. Zum einen dient es als eine Methode zur Erschliessung zentraler Konzepte oder Themen in umfangreichen Textsammlungen, zum anderen kann es auch als exploratives Werkzeug für stilistische und formale Analysen eingesetzt werden. Topic Modeling bildet den Oberbegriff für eine Gruppe von Verfahren, welche statistische Methoden mit Ansätzen maschinellen Lernens kombinieren. Diese funktionieren zwar nach ähnlichen Grundprinzipien, unterscheiden sich aber bezüglich der zugrundeliegenden mathematischen Modelle. [8]

Der in der Forschung am weitesten verbreitete Topic-Modeling-Ansatz ist Latent Dirichlet Allocation (LDA). Ausgehend von der Annahme, dass eine begrenzte Zahl von Wörtern oder Tokens[9] in Textsegmenten (z.B. Absatz, Doku-

5 Weiterführende Informationen zu Euro-Climhist finden sich auf der Website. Vgl. https://www.euroclimhist.unibe.ch, Stand: 30.06.2023.

6 Da vor 1676 Wetterbeobachtungen äusserst selten vorkommen, wurden diese Jahre nicht berücksichtigt, ebenso wie die Beobachtungen seines temporären Stellvertreters ausgeschlossen wurden.

7 Vgl. Fechne/Weiss, Einsatz Topic Modeling, Kap. 1.2; Schöch, Topic Modeling Genre, Abs. 2.

8 Vgl. Lamba/Madhusudhan, Text Mining, S. 105-107, S. 113-114; Blei, Probabilistic Topic Modeling, S. 77.

9 Tokens sind durch Leerschläge voneinander abgetrennte Elemente. In der Regel sind dies Wörter, es können aber auch Zahlen, Abkürzungen oder – vor allem beim Vorhandensein von Bindestrichen – Wortteile sein. Vgl. Lamba/Madhusudhan, Text Mining, S. 79-81. In der vorliegenden Arbeit werden Token, Begriff, Wort oder Terminus synonym verwendet.

ment, Brief, Tweet usw.) häufig zusammen auftreten, wird die Häufigkeit derselben ermittelt und mit der Auftretenshäufigkeit anderer Tokens in demselben Segment verglichen. Die Tokens werden zu Topics gebündelt und die Annahmen über die Wahrscheinlichkeiten zu deren Zusammensetzung in einem iterativen Verfahren verfeinert. Die Reihenfolge der Tokens innerhalb der Texteinheit spielen keine Rolle und einzelne Tokens können auch Bestandteil mehrerer Topics sein. Im Weiteren wird die Wahrscheinlichkeit, mit welcher die einzelnen Topics in den Segmenten vorkommen, prozentual berechnet, wodurch die zentralen Themenkomplexe in den einzelnen Texteinheiten sichtbar gemacht werden können.[10]

Obwohl die Begriffe teilweise synonym verwendet werden, entsprechen Topics nicht dem alltagssprachlichen Verständnis von Themen, da die Wortketten eine Interpretation erfordern. Sie können je nach Korpus zwar abstrakte Themenbegriffe (z.B. Politik) enthalten, aber auch ausschliesslich aus Wörtern (z.B. Wahl, abstimmen, Kandidatin, Rede usw.) bestehen, anhand derer im Idealfall das übergeordnete Thema abstrahiert werden kann. Es ist somit eine der wesentlichen Aufgaben der Forschenden, die Bedeutung der Topics anhand der Wortketten zu interpretieren und ihnen einen Überbegriff oder ein Thema zuzuordnen. Da die Algorithmen die Topics ausschliesslich aus den Tokens im vorhandenen Text zusammensetzen, ist keine vorgängige Annotation und keine Trainingsphase erforderlich. Das Verfahren lässt sich ohne vorgängige Annotation oder Trainingsphase gattungs- und sprachunabhängig anwenden und gestaltet sich vom Modellierungsprozess her simpel.[11] Aufgrund verschiedener optional wählbarer Parameter erfordert eine zielführende Anwendung jedoch die Konzeption eines Workflows, der auf den jeweiligen Korpus und die Bedürfnisse zugeschnitten und iterativ optimiert werden muss.[12]

LDA und andere Topic-Modeling-Ansätze sind Teil des grösseren Bereichs der probabilistischen Modellierung und ermitteln versteckte Strukturen über einen generativen Prozess, welcher sowohl sichtbare als auch versteckte Variablen beinhaltet. Die Konsequenz davon ist, dass die Topics auch bei Verwendung der identischen Datengrundlagen und Parameter bei jeder Anwendung

10 Vgl. Blei, Probabilistic Topic Modeling, S. 77-79; Graham/Milligan/Weingart, Exploring Big Data, S. 117-118; Fechne/Weiss, Einsatz Topic Modeling, Kap. 1.1; Hodel, Supervised and Unsupervised, S. 162-164; Hodel/Möbus/Serif, Inferenzen und Differenzen, S. 183-184.

11 Vgl. Blei, Probabilistic Topic Modeling, S. 77-79; Graham/Milligan/Weingart, Exploring Big Data, S. 119-120; Schöch, Topic Modeling Genre, Abs. 13-14; Wehrheim, Economic History, S. 89-92; Hodel, Supervised and Unsupervised, S. 164.

12 Vgl. Schöch, Topic Modeling Genre, Abs. 16, 20;

unterschiedlich ausgegeben werden. Somit sind die Resultate – auch wenn sie in ähnlicher Form erscheinen – nicht eins zu eins reproduzierbar.[13] Im Weiteren kann der Umstand, dass der Algorithmus nur die vorhandenen Tokens berücksichtigt, dazu führen, dass die Topics weniger inhaltliche Muster als vielmehr die formale Gestaltung der Textsegmente, wie etwa stilistische oder orthografische Eigenheiten, abbildet. Je nach Erkenntnisinteresse stellt dies einen Verzerrungsfaktor dar, der insbesondere bei heterogenen Korpora auftritt. Umgekehrt bedeutet dies, dass sich Topic Modeling vor allem für die thematische Exploration von homogenen Textsammlungen gut eignet.[14]

Für den Modellierungsprozess existieren zahlreiche Tools und Programmpakete in unterschiedlichen Programmiersprachen und Umgebungen. Allen diesen Tools ist gemeinsam, dass sie – trotz des gemeinsamen Labels Topic Modeling – unterschiedliche Resultate hervorbringen. Dies hängt einerseits mit den bereits erwähnten Zufallsvariablen zusammen und ist anderseits auf die Einbindung unterschiedlich konfigurierter Algorithmen in den Tools zurückzuführen.[15] Der in der vorliegenden Arbeit verwendete „Werkzeugkasten“ MAchine Learning for LanguagE (MALLET) wurde 2002 erstmals veröffentlicht und gilt als solides Topic-Modeling-Toolkit, weshalb es vor allem in geisteswissenschaftlichen Studien häufig benutzt wird.

1.3 Forschungsüberblick zu Topic Modeling

In den letzten zwei Jahrzehnten wurden im Zusammenhang mit Topic Modeling zahlreiche Studien von Forschenden aus unterschiedlichen Disziplinen verfasst, wobei beispielsweise allein Pooja Kherwa und Poonam Bansal in einer Metastudie rund 300 zwischen 2003 und 2018 erschienene Aufsätze zu Topic Modeling untersuchten.[16] Allgemein lassen sich bezüglich der Ausrichtung wissenschaftlicher Studien im Zusammenhang mit Topic Modeling drei Tendenzen erkennen: Erstens gibt es Aufsätze aus dem Informatikbereich, die sich schwerpunktmässig mit den zugrundeliegenden statistischen und technischen Eigenheiten und Entwicklungspotenzialen von Topic Modeling auseinandersetzen. Im Gegensatz dazu sind Forschende aus geisteswissenschaftlichen Disziplinen in der Regel eher anwendungs- und resultatorientiert. Eine dritte Kategorie bil-

13 Vgl. Blei, Probabilistic Topic Modeling, S. 79-80; Graham/Milligan/Weingart, Exploring Big Data, S. 157; Schöch, Topic Modeling Genre, Abs. 14.
14 Vgl. Fechne/Weiss, Einsatz Topic Modeling, Kap. 1.2.
15 Vgl. Graham/Milligan/Weingart, Exploring Big Data, S. 130, 157.
16 Vgl. Kherwa/Bansal, Topic Modeling.

den diejenigen Aufsätze, welche sich mit Fragen rund um die epistemologischen Konsequenzen des Einbezugs digitaler Methoden in Disziplinen, die traditionell andere Wege der Erkenntnisgewinnung beschreiten, auseinandersetzen.

Aus eher technischer Sicht bildet der 2003 erschienene und häufig zitierte Aufsatz von David Blei, Andrew Ng und Michael I. Jordan, in welchem Terminologie, Grundlagen und Potenziale von LDA erstmals beschrieben wurden, einen wesentlichen Ausgangspunkt.[17] Diese erstmalige Skizzierung von LDA wurde unter dem Hintergrund der Forschung an unterschiedlichen Methoden zur effizienten Organisation und Auffindbarkeit von Informationen in grossen Datenmengen entwickelt.[18] LDA entstand somit gleichzeitig mit sowie in Anlehnung und Abgrenzung zu anderen Topic-Modeling-Ansätzen.[19] Da diverse Elemente in den folgenden Jahren weiterentwickelt wurden, fasste Blei den Stand der Forschung sowie die möglichen zukünftigen Richtungen im Jahr 2012 zusammen.[20] Topic Modeling als Methode für die algorithmische Durchleuchtung unstrukturierter Textdaten wird unter anderem auch als Anwendungsform von Text Mining klassifiziert. Während sich viele Übersichtswerke in diesem Bereich vornehmlich auf statistische und theoretische Aspekte beschränken, richtet sich etwa das 2022 publizierte Buch *Text Mining for Information Professionals* an ein anwendungsorientiertes Publikum im Bibliotheks- und Informationsbereich. In Bezug auf Topic Modeling werden dabei potenzielle Workflows und Anwendungsszenarien, verfügbare Tools sowie konkrete Projekte vorgestellt.[21]

Als frühes Beispiel einer Anwendung von Topic Modeling in den Geisteswissenschaften ist eine 2006 von Newman und Block publizierte Studie zu nennen. Hierin wurden die prägenden Topics in der zwischen 1728 und 1800 erschienen Zeitung *Pennsylvania Gazette* identifiziert und deren Auftretenswahrscheinlichkeit im zeitlichen Längsschnitt analysiert. Dabei wurde insbesondere das Potenzial von Topic-Modeling-Ansätzen für die Bearbeitung geschichtswissenschaftlicher Fragestellungen im Kontext der wachsenden Menge an digital zugänglichen historischen Dokumenten thematisiert.[22] Mit *Mining the Dispatch* wurden die Resultate eines 2011 abgeschlossenen Projekts veröffentlicht, welches ebenfalls auf einer Anwendung von Topic Modeling auf Zeitungsartikel

17 Vgl. Blei/Ng/Jordan, Latent Dirichlet Allocation.
18 Vgl. Newman/Block, Probabilistic Topic, S. 753-754.
19 Vgl. dazu beispielsweise Wallach/Mimno/McCallum, Rethinking LDA.
20 Vgl. Blei, Probabilistic Topic Models.
21 Vgl. Lamba/Madhusudhan, Text Mining.
22 Vgl. Newman/Block, Probabilistic Topic.

gründet. Anhand der Zeitung *Daily Dispatch* wurden die Kontinuitäten und Veränderungen im sozialen und politischen Leben der Stadt Richmond kurz vor und während des amerikanischen Bürgerkrieges (1861-1865) erforscht.[23]

Fand der Diskurs rund um Topic Modeling bis dahin vor allem in Blogs, Präsentationen und ähnlichen Formaten statt, erfuhr der Ansatz zu Beginn der 2010er Jahre in den digital arbeitenden Geisteswissenschaften stärkere Beachtung. Um Topic Modeling in den Digital Humanities zu fördern und gleichzeitig eine kritische Methodendiskussion anzustossen, widmeten die Editoren des *Journal of Digital Humanities* dem Thema Topic Modeling im Jahr 2012 eine vollständige Ausgabe. Diese enthielt neben diversen Einführungen auch Aufsätze zu Anwendungsbeispielen aus unterschiedlichen Disziplinen und ein Kapitel zu Topic-Modeling-Tools.[24] Im Weiteren evaluierte Matthew L. Jockers 2013 in der Monografie *Macroanalysis* Topic Modeling und andere digitale Werkzeuge im Hinblick auf deren Einsatzpotenzial in der Literaturwissenschaft, wobei er für eine stärkere Einbindung makroanalytischer Methoden als komplementären Zugangsweg zum traditionellen Close Reading plädierte.[25] In ähnlicher Weise argumentierten auch Graham, Milligan und Weingart, welche 2015 mit der Monografie *Exploring big historical data: The Historian's Macroscope* das geschichtswissenschaftliche Pendant zu Jockers Werk publizierten. Neben der Vermittlung von praktischem Wissen zum Umgang mit Daten und Tools reflektierten sie die Einflüsse digitaler Methoden auf das Selbstverständnis und die Arbeitsweise der Geschichtswissenschaften.[26]

Im deutschsprachigen Raum finden sich ab Mitte der 2010er Jahre zunehmend Anwendungen von Topic Modeling, wobei insbesondere die Genre-Forschung von Christof Schöch erwähnenswert ist. Mit Hilfe von Topic Modeling gelang es ihm, französische Dramen aus dem Zeitraum von 1630 und 1789 den Subkategorien Tragödie, Komödie und Tragikomödie zuzuweisen und signifikante Unterschiede innerhalb der Subkategorien aufzuzeigen. Abgesehen von den Resultaten sind insbesondere seine Untersuchungen zu den Auswirkungen unterschiedlicher Konfigurationsoptionen beim Modellierungsprozess wesentlich.[27] Peter Andorfer untersuchte im Jahr 2017 eine Briefkorrespondenz aus dem 19. Jahrhundert mit Topic Modeling und sah seine Resultate als Beleg, dass der Ansatz auch ohne eingehendere statistische Kenntnisse erfolgreich ge-

23 Vgl. Nelson, Richmond Daily.
24 Vgl. Meeks/Weingart, Digital Humanities.
25 Vgl. Jockers, Macroanalysis.
26 Vgl. Graham/Milligan/Weingart, Exploring Big Data.
27 Vgl. Schöch, Topic Modeling Mallet; Schöch, Topic Modeling Genre.

nutzt werden kann.[28] Als Reaktion darauf plädierten Martin Fechne und Andreas Weiss dafür, dass zumindest die Konsequenzen der Benutzung von Algorithmen abgeschätzt werden sollten. Dennoch erachteten die Autoren Topic Modeling als geeigneten Ansatz für eine breitere Nutzung in den Geistes- und Geschichtswissenschaften.[29]

In der 2019 erschienenen Studie mit wirtschaftsgeschichtlichem Schwerpunkt erprobte Lino Wehrheim den gezielten Einsatz von Topic Modeling in einer geschichtswissenschaftlichen Subdisziplin. Zu diesem Zweck analysierte er 2'675 Artikel, welche zwischen 1941 und 2016 im *Journal of Economic History* publiziert wurden. Anhand ausgewählter Topics konnte er zeigen, dass sich seine Resultate mit den auf traditionellen Methoden basierenden Erkenntnissen der Forschung vergleichen lassen.[30] In einer 2022 erschienen Studie verglichen Tobias Hodel, Dennis Möbus und Ina Serif auf Basis von drei historischen Korpora die Resultate der beiden in den Digital Humanities und Informatik vorrangig genutzten Topic-Modeling-Engines MALLET und Gensim. Im Zentrum der Studie stand eine kritische Auseinandersetzung mit theoretischen und methodischen Aspekten von Topic Modeling, wobei einerseits die Einflüsse unterschiedlicher Parameterkonfigurationen untersucht und andererseits die Eignung des Zusammenspiels von Close und Distant Reading als neue Form der Heuristik in den Geschichtswissenschaften diskutiert wurde. Die praktische Umsetzung ergab, dass sich bereits mit wenig Aufwand Themenfelder in unstrukturierten Korpora finden lassen und auch der Nachvollzug von Abschreibeprozessen möglich ist.[31]

1.4 Erkenntnisinteresse und Aufbau

Die vorliegende Arbeit verfolgt in erster Linie eine anwendungs- und resultatorientierte Stossrichtung und widmet sich den Fragen, inwiefern sich Topic Modeling für die Analyse der Wetterbeobachtungen von Pater Joseph Dietrich eignet und welche Ausgangspunkte für weitere Analysen möglich sind. Diese Fragen werden zwar unter dem Hintergrund der Erkenntnisinteressen der historischen Klimaforschung betrachtet, sollen aber auch allgemein Aufschluss über die Potenziale und Grenzen der Anwendung von Topic Modeling auf frühneuzeitliche Texte geben, womit sie für ein breiteres Publikum im Archiv- und

28 Vgl. Andorfer, Turing Test.
29 Vgl. Fechne/Weiss, Einsatz Topic Modeling.
30 Vgl. Wehrheim, Economic History.
31 Vgl. Hodel/Möbus/Serif, Inferenzen und Differenzen.

Informationsbereich interessant sein dürften. Aufgrund der vielfältigen Realisierungsmöglichkeiten stellt sich im Weiteren die Frage, wie sich der Topic-Modeling-Prozess zielführend anwenden lässt. Entsprechend ist die Methode selbst Gegenstand kritischer Betrachtung, wobei vor allem der Einfluss von Parametereinstellungen und anderen Entscheidungen reflektiert werden soll.

Aufgrund der letztgenannten Frage wird die Methodik nicht wie allgemein üblich in der Einleitung beschrieben, sondern in einem eigenen Kapitel behandelt, in welchem die drei zentralen Arbeitsschritte beim Topic Modeling aufgeführt sind. Hier wurden auch bereits Modellierungsprozesse umgesetzt, um beispielsweise Aufschluss über den Einfluss bestimmter Parameterkonfigurationen zu erhalten. Im nächsten Kapitel folgt die eigentliche Analyse, die sich nach Art der Zusammensetzung der Datengrundlage in drei Unterkapitel gliedert. Im Rahmen der Untersuchung werden einerseits Elemente des Modellierungsprozesses kritisch reflektiert und andererseits die Resultate im Hinblick auf ihre Aussagekraft sowie das Potenzial für weiterführende Untersuchungen bewertet. Dazu werden im Sinne des Scalable Reading[32] sowohl allgemeine Tendenzen zur Gesamtheit der Daten als auch konkrete Quellenstellen miteinander in Beziehung gesetzt. Im Fazit werden die Resultate in einen breiteren Kontext gestellt und Möglichkeiten für weiterführende Forschungsansätze skizziert.

2 Methode

2.1 Preprocessing

Jede Form der computergestützten Textanalyse bedingt eine vorgängige Aufbereitung des gewählten Textkorpus, wobei je nach Ausgangslage, Methode, Anwendungsbereich und Anforderungen ein unterschiedlicher Grad an Manipulation gewählt werden kann. So gehört die Entfernung von Satzzeichen und die systematische Ersetzung von Gross- durch Kleinbuchstaben (Case Normalization) in der Regel zum Mindestmass an Normalisierung und wird in der Java-Version von MALLET standardmässig umgesetzt. Eine stärkere Normalisierung der Texte, bei der die einzelnen Wörter auf ihren Wortstamm oder den Kern ihrer Bedeutung reduziert werden, kann mit Hilfe von Algorithmen zum Stemming oder der Lemmatisierung vorgenommen werden.[33] Aufgrund der inkon-

32 Der Anglist Martin Mueller schlug den Begriff „Scalable Reading“ als Bezeichnung für einen Ansatz vor, der Distant- und Close-Reading-Verfahren verbindet. Vgl. Viehhauser, Mittelalterliche Texte, S. 32.

33 Vgl. Lamba/Madhusudhan, Text Mining, S. 79-85.

sistenten Orthografie und der sprachlichen Eigenheiten des Autors wurde in der vorliegenden Arbeit auf die Anwendung einer automatisierten Sprachnormalisierung verzichtet und nur minimale händische Eingriffe vorgenommen. So wurden Abkürzungen bereits bei der ansonsten diplomatischen Transkription soweit möglich ausgeschrieben. Zudem wurden heute nicht mehr gebräuchliche Zeichen (æ, œ, ë, v̈, ÿ) normalisiert, wobei das anlautende „v“ anstelle von „u“[34] und das Scharf-s („ß“) beibehalten wurden. Die Getrennt- und Zusammenschreibung wurde bei der Transkription möglichst vorlagengetreu abgebildet.

Während in den vorhin erwähnten Studien von Schöch und Jockers bestimmte Wortarten bereits auf Ebene der Textvorbereitung entfernt wurden, ist es auch möglich, bestimmte Wörter oder Wortarten durch das Einbinden sogenannter Stoppwort-Listen erst beim Modellierungsvorgang auszuschliessen. Dieser Ansatz ist einfach zu realisieren und findet deshalb in den meisten Studien Verwendung, wobei in der Regel Funktionswörter wie Artikel oder Präpositionen, die keine oder wenig inhaltliche Bedeutung enthalten, ausgegrenzt werden.[35] Da beim Topic Modeling der Text in Tokens unterteilt wird, werden auch semantisch zusammenhängende Wortpaare oder -gruppen (z.B. „Heiliger Stuhl“ oder „Vereinigte Staaten“) getrennt. Diesem Effekt kann ebenfalls entweder im Zuge der Modellierung oder bereits auf Ebene der Textvorbearbeitung begegnet werden. So besteht beispielsweise ein pragmatischer Ansatz darin, zusammengehörige Wortpaare (z.B. „heilige_stuhl“) im Rahmen des Preprocessing zu verketten.[36]

Für die vorliegende Studie wurden die Zusammengehörigkeit von Temperaturbegriffen mit negierenden oder qualifizierenden Termini (z.B. „nit_kalt“, „grimmig_kalt“, „kein_Schnee“, „vill_Schnee“) auf Textebene zu N-Grammen zusammengeführt. Diese häufig auftretenden Wortpaare oder -gruppen wurden vorweg mit Hilfe von AntConc[37] identifiziert. Trotz der vorgenommenen Auszählungen basierte die Wahl weitgehend auf subjektiven Gesichtspunkten. Dasselbe gilt auch für die Stoppwort-Liste, welche im Hinblick auf den Ausschluss von Funktionswörtern erstellt und im Rahmen der unzähligen Testläufe fortwährend modifiziert wurde. Für alle in der Arbeit behandelten Model-

34 Dietrich schrieb beispielsweise „vnd“ statt „und“. Im Wortinneren verwendete er das heutige „u“.

35 Vgl. Wallach/Mimno/McCallum, Rethinking LDA, S. 1; Jockers, Macroanalysis, S. 131; Graham/Milligan/Weingart, Exploring Big Data, S. 86; Lamba/Madhusudhan, Text Mining, S. 85.

36 Vgl. Mimno, Using Phrases.

37 AntConc wurde von Laurence Anthony entwickelt und eignet sich besonders gut für die Analyse von Wortkonkordanzen. Vgl. Graham/Milligan/Weingart, Exploring Big Data, S. 79-81.

lierungsprozesse wurde dieselbe Stoppwort-Liste verwendet. Neben der sprachlichen Vorbearbeitung der Texte bildet die Segmentierung des Korpus in Einzeldokumente ein zentrales Element. Während die Bildung der Topics eine gewisse Mindestmenge an Dokumenten erfordert, wird ab einer gewissen Anzahl die Performanz stärker von deren Umfang als deren Menge beeinflusst. Die einzelnen Dokumente sollten nicht zu umfangreich sein, benötigen aber eine gewisse Mindestlänge.[38] Dabei bestehen die Möglichkeiten, die Texte entlang gegebener struktureller Einheiten, wie beispielsweise Kapitel, Absätze usw., zu separieren oder sie nach einer vordefinierten Zahl an Wörtern zu unterteilen.[39]

Während in vielen Studien die Segmentierung unter dem Hintergrund ihres Einflusses auf die Resultate thematisiert wird, bildet sie in der vorliegenden Arbeit einen integralen Bestandteil der Analyse. In einer ersten Art der Segmentierung wurden die Daten nicht chronologisch, sondern in Bezug zum jährlichen Zyklus angeordnet, sprich unabhängig vom Jahr entlang den Monaten in zwölf Segmente zusammengeführt. Dadurch wird eine synchrone Betrachtung ermöglicht, die stärker inhaltliche Gemeinsamkeiten in den Segmenten statt Veränderungen über den Gesamtzeitraum aufzeigen soll. Die Schwäche dieser Art der Segmentierung liegt darin, dass ortsspezifische Eigenheiten nicht berücksichtigt werden. Zu diesem Zweck wurden die Daten in einer zweiten Art der Segmentierung in Einheiten zu den einzelnen Ortschaften und den Jahreszeiten zusammengeführt, womit beispielsweise ein Vergleich des Winters in Einsiedeln mit demjenigen in Freudenfels vorgenommen werden kann. Um Hinweise auf Konstanten und Veränderungen der Beobachtung über den gesamten Zeitraum zu erhalten, wurden die Daten in einer dritten Art der Segmentierung in Einheiten pro Jahr zusammengeführt.

2.2 Modellierungsprozess

Der eigentliche Modellierungsprozess erfordert die Definition der Anzahl Topics, welche vom Algorithmus gebildet werden sollen. Letzteres ist notwendig, weil das System – zumindest nicht ohne komplementäre Wege zu deren statistischen Bestimmung[40] – die optimale Anzahl an Topics nicht im Vornherein fest-

38 Vgl. Tang et al., Limiting Factors, S. 196.
39 Vgl. Graham/Milligan/Weingart, Exploring Big Data, S. 117.
40 Es ist auch möglich, die ideale Anzahl an Topics annäherungsweise zu berechnen. Hodel, Möbus und Serif entwickelten beispielsweise eine Metrik, deren Resultat als Indikator für die Trennschärfe der Topics genutzt werden kann. Vgl. Hodel/Möbus/Serif, Inferenzen und Differenzen, S. 194-195. Auch Wehrheim stützte sich für die Bestimmung Anzahl Topics auf Berechnungen mit einer eigenen Metrik. Vgl. Wehrheim, Economic History, S. 113-114.

setzen kann. Allerdings ist dies auch für Forschende schwierig, da sich die „optimale“ Anzahl Topics nur iterativ ermitteln lässt.[41] Im Unterschied zu den anderen Studien, welche in der Regel 50 Topics und mehr verwendeten, ergaben die eigenen Tests, dass bereits bei der geringen Zahl von 5 Topics interpretierbare Resultate vorlagen. Dies ist einerseits mit der allgemein geringen Datenmenge und andererseits mit deren hohen inhaltlichen Homogenität zu begründen. Bei mehr als 30 Topics zeigte sich, dass lediglich die Zahl der wenig aussagekräftigen Topics zunahm. Im Weiteren konnte festgestellt werden, dass für den Bereich dazwischen je nach Modell unterschiedliche Effekte deutlicher erkennbar werden. Diese Erkenntnis führte zu der Annahme, dass es nicht eine ideale Anzahl an Topics gibt, sondern dass es vielmehr einen Bereich gibt, in welchem nutzbare Resultate vorhanden sind. Bei mehreren Modellen wird so einerseits die Zahl der Perspektiven auf die Daten erhöht und andererseits der Einfluss dieses Parameters besser erkennbar. Entsprechend wurden in der vorliegenden Arbeit für jede Art der Segmentierung Modellierungsprozesse für 5, 10, 15, 20, 25 und 30 Topics durchgeführt und visualisiert.

Während die Definition der Anzahl auszugebender Topics unabdingbar ist, bestehen weitere Parameter, mit denen der Modellierungsprozess optional beeinflusst werden kann. So kann beispielsweise die Anzahl an Wiederholungen, die der Algorithmus zur Verfeinerung der Zusammensetzung der Topics durchführt, gewählt werden. Gemäss Schöch führt eine hohe Zahl an Wiederholungen dazu, dass sich die Ergebnisse unterschiedlicher Modellierungsprozesse (mit denselben Parametern) weniger voneinander unterscheiden.[42] Im Weiteren kann durch die Modifikation der sogenannten Hyperparameter Alpha und Beta das Verteilungsprofil der Wörter innerhalb der Topics sowie dasjenige der Topics innerhalb der Dokumente beeinflusst werden. Standardmässig ist das Modell so konfiguriert, dass alle Topics über den gesamten Korpus mit derselben Wahrscheinlichkeit vorkommen, wobei sich lediglich die Auftretenswahrscheinlichkeit in den einzelnen Dokumenten unterscheidet. Durch Anpassungen an den Hyperparametern wird erreicht, dass einige Topics über den Gesamtkorpus gesehen häufiger vorkommen dürfen als andere. Mit dem Optimierungsintervall wird definiert, wie gross die Abweichung von der standardmässig flachen Wahrscheinlichkeitsverteilung ist. Eine sehr starke Optimierung führt zu einer starken Differenz zwischen einigen wenigen Topics, die mit ho-

41 Vgl. Hodel, Supervised and Unsupervised, S. 164.
42 Vgl. Schöch, Topic Modeling Genre, Abs. 14, 20-21.

her Wahrscheinlichkeit vorkommen, und den Übrigen mit tiefer Auftretenswahrscheinlichkeit.[43]

Der Einfluss der Hyperparameter wurde in der vorliegenden Arbeit auf Basis des Datensatzes mit der Segmentierung nach Monaten kumuliert eruiert, indem mehrere Modellierungsprozesse mit unterschiedlichen Kombinationen bezüglich der Anzahl an Topics, der Anzahl Iterationen und des Optimierungsintervalls durchgeführt wurde. Im Hinblick auf die Anzahl Iterationen zeigte sich, dass die Anwendung mit 6'000 Iterationen[44] im Vergleich zur Standardkonfiguration von 400 Iterationen wesentlich zu einer Stabilisierung beitrug, sprich die Resultate auch bei mehrmaliger Anwendung des Modellierungsprozesses ähnlich ausfielen. Die Tests mit den Optimierungsintervallen 0, 10, 50, 100, 500, 1'000 und 2'000 ergaben, dass mit dem Wert 50 die differenziertesten Topics abgebildet wurden. Alle Modellierungsprozesse in der vorliegenden Studie wurden mit dem Optimierungsintervall 50 und mit 6'000 Iterationen durchgeführt.

2.3 Postprocessing

Der Output des Modellierungsprozesses gestaltet sich je nach verwendetem Tool unterschiedlich. Bei der Verwendung der Java-basierten Engine MALLET werden die Resultate in Textdateien ausgegeben, wobei mittels eines Befehls die erwünschten Outputs definiert werden können. Diese Outputs sind – vor allem bei einer grossen Anzahl an Topics – für Menschen kaum lesbar, weshalb im Rahmen des Postprocessing weitere Schritte zur Repräsentation, Interpretierbarkeit und Selektion vorgenommen werden müssen. In der Regel werden die Resultate mit Hilfe von Visualisierungen zugänglich gemacht. Unabhängig von der gewählten Vorgehensweise ist zu beachten, dass durch Visualisierungen lediglich Aspekte der vorhandenen Resultate dargestellt und somit allfällige Fehler im Arbeitsprozess zu Verzerrungen führen können.[45]

In der vorliegenden Arbeit wurden die mit Hilfe von MALLET erzeugten Rohdaten in mehreren Schritten verarbeitet. Für die Visualisierung der Ergebnisse wurde die webbasierte Plattform Observable genutzt, wo basierend auf bereits veröffentlichten Notebooks eine auf die Bedürfnisse der vorliegenden Arbeit hin individualisierte Vorlage konzipiert wurde. Diese Notebooks enthal-

43 Vgl. Schöch, Topic Modeling Mallet.
44 Für die Wahl des Werts dienten als Orientierungspunkt die Analysen von Schöch, der ebenfalls mit 6’000 Iterationen arbeitete. Vgl. Schöch, Topic Modeling Genre, Abs. 19.
45 Vgl. Lamba/Madhusudhan, Text Mining, S. 107-108.

ten in kompakter Form Heatmaps und Listen zu den Topics sowie den Auftretenswahrscheinlichkeiten und der Frequenz der einzelnen Tokens. Zwecks Vergleichbarkeit wurde für die Abbildung des Farbverlaufs in den Heatmaps immer dieselbe nichtlineare Skala (Quadratwurzel-Skala) verwendet.[46] Zudem wurden interaktive Elemente zur Individualisierung der Gestalt der Heatmaps (Farbwahl, Darstellungsbreite, Sortierung) integriert. Im Sinne der Transparenz wurden in jedem Notebook auch die Listen zu den verwendeten Stoppwörtern und N-Grammen sowie Informationen zu den bei der Modellierung verwendeten Parametern aufgeführt. Die Informationen zu den Grundlagen und Vorlagen der Visualisierung sind im Anhang der Notebooks ersichtlich.[47]

Um bestimmte Tendenzen, welche in den Heatmaps erkennbar wurden, eingehender zu studieren, wurden Begleitanalysen mit Voyant Tools[48] durchgeführt. Hierbei wurde ausschliesslich die Analysefunktion für die Berechnung der relativen Frequenz ausgewählter Begriffe in den jeweiligen Einheiten (z.B. Ortschaft oder Jahr) genutzt. Diese gibt Aufschluss darüber, wie häufig ein Wort im Verhältnis zur Textmenge in der jeweiligen Einheit vorkommt, womit sich auch Einheiten mit unterschiedlichen Textmengen miteinander vergleichen lassen. Es wurden hierfür dieselben Stoppwort-Listen verwendet wie beim Topic Modeling.

3 Analyse

3.1 Segmentierung pro Monat kumuliert

Als Ausgangspunkt für die Analyse wurden die Resultate[49] zur Segmentierung, welche die Daten kumuliert pro Monat gliedert, gewählt. Es lassen sich anhand dieser Art der Datenzusammenstellung bestimmte allgemeine Tendenzen, die

46 Bei der Quadratwurzel-Skala handelt es sich um eine nichtlineare Skala, die grössere Zahlen in kleinere Bereiche komprimiert, damit die Differenzen in kleineren Wertebereichen trennschärfer abgebildet werden. Vgl. Wilke, Datenvisualisierung, S. 16-20.

47 Es handelt sich hierbei um die Outputs und nicht das eigentliche Datensample. Weil die Wetterdaten im Rahmen einer anderen Arbeit veröffentlicht werden sollen, wird das Datensample noch nicht öffentlich zur Verfügung gestellt.

48 Voyant Tools ist eine webbasierte Textanalyseplattform für die Geisteswissenschaften. Das Tool kombiniert in einfacher Art und Weise verschiedene Analysefunktionen und Visualisierungsformen. Vgl. https://voyant-tools.org, Stand: 30.06.2023.

49 Es werden hier nur diejenigen Heatmaps abgebildet, deren Topics im Rahmen der Analyse eingehender beschrieben werden. Für eingehendere Informationen zur Zusammensetzung der Topics, deren prozentualen Auftretenswahrscheinlichkeiten sowie weiteren Modellen sind die entsprechenden Notebooks zu sichten.

auch für die späteren Analysen massgeblich sind, exemplarisch erläutern. Das Augenmerk wird hier vor allem auf die Implikationen einer variierenden Anzahl an Topics gelegt. Ein wesentliches Element der Analyse bildet die kritische Reflexion des Wechselspiels zwischen den Farbverläufen der Heatmaps, welche die Auftretenswahrscheinlichkeiten über die Gesamtheit der Topics abbilden, und die genauere Betrachtung der inneren Zusammensetzung der einzelnen Topics. Dies wird zunächst am Beispiel der Modellierung von 5 Topics (Abb. 1, links) illustriert. Hier zeigt sich, dass Topic 5 über alle Monate hinweg eine sehr hohe Auftretenswahrscheinlichkeit (40-73%) aufweist, wobei vor allem für die Monate April bis Oktober erhöhte Werte erkennbar sind.[50] Der Grund dafür ist, dass es sich aus generell sehr häufigen und in allen Monaten vorkommenden Begriffen wie „wetter“, „himmel“, „sonne“, „gewülk“, „luft“, „wind“ usw. zusammensetzt. Da weitere Termini („sonne“, „warm“, „schön“, „schöner“, „scheinte“) nicht ausschliesslich, aber tendenziell eher wärmebezogen sind, erklärt sich die erhöhte Auftretenswahrscheinlichkeit in den Übergangs- und Sommermonaten.

Im Gegensatz dazu weist Topic 1 einen klaren Schwerpunkt in den Wintermonaten (55-58%) und teilweise auch in den Übergangsmonaten April (28%) und Oktober (13%) auf und beinhaltet viele kältebezogene Wörter wie „schnee“, „kälte“ und „sehr_kalt“. Bei Topic 2 liegt der farbliche Akzent in der Heatmap auf den Monaten April bis Juni (11-16%), was einen Bezug zum Frühling suggeriert. Während Topic 1 und 5 vor allem aus Tokens mit direktem Bezug zu Wetterphänomenen bestehen, offenbart sich die Abgrenzung des Frühlings-Topics neben Wetterbegriffen („vngewitter“, „tundern“) vor allem über Wörter zur Vegetationsentwicklung („bluest“, „grüene“, „buechen“, „wachßen“) und zu lebensweltlichen Aspekten („procession“, „spazieren“, „kirchen“, „mangel“). Bei Topic 3 dominieren Begriffe mit Bezug zur Landwirtschaft („heüw“, „frucht“, „korn“, „ernd“, „veld“, „roggen“) und zu Witterungsverhältnissen, die sich tendenziell negativ auf die Ernte oder Erntepraxis („schaden“, „vngewitter“, „hagel“, „plazreegen“) auswirken. Entsprechend ist eine erhöhte Auftretenswahrscheinlichkeit in den Monaten Juni bis August (11-31%) und teilweise im September (8%) erkennbar, wobei der einzige Temperaturbegriff „hiz“ ebenfalls bestens zu diesem Sommer-Topic passt. Topic 4 wird geprägt vom Weinbau, weshalb der farbliche Akzent auf den Monaten September (14%) und Oktober (24%) liegt. Neben den allgemeinen Begriffen „trauben“, „wein“, „herpst“ und „reeben“ finden sich in diesem Topic zahleiche Andeutungen auf die Erntepra-

50 Die farblichen Unterschiede sind bei Topic 5 in der Heatmap weniger deutlich wahrnehmbar, weil die gewählte nichtlineare Farbskala besonders Differenzen in tieferen Segmenten hervorhebt.

xis („wimmlen“, „gelten“, „eimer“) sowie die Standorte der Reben („pefiken“, „vechtnauw“). Zudem kommen häufiger Tokens mit Kältebezug („nicht_kalt“, „kalt“, „reifen“) vor, welche im Kontext der Traubenlese wichtig waren.

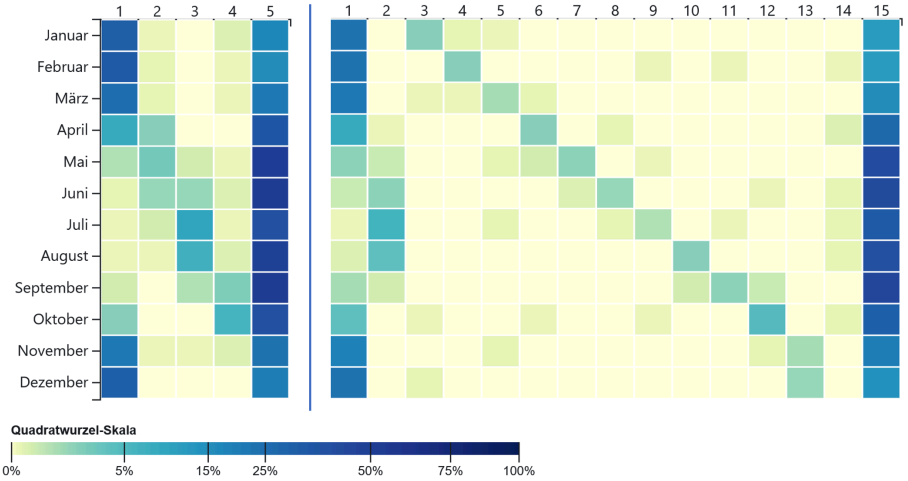

Abb. 1. Segmentierung pro Monat kumuliert, Modelle 5 und 15. Die oben abgebildeten Zahlen stehen für das jeweilige Topic. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, Stand: 30.06.2023.

Bereits bei fünf Topics zeigt sich somit in der Heatmap ein Muster, welches einigermassen adäquat und trennscharf die Jahreszeiten abbildet. Diese Differenzierung ergibt sich inhaltlich nicht allein über direkte Wetterbeschreibungen, sondern auch über Begriffe zur Vegetation, Landwirtschaftspraxis und zur monastischen Lebenswelt, welche von der Witterung beeinflusst und somit indirekt dazu in Bezug stehen. Bei der Betrachtung der Heatmaps zu Modellen mit einer höheren Anzahl an Topics wird allgemein eine Tendenz zur stärkeren Ausdifferenzierung ersichtlich. Neben Wortketten mit jahreszeitlicher Ausprägung bilden sich zunehmend Topics, die vor allem in einem einzelnen Monat eine erhöhte Auftretenswahrscheinlichkeit aufweisen. In Modell 15 (Abb. 1, rechts) zeichnet sich beispielsweise ein Winter-Topic (Topic 1, 48-49%), ein Sommer-Topic (Topic 2, 12-25%), ein Topic mit allgemein tiefer (Topic 14, 0-3%) und eines mit generell hoher Auftretenswahrscheinlichkeit (Topic 15, 34-66%) ab. Neben Topic 13, welches erhöhte Werte für die Monate November (9%) und Dezember (11%) abbildet, heben die Topics 3 bis 12 jeweils einen anderen Monat durch höhere Wahrscheinlichkeiten (8-21%) hervor.

Eine eingehendere Betrachtung der Zusammensetzung der einzelnen Topics gibt Aufschluss darüber, welche inhaltlichen Hintergründe für die Betonung der einzelnen Monate in Topic 3 bis 12 massgeblich sind. In mehreren Topics erscheinen religiöse Feiertage, die an ein bestimmtes Datum oder – bei Abhängigkeit vom Osterzyklus – an einen Zeitraum gebunden sind. So beziehen sich die häufigsten Tokens in Topic 3 auf das Meinradsfest („meinradi“) im Januar, in Topic 5 (März) auf dasjenige zu Ehren des Heiligen Benedikt („benedicti“), und Topic 6 auf die häufig im April stattfindende Osterfeier („oster“ und auch „ostern“, „hochheilige“, „ostermonntag“). Daneben finden sich in letztgenanntem Topic viele Vegetationsbegriffe („grüene“, „bluest“, „grüen“, „weiden“, „gruenen“, „grüenen“, „kirschi“) und indirekte Witterungsindikatoren („reifen“, „wäßerig“, „schnees“) sowie ein Token („spazieren“) zum klösterlichen Alltag.

Während der Frühlingsbezug in Topic 6 deutlich erkennbar ist, gestaltet sich die inhaltliche Entschlüsselung anderer Topics als schwieriger. Topic 3, das mit erhöhter Wahrscheinlichkeit im Januar auftritt, enthält beispielsweise kältebezogene Tokens wie „byßwind“, „kälterer“ und „pik“ (Raureif). Auch das Wort „zwechtenen“ (Schneewehe) und das damit in Zusammenhang stehende Wort „verwähet“ weist auf den Winter hin. Eine Sichtung des Quellentextes ergab, dass der Autor den letztgenannten Begriff ausschliesslich dann verwendete, wenn der Wind die Strassen mit Schnee überdeckte oder zu Schneewehen formte. In diesem Kontext benutzte er teilweise das ebenfalls im Topic enthaltene Wort „haufen“, welches aber auch in anderen Bedeutungszusammenhängen vorkommen konnte. Die Öffnung der schneebedeckten Strassen oblag den Knechten, die neben der Benutzung von Schaufeln in der Regel Ochsen oder Pferde einsetzten. Entsprechend kommen die Begriffe „rosßen“, „knecht“ und „ochßen“ im Topic ebenfalls vor. Insgesamt bezieht sich das Topic somit auf jahreszeitlich bedingte Einschränkungen der Mobilität und die damals gängigen Praktiken zu deren Überwindung.

Auch wenn die Heatmap von Modell 15 aufgrund der hohen Trennschärfe und des klaren Musters auf den ersten Blick eine vermeintlich klare Aussage suggeriert, lässt sich die Zusammensetzung der einzelnen Topics nicht immer so einfach entschlüsseln und erfordert entsprechend im Sinne des Scalable Readings in gewissen Fällen eine Sichtung des zugrundeliegenden Quellentextes. Dieser unter dem Hintergrund inhaltlicher Aspekte entstandene Eindruck lässt sich teilweise auch damit begründen, dass in vielen Topics von Modell 15 Tokens mit einer niedrigen Frequenz und reduzierter Aussagekraft vorkommen. Es handelt sich hierbei um einen Effekt, der in Zusammenhang mit der Wahl der Anzahl an Topics steht. So führt die erwähnte Ausdifferenzierung der Mo-

delle mit steigender Anzahl an Topics allgemein dazu, dass die Auftretenswahrscheinlichkeit einzelner Topics abnimmt und vermehrt Begriffe mit niedrigerer Frequenz vorkommen.

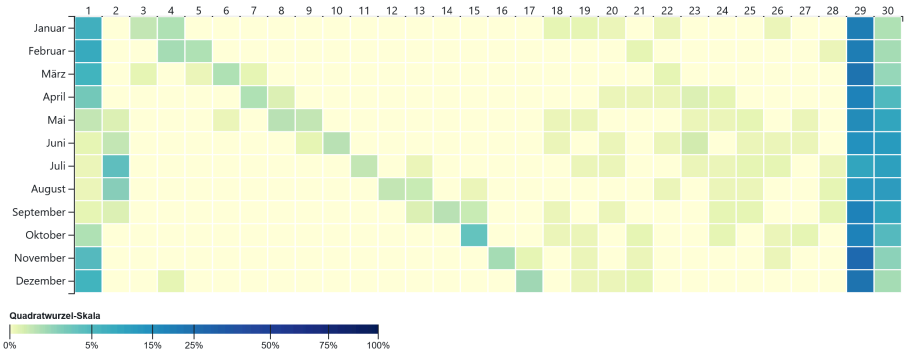

Abb. 2. Segmentierung pro Monat kumuliert, Modell 30. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, Stand: 30.06.2023.

Eine höhere Anzahl an Topics führt dazu, dass die Auftretenswahrscheinlichkeiten sich über mehr Felder verteilen und entsprechend abnehmen, was sich in den Heatmaps visuell durch ein Ausbleichen erkennbar ist. Neben diesem Effekt führt eine höhere Anzahl an Topics auch dazu, dass tendenziell mehr Topics mit einer äusserst tiefen Auftretenswahrscheinlichkeit und einer geringeren Trennschärfe generiert werden. Ein Vorkommen dieser Topics ist zwar in mehreren Monaten möglich, allerdings liegt die Wahrscheinlichkeit bei den einzelnen Monaten durchgehend unter fünf Prozent. Während dies bei Modell 15 lediglich auf Topic 14 zutrifft, sind es bei Modell 20 deren drei (Topics 17-19), bei Modell 25 sieben (Topics 17-23) und bei Modell 30 (Abb. 2) insgesamt elf (Topics 18-28). Diese Zahlen unterstützen den Eindruck, welche eine oberflächliche Betrachtung der Heatmaps vermittelt. Abgesehen von der Zunahme der wenig trennscharfen und wahrscheinlichen Topics verändert sich das grundlegende Muster von Modell 15 bis 30 nur marginal.

Eine eingehendere Untersuchung der internen Topic-Zusammensetzungen ergibt, dass sich diese unspezifischen Topics aus Tokens mit geringer Wortfrequenz bilden. Die beiden häufigsten Begriffe in Topic 14 von Modell 15 sind „ohrt“ und „nunn“, welche in der Kombination mit den übrigen Tokens lediglich sechsmal vorkommen; die weiteren Termini erscheinen zwischen drei- und fünfmal. Bereits minimale Einflüsse wie Verschreibungen des Autors, Transkriptionsfehler, unerwünschte Effekte bei der Tokenisierung und Ähnliches

können in dieser Grössenordnung zu Verschiebungen in der Zusammensetzung führen. Da beim Modellierungsprozess alle Tokens einem oder mehreren Topics zugeordnet werden, bilden die für die Analyse berücksichtigten 20 Begriffe zudem nur diejenigen mit der höchsten Frequenz innerhalb der Topics ab. Die Betrachtung der gesamten Liste zu Topic 14 zeigt, dass es neben den fünf aufgeführten Tokens mit dreimaligem Auftreten noch 31 weitere Begriffe gibt, die in diesem Verbund ebenfalls dreimal vorkommen. Obwohl die Anordnung auf statistischen Prinzipien beruht, entbehrt diese hinsichtlich der geringen Grössenordnung nicht eines gewissen Masses an Arbitrarität.

Auch bei anderen Modellen fällt auf, dass die erwähnten Topics mit geringer Auftretenswahrscheinlichkeit und Trennschärfe aus Tokens mit tiefer Frequenz bestehen, so beispielsweise bei den Topics 18 bis 28 in Modell 30. Am ehesten nutzbar erscheint hier Topic 23, das in den Monaten April bis Juli zu einem bis vier Prozent wahrscheinlich ist. Anhand der Begriffe „bluest“ und „früchten“ lässt sich dezidiert auf das Thema Vegetationsentwicklung im Frühling schliessen, wobei „kein_tropfen“, „watten“ (infolge nasser Wege) und „gestrahlet“ (Blitz) auf Wetter- und Witterungsbedingungen hinweisen. Die genannten Tokens befinden sich in Gesellschaft mit schwierig zu deutenden Funktionswörtern wie „här“ und „hinzu“, Adjektiven wie „hochheilig“ oder Substantiven wie „praelat“ und „zulauf“. Auch wenn der Nutzen solcher Topics mit geringerer Auftretenswahrscheinlichkeit aus inhaltlicher Sicht fragwürdig ist, bieten sie dennoch einen Gradmesser für die Einschätzung, ab wie vielen Topics eine zu starke Ausdifferenzierung erfolgt. Während die Modelle bis 20 Topics überwiegend trennscharfe Auftretenswahrscheinlichkeiten abbilden, werden in den Modellen ab 25 Topics überwiegend Topics mit äusserst geringen Werten kreiert. Allerdings sind Topics mit geringerer Wortfrequenz ebenfalls in Modellen, die trennscharfe Muster aufweisen, möglich. Aus diesem Grund müssen für eine Beurteilung eines Modells und eines Topics immer auch die Wortfrequenzen berücksichtigt werden. Trotz Differenzen bei den unterschiedlichen Formen der Segmentierungen unterschiedlich ausfallen, deuteten die bisherigen Ergebnisse darauf hin, dass bei der geringen Datenmenge Modelle zwischen fünf und 30 Topics einen sinnvollen Bereich abdecken.

Aus inhaltlicher Perspektive zeigen die bisherigen Resultate, dass mit der gewählten Art der Segmentierung Differenzen zwischen Jahreszeiten und Monaten herausgearbeitet werden konnten. Diese gründen allerdings nicht nur auf witterungsbedingten Unterschieden, sondern konstituieren sich auch über landwirtschaftliche oder rituelle Praktiken sowie über indirekt mit der Witterung in Bezug stehenden Indikatoren, wie beispielsweise der Vegetationsent-

wicklung. Je nach Modell prägen diese Bereiche die Topics in unterschiedlichem Masse. Allerdings lässt sich annehmen, dass die Zusammensetzung der Topics bis zu einem gewissen Grad auch von ortsspezifischen Faktoren abhängt. Dies zeigt sich etwa am erwähnten Beispiel mit den Tokens zu kirchlichen Feiertagen, welche nicht nur, aber schwerpunktmässig in Einsiedeln gefeiert wurden. Ein weiterer Hinweis darauf ist, dass das Winter-Topic, welches sich in allen Modellen an erster Stelle befindet, auch erhöhte Auftretenswahrscheinlichkeiten von Oktober bis Mai aufweist. Dies passt tendenziell eher zu den klimatischen Bedingungen in Einsiedeln als denjenigen in den tiefergelegenen Gebieten, weshalb im Folgenden auf Basis einer anderen Art der Segmentierung ortsspezifische Unterschiede analysiert werden.

3.2 Segmentierung pro Beobachtungsort und Jahreszeit kumuliert

Da bei der vorliegenden Art der Segmentierung mit den Beobachtungsorten und Jahreszeiten zwei Attribute kombiniert werden, wird die Lesbarkeit der Heatmaps erschwert. Als Hilfsmittel für die Analyse wurden die Heatmaps deshalb in der Form konfiguriert, dass bei jedem Modell zwischen zwei Arten der Sortierung gewählt werden kann. So können die Resultate einerseits nach Ortschaften angeordnet werden, womit besser erkennbar ist, welche Topics sich vordergründig auf die Bedingungen und Praktiken in einem bestimmten Gebiet beziehen. Mit der Sortierung nach Jahreszeiten werden hingegen Topics, die ortsübergreifende Gemeinsamkeiten in bestimmten Jahreszeiten aufzeigen, hervorgehoben, was Abbildung 3 veranschaulicht.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.16

Abb. 3. Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 5. Die linke Heatmap bildet die Sortierung nach Beobachtungsort ab, die rechte Darstellung die Sortierung nach Jahreszeit. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, Stand: 30.06.2023.

Auch bei der vorliegenden Art der Segmentierung zeigen sich bei den Modellen diverse Unterschiede, weil sich bestimmte Effekte erst ab einer gewissen Zahl an Topics zeigen. Während bei Modell 5 tendenziell Topics mit allgemein hoher Auftretenswahrscheinlichkeit in mehreren Einheiten vorkommen, werden mit höherer Anzahl zunehmend Topics gebildet, die spezifisch auf einzelne Felder zutreffen. In Modell 20 (Abb. 4) weist Topic 13 beispielsweise eine hohe Wahrscheinlichkeit (15%) für den Herbst in Einsiedeln und eine geringe (1%) für den dortigen Sommer auf. Das Topic enthält einerseits wetterbezogene Begriffe wie „vöhn“, „sehr_kalt“, „reifen“ und „sehr_warm“, die vor allem bezüglich der Temperatur auf das mögliche Spektrum in dieser Übergangszeit hinweisen. Die Tokens „solemnitet“, „priester“, „recreation“, „jahrzeit“, „engelweyhung“ und „recreationem“ widerspiegeln hingegen Aspekte des monastisch-rituellen Lebens, die vordergründig in der Gemeinschaft des Mutterklosters Einsiedeln intensiv zelebriert wurden. Je nach Modell ist es somit möglich, orts- und jahreszeiten-

spezifische Eigenheiten in den Topics sichtbar zu machen. Der beschriebene Effekt zeigt sich allerdings erst in verstärkter Form ab Modell 15.[51]

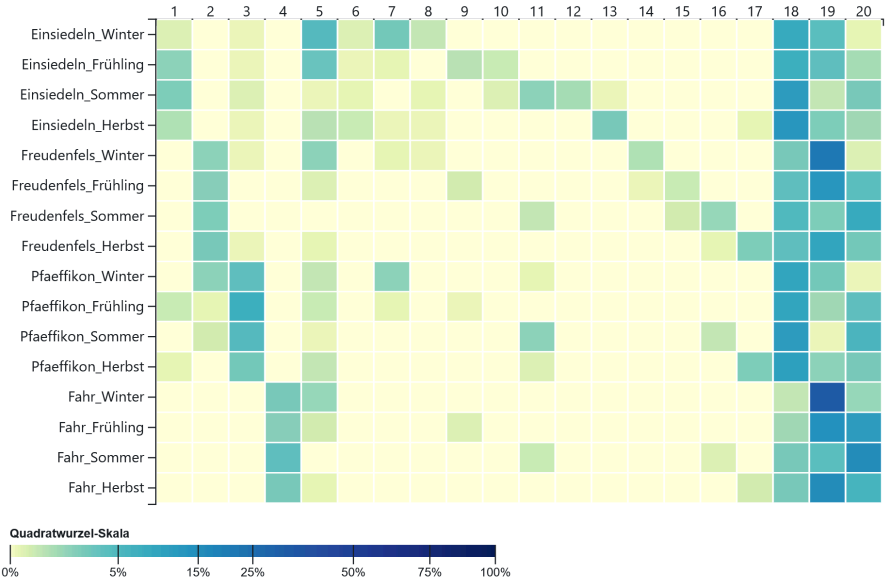

Abb. 4. Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 20. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, Stand: 30.06.2023.

Im Gegensatz zur vorherigen Analyse, wo die Unterschiede zwischen den Modellen stark im Zentrum standen, werden im Folgenden andere Tendenzen untersucht. Es geht weniger um die Frage, welches Modell welches Muster wie stark abbildet, als vielmehr darum, welche Effekte sich modellübergreifend zeigen und inwiefern sich daraus Schlüsse zum Einfluss der Beobachtungsorte auf die Zusammensetzung der Topics ziehen lassen. Im Wesentlichen konnten bezüglich dieser Frage drei relevante Tendenzen ausgemacht werden. Erstens werden Topics gebildet, die erhöhte Auftretenswahrscheinlichkeiten für einen bestimmten Beobachtungsort aufweisen, was bei den ersten vier Topics in Modell 20 gut ersichtlich ist. Topic 1 zeigt erhöhte Werte für die Jahreszeiten Frühling (12%), Sommer (14%) und Herbst (14%) sowie eine niedrige Wahrscheinlichkeit für den Winter (3%) in Einsiedeln. Letzteres ist damit zu begründen, dass die Begriffe vor allem Witterungsverhältnisse („tunderwetter“, „regenwet-

51 Das beschriebene Topic 13 in Modell 20 weist hohe Ähnlichkeiten mit Topic 9 in Modell 15 auf. In Modell 10 konnte keine Entsprechung gefunden werden.

ter“, „frischer“) sowie Landwirtschafts- und Klosterpraktiken im Sommer und in den Übergangsmonaten abbilden. Der Bezug zu Einsiedeln offenbart sich hier vor allem über die beiden letztgenannten Kategorien. So stehen die Wörter einerseits mit prägenden Elementen des rituellen Lebens („procession“, „vesper“, „volk“, „gottshauß“, „kloster“) in Zusammenhang, andererseits weisen sie auf die in Einsiedeln vorherrschende Landwirtschaftspraxis der Viehwirtschaft („heüw“, „veych“, „graß“, „matten“, „feld“) hin.

In Topic 2, das erhöhte Werte (12-15%) für Freudenfels aufweist, kommen nur wenige landwirtschaftsbezogene Tokens („haber“, „reeben“, „veld“), dafür viele regionale Ortsbegriffe („eschenz“, „cell“, „sonnenberg“, „clingenzell“) vor. Auffallend häufig sind hier Wörter im Zusammenhang mit den Windverhältnissen („still“, „wähete“, „rüehwig“, „wind“) und der Windrichtung („vnderluft“, „oberluft“). Topic 3 zeigt eine hohe Auftretenswahrscheinlichkeit (16-27%) für Pfäffikon, setzt sich allerdings aus Tokens mit tendenziell geringer Frequenz zusammen, was grösstenteils mit der geringen Datenmenge zu diesem Beobachtungsort begründet werden kann. Auch hierin dominieren Begriffe, die sich insbesondere auf die Umgebung und Lebenswelt („schif“, „bach“, „see“, „vfnauw“) beziehen. Das Schloss am Zürichsee lag nämlich neben einem Bach, wo Fischzucht („weyer“) betrieben wurde. Die Begriffe „brüel“ und „einsidlen“ deuten auf die Nähe und enge Verbindung zum Mutterkloster hin. Die Tokens in Topic 4, das ausschliesslich Wahrscheinlichkeiten (13-19%) für das Frauenkloster Fahr aufweist, heben ebenfalls das dortige geografische und klösterliche Umfeld („zürrich“, „limmet“, „klosterfrauwen“) hervor. Daneben finden sich aber auch viele Wörter mit Witterungsbezug („sonnenscheiniger“, „trüeber“, „milter“, „sonnenscheinig“, „sehr_heisßer“, „bezogener“, „tröpflen“, „gewulket“).

Da die beschriebenen Topics vordergründig lebensweltliche und geografische Charakteristiken der einzelnen Beobachtungsorte abbilden, welche problemlos auch mit anderen Methoden erschlossen werden können, stellt sich die Frage nach dem Nutzen. Dieser liegt darin, dass sich beim Vergleich der standortspezifischen Topics unerwartete Muster zeigen, wobei weiterführende Analyse zu neuen Erkenntnissen führen können. Dies wird im Folgenden am Beispiel der auffallend häufigen Begriffe mit Bezug zu Wind in Topic 2 verdeutlicht. Zu diesem Zweck wurden die Daten der Beobachtungsorte mit Voyant Tools, welches erweiterte Möglichkeiten der Textanalyse bietet, untersucht. Anhand der relativen Frequenz (Abb. 5) der im Topic vorkommenden Begriffe „still“, „wähete“, „rüehwig“, „wind“, „vnderluft“ und „oberluft“ zeigt sich, dass alle diese Wörter im Verhältnis zur Datenmenge pro Beobachtungsort in Freudenfels

durchgehend häufiger vorkommen als in den Beschreibungen zu den anderen Orten.

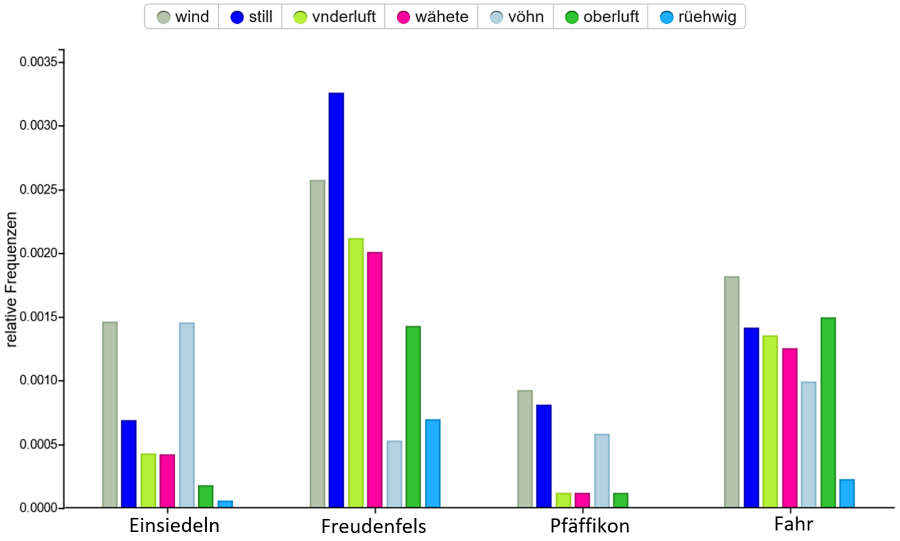

Abb. 5. Relative Frequenzen der Begriffe „wind“, „still“, „vnderluft“, „wähete“, „vöhn“, „oberluft“, und „rüehwig“ pro Beobachtungsstandort. Die Darstellung wurde mit Voyant Tools erstellt.

Der Befund weist darauf hin, dass Dietrich die entsprechenden Begriffe in Freudenfels im Vergleich zu anderen Ortschaften häufiger verwendete. Dies könnte damit begründet werden, dass er in Freudenfels aufgrund der exponierten Lage des Schlosses auf einem Hügel den Wind besser wahrnahm und so auch die Windrichtung genauer bestimmen konnte. Dass er in Freudenfels auch viel das Fehlen von Wind („still“) beschrieb, zeugt ebenfalls von einer höheren Sensibilität für Luftbewegungen. Somit ist es möglich, dass er seinen Dokumentationsstil auf die jeweiligen Begebenheiten anpasste, was im Hinblick auf Analysen zu den Witterungsbedingungen auf Basis seiner Angaben relevant sein kann. In diesem Zusammenhang ist eine tieferreichende Auseinandersetzung mit der Begriffsverwendung des Autors notwendig. So verfügte er über ein breites Vokabular für die Beschreibung der Windrichtung. Wird beispielsweise der Südwind („vöhn“) bei der Analyse in Voyant Tools mitberücksichtigt, ergibt sich die höchste relative Frequenz für Einsiedeln und die tiefste für Freudenfels. Dies ist ein Hinweis darauf, dass die Differenzen nicht nur auf eine temporäre und ortsabhängige Sensibilität, sondern auch auf unterschiedliche lokale Windverhältnisse zurückzuführen ist. Ohne dass an dieser Stelle die Analyse vertieft wird,

kann suggeriert werden, dass die Topics Muster andeuten können, deren Entschlüsselung mit anderen Methoden gewinnbringend sein kann.

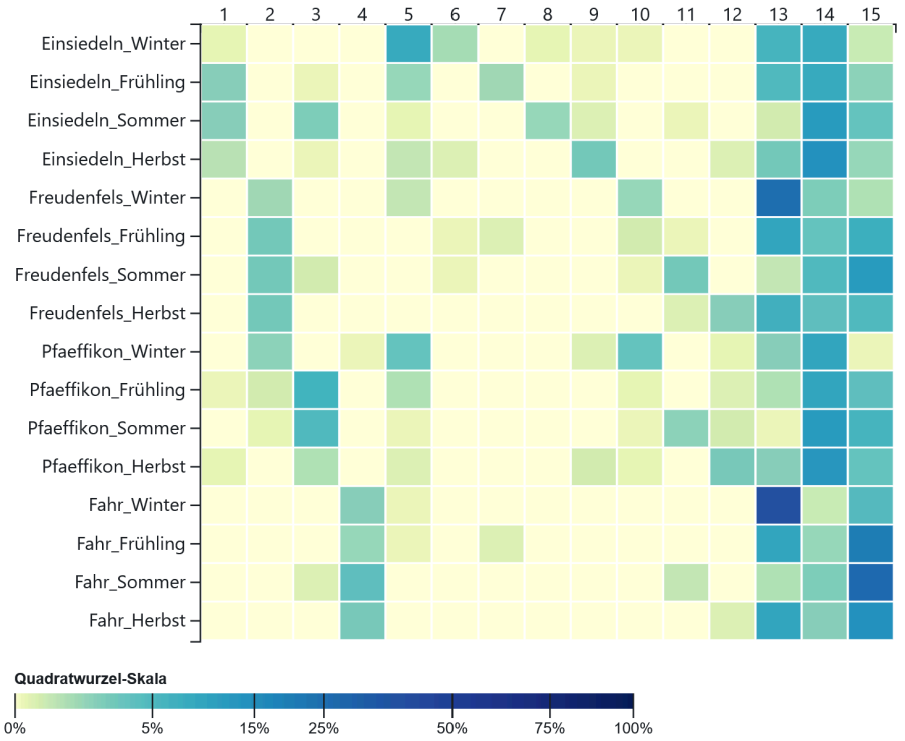

Abb. 6. Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 15. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, Stand: 30.06.2023.

Nach diesem Exkurs werden wiederum die allgemein erkennbaren Tendenzen bei der Modellierung der vorliegenden Art der Segmentierung thematisiert. Bei den behandelten Topics fällt im Weiteren auf, dass diese teilweise nicht nur für einen, sondern mehrere Orte erhöhte Auftretenswahrscheinlichkeiten aufweisen. Topic 2 in Modell 20, das erhöhte Werte für Freudenfels abbildet, ist beispielsweise auch in Pfäffikon im Winter zu einem gewissen Grad (12%) wahrscheinlich. Derselbe Effekt zeigt sich in Topic 5 von Modell 15 (Abb. 6), das höhere Auftretenswahrscheinlichkeiten in Einsiedeln (Winter 28%, Frühling 11%), Pfäffikon (Winter 18%, Frühling 8%) und zum Teil in Freudenfels (Winter 6%, Frühling 0%) abbildet, während die Werte für Fahr (Winter 1%, Frühling 1%) vernachlässigbar sind. Das Topic enthält Schreibvariationen der Begriffe „kalt“

(„kälte“, „kelte“) und „Schnee“ („schnee“, „schnees“, „schneelin“), was zwar den schwerpunktmässigen Bezug zum Winter, nicht aber die teilweise grossen Differenzen zwischen den Standorten erklärt.

Bei der Betrachtung der übrigen Tokens zeigt sich, dass es mehrere Wörter gibt, die in Bezug zu Transport und Mobilität („schlitten“, „strasßen“, „ochßen“, „mennweeg“ „weeg“) gesetzt werden können. Da die Versorgung des auf 880 Metern gelegenen Klosters mit den Erzeugnissen aus der unmittelbaren Umgebung keineswegs gedeckt werden konnte, wurden Lebensmittel und andere Güter häufig von den Aussenstationen nach Pfäffikon und von dort nach Einsiedeln geführt. Im Winter und teilweise auch im Frühling konnten die von Pferden oder Ochsen gezogenen Waren auf Schlitten verladen werden, was im Vergleich zu Wagen eine einfachere Transportmöglichkeit darstellte. Unter diesem Hintergrund lässt sich die Schlussfolgerung ziehen, dass die äusserst häufig auftretenden Varianten von „schnee“ weniger einen Hinweis auf Witterungsphänomene im Winter liefern, als vielmehr im Zusammenhang mit Transportpraktiken zu lesen sind. Damit erklären sich auch die tieferen Werte für Freudenfels, wo nur sporadisch Waren abgeholt wurden, und dem Kloster Fahr, wo der Gütertransport in anderer Form organisiert wurde. Im Topic zeigen sich somit nicht nur ortsspezifische Bräuche im Kontext einzelner Jahreszeiten, sondern gebietsübergreifende und teilweise miteinander in Beziehung stehende Praktiken. Es erfordert jedoch ein gewisses Mass an Hintergrundwissen, diese Praktiken anhand einzelner Tokens erkennen zu können.

Diese Erkenntnisse sind vor allem im Hinblick auf die Interpretation eines dritten Musters, welches bei der vorliegenden Art der Segmentierung erkennbar ist, relevant. Es handelt sich hierbei um diejenigen Topics, die in einer bestimmten Jahreszeit für alle Beobachtungsorte eine erhöhte Auftretenswahrscheinlichkeit aufweisen und sollte – so zumindest die Erwartung – vordergründig Witterungsphänomene abbilden. Dieses Muster wird in den Heatmaps bei einer Sortierung nach Jahreszeiten besser sichtbar. Bei Topic 3 in Modell 10 (Abb. 7) treten erhöhte Werte an allen vier Beobachtungsorten in der Jahreszeit Winter (Einsiedeln 31%, Freudenfels 23%, Pfäffikon 16%, Fahr 17%) auf. Es setzt sich weitgehend aus Begriffen zur Temperatur („kälte“, „kalt“, „sehr_kalt“, „sehr_kalter“, „milt“, „milter“, „kalter“, „nit_sonders_kalt“) zusammen, enthält aber auch weitere Wörter im direkten oder indirekten Zusammenhang mit der Witterung und Himmelsbedeckung („schnee“, „vöhn“, „hell“ „wind“, „schneelin“, „heller“, „hellem“, „schneyen“) sowie der Mobilität („strasßen“, „strasß“).

Abgesehen von der letztgenannten Kategorie handelt es sich um ein Topic, dass wenig von lebensweltlichen Aspekten wie der Landwirtschaft, Transportpraktiken oder rituellen Handlungen beeinflusst wird und somit im Hinblick auf die Witterungsverhältnisse aufschlussreich ist. Die Differenzen bei den Auftretenswahrscheinlichkeiten des Topics an den verschiedenen Ortschaften weisen auf bestimmte regionale Tendenzen hin, sollten aufgrund der ungleichen Datenmengen zu den einzelnen Ortschaften, diversen Einflussfaktoren beim Modellierungsprozess und individuellen Eigenheiten des Dokumentationsstils des Autors aber nicht als absoluter Gradmesser verstanden werden.[52]

52 Das hier behandelte Topic zeigt sich in leicht veränderter Zusammensetzung auch in anderen Modellen, wobei die Auftretenswahrscheinlichkeiten variieren. Individuelle Eigenheiten des Dokumentationsstils können beispielsweise Veränderungen in der Wortwahl sein, die sich über die Zeit ergeben. Dieser Aspekt wird in den nachfolgenden Analysen thematisiert.

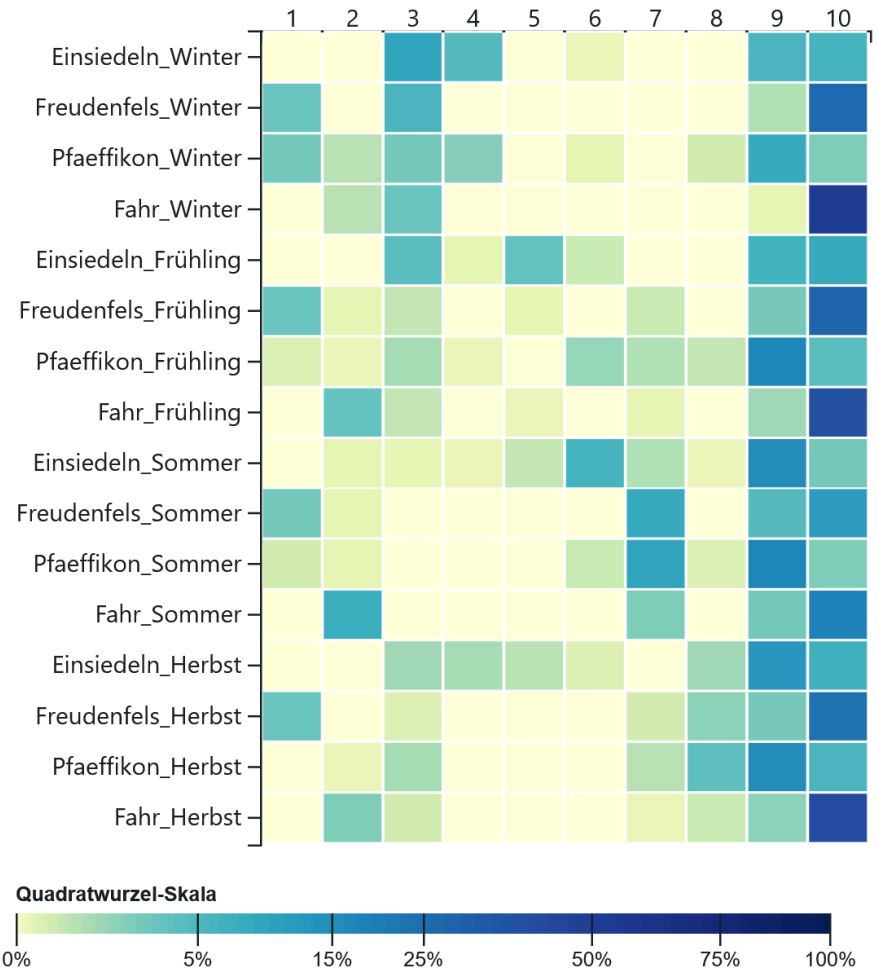

Abb. 7. Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 10. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, Stand: 30.06.2023.

Der hohe Wert für Einsiedeln mag zum Teil mit den dortigen klimatischen Bedingungen erklärt werden, die abgebildeten Zahlen sind jedoch das Produkt unterschiedlicher Faktoren und es ist nicht gänzlich auszuschliessen, dass bestimmte ortsspezifische Ausprägungen das Resultat verzerren. Dennoch zeigt das Muster in der Heatmap einige Tendenzen, die auf eine Eignung für die Unterscheidung regionaler klimatischer Bedingungen sprechen. So weist das Topic auch in den Jahreszeiten Frühling (Einsiedeln 20%, Freudenfels 6%, Pfäffi-

kon 9%, Fahr 6%) und Herbst (Einsiedeln 10%, Freudenfels 3%, Pfäffikon 9%, Fahr 4%) Auftretenswahrscheinlichkeiten auf. Hierin widerspiegelt sich die Tatsache, dass Schnee und Kälte bis weit in den Frühling und ab dem späteren Herbst zu jener Zeit allgemein keine Seltenheit und speziell in Einsiedeln normal waren.

Diese nach Orten differenzierte Aufschlüsselung lässt sich zu den Analysen, die ihm Rahmen der vorherigen Art der Segmentierung vorgenommen wurden, in Bezug setzen. Hier fiel auf, dass sich das in allen Modellen an erster Stelle befindliche Winter-Topic jeweils auch auf die Übergangsmonate erstreckte. Obwohl in der vorliegenden Untersuchung aus Gründen der Lesbarkeit Jahreszeiten statt Monate als Einheiten gewählt wurden, zeigt sich im Allgemeinen derselbe Effekt. Allerdings tritt dieser in Einsiedeln wesentlich stärker hervor als in den anderen Orten. Dies wird höchstwahrscheinlich einer der Gründe sein, warum sich das Winter-Topic in der vorherigen Analyse auch stärker in den Übergangsmonaten abzeichnete.

Die beiden Beispiele zeigen, dass sich auf Basis von Dietrichs Wetterbeobachtungen mit Hilfe von Topic Modeling ortsspezifische Heatmaps erzeugen lassen, welche grosse Ähnlichkeiten zu den Diagrammen mit durchschnittlichen Monatstemperaturen auf Grundlage von Messungen aufweisen. Allerdings ist zu beachten, dass sich in den bisherigen Analysen vor allem Topics mit einem klaren Bezug zu Kälte bildeten. Durch diese können zwar die Temperaturverläufe in den Winter- und teilweise auch in den Übergangsmonaten ansatzweise nachgezeichnet werden, die Sommermonate treten allerdings nur aufgrund des Fehlens von Kälte in Erscheinung. Für eine stärkere Differenzierung bräuchte es folglich ein Topic, das sich vornehmlich auf die Wärme bezieht. Bei der vorliegenden Datengrundlage artikulieren sich die sommerbezogenen Topics jedoch stärker über landwirtschaftliche und kulturelle Praktiken, womit sie sich nur bedingt für witterungsspezifische Analysen eignen.

An dieser Stelle werden die geäusserten Überlegungen nicht weiter vertieft. Die vorliegende Analyse hat gezeigt, dass ortsspezifische Faktoren einen starken Einfluss haben können, wobei sich dieser in unterschiedlicher Form artikulieren kann. Da die Daten jeweils nach Monaten oder Jahreszeiten kumuliert wurden, fehlte bis jetzt die zeitliche Perspektive. Entsprechend konnte der Einfluss bestimmter Faktoren, wie beispielsweise Änderungen im Dokumentationsstil des Autors, nur vermutet werden. Aus diesem Grund werden in der nachfolgenden Art der Segmentierung Tendenzen auf zeitlicher Ebene untersucht.

3.3 Segmentierung pro Jahr über den Gesamtzeitraum

Bei der Segmentierung pro Jahr werden weniger monatliche oder jahreszeitliche Unterschiede als vielmehr Entwicklungen über den gesamten Zeitraum sichtbar gemacht. Bereits in der
Heatmap von Modell 5 (Abb. 8) zeigen sich deutlich erkennbare Muster. So weist Topic 3 erhöhte Auftretenswahrscheinlichkeiten für diejenigen Jahre auf, in denen sich Dietrich in Freudenfels aufhielt, und für seine kurze Präsenzzeit in Pfäffikon. Die Übereinstimmungen der beiden Standorte lassen sich über Begriffe zur Lage („see“) und Mobilitätspraxis („schif“, „reitete“) sowie zum Weinbau („reeben“, „wimmlen“, „reebstok“) erklären. Weitere Wörter betreffen die Landwirtschaftspraxis („haber“, „veld“, „garben“) und die Windverhältnisse („still“, „wähete“, „luft“) in Freudenfels, was die bereits in der vorangegangenen Analyse aufgezeigte spezifische Prägung von Elementen der Witterung oder Art der Witterungsbeschreibung an diesem Ort unterstreicht. Topic 1 weist eine äusserst hohe Auftretenswahrscheinlichkeit in den Jahren 1678 bis 1681 (42-70%) auf, welche von 1682 (28%) bis 1689 (1%) kontinuierlich abnimmt und in den folgenden Jahren bei null verharrt. Zwischen 1683 und 1688 kommt hingegen Topic 2 mit höherer Wahrscheinlichkeit vor, wobei sich danach immer wieder Jahre mit tieferen Werten zeigen. Diese Lücken fallen grösstenteils auf diejenigen Zeiträume, in denen sich der Autor ausserhalb von Einsiedeln aufhielt. Somit handelt es sich bei den Topics 1 und 2 um Wortketten mit starkem Bezug zum Beobachtungsort Einsiedeln, die den Beobachtungszeitraum in zwei Teile untergliedern und sich in den 1680er Jahren sukzessive ablösen.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.16

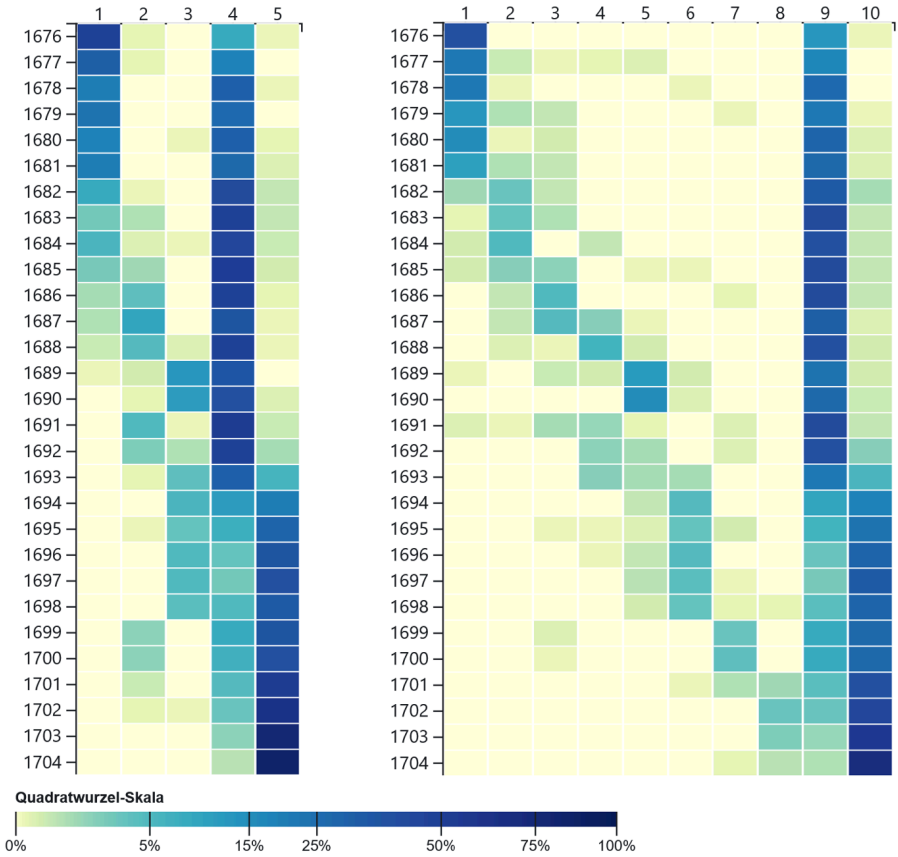

Abb. 8. Segmentierung pro Jahr über den Gesamtzeitraum, Modelle 5 und 10. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, Stand: 30.06.2023.

Inhaltlich ist vor allem Topic 1 schwer zu interpretieren, was unter anderem an den vielen Begriffen mit tiefer Frequenz liegt. Die Ursache für dieses komplementäre Clustering scheint allerdings weniger mit inhaltlichen als vielmehr mit formalen Aspekten zusammenzuhängen. Diese lassen sich durch einen zeitlichen Vergleich der Schreibweise einzelner Begriffe nachvollziehen, wofür im vorliegenden Fall mit Hilfe von Voyant Tools[53] die relativen Frequenzen (Abb. 9) ausgewertet wurden. Für den Begriff „Vieh“ benutzte Dietrich im Zeitraum

53 Aufgrund seines Todes schrieb Dietrich im Jahr 1704 nur bis zum 19. März, weshalb die Werte für dieses Jahr in den mit Voyant Tools erstellten Grafiken nicht berücksichtigt wurden.

zwischen 1681 und 1685 ausschliesslich die Schreibvariante „veych“, bevor er in letztgenanntem Jahr dazu überging, auch die Variante „vych“ zu verwenden. Die erstgenannte Form kommt zwar bis 1687 vor, erscheint aber ab 1686 weniger häufig als die zweite, welche bis zum Ende des Tagebuchs anzutreffen ist. Dieselbe Tendenz zeigt sich bei den beiden Schreibvarianten „continuirt“ (Topic 1) und „continuiert“ (Topic 2), die der Autor häufig im Zusammenhang mit der Beschreibung gleichbleibender Witterung benutzte. Während die erste Form bis 1685 vorherrschend ist und 1688 zum letzten Mal vorkommt, setzt die Verwendung der zweiten ab 1685 ein, wobei sie Dietrich ab 1689 ausschliesslich benutzte. Auffallend ist auch die Schreibweise „regen“ in Topic 1, welche nicht in Topic 2, sondern in den Topics 4 und 5 als „reegen“ sehr häufig auftaucht. Hier findet der Übergang von einer zur anderen Schreibweise bereits früher statt. Die Variante „regen“ ist nämlich bis 1680 vorherrschend, und kommt danach trotz weiterer Verwendung im Vergleich zur gedehnten Form deutlich weniger oft vor.

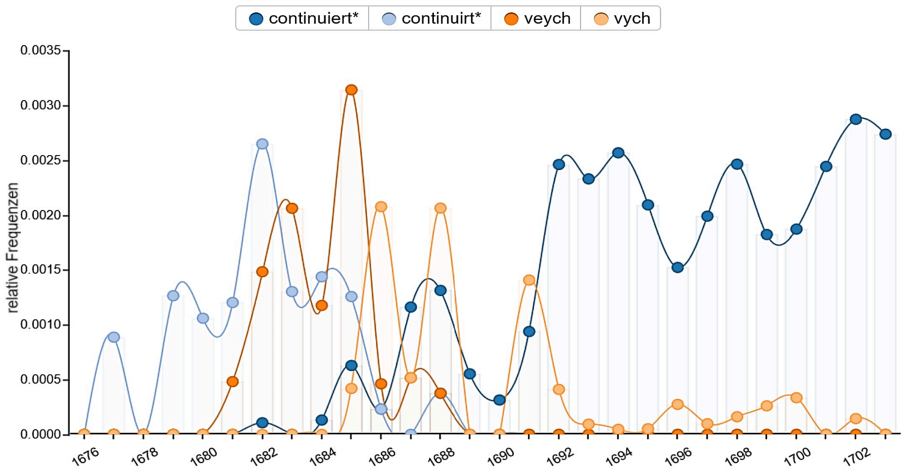

Abb. 9. Relative Frequenzen der Begriffe „contiuiert*“, „continuirt*“, „veych“, und „vych“ über den Gesamtzeitraum. Die Sterne bedeuten, dass alle möglichen Deklinationsformen berücksichtigt wurden. Die Darstellung wurde mit Voyant Tools erstellt.

Anhand der Beispiele lässt sich somit ansatzweise eine Veränderung der Orthografie erkennen. Obwohl sich Mitte der 1680er Jahre ein Übergang erkennen lässt, artikuliert sich die adaptierte Schreibpraxis nicht in einem klaren Bruch, sondern in einem Übergangsprozess mit teilweise parallel existierenden Schreibvarianten. Dieser Befund basiert hier zwar auf wenigen Beispielen, stimmt jedoch mit bereits vorweg gemachten Feststellungen überein. So fiel bei der Lektüre des Tagebuchs auf, dass Dietrich bis 1684 ausschliesslich die Form

„vnd“ und später nur noch die Variante „vnnd“ verwendete. Als Übergangspunkt lässt sich seine Reise an die Frankfurter Büchermesse ausmachen. Neben diesem Einzelereignis wird ihn wohl auch seine Tätigkeit als Direktor der Einsiedler Stiftsdruckerei (1674-1680) zu Reflexionen zur Orthografie angestachelt haben.

Das Beispiel unterstreicht, dass mit Hilfe von Topic Modeling nicht Themen im engeren Sinn, sondern Muster generiert werden, die teilweise erst durch den Einsatz komplementärer Ansätze verständlich werden und als Ausgangspunkt für weiterführende Analysen dienen können. Im vorliegenden Fall offenbarten sich interessante Muster, die vor allem unter dem Hintergrund sprachwissenschaftlicher Untersuchungen zur individuellen Schreibpraxis am Übergang vom Früh- zum Neuhochdeutschen und vor der allgemeinen Sprachnormierung interessant sind, wobei auch dialektologische Gesichtspunkte eine Rolle spielen. An dieser Stelle werden die genannten Aspekte nicht vertieft behandelt, sondern die methodischen Grundlagen, die zu diesem Ergebnis geführt haben, kritisch reflektiert. Dass die beschriebenen Muster in dieser Form erkennbar sind, ist unter anderem auf das gewählte Optimierungsintervall zurückzuführen. Dadurch werden tendenziell trennscharfe Topics generiert, deren Vorkommen stärker über die Bezüge zwischen den einzelnen Tokens als über die Auftretenswahrscheinlichkeit des Topics im Gesamtkorpus begründet wird. Obwohl die Wortfrequenzen der Tokens in Topic 1 und 2 überschaubar bleiben, scheint zwischen den Tokens ein statistisch starker Zusammenhang zu bestehen, weshalb die beiden Topics bereits bei Modell 5 erscheinen.[54]

Abgesehen von der Erkenntnis, dass mit Hilfe eines starken Optimierungsintervalls die Bildung distinktiver Muster begünstigt wird, ist für weiterführende formale Analysen die Wahl der Stoppwörter zu überdenken. Im vorliegenden Fall wurden viele Funktionswörter in mehreren möglichen Schreibvarianten ausgeschlossen, weshalb nur spekuliert werden kann, ob sich die Muster bei einer weniger aggressiven Ausschlusspraxis überhaupt bemerkbar gemacht hätten oder ob sie noch stärker hervorgetreten wären. Im letztgenannten Fall wären idealerweise noch mehr Begriffe im Zusammenhang mit der beschriebenen Tendenz aufgetaucht. Auch wenn Stichwortabfragen ohne vorheriges Topic Modeling durchführbar sind, können die Topics – wie im beschriebenen Beispiel – darauf hinweisen, bei welchen Begriffen ein Zusammenhang zwischen Schreibvarianten und zeitlichen Veränderungen möglich sind.

54 Die geringe Datenmenge dürfte der Grund sein, weshalb die Auftretenswahrscheinlichkeit in den frühen Jahren dermassen hoch ist.

Die Analysemöglichkeiten bei der vorliegenden Art der Segmentierung erschöpfen sich nicht nur auf Aspekte der Orthografie. So erscheint beispielsweise in Topic 1 der Begriff „manns_gedenken“ (Menschengedenken), welchen Dietrich für die Betonung der Schwere von Naturkatastrophen und ausserordentlichen Witterungsphänomenen sowie damit in Zusammenhang stehenden Erscheinungen (z.B. Auswirkungen auf Ernteerträge, die Preisentwicklung, Pegelstände von Gewässern oder die Schneehöhe) benutzte. Die Redewendung „seit Menschgedenken“ suggeriert auf den ersten Blick, dass ein für einen Zeitraum von mehreren Jahrzeiten einmaliges und somit extremes Ereignis stattfand. Obwohl es in einzelnen Fällen zutreffen kann, verwies Pfister darauf, dass das menschliche Erinnerungsvermögen in Bezug auf Witterungsphänomene in vormoderner Zeit eher kurz war und sich anhand von Quellenvergleichen zeigen lässt, dass häufig innerhalb weniger Jahre vergleichbare Ereignisse mit der Zuschreibung „seit Menschgedenken“ versehen wurden, womit sich deren Aussagekraft relativiert.[55]

Der Begriff „manns_gedenken“ kommt – neben den beiden einzeln auftretenden Varianten „mans gedenken“ und „mans gedenkhen“ – im Tagebuch bis ins Jahr 1687 insgesamt 18-mal vor, findet danach aber keine Verwendung mehr. Das Fehlen des Begriffs nach 1687 ist hier nicht auf das Ausbleiben der weiterhin auftretenden und von Dietrich beschriebenen Extremen zurückzuführen, womit es naheliegt, dass der Autor diesen Begriff aus anderen Gründen nicht mehr verwendete. Ausgehend von diesem Hinweis wäre es aus Perspektive der Wissensgeschichte zum Klima interessant zu erörtern, inwiefern es sich bei dieser Anpassung des Dokumentationsstils um eine bewusste Entscheidung des Autors handelte, ob diese auf einer veränderten Naturwahrnehmung zurückzuführen ist und inwiefern sich die Hintergründe dazu erschliessen lassen. Abgesehen davon lässt sich anhand dieses einen Beispiels auch die Bedeutung von N-Grammen zeigen. Da der Autor den Terminus „gedenken“ in der Regel als Verb benutzte, wäre der hier geschilderte Zusammenhang ohne die vorherige Verknüpfung nicht erkennbar gewesen.

Im Weiteren sind auch die Topics 4 und 5 im Zusammenhang mit einem veränderten Schreibstil zu lesen. Topic 4 weist – abgesehen vom Jahr 1676 (28%) – bis 1693 durchgehend äusserst hohe Auftretenswahrscheinlichkeiten (42-73%) auf, welche danach bis 1703 grösstenteils auf einem hohen Niveau (12-33%) verbleiben, aber tendenziell rückläufig sind. Dahingegen sind die Werte (0-9%) von Topic 5 bis 1692 eher niedrig, erreichen danach aber sukzessive eine hohe Stufe

55 Vgl. Pfister, Wetternachhersage, S. 36.

(24-87%). Somit ist auch hier in Bezug auf den zeitlichen Verlauf bis zu einem gewissen Grad ein komplementärer Charakter der beiden Topics erkennbar, wobei der Bruch bei den Auftretenswahrscheinlichkeiten im Jahr 1693 zu verorten ist. Es handelt sich exakt um dasjenige Jahr, in welchem Dietrich von einer unregelmässigen zu einer täglichen Tagebuchführung überging. Daraus lässt sich die These ableiten, dass mit diesem Übergang auch eine Veränderung des Schreibstils einherging.

Obwohl sich die beiden Topics durchgehend aus eher allgemeinen Begriffen zusammensetzen, lassen sich bei einer genaueren Analyse der Wortfrequenzen in den entsprechenden Zeiträumen Indizien für die Unterstreichung der geäusserten These finden. Während der Begriff „wetter“ in Topic 4 häufiger vorkommt als „himmel“, verhält es sich bei Topic 5 umgekehrt. Die Darstellung der relativen Frequenzen (Abb. 10) zeigt, dass der Schnitt- oder Übergangspunkt exakt auf das Jahr 1693 fällt. Eine signifikante Zunahme bei der relativen Frequenz zeigt sich ab 1693 in Topic 5 auch bei den Wörtern „sonne“, „gewülk“ und „sonnenschein“ sowie teilweise auch bei „hell“ und „heller“. Dies weist darauf hin, dass sich Dietrich mit dem Übergang zur täglichen Berichterstattung bei der Beschreibung des Wetters stärker am atmosphärischen Zustand orientierte.[56] Zugleich nahm die Zahl der mit dem eher unspezifischen Begriff „wetter“ eingeleiteten Beobachtungen ab. Somit scheint der Übergang zum täglichen Tagebuchführen auch bis zu einem gewissen Grad mit einer spezifischeren Beobachtung und Beschreibung des Wetters einherzugehen, was auch mit dem subjektiv gewonnen Eindruck bei der Transkription der Wettereinträge übereinstimmt.

Obwohl die Beispiele die These, dass mit dem Übergang zum täglichen Schreiben auch Veränderungen beim Beobachtungsstil einherging, untermauern, ist ein weiterer Faktor für die Ausprägung der Auftretenswahrscheinlichkeiten in den beiden Topics zu berücksichtigen. So bezieht sich Topic 4 auf den Zeitraum, in welchem sich der Autor grösstenteils in Einsiedeln aufhielt. Darauf weisen insbesondere die häufig vorkommenden Tokens „schnee“ und „heüw“ hin. Topic 5 umfasst hingegen die Periode der häufigen Standortwechsel, wobei sich insbesondere der längere Aufenthalt in Freudenfels anhand des vermehr-

56 Ein Grund hierfür mag sein, dass der atmosphärische Zustand jeweils einfach mit einem Blick nach oben ermittelt werden konnte. Dahingegen ist es ohne Messungen und statistische Mittel schwierig, einen Tag unter der Berücksichtigung der natürlichen Schwankungen im Tagesablauf und seiner jahreszeitlichen Verortung als kalt, warm oder durchschnittlich zu klassifizieren.

ten Auftauchens und der höheren Frequenz von Wörtern mit Bezug zum Wind („luft“, „vnderluft“, „still“, „wind“, „wähete“) bemerkbar macht.

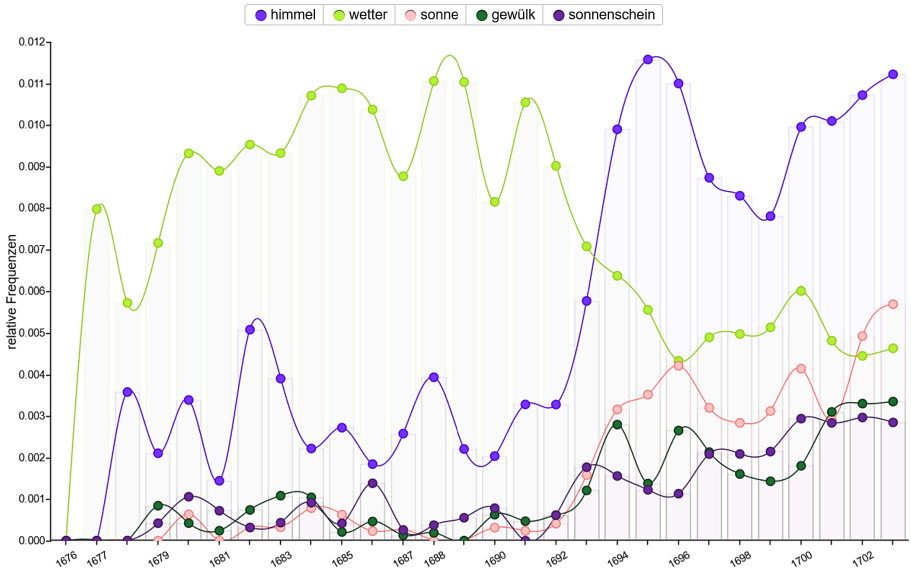

Abb. 10. Relative Frequenzen der Begriffe „himmel*“, „wetter*“, „sonne“, „gewülk“ und „sonnenschein“ über den Gesamtzeitraum. Die Darstellung wurde mit Voyant Tools erstellt.

Da die Modelle mit einer höheren Anzahl an Topics einen grösseren Differenzierungsgrad aufweisen, lassen sich mit ihnen einige der beschriebenen Effekte eingehender analysieren, was hier an einem Beispiel veranschaulicht wird. So existieren in Modell 10 (Abb. 8) zwei Topics, die erhöhte Auftretenswahrscheinlichkeiten für diejenigen Zeiträume aufweisen, in denen sich der Autor grösstenteils in Freudenfels aufhielt. Während bei Topic 5 äusserst hohe Werte (34-40%) für die Jahre 1689 bis 1690 erkennbar sind, zeigt sich dieses Muster bei Topic 6 teilweise für 1693 (9%) und stärker für die Periode von 1694 bis 1698 (18-21%). Dahingegen ist die Auftretenswahrscheinlichkeit (3-4%) von Topic 6 für die Jahre 1689 und 1690 gering. Dies deutet darauf hin, dass ein grundsätzlicher Unterschied zwischen den Beschreibungen des ersten und der späteren Aufenthalte in Freudenfels besteht. Bei der Gegenüberstellung der Tokens in den beiden Topics fällt auf, dass in Topic 5 viele Wörter mit Bezug zur Landwirtschaftspraxis und zu konkreten Ortschaften auftreten, aber keines zum Wind zu finden ist. Im Gegensatz dazu kommen windbezogene Tokens in Topic 6 („still“, „wähete“, „luft“, „rüehwig“, „vnderluft“, „oberluft“) häufig und

teilweise in hoher Frequenz vor, während Landwirtschaftsbegriffe tendenziell seltener sind. Zudem enthält Topic 6 einige Begriffe zum atmosphärischen Zustand („schön“, „himel“ (sic!), „hell“, „bedektem“), wohingegen diese oder ähnliche Wörter in Topic 5 fehlen. Dies ist ein Beleg dafür, dass der Bruch um 1693 weniger auf ortsspezifische Einflüsse als vielmehr auf den beschriebenen Stilwechsel zurückzuführen ist.

Insgesamt zeigt sich, dass bei der diachronen Betrachtung Muster in Erscheinung treten, die sich im Hinblick auf die Veränderungen der Orthografie des Autors als aufschlussreich erweisen und als Ausgangspunkt für weiterführende sprachwissenschaftliche Analysen dienen können. Dies ist nur deshalb möglich, weil die Texte nicht vorweg normalisiert wurden. Abgesehen davon offenbaren die Modelle auch Veränderungen im Schreibstil, wobei insbesondere der Übergang zur täglichen Berichterstattung deutlich hervortritt. Am Beispiel der Verwendung des Begriffs „manns_gedenken“ konnte exemplarisch vorgeführt werden, dass sich über die Topics und weiterführende Recherchen auch Hinweise zu wissensgeschichtlichen Aspekten im Zusammenhang mit Wetterextremen und Naturkatastrophen herstellen lassen. Ebenso lassen sich ortsspezifische Charakteristika wie der starke Windbezug in Freudenfels erschliessen.

4 Fazit und Ausblick

Im Rahmen der vorliegenden Arbeit wurde erörtert, inwiefern sich die Wetterbeobachtungen von Pater Joseph Dietrich mit Topic Modeling analysieren lassen, welche Ausgangspunkte sich für weiterführende Analysen ergeben und wie der Modellierungsprozess für eine erfolgreiche Anwendung konfiguriert werden muss. Insgesamt erfordert eine Anwendung von Topic Modeling eine Vielzahl an Entscheidungen, deren methodische Begründung und transparente Vermittlung im Hinblick auf wissenschaftliche Ansprüche eine grosse Herausforderung darstellt, weshalb die Was-wäre-wenn-Frage eine ständige Begleiterin beim Prozess ist. Dabei ist zu beachten, dass mit Hilfe von Topic Modeling keine direkten oder impliziten Wahrheiten generiert werden, sondern dass die Resultate nur Annäherungen abbilden, deren Sinn und Nutzen sich erst durch weiterführende Analysen erschliessen. Eine zielführende Anwendung von Topic Modeling bedingt somit eine iterative Auseinandersetzung mit dem Einfluss zugrundeliegender Modellierungsoptionen und der Interpretation der generierten Outputs.

Um diesem doppelten Anspruch gerecht zu werden, wurde in der vorliegenden Arbeit ein Ansatz gewählt, der beide Bereiche zusammenführt. So wurde im Gegensatz zu vielen anderen Studien die Zahl der zu modellierenden Topics weder mathematisch berechnet noch auf einen fixen Wert festgesetzt, sondern für jede Art der Segmentierung in einem vordefinierten Bereich ausgegeben, was einen direkten Vergleich der Veränderungen bei einer unterschiedlichen Zahl an Topics ermöglichte. Aus arbeitsorganisatorischer Sicht erwiesen sich in diesem Zusammenhang die individualisierbaren Notebooks in Observable als nützliches Instrument für die vergleichende Darstellung der vielen Modelle. Es zeigte sich, dass sich bei einer höheren Anzahl an Topics tendenziell eine Ausdifferenzierung in Form erhöhter Wahrscheinlichkeiten für einzelne Einheiten ergibt und dass ab einer gewissen Zahl vornehmlich Topics mit allgemein geringer Auftretenswahrscheinlichkeit und Trennschärfe generiert werden. Trotz dieses Effekts bleiben bestimmte Muster über die Modelle hinweg ähnlich, weshalb sich prinzipiell alle Modelle für weiterführende Interpretationen eigneten.

Da die vorliegende Datengrundlage einerseits stark chronologisch gegliedert und andererseits inhaltlich auf zyklische Phänomene im Jahresablauf ausgerichtet war, ergab sich die Möglichkeit einer doppelten Betrachtungsweise, die in Form synchroner und diachroner Segmentierungen für den Modellierungsprozess nutzbar gemacht wurde. Im Gegensatz zu anderen Studien wurde die Segmentierung nicht als zwingend zu definierender Parameter, sondern als Chance für eine multiperspektivische Analyse, mit Hilfe derer gezielt bestimmte Aspekte hervorgehoben werden können, verstanden. Während bei den synchronen Arten der Segmentierung der Fokus stärker auf den monatlichen, jahreszeitlichen und ortsabhängigen Bedingungen lag, wurde die diachrone Datengrundlage stärker für die Betrachtung individueller und längerfristiger Phänomene genutzt.

Diese Vorgehensweise ermöglichte eine Herausarbeitung der inhaltlichen, orthografischen und stilistischen Elemente, welche die sichtbaren Tendenzen bei den jeweiligen Modellen prägten. Gleichzeitig bildeten sie auch potenzielle Verzerrungsfaktoren auf die Resultate bei anderen Arten der Segmentierungen. So konnte beispielsweise anhand der Segmentierung pro Ortschaft und Jahreszeit der Einfluss ortsspezifischer Eigenheiten aufgezeigt werden. Dieser beeinflusste auch die Ergebnisse bei der kumulierten Segmentierung pro Monat, was dort allerdings nur bedingt nachvollziehbar war. Dass sich durch die unterschiedlichen Arten der Segmentierung zusätzliche Perspektiven erge-

ben, konnte exemplarisch anhand des Windbezuges in Freudenfels illustriert werden.

Die inhaltliche Interpretation der Heatmaps und der Zusammensetzung der Topics wurden im Sinne des Scalable Readings auf unterschiedlichen Ebenen umgesetzt. So führte die Frage nach der Bedeutung und dem Verwendungszweck einzelner Tokens immer wieder zurück zum Quellentext. Hierbei erwies es sich als Vorteil, dass die Entstehungszusammenhänge des Tagebuchs, die biografischen Hintergründe des Autors sowie die landwirtschaftlichen und kulturellen Rahmenbedingungen bereits bekannt waren. Im Weiteren zeigte sich, dass sich verschiedene orts- und zeitspezifische Differenzen mit anderen Methoden des Distant Reading eingehender untersuchen lassen. In diesem Zusammenhang erwiesen sich die Berechnungen und Visualisierungen der relativen Frequenzen mit Voyant Tools als einfacher und effizienter Ansatz, um Thesen, die ausgehend von den Topic-Modeling-Resultaten formuliert wurden, weiterzuverfolgen.

Insgesamt zeigte sich, dass die Anwendung von Topic Modeling auf die Wetterbeobachtungen von Pater Joseph Dietrich trotz der geringen Datenmenge eine Vielzahl an Ansatzpunkten für weiterführende Untersuchungen boten, insbesondere im Zusammenhang mit der Schreibpraxis des Autors sowie im Bereich sprachlicher und stilistischer Phänomene. Wie am Beispiel des Begriffs „manns_gedenken“ aufgezeigt wurde, können sich hieraus auch relevante Ergebnisse im Hinblick auf Fragen der Wissensgeschichte zu Klima und Naturkatastrophen ergeben. Weitere Resultate, wie beispielsweise die Ähnlichkeit der mit Topic Modeling erzeugten Heatmaps mit denjenigen zu durchschnittlichen Monatstemperaturen, weisen auf weitere Potenziale für die historische Klimaforschung hin. Allerdings ist hierbei zu beachten, dass die mit Topic Modeling erzeugten Resultate trotz vermeintlicher Ähnlichkeit auf anderen Informationen als lediglich derjenigen zur Temperatur aufbauen und so Verzerrungen und Fehlinterpretationen möglich sind.

Im Hinblick auf die Einsatzpotenziale von Topic Modeling für die historische Klimaforschung wäre es in einer nachfolgenden Studie interessant zu ermitteln, inwiefern sich die in der vorliegenden Arbeit ermittelten Kälte-Topics für das Auffinden von Extremen eignen. Dazu wird der Ansatz, die Daten in Segmente mit unterschiedlicher zeitlicher Auflösung zu unterteilen, als zielführend erachtet. So wurden bereits testweise Modellierungsprozesse mit einer chronologischen Segmentierung pro Monat angestossen. Es zeigte sich, dass sich die bereits erwähnten Einflussfaktoren wie Witterung, kulturelle und landwirtschaftliche Praktiken, aber auch stilistisch-orthografische Eigenheiten in ei-

nem komplexen Wechselspiel offenbaren, was eine differenziertere Betrachtung erfordert. Während Einzelereignisse wie Naturkatastrophen in den bisherigen Analysen aufgrund ihres beschränkten zeitlichen Wirkungsradius in den Topics nicht hervortraten, wiesen Tests darauf hin, dass sich diese bei einer Segmentierung pro Tag deutlicher zeigen. Dies dürfte vor allem im Hinblick auf Fragen der Auffindbarkeit und im Kontext der Klimafolgen- und Naturkatstrophenforschung von Bedeutung sein.

Im Weiteren wäre es möglich, die Anwendung von Topic Modeling auf den Text des gesamten Einsiedler Kloster-Tagebuchs auszuweiten und die Bandbreite der möglichen Themen so zu erhöhen. Abgesehen davon können weitere Anwendungsbeispiele in den Geisteswissenschaften generell dabei helfen, digitale Methoden stärker in den einzelnen Disziplinen zu verankern. Da unter dem Hintergrund der zunehmenden Digitalisierung und der Verfügbarkeit zuverlässigerer Verfahren der automatischen Texterkennung vermehrt vormoderne Quellen in digitaler Form zugänglich sind, wären die Resultate auch für Archive und Spezialbibliotheken interessant. Insbesondere für wissenschaftliche Bibliotheken, die ihr Angebot in den Bereichen der Vermittlung und der Forschungsunterstützung stetig ausbauen, könnten Ansätze wie Topic Modeling in Zukunft relevant sein.

Bibliographie

Forschungsliteratur

Andorfer, Peter: Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich, in: Zeitschrift für digitale Geisteswissenschaften, 2017. Online: https://zfdg.de/2017_002, Stand: 30.06.2023.

Blei, David M.: Probabilistic Topic Models, in: Communications of the ACM 55/4, 2012, S. 77-84. Online: https://dl.acm.org/doi/10.1145/2133806.2133826, Stand: 30.06.2023.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I.: Latent Dirichlet Allocation, in: Journal of Machine Learning Research 3, 2003, S. 993-1022. Online: https://dl.acm.org/doi/10.5555/944919.944937, Stand: 30.06.2023.

Fechne, Martin; Weiss, Andreas: Einsatz von Topic Modeling in den Geschichtswissenschaften. Wissensbestände des 19. Jahrhunderts, in: Zeitschrift für digitale Geisteswissenschaften 2, 2017. Online: https://zfdg.de/2017_005, Stand: 30.06.2023.

Graham, Shawn; Milligan, Ian; Weingart, Scott: Exploring Big Historical Data. The Historian's Macroscope. London 2015.

Henggeler, Rudolf: Professbuch der Fürstl. Benediktinerabtei U. L. Frau zu Einsiedeln. Festgabe zum tausendjährigen Bestand des Klosters. Einsiedeln 1934 (Monasticon-Benedictinum Helvetiae 3).

Hodel, Tobias: Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities, in: Archives, Access and Artificial Intelligence. Working with Born-digital and Digitized Archival Collections. Bielefeld 2022 (Digital Humanities Research 2), S. 157-177. Online: https://doi.org/10.48350/169050, Stand: 30.06.2023.

Hodel, Tobias; Möbus, Dennis; Serif, Ina: Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora, in: Gerlek, Selin; Kissler, Sarah; Mämecke, Thorben; Möbus, Dennis (Hg.): Von Menschen und Maschinen. Mensch-Maschine-Interaktionen in digitalen Kulturen. Hagen 2022 (Digitale Kultur 1), S. 181-205. Online: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001838, Stand: 30.06.2023.

Jockers, Matthew L.: Macroanalysis. Digital Methods & Literary History. Urbana 2013.

Kherwa, Pooja; Bansal Poonam: Topic Modeling: A Comprehensive Review, in: EAI Endorsed Transactions on Scalable Information Systems 7/24, 2020. Online: https://eudl.eu/doi/10.4108/eai.13-7-2018.159623, Stand: 30.06.2023.

Lamba, Manika; Madhusudhan, Margam: Text Mining for Information Professionals. An Uncharted Territoriy. Cham 2022.

Meeks, Elijah; Weingart, Scott B.: The Digital Humanities Contribution to Topic Modeling, in: Journal of Digital Humanities 2/1, 2012. Online: http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling, Stand: 30.06.2023.

Mimno, David: Using Phrases in Mallet Topic Models, 2015. Online: http://www.mimno.org/articles/phrases, Stand: 30.06.2023.

Nelson, Robert K.: Richmond Daily Dispatch, 1869-1865 and Mining the Dispatch, in: Journal of American History 99/1, 2012, S. 386-388. Online: https://doi.org/10.1093/jahist/jas157, Stand: 30.06.2023.

Newman, David J.; Block, Sharon: Probabilistic Topic Decomposition of an Eighteenth Century American Newspaper. In Journal of the American Society for Information Science and Technology 57/6, 2006, S. 753-767. Online: https://dl.acm.org/doi/10.5555/1124169.1124187, Stand: 30.06.2023.

Pfister, Christian: Wetternachhersage. 500 Jahre Klimavariationen und Naturkatastrophen (1496-1995). Bern 1999.

Schöch, Christof: Topic Modeling with Mallet. Hyperparameter Optimization, in: The Dragonfly's Gaze. Computational Analysis of Literary Texts, 2016. Online: https://dragonfly.hypotheses.org/1051, Stand: 30.06.2023.

Schöch, Christof: Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama, in: Digital Humanities Quarterly 11/2, 2017. Online: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html, Stand: 30.06.2023.

Tang, Jian; Meng, Zhaosi; Nguyen, Xuanlong; Mei, Qiaozhu; Zhang, Ming: Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis, in: Proceedings of the 31st International Conference on Machine Learning 32/1, 2014, S. 190-198. Online: https://proceedings.mlr.press/v32/tang14.html, 31.08.2022. (=Tang et al., Limiting Factors)

Viehhauser, Gabriel: Mittelalterliche Texte als Modellierungsaufgabe, in: Fischer, Martin (Hg.): Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen. Bamberg 2020 (Bamberger interdisziplinäre Mittelalterstudien 15), S. 15-50.

Wallach, Hanna; Mimno, David; McCallum, Andrew: Rethinking LDA. Why Priors Matter, in: Advances in Neural Information Processing Systems 22, 2009. Online: https://papers.nips.cc/paper/2009/hash/0d0871f0806eae32d30983b62252da50-Abstract.html, Stand: 30.06.2023.

Wehrheim, Lino: Economic History Goes Digital. Topic Modeling the Journal of Economic History, in: Cliometrica, Journal of Historical Economics and Econometric History 13/1, 2019, S. 83-125. Online: https://doi.org/10.1007/s11698-018-0171-7, Stand: 30.06.2023.

Wilke, Claus O.: Datenvisualisierung – Grundlagen und Praxis: Wie sie aussagekräftige Diagramme und Grafiken gestalten. Heidelberg 2020.

Observable-Notebooks

Segmentierung pro Monat kumuliert. Online: https://observablehq.com/@lheinzmann/tm_monate_kumuliert, Stand: 30.06.2023.

Segmentierung pro Beobachtungsort und Jahreszeit kumuliert. Online: https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, Stand: 30.06.2023.

Segmentierung pro Jahr über den Gesamtzeitraum. Online: https://observablehq.com/@lheinzmann/tm_jahre_gesamtzeitraum, Stand: 30.06.2023.