Universitäten Bern und Lausanne

Master of Advanced Studies in Archival, Library and Information Science (MAS ALIS)

Studiengang 2020-2022

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

Masterarbeit in

Digital Humanities

Betreut von

Prof. Dr. Tobias Hodel

Vorgelegt von

Lukas Heinzmann

lukas.heinzmann@protonmail.com

lukas.heinzmann@unibe.ch

Bern, 31. August 2022

Abstract

Obwohl computergestützte Textanalyseverfahren ein erhebliches Potenzial für die Erforschung grosser Textmengen bergen, fanden sie bisher in den Geisteswissenschaften mit Ausnahme einiger anwendungsorientierter Spezialgebiete wenig Anklang. Der Beitrag will am Beispiel einer Anwendung von Topic Modeling auf die Wetterbeobachtungen des Einsiedler Paters Joseph Dietrich (1645-1704) aufzeigen, inwiefern sich algorithmische Zugänge für die Analyse frühneuzeitlicher Texte eignen und welche Faktoren für eine erfolgreiche Umsetzung zu berücksichtigen sind. Es zeigte sich, dass sich mit Hilfe von Topic Modeling einerseits inhaltliche Erkenntnisse gewinnen liessen und andererseits orthografische und stilistische Eigenheiten herausgearbeitet werden konnten, welche sich als Ausgangspunkt für weiterführende Analysen anbieten.

Selbstständigkeitserklärung

„Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Art. 36 Abs. 1 Buchst. r des Gesetzes über die Universität Bern und Art. 69 des Universitätsstatuts zum Entzug des aufgrund dieser Arbeiten verliehenen Titels berechtigt ist."

Datum	Unterschrift
31. August 2022	

Inhalt

1. Einleitung 3
   1.1. Kontext und Sample 4
      1.1.1. Autor und Werk 4
      1.1.2. Datensample 6
   1.2. Topic Modeling 8
   1.3. Forschungsüberblick zu Topic Modeling 11
   1.4. Erkenntnisinteresse und Aufbau 18
2. Methode 20
   2.1. Preprocessing 20
      2.1.1. Forschungspositionen 20
      2.1.2. Realisierung 22
   2.2. Modellierungsprozess 24
      2.2.1. Forschungspositionen 24
      2.2.2. Realisierung 26
   2.3. Postprocessing 28
      2.3.1. Forschungspositionen 28
      2.3.2. Realisierung 29
3. Analyse 33
   3.1. Segmentierung pro Monat kumuliert 33
   3.2. Segmentierung pro Beobachtungsort und Jahreszeit kumuliert 42
   3.3. Segmentierung pro Jahr über den Gesamtzeitraum 49
4. Fazit und Ausblick 57
5. Bibliografie 61
   5.1. Forschungsliteratur 61
   5.2. Observable-Notebooks 64
6. Anhang 65
   6.1. Abbildungsverzeichnis 65
   6.2. Stopwords-Liste 65
   6.3. N-Gramme 67
   6.4. Topics und Wortfrequenzen 67
      6.4.1. Segmentierung pro Monat kumuliert 67
      6.4.2. Segmentierung pro Beobachtungsort und Jahreszeit 74
      6.4.3. Segementierung pro Jahr über den Gesamtzeitraum 77

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. Umgekehrt kommen – mit Ausnahme von wenigen Spezialgebieten – für die weiteren Arbeitsschritte in der Regel überwiegend erprobte Methoden der jeweiligen Disziplin zur Anwendung. Insbesondere texttechnologische Verfahren 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 dass sie sich dabei von der eigenen epistemologischen Basis trennen müssen. Der Einsatz digitaler Methoden führt nämlich nicht zu einer Veränderung der grundlegenden Fragen einer Disziplin, sondern eröffnet neue Perspektiven und Zugangswege zur Beantwortung derselben. Da die von Algorithmen produzierten Resultate jedoch keine direkten Wahrheiten abbilden, bleibt es die Aufgabe von Geisteswissenschaftlerinnen und Geisteswissenschaftlern, diese unter dem eigenen disziplinären Hintergrund zu interpretieren und nutzbar zu machen.[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 einer Dissertation, bei welcher der Fokus auf textgenetischen und klimageschichtlichen Fragen liegt.[2] Aufgrund verschiedener individueller Faktoren und externer Einflüsse ist das Tagebuch gekennzeichnet von Veränderungen und Brüchen formaler, inhaltlicher und stilistischer Natur. Diese konnten zwar bis zu einem gewissen Grad dekonstruiert werden, allerdings zeigte sich, dass dies allein

[1] Es handelt sich um die Kernaussagen in der Einleitung des Werks *Exploring Big Historical Data. The Historian's Macroscope* von Graham, Milligan und Weingart. Vgl. Graham, Milligan, Weingart 2015: 31-33. Die Zitierweise in der vorliegenden Arbeit folgt weitgehend den Richtlinien für die Erstellung schriftlicher Arbeiten an der Abteilung für Wirtschafts-, Sozial- und Umweltgeschichte der Universität Bern. Vgl. dazu https://www.hist.unibe.ch/ueber_uns/abteilungen/wirtschafts__sozial__und_umweltgeschichte/index_ger.html, 31.08.2022. Websites werden in der Bibliografie erfasst, insofern auf ihre Inhalte Bezug genommen wird; weiterführende Links zählen nicht dazu. Zahlen in Zusammenhang mit statistischen Werten oder Modellen werden nicht ausgeschrieben. Im Weiteren wurde auf eine Kursivierung der zahlreichen englischsprachigen Begriffe verzichtet.

[2] Die Dissertation wurde noch nicht fertiggestellt und entsprechend noch nicht publiziert.

mit Close Reading[3] schwierig ist. Durch den Einbezug quantitativer Methoden[4] gelang es zwar, bestimmte Aspekte besser fassbar zu machen zu machen, diese griffen aber hinsichtlich der Analyse stilistischer und orthografischer Eigenheiten zu kurz.

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 Extrakte in Form von Natur- und Wetterbeobachtungen, welche zum Zweck einer klimageschichtlichen Analyse extrahiert und transkribiert wurden, verwendet.[5] Die Beschränkung auf ein inhaltlich und formal weitgehend homogenes Korpus soll eine differenziertere Betrachtungsweise ermöglichen. Bevor die Forschungsfragen im letzten Kapitel der Einleitung konkretisiert werden, folgen genauere Informationen zu Autor, Werk und Datensample sowie zum Topic-Modeling-Ansatz und den in diesem Zusammenhang erschienen Forschungsarbeiten. Unter diesem Hintergrund lässt sich das Erkenntnisinteresse besser skizzieren.

1.1. Kontext und Sample

1.1.1. Autor und Werk

Ludwig Dietrich war der Sohn des Rapperswiler Schultheissen Johann Peter Dietrich (1611-1681), welcher unter anderem ein Tagebuch über die Belagerung der Stadt Rapperswil durch die Zürcher (1656) verfasst hatte.[6] Nach seiner Zulassung zum Noviziat im Jahr 1660 legte er 1662 im Kloster Einsiedeln die Profess ab und nahm den Vornamen Joseph an. 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

[3] Der Terminus „Close Reading“ bezeichnet das Lesen von Texten und bildet den Gegenpart zum Begriff „Distant Reading“, welchen der amerikanische Komparatist Franco Moretti um die Jahrtausendwende prägte. Es handelt sich hierbei um einen Überbegriff für alldiejenigen Ansätze, die sich auf eine Erschliessung grosser Textmengen richten, ohne dass diese gelesen werden. Entsprechend wird dieser Terminus häufig im Zusammenhang mit digitalen Methoden verwendet. Vgl. Viehhauser 2020: 24-25.

[4] Die quantifizierenden Methoden bestanden vor allem in Auszählungen. So wurden beispielsweise die Anzahl der beschriebenen Tage gezählt und in Relation zu deren effektiven Zahl sowie zur Anzahl der geschriebenen Seiten gestellt. Dadurch liess sich ansatzweise nachvollziehen, wie sich die Frequenz und die Textmenge über den Gesamtzeitraum verändern, was wiederum Rückschlüsse auf die Schreibpraxis des Autors zuliess.

[5] Das gesamte Tagebuch wird im Rahmen eines digitalen Editionsprojekts aufbereitet. Zum Zeitpunkt der vorliegenden Arbeit war die Transkription noch nicht vollständig abgeschlossen. Zum digitalen Editionsprojekt vgl. http://www.dietrich-edition.unibe.ch, 31.08.2022

[6] Die Biografie des Autors wird in der noch nicht veröffentlichten Dissertation ausführlich beschrieben. Die hier aufgeführten Informationen entsprechen der Charakterisierung von Pater Rudolf Henggeler in seinem 1934 erschienenen Professbuch. Vgl. Henggeler 1934: 325-328. Diese ist auch auf der Website des Klosterarchivs Einsiedeln auffindbar. Vgl. http://www.klosterarchiv.ch, 31.08.2022.

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.

Aufenthaltszeitraum	**Aufenthaltsort**
21.01.1662 – 26.11.1688	Kloster Einsiedeln
26.11.1688 – 07.12.1690	Schloss Freudenfels
07.12.1690 – 28.07.1692	Kloster Einsiedeln
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

Neben zahlreichen handschriftlichen Dokumenten im Zusammenhang mit Dietrichs administrativen Tätigkeiten ist im Klosterarchiv Einsiedeln ein grösstenteils von seiner Hand stammendes Tagebuch überliefert. Dieses wurde von Pater Friedrich Helmlin 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 der einzelnen Bücher zwischen 388 und 952 beschriebene Seiten beträgt. Ebenso stark variiert der zeitliche Bezugsrahmen, der von 13 Monaten bis sechseinhalb Jahre pro Band reicht. Dietrich führte das Tagebuch auch während seiner Aufenthalte in den Aussenstationen fort, weshalb sich zwölf Bände auf Einsiedeln, vier auf Freudenfels sowie je einer auf Pfäffikon und einer auf Fahr beziehen. Da er teilweise die Ereignisse nach seiner Rückkehr von den Aussenstationen rekonstruierte und zeitweise Pater Sebastian Reding (1667-1724) als Stellvertreter in Einsiedeln fungierte, überlagern sich bestimmte Bände.

Insgesamt kann konstatiert werden, dass die Entstehung des Tagebuchs nicht linear verlief, sondern aufgrund seiner häufigen Ortswechsel und diversen externen Faktoren viele formale und inhaltliche Unregelmässigkeiten oder Besonderheiten aufweist. 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, welcher mit dem Beginn der täglichen Berichterstattung zum Wetter einherging. 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,[7] sondern hauptsächlich um narrative Beschreibungen der gefühlten Temperatur, der Intensität und Dauer von Niederschlägen, Windstärke- und Richtung, Himmelsbedeckung usw. Obwohl das Tagebuch viele Beobachtungen dieser Art enthält, handelt es sich nicht um ein ausschliesslich dem Wetter gewidmetes Tagebuch. Das Wetter ist ein Thema neben anderen und findet teilweise nur indirekt im Kontext der Schilderungen landwirtschaftlicher oder kultureller Praktiken im Kloster Erwähnung. Im Hinblick auf die Versorgung waren witterungsbedingte Einflüsse auf die Vegetationsentwicklung, Qualität und Quantität der Ernte sowie die Preisentwicklung für Dietrich von Interesse. Das Wetter konnte aber auch ein Thema im Zusammenhang mit der Durchführung von Prozessionen oder der Mobilität sein, vor allem wenn es hierbei Einschränkungen gab. Im Weiteren stellten witterungsbedingte Naturkatastrophen wie Lawinen, Stürme, Dürren oder Überschwemmungen für die Zeitgenossen prägende Ereignisse dar.

Die Beobachtungen von Dietrich und anderen Zeitgenossen erscheinen somit häufig in engem Bezug zu ihrer Lebenswelt und ihrem Denkhorizonts, weshalb sie unter diesem Hintergrund zu interpretieren sind. Die historische Klimageschichte, welche sich auf schriftliche und bildliche Dokumente stützt und diese mit Hilfe klimatologischer und geschichtswissenschaftlichen Methoden auswertet, trägt diesem Umstand Rechnung, indem sie drei miteinander verzahnte Bereiche unterscheidet: Während sich die Klimarekonstruktion auf die Nachbildung des Witterungs- und Klimaverlaufs vor der Entstehung meteorologischer Messnetze im späten 19. Jahrhundert fokussiert, widmet sich die historische Klimafolgenforschung der Verletzlichkeit von Wirtschaft und Gesellschaft für Witterungsextreme und witterungsbedingte Katastrophen. Der letzte Bereich, die Wissensgeschichte, bezieht sich auf die gesellschaftliche Wahrnehmung

[7] Erste instrumentelle Messungen wurden um die Wende zum 17. Jahrhundert in Italien durchgeführt. Der Universalgelehrte Johann Jakob Scheuchzer (1672-1733) war nach heutigem Wissensstand der erste, welcher meteorologische Messungen in der Schweiz (ab 1708) vornahm. Vgl. Pfister 1999: 26-27.

und Interpretation von Witterungs- und Klimaphänomenen.[8] Um klimageschichtliche relevante Informationen aus Dokumenten systematisch erfassen und auswerten zu können, wurde in Bern ab den 1970er Jahren die Datenbank Euro-Climhist aufgebaut. Diese macht Daten zu Wetterereignissen vom 16. bis ins 19. Jahrhundert sowie deren Folgen für Mensch und Umwelt zugänglich. Lag der Schwerpunkt ursprünglich auf dem Gebiet der heutigen Schweiz, enthält Euro-Climhist mittlerweile Daten zu mehreren europäischen Ländern.[9]

Damit die Witterungsdaten über unterschiedliche Suchparameter abgefragt werden können, werden sie bei Euro-Climhist nach einem vorgegebenen Schema erfasst. Grundlegende Elemente sind dabei die Datierung und Lokalisierung der einzelnen Informationen sowie deren Quellennachweis. Im Weiteren werden sie nach ihrer Art in sechs Hauptkategorien unterteilt. So wird etwa zwischen deskriptiven Daten, welche Beschreibungen zu Witterung und direkten witterungsbedingten Konsequenzen umfassen, und biophysischen Proxydaten unterschieden. Letztere beinhalten biologische Prozesse, wie beispielsweise das Pflanzenwachstum, oder physikalische Indikatoren wie die Eisbedeckung von Gewässern, anhand derer sich Rückschlüsse auf Witterungsverläufe ziehen lassen. Ebenfalls indirekt mit dem Wetter hängen Wirtschaftsdaten und soziopolitische Daten (Wetterprozessionen, Bittgottesdienste, Verfolgung von Minderheiten) zusammen.[10]

Im Hinblick auf eine klimageschichtliche Auswertung und eine spätere Übertragung in Euro-Climhist wurden die Wetterbeobachtungen von Pater Joseph Dietrich händisch transkribiert und gemäss den Vorgaben der genannten Datenbank in einer Excel-Tabelle strukturiert erfasst. Die Beschreibungen wurden dazu jeweils einem konkreten Datum und Ort zugeordnet und in einer Zeile zusammen mit weiteren Metainformationen abgebildet. Inhaltlich wurden Informationen, die entweder direkt oder indirekt mit der Witterung in Beziehung standen, extrahiert, ebenso wie Schilderungen über die Folgen von extremen Witterungsphänomenen und Naturkatastrophen. Aufgrund des narrativen Stils war eine genaue Abgrenzung der Themen nicht immer eindeutig machbar, weshalb die einzelnen Einträge viele Informationen im Zusammenhang mit der Denk- und Lebenswelt des Autors enthalten und von der Länge her stark variieren.

Für das Datensample in der vorliegenden Arbeit wurden bestimmte Datensätze nicht berücksichtigt. Dies betrifft Wetterinformationen, die entweder von Dietrichs Stellvertreter Reding erfasst wurden oder die Dietrich bei der Rekonstruktion der Bände nachträglich verzeichnete.

[8] Vgl. Mauelshagen 2010: 20.
[9] Weiterführende Informationen zu Euro-Climhist finden sich auf der Website. Vgl. https://www.euroclim-hist.unibe.ch, 31.08.2022.
[10] Für genauere Beschreibungen der einzelnen Kategorien vgl. https://www.euroclim-hist.unibe.ch/de/dbsuche, 31.08.2022.

Ersteres ist damit zu begründen, dass sich Redings Schilderungen sowohl vom Stil als auch von der Orthografie erheblich von Dietrichs Schreibpraxis unterscheiden, was sich auf das Resultat auswirken würde.[11] Der Ausschluss der nachträglich rekonstruierten Teilbände ist damit zu begründen, dass das Vorhandensein mehrerer Einträge pro Datum zu einer Verzerrung der Resultate hätte führen können. Berücksichtigt wurden somit nur Datensätze von Dietrichs Hand, die er zeitnah und am jeweiligen Ort verfasst hatte.

1.2. Topic Modeling

Topic Modeling ist – vereinfacht ausgedrückt – eine computergestützte Methode, mit Hilfe derer zusammenhängende Informationen in grossen Datenmengen sichtbar gemacht werden können.[12] Obwohl Topic Modeling in der Forschungsliteratur häufig in Bezug mit der Analyse von Texten thematisiert wird, lässt es sich auch auf andere Informationstypen, wie biologische Daten (DNS), Bilder, Musiknoten usw., anwenden. Aufgrund des breiten Spektrums an Anwendungsmöglichkeiten kommt Topic Modeling in vielen Forschungsfeldern wie der Bioinformatik, Medizin, Computerwissenschaften sowie diversen Gebieten der Sozial- und Geisteswissenschaften und anderen Disziplinen zum Einsatz. In Bezug auf Texte und Textsorten ist der Ansatz ebenso breit anwendbar, wobei sich generell zwei übergeordnete Verwendungszecke ausmachen lassen. Zum einen stellt es eine Methode dar, mit Hilfe derer zentrale Konzepte oder Themen in umfangreichen Textsammlungen erschlossen werden können. Deshalb ist der Ansatz für die Bereiche Information Retrieval und Text Mining, welche sich auf das Auffinden von Informationen in unstrukturierten Korpora fokussieren, relevant. Da Topic Modeling nicht nur Themen, sondern allgemein versteckte Strukturen aufdeckt, ist es andererseits auch als explorativ Werkzeug für stilistische und formale Analysen einsetzbar.[13]

Es handelt sich bei Topic Modeling nicht um ein einzelnes Verfahren, sondern um einen 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.[14] Der in der Forschung am weitesten verbreitete und gemäss dem Begründer David Blei auch simpelste Topic-Modeling-Ansatz ist Latent Dirichlet Allocation (LDA). Ausgehend von der Annahme,

[11] Vorgängige Tests mit den Daten beider Autoren zeigten, dass die Unterschiede in Stil und Orthografie in den Topics erkennbar waren. Da von Redings Hand vergleichsweise wenig Text vorhanden ist, bildeten die ihm zugewiesenen Topics stärker formale Gesichtspunkte als inhaltlich interpretierbare Muster ab.
[12] Vgl. Fechner, Weiss 2017: Kap. 1.2; Schöch 2017: Abs. 2.
[13] Vgl. Lamba, Madhusudhan 2022: 105-107, 113-114; Blei 2012: 77; Unkel 2020: Kap. 21.1.
[14] Vgl. Unkel 2020: Kap. 21.1.

dass eine begrenzte Zahl von Wörtern oder Tokens[15] in Textsegmenten (z.B. Absatz, Dokument, Brief, Tweet usw.) häufig zusammen auftreten, zählen die Algorithmen deren Häufigkeit und vergleichen sie mit der Auftretenshäufigkeit anderer Tokens in demselben Segment. Die Tokens werden zu Topics gebündelt und deren Zusammensetzung in einem iterativen Verfahren, im Zuge dessen sich die Annahmen über die Wahrscheinlichkeiten verändern, verfeinert. Die Reihenfolge der Tokens innerhalb der Texteinheit spielen keine Rolle und einzelne Tokens können – ausgehend von der Annahme, dass Texte in der Regel mehrere Themen enthalten – auch Bestandteil unterschiedlicher 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.[16]

Obwohl die Begriffe teilweise synonym verwendet werden, entsprechen Topics nicht dem alltagssprachlichen Verständnis von Themen, da die Wortketten einer 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 interpretatorisch zu erschliessen 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[17] erforderlich. Das Verfahren lässt sich somit nicht nur gattungs- und sprachunabhängig[18] anwenden, sondern gestaltet sich auch vom eigentlichen Modellierungsprozess her simpel. Im Minimalfall ist neben einem Textkorpus nur eine Instanz zur Aktivierung des Algorithmus und die Festsetzung der Anzahl der zu produzierenden Topics erforderlich.[19] Daneben gibt es weitere optionale Parameter zur

[15] Beim Topic Modeling gelten Tokens als Elemente, die durch Leerschläge voneinander abgetrennt sind. In der Regel sind dies Wörter, es können aber auch Zahlen, Abkürzungen oder – vor allem beim Vorhandensein von Bindestrichen – Wortteile sein. Vgl. Lamba, Madhusudhan 2022: 79-81. Da in der vorliegenden Arbeit Zahlen und Abkürzungen durch die Stopwords-Liste von der Analyse ausgeschlossen wurden, kommen überwiegend ganze Wörter vor. Aus diesem Grund wird Token in der Folge synonym mit Begriff, Wort oder Terminus verwendet.

[16] Vgl. Blei 2012: 77-79; Graham, Milligan, Weingart 2015: 117-118; Fechne, Weiss 2017: Kap. 1.1; Hodel 2022: 162-164; Hodel, Möbus, Serif 2022: 183-184.

[17] Im Bereich Machine Learning wird zwischen überwachten und unüberwachten Lernmethoden unterschieden. Da LDA keine Trainingsphase erfordert, wird es allgemein zu letzter Kategorie gezählt. Es existieren allerdings auch Erweiterungen für LDA, durch welche etwa durch Einbindung eines Wörterbuchs den Prozess der Topic-Bildung in eine bestimmte Richtung gelenkt werden kann. Vgl. Lamba, Madhusudhan 2022: 15-17, 113.

[18] Obwohl sich Topic Modeling grundsätzlich gattungs- und sprachunabhängig anwenden lässt, können die Resultate durch sprach-, text- oder korpusbedingte Eigenheiten beeinflusst werden. Teilweise lassen sich diese Effekte durch Erweiterungen des LDA-Modells, zum Beispiel für den Umgang mit mehrsprachigen Texten, ausgleichen. Vgl. Wehrheim 2019: 92-93.

[19] Vgl. Blei 2012: 77-79; Graham, Milligan, Weingart 2016: 119-120; Schöch 2017: Abs. 13-14; Wehrheim 2019: 89-92; Hodel 2022: 164.

Beeinflussung des Modellierungsprozesses und Resultats. Obwohl die algorithmische Modellierung simpel ist, erfordert eine zielführende Anwendung von Topic Modeling aufgrund von Unabwägbarkeiten im Hinblick auf die Konfiguration der verschiedenen Parameter einen elaborierten Workflow, der auf den jeweiligen Korpus und die Bedürfnisse zugeschnitten und iterativ optimiert werden muss.[20]

Der Workflow und die Individualisierungsmöglichkeiten werden im Rahmen der Beschreibung zur methodischen Realisierung eingehender thematisiert, weshalb an dieser Stelle lediglich auf einige Implikationen, welche allgemein für die Arbeit mit Topic Modeling und das Verständnis der bisherigen Forschung relevant sind, hingewiesen wird. 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 Parametern bei jeder Anwendung unterschiedlich ausgegeben werden. Somit sind die Resultate – auch wenn sie in ähnlicher Form erscheinen – nicht eins zu eins reproduzierbar.[21] 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. Dies kann für die Bearbeitung stilometrischer, sprachwissenschaftlicher oder ähnlich ausgerichteter Forschungsfragen interessant sein. Bei Anwendungen, die stärker auf eine inhaltliche Erschliessung abzielen, stellt dies einen möglichen Verzerrungsfaktor dar, der insbesondere bei heterogenen Korpora auftreten kann. Umgekehrt bedeutet dies, dass sich Topic Modeling vor allem für die thematische Exploration von homogenen Textsammlungen gut eignet.[22]

Für den Modellierungsprozess existieren zahlreiche Tools und Programmpakete in unterschiedlichen Programmiersprachen und Umgebungen.[23] 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 darauf zurückzuführen, dass die Tools unterschiedlich konfigurierte Algorithmen anwenden.[24] Der in der vorliegenden Arbeit verwendete „Werkzeugkasten“ MAchine Learning

[20] Vgl. Schöch 2017: Abs. 16, 20;
[21] Vgl. Blei 2012: 79-80; Graham, Milligan, Weingart 2015: 157; Schöch 2017: Abs. 14.
[22] Vgl. Fechne, Weiss 2017: Kap. 1.2; Unkel 2020: Kap. 21.1.
[23] Zwischen 2002 und 2010 wurden zahlreiche Tools für die Anwendung von *Topic Modeling* veröffentlicht, die allerdings häufig nicht weiterentwickelt wurden. Insbesondere für die Programmiersprachen R und für Python sind verschiedene Programmpakete verfügbar. Für eine Übersicht zu älteren Tools vgl. Chappelier 2011: 213. Für eine aktuelle Zusammenstellung vgl. Lamba, Madhusudhan 2022: 108-109. David Blei führt zudem auf seiner Website eine Liste zu LDA-Software. Vgl. https://www.cs.columbia.edu/~blei/topicmodeling_software.html, 31.08.2022.
[24] Vgl. Graham, Milligan, Weingart 2015: 130, 157.

for LanguagE (MALLET) wurde an der Universität Massachusetts Amherst von einem Team rund um Andrew McCallum entwickelt und erstmals 2002 veröffentlicht. Ursprünglich als Toolkit für die linguistische Datenverarbeitung (im Englischen Natural Language Processing, kurz NLP) konzipiert, erweiterte der seit 2005 im Team arbeitende und aktuell für MALLET verantwortliche David Mimno das Tool um eine LDA-basierte Topic-Modeling-Funktion. MALLET gilt als solides Topic-Modeling-Toolkit, welches vor allem in geisteswissenschaftlichen Studien häufig benutzt wurde und dank der Betreuung von Mimno weiterhin funktionsfähig ist. Als Java-basierte Open-Source-Software kann MALLET in andere Umgebungen[25] eingebunden oder um zusätzliche Komponenten erweitert werden.[26] So ist MALLET beispielsweise auch Bestandteil des GUI Topic Modeling Tool (GTMT), welches eine vereinfachte Anwendung von Topic Modeling über eine grafische Benutzeroberfläche bietet.[27]

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. Um einen Eindruck auf die Bandbreite zu vermitteln, wird hier exemplarisch auf die 2020 erschienene Metastudie von Kherwa und Bansal verwiesen. In dieser wurden rund 300 zwischen 2003 und 2018 erschienene Aufsätze zu Topic Modeling untersucht, wobei die unterschiedlichen methodischen Ansätze klassifiziert und auf Anwendungsbeispiele in diversen Bereichen hingewiesen wurde.[28] Aufgrund der zahlreichen Anwendungsfelder sowie der Unterschiede bezüglich Untersuchungsgegenstand, Schwerpunktsetzung und Erkenntnisinteresse ist es im Rahmen der vorliegenden Arbeit erforderlich, den Forschungsüberblick auf ausgewählte Bereiche zu beschränken.

Allgemein lassen sich bezüglich der Ausrichtung wissenschaftlicher Studien im Zusammenhang mit Topic Modeling drei Tendenzen erkennen: Auf der einen Seite gibt es Aufsätze aus dem Informatikbereich, die sich schwerpunktmässig mit den zugrundeliegenden statistischen und technischen Eigenheiten von Topic Modeling auseinandersetzen. Obwohl sie häufig auch Texte als Untersuchungsbasis verwenden, erfolgt die Bewertung weniger aufgrund inhaltlicher

[25] David Mimno hat beispielsweise einen Wrapper für die Einbindung von MALLET in R programmiert, womit sich einerseits der Workflow abbilden und andererseits der von MALLET generierte Output leichter weiterverarbeiten und visualisieren lässt. R ist eine Programmumgebung, die insbesondere für die Bearbeitung und statistische Auswertung grosser Datenmengen konzipiert ist. Vgl. Graham, Milligan, Weingart 2015: 149, 151-156. Für den Download des Wrappers vgl. https://github.com/mimno/RMallet, 31.08.2022. Für weiterführende Informationen zu Topic Modeling mit R vgl. Jockers, Thalken 2020: 211-236.
[26] Vgl. Graham, Milligan 2012; Jockers 2013: 124.
[27] Für weitere Hintergrundinformationen zu GTMT vgl. Graham, Milligan, Weingart 2015: 121-126. Für den Download des Tools vgl. https://code.google.com/archive/p/topic-modeling-tool/, 31.08.2022.
[28] Vgl. Kherwa, Bansal 2020.

als vielmehr auf Basis mathematischer Kriterien und zielt auf die Optimierung und Weiterentwicklung des Ansatzes ab. Im Gegensatz dazu sind Forschende aus anderen Disziplinen in der Regel eher anwendungs- und resultatorientiert und setzen die mit Topic Modeling erzielten Ergebnisse in Vergleich zu den mit traditionellen Methoden erarbeiteten Resultaten. Die erkenntnisleitende Frage ist dabei nicht wie, sondern ob und inwiefern sich Topic Modeling für die eigenen disziplinären Bedürfnisse nutzen lässt. Eine dritte Kategorie bilden 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.

Die beschriebene Kategorisierung ist keineswegs trennscharf, da viele Studien zu mehreren Aspekten Stellung beziehen. Dennoch eignet sich die Unterscheidung als Hilfsinstrument zur Eruierung der Ausrichtung der einzelnen Forschungstitel und deren zugrundeliegenden Erkenntnisinteressen. Aufgrund des anwendungsorientierten Schwerpunkts der vorliegenden Arbeit liegt der Fokus des Forschungsüberblicks ebenfalls auf Aufsätzen dieser Art, wobei vor allem Studien aus den digital arbeitenden Geisteswissenschaften oder Digital Humanities und im Besonderen mit geschichts- und literaturwissenschaftlichem Hintergrund vorgestellt werden. Im Zusammenhang mit diesen Disziplinen werden vereinzelt auch epistemologische Aspekte erwähnt, ebenso wie es für die Beschreibung des Ursprungs von Topic Modeling unumgänglich ist, auf Aufsätze aus der Informatik zu verweisen.

Obwohl es sich nicht um das älteste Verfahren handelt, ist LDA im Bereich der Digital Humanities der am weitesten verbreitete Ansatz. Aufgrund seiner Popularität und dem Umstand, dass viele Topic-Modeling-Tools darauf aufbauen, wird Topic Modeling häufig mit der Anwendung von LDA gleichgesetzt.[29] Entsprechend wird der 2003 erschienene Aufsatz von David Blei, Andrew Ng und Michael I. Jordan, in welchem Terminologie, Grundlagen und Potenziale von LDA erstmals beschrieben wurden, in der Forschungsliteratur häufig als Ausgangspunkt von Topic Modeling bezeichnet.[30] Diese erstmalige Skizzierung von LDA entstand allerdings unter dem Hintergrund der Forschung an unterschiedlichen Methoden zur effizienten Organisation und Auffindbarkeit von Informationen in grossen Datenmengen.[31] LDA entstand somit

[29] Vgl. Graham, Milligan, Weingart 2016: 119.
[30] Vgl. Blei, Ng, Jordan 2003. Insbesondere David Blei, der aktuell als Professor für Statistik und Computerwissenschaften an der Columbia University amtet, veröffentlichte in den letzten knapp 20 Jahren zahlreiche wissenschaftliche Artikel im Zusammenhang mit probabilistischen Ansätzen und Machine Learning. Weiterführende Informationen zu David Bleis Forschungs- und Lehrtätigkeiten sowie eine Publikationsliste finden sich auf der Website der Columbia University. Vgl. http://www.cs.columbia.edu/~blei/, 31.08.2022.
[31] Eine wesentliche Grundlage bildete der von Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer und Richard Harshman (1990) vorgestellte Ansatz Latent Semantic Indexing (LSI). Dieser war ursprünglich auf die Indexierung im Rahmen von Information Retrieval ausgerichtet, wurde später aber verschiedentlich modifiziert. So erweiterte etwa Thomas Hofmann (1999) den Ansatz um

gleichzeitig mit sowie in Anlehnung und Abgrenzung zu anderen Topic-Modeling-Ansätzen, wobei diverse Elemente des Verfahrens in den nachfolgenden Jahren weiterentwickelt wurden.[32] David Blei fasste den Stand der Forschung sowie die möglichen zukünftigen Richtungen im Jahr 2012 zusammen. Der Aufsatz wird in der Forschungsliteratur aufgrund seiner Verständlichkeit häufig als Einstiegslektüre zu LDA empfohlen.[33]

Da Topic Modeling für die algorithmische Durchleuchtung grosser Textmengen, welche sich als unstrukturierte Daten nur schwer Computern erschliessen lassen, konzipiert wurde, wird es als Anwendungsform von Text Mining klassifiziert und entsprechend auch in der Überblicksliteratur zu diesem Bereich beschrieben. So werden die theoretischen Grundlagen von LDA im 2011 erschienenen französischsprachigen Werk *Modèles statistiques pour l'accès à l'information textuelle* eingehend erläutert und die Unterschiede zu anderen Topic-Modeling-Ansätzen skizziert.[34] In ähnlicher Form wird das Thema auch im Titel *Mining Text Data* aus dem Jahr 2012 behandelt.[35] Während die beiden Werke sich vornehmlich auf statistische und theoretische Aspekte beschränken, richtet sich das 2022 publizierte Buch *Text Mining for Information Professionals* an ein anwendungsorientiertes Publikum im Bibliotheks- und Informationsbereich. In Bezug auf Topic Modeling werden vor allem potenzielle Anwendungsszenarien, verfügbare Tools und mögliche Workflows sowie konkrete Projekte vorgestellt. Aufgrund der Aktualität, den systematischen Zusammenstellungen von Forschungsarbeiten zu diversen Textsorten und den vielen weiterführenden Verweisen bildet dieses Werk einen ausgezeichneten Ausgangspunkt für die vertiefte Auseinandersetzung mit dem Ansatz.[36]

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.[37] Mit *Mining the Dispatch* wurden die Resultate eines 2011 abgeschlossenen Projekts veröffentlicht, welches ebenfalls auf einer Anwendung von Topic

probabilistische Elemente und begründete die Methode Probabilistic Latent Semantic Analysis (LSA). Vgl. Newman, Block 2006: 753-754.

[32] Vgl. dazu beispielsweise Steyvers, Griffiths 2007; Wallach, Mimno, McCallum 2009. David Mimno, der unter anderem zusammen mit Hanna Wallach und Andrew McCallum diverse Beiträge zum Thema verfasste, sammelte bis 2011 bibliografische Titel mit Schwerpunkt auf technisch-methodische Aspekte von Topic Modeling. Vgl. https://mimno.infosci.cornell.edu/topics.html, 31.08.2022.

[33] Vgl. Blei 2012.

[34] Vgl. Chappelier 2011.

[35] Vgl. Crain et al. 2012.

[36] Vgl. Lamba, Madhusudhan 2022.

[37] Vgl. Newman, Block 2006.

Modeling auf Zeitungsartikel 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.[38]

Abgesehen von einzelnen Publikationen fand der Diskurs rund um Topic Modeling in den digital arbeitenden Geisteswissenschaften bis dahin zu einem grossen Teil in Blogs, Präsentationen und ähnlichen Formaten statt.[39] Während beispielsweise Ted Underwood eine einfache Einführung zu Topic Modeling verfasste, schrieb Clay Templeton eine Übersicht zu Topic Modeling in den Geisteswissenschaften und Scott Weingart skizzierte neben der Entstehung von Topic Modeling das Potenzial für Folgeauswertungen in Form von Netzwerkanalysen.[40] Im Weiteren dokumentierte Cameron Blevins eine praktische Anwendung von Topic Modeling mit dem zwischen 1785 und 1812 verfassten Tagebuch von Martha Ballard (1735-1812).[41] 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.[42]

Es war ein Kernanliegen der erwähnten Ausgabe, die Informatik und die Geisteswissenschaften einander näherzubringen, weshalb sowohl technische Hintergründe erläutert als auch die Potenziale und Grenzen für die praktische Anwendung betont wurden. Dass die Funktionsweise von und der Umgang mit Tools ein grosses Thema war, ist damit zu begründen, dass die Einbindung von Tools in den geisteswissenschaftlichen Forschungsprozess gespalten aufgenommen wurde. Den Vorbehalten von traditionell arbeitenden Forschenden standen dabei teilweise die zu optimistische Erwartungen der Aufgeschlossenen gegenüber. Insbesondere Benjamin M. Schmidt zeigte in seinem Artikel anhand konkreter Beispiele, wie fehlendes Verständnis und eine unkritische Adaption der Resultate von Topic Modeling zu falschen Schlussfolgerungen führen kann.[43] Abgesehen von mangelndem Hintergrundwissen zu mathematisch-technischen Grundlagen erweist sich auch die praktische Handhabung von Topic-Modeling-Tools als Hürde. Shawn Graham und Ian Milligan verfassten deshalb einerseits im *Journal of Digital Humanities* einen kurzen Artikel zum Tool MALLET und veröffentlichten andererseits im Jahr 2012 zusammen mit Scott Weingart auf dem Portal *programminghistorian* eine auch

[38] Vgl. Nelson 2012. Die Resultate des Projekts sind auf einer Website, welche im November 2020 aktualisiert wurde, zugänglich. Vgl. https://dsl.richmond.edu/dispatch, 31.08.2022.
[39] Vgl. Cohen, Troyano 2012; Meeks, Weingart 2012.
[40] Vgl. Templeton 2011; Underwood 2012; Weingart 2012.
[41] Vgl. Blevins 2010.
[42] Vgl. Cohen, Trayono 2012; Meeks, Weingart 2012.
[43] Vgl. Schmitt 2012.

für Personen ohne Informatikkenntnisse verständliche Schritt-für-Schritt-Anleitung.[44] Während sich diese Anleitung auf die Benutzung von MALLET in der Kommandozeile bezieht, entstanden bereits seit 2009 Tools mit grafischen Benutzeroberflächen.[45]

Es lässt sich konstatieren, dass in den Geisteswissenschaften zu Beginn der 2010er Jahre ein gesteigertes Interesse[46] rund um das Thema Topic Modeling erwuchs, wobei einerseits wesentliche Grundlagen bezüglich der praktischen Anwendung von Topic Modeling geschaffen und andererseits epistemologische Fragen im Zusammenhang mit der Nutzung von Tools in den Geisteswissenschaften diskutiert wurden. Dabei zeigte sich auch, dass sich die einzelnen Disziplinen aufgrund der unterschiedlichen Erkenntnisinteressen, Untersuchungsgegenstände und Denktraditionen mit jeweils anderen Fragen auseinandersetzten. Während in den Geschichtswissenschaften der Fokus stärker auf Themen und Diskursen im zeitlichen Verlauf lag, nutzte beispielsweise die Literaturwissenschaftlerin Lisa Rhody Topic Modeling für die Analyse metaphorischer Sprachmuster in nicht-fiktionalen Texten.[47] In der Folge erschienen sowohl in der Literatur- als auch in der Geschichtswissenschaft Werke, welche sich allgemein mit der Frage nach den Möglichkeiten der zunehmend digitalen Forschungsmethoden und den Auswirkungen für die jeweilige Disziplin auseinandersetzten.

So wies Matthew L. Jockers 2013 in der Monografie *Macroanalysis* anhand mehrerer Beispiele auf das Potenzial von Topic Modeling hin. Dabei fokussierte er sich nicht nur auf die Erkennung von Themen, sondern versuchte auch anhand der Muster Rückschlüsse auf die Herkunft und das Geschlecht von Autorinnen und Autoren zu ziehen. Neben Topic Modeling evaluierte Jockers weitere digitale Werkzeuge für die Literaturwissenschaft. Er kam zum Schluss, dass Close Reading angesichts der neuen digitalen Möglichkeiten und der verfügbaren Datenmengen nicht mehr als einzige Methode genügt, weshalb er für die stärkere Einbindung makroanalytischer Zugänge plädierte. Diese sollten traditionelle Methoden nicht konkurrenzieren oder ersetzen, sondern komplementäre Zugangswege eröffnen.[48] 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

[44] Vgl. Graham, Milligan 2012; Graham, Weingart, Milligan 2012.
[45] Die Stanford Natural Language Processing Group pulizierte 2009 eine erste Version der Stanford Topic Modeling Toolbox, welche auf der Programmiersprache Scala aufbaute und LDA verwendet. Vgl. Brett 2012. Für den Download: https://downloads.cs.stanford.edu/nlp/software/tmt/tmt-0.4, 31.08.2022.
[46] Meeks und Weingart sprachen ebenfalls von einer „explosion of interest" seit 2010. Vgl. Meeks, Weingart 2012.
[47] Vgl. Rhody 2012.
[48] Vgl. Jockers 2013.

Daten und Tools reflektierten sie die Einflüsse digitaler Methoden auf das Selbstverständnis und die Arbeitsweise der Geschichtswissenschaften.[49]

Während sich die geisteswissenschaftliche Auseinandersetzung mit Topic Modeling bis dahin vornehmlich auf den englischsprachigen Raum fokussierte, fand der Ansatz vermehrt auch in anderen Sprachgebieten Anklang. So untersuchten beispielsweise Ingrid Falk, Delphine Bernhard und Christophe Gérard mit Hilfe von Topic Modeling die Veränderung der semantischen Bedeutung in der Verwendung des französischen Begriffs „quenelle“ auf Basis diverser Zeitungen im Zeitraum von 1987 bis 2014.[50] Neben dieser spezifischen linguistischen Studie setzte sich Guillaume Carbou mit epistemologischen Fragen bezüglich der Rolle statistischer Textauswertung – darunter auch Topic Modeling – in den Digital Humanities auseinander. Dabei betonte er die Wichtigkeit, sich der erkenntnistheoretischen Grundlagen der eigenen Forschungsarbeit bewusst zu sein und warnte vor einer übereilten Anwendung textstatistischer Verfahren ohne das Vorhandensein basaler Kenntnisse von Theorien zur Interpretation von Texten.[51]

Im deutschsprachigen Raum finden sich ab Mitte der 2010er Jahre zunehmend Anwendungen von Topic Modeling. In diesem Zusammenhang ist vor allem die Genre-Forschung von Christof Schöch, aktuell Professor für Digital Humanities an der Universität Trier, erwähnenswert. Er nutzte Topic Modeling unter anderem dazu, französische Dramen aus dem Zeitraum von 1630 und 1789 zu analysieren. Dabei war es ihm möglich, die die Dramen aufgrund der Auftretenswahrscheinlichkeiten der Topics den Subkategorien Tragödie, Komödie und Tragikomödie zuzuweisen und signifikante Unterschiede innerhalb der Subkategorien aufzuzeigen. Neben den Resultaten ist vor allem der Umstand, dass er die Auswirkungen unterschiedlicher Konfigurationsoptionen beim Modellierungsprozess in die Untersuchungen miteinbezog, hervorzuheben. Es gelang ihm, die – für viele Geisteswissenschaftlerinnen und Geisteswissenschaftler schwierig verständlichen – Resultate aus dem informatischen Bereich für die eigene Forschung nutzbar zu machen.[52]

[49] Vgl. Graham, Milligan, Weingart 2015.
[50] Vgl. Falk, Bernhard, Gérard 2014. Trotz intensiver Suche konnten im französischsprachigen Raum allgemein nur wenige geeignete Studien mit Schwerpunkt auf die Anwendung von Topic Modeling gefunden werden, weshalb in der vorliegenden Arbeit der Fokus auf englisch- und deutschsprachigen Werken liegt.
[51] Vgl. Carbou 2017.
[52] Vgl. Schöch 2016; Schöch 2017. Neben dem Verfassen weiterer Arbeiten zu Topic Modeling präsentierte Schöch den Ansatz auch an Workshops. Aktuell betreut er auch mehrere Hochschularbeiten zu Topic Modeling. Für eine Übersicht vgl. https://christof-schoech.de/tag/method-topic-modeling, 31.08.2022.

Allerdings konnten auch Forschende, welche aus ihren Unsicherheiten bezüglich den Grundlagen des Algorithmus kein Geheimnis machten, mit Hilfe von Topic Modeling vielversprechende Resultate erzielen, was zwei im Jahr 2017 erschienene Studien mit geschichtswissenschaftlichem Hintergrund zeigen. Peter Andorfer widmete sich der Frage, ob und inwiefern die mit Topic Modeling erzeugten Wortketten einer manuellen Verschlagwortung überlegen sind, und veranschaulichte dies am Beispiel einer Briefkorrespondenz aus dem 19. Jahrhundert. Obwohl der Autor die automatisierte Erschliessung zunächst als „bessere" Methode darstellte, relativierte er diese Ansicht zum Schluss mit dem Verweis auf verschiedene offengebliebene Fragen. Die ebenfalls formulierte Frage, ob Topic Modeling auch ohne eingehendere statistische Kenntnisse umgesetzt werden kann, bejahte er einerseits mit dem Verweis auf einige der Resultate, andererseits äusserte er jedoch auch Zweifel an deren Repräsentativität.[53] Martin Fechne und Andreas Weiss kritisierten Andorfers Schlussfolgerungen und plädierten dafür, dass trotz eines mangelnden Verständnisses der Algorithmen und Werkzeuge zumindest die Konsequenzen von deren Benutzung abgeschätzt werden sollten. Für die Analyse zweier Projekte mit Wissensbeständen aus dem 19. Jahrhundert setzten sie auf ein kombiniertes Verfahren manueller und automatisierter Vergabe von Topics, indem sie iterativ mit Training- und Testsets arbeiteten und auf das jeweilige Korpus sowie die Zielsetzung hin zugeschnittene Workflows entwickelten. Aufgrund vielversprechender Resultate erachteten die Autoren Topic Modeling als geeigneten Ansatz für eine breitere Nutzung in den Geistes- und Geschichtswissenschaften.[54]

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 und somit ein grosses Potenzial für weiterführende Anwendungen von Topic Modeling besteht.[55] Auch Claus Boye Asmussen und Charles Møller nahmen in einem 2019 erschienenen Aufsatz wissenschaftliche Publikationen in den Fokus, verfolgten jedoch einen stärker praxisorientierten Ansatz. Das Ziel ihrer Arbeit bestand nämlich darin, auf Basis von LDA einen Workflow zu konzipieren, der sich allgemein zum Zweck einer breitangelegten Literaturrecherche eignet. Anhand eines Korpus von 650 Papers zu Zeitungen, Pressetexten, Reden, Tweets sowie Blog- und Forumsposts zeigten sie, dass sich mit Hilfe von Topic Modeling in kurzer Zeit ein guter Überblick über eine grosse Textmenge ver-

[53] Vgl. Andorfer 2017.
[54] Vgl. Fechne, Weiss 2017.
[55] Vgl. Wehrheim 2019.

schaffen lässt. Die Autoren erachteten diese Vorgehensweise zudem im Vergleich zu herkömmlichen und häufig auf subjektiven Kriterien beruhenden Auswahlverfahren, als transparenter, da mit dem von ihnen entwickelten Verfahren der Prozess reproduziert werden kann.[56]

In einer jüngst erschienen Studie verglichen Tobias Hodel, Dennis Möbus und Ina Serif die Resultate der beiden sowohl in den Digital Humanities als auch in der Informatik vorrangig genutzten Topic-Modeling-Engines MALLET und Gensim, wobei sie als Grundlage drei unterschiedliche historische Korpora verwendeten. 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 liessen und auch der Nachvollzug von Abschreibeprozessen möglich ist. Der Aufsatz ist ein ausgezeichnetes Beispiel dafür, wie sich eine inhaltsorientierte Anwendung mit Reflektionen sowohl technischer als auch epistemologischer Natur verbinden lassen.[57]

1.4. Erkenntnisinteresse und Aufbau

In den vorherigen Unterkapiteln wurden die Grundprinzipien von Topic Modeling erläutert und anhand ausgewählter Studien ein Überblick über die bisherigen Forschungstendenzen gewährt. Auf Basis dieser Ausführungen fällt es leichter, das eigene Erkenntnisinteresse zu artikulieren und in der aktuellen Forschungslandschaft zu verorten. Die vorliegende Arbeit verfolgt in erster Linie eine anwendungs- und resultatorientierte Stossrichtung und widmet sich der 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 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

[56] Vgl. Asmussen, Møller 2019.
[57] Vgl. Hodel, Möbus, Serif 2022.

Arbeitsschritte beim Topic Modeling aufgeführt sind. Dabei werden jeweils zuerst die theoretischen und praktischen Erkenntnisse aus der Forschungsliteratur erwähnt, bevor die eigene Realisierungsform charakterisiert und begründet wird. Zu diesem Zweck 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[58] 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.

[58] Der Anglist Martin Mueller schlug den Begriff „Scalable Reading“ als Bezeichnung für einen Ansatz vor, der Distant- und Close-Reading-Verfahren verbindet, indem wie mit einem Zoomobjektiv zwischen Mikro- und Makroperspektive hin und her bewegt wird. Vgl. Viehhauser 2018: 32.

2. Methode

2.1. Preprocessing

2.1.1. Forschungspositionen

Nachdem ein Korpus für die Analyse ausgewählt wurde, bedingt jede Form der computergestützten Textanalyse eine vorgängige Aufbereitung der Texte, wobei diese je nach Ausgangslage, Methode, Anwendungsbereich und Anforderungen unterschiedlich ausfallen kann. Neben der Konvertierung der Daten in das von der jeweiligen Instanz benötigte Format bestehen beim Topic Modeling die beiden wesentlichen Schritte in der natürliche Sprachverarbeitung (Natural Language Processing) des Textes und dessen Unterteilung in einzelne Segmente (oder Dokumente). Die sprachliche Vorbearbeitung, welche in den Bereichen Data Mining und Information Retrieval eine zentrale Rolle spielt, kann unterschiedliche Ausprägungen der Manipulation des Textes beinhalten. So gehört die Entfernung von Satzzeichen und die systematische Ersetzung von Gross- durch Kleinbuchstaben (Case Normalization) in der Regel zum Mindestmass an Normalisierung. Eine stärkere Normalisierung kann durch Algorithmen zum Stemming oder der Lemmatisierung vorgenommen werden. Beim Stemming werden die Suffixe der Wörter entfernt. So würden beispielsweise die beiden Formen „lachen“ und „lachte“ auf den Wortstamm „lach“ reduziert. Die Lemmatisierung geht einen Schritt weiter, indem sie den Kontext eines Wortes berücksichtigt und so auch stark abweichende Ausprägungen, wie beispielsweise „besser“ als Variante von „gut“, auf den Kern der Bedeutung zurückführt. Aufgrund regelspezifischer Differenzen sind Stemming- und Lemmatisierungsalgorithmen sprachabhängig, wobei sie je nach Konfiguration unterschiedlich stark normalisieren.[59]

Während die Entfernung von Satzzeichen und die Normalisierung der Gross- und Kleinschreibung beim Topic Modeling unumstritten ist und deshalb auch in der Java-Version von MALLET standardmässig umgesetzt wird, gestaltet sich die weitere Verarbeitung in der Forschung unterschiedlich. Tendenziell scheinen Forschende mit geschichtswissenschaftlichem Hintergrund in dieser Hinsicht jüngst zurückhaltender zu sein als diejenigen aus der Linguistik oder Literaturwissenschaft, was – abgesehen von grundsätzlichen Bedenken hinsichtlich einer möglichen Manipulation der historischen Quellen[60] – auch auf die Schwierigkeit der Normalisierung von Texten mit inkonsistenter Orthografie zurückzuführen sein dürfte.[61] So verzichte-

[59] Vgl. Lamba, Madhusudhan 2022: 79-85.
[60] Gemäss Hodel, Möbus und Serif wird unter diesem Hintergrund das Thema „data cleaning“ in den Digital Humanities aktuell kritisch diskutiert. Vgl. Hodel, Möbus, Serif 2022: 188.
[61] Obwohl Tools für die automatische Normalisierung von vormodernen Texten existieren, erfordert eine erfolgreiche Umsetzung ein aufwändiges Training. Vgl. Hodel, Möbus, Serif 2022: 199.

ten Andorfer, Fechne und Weiss sowie Hodel, Möbus und Serif auf Stemming und Lemmatisierung.[62] Dahingegen nahm Schöch vorgängig eine Lemmatisierung vor, indem er ein eigens für älteres Französisch entwickeltes Sprachmodell verwendete. Dieses nutzte er auch im Rahmen einer Parts-of-Speech Taggings (POS) für die Ermittlung der Wortarten der Lemmata. Mit dem Ziel, nur sinntragende Wortkategorien zu berücksichtigen, wurden alle Wörter ausser Nomen, Verben, Adjektive und Adverbien vorweg entfernt.[63] Jockers ging einen Schritt weiter, indem er – nach der Identifizierung der Wortarten mittels POS – alle Wortarten ausser Nomen eliminierte.[64]

Der Ausschluss bestimmter Wörter oder Wortarten kann entweder durch vorgängiges Löschen auf Ebene des zugrundeliegenden Textes oder erst beim Modellierungsvorgang vorgenommen werden. Letzteres bedingt das vorgängige Erstellen einer Liste mit Begriffen (Stopwords), welche beim Topic-Modeling-Prozess ignoriert werden sollen. In der Regel sind dies Funktionswörter wie Artikel oder Präpositionen, die keine oder wenig inhaltliche Bedeutung enthalten. Die Benutzung von Stopwords-Listen bietet einige praktische Vorteile. So kann die Stopwords-Liste unabhängig von der Ermittlung von Wortarten iterativ auf den jeweiligen Korpus zugeschnitten werden oder es können mehrere Stopwords-Listen für dieselbe Textgrundlage getestet werden. Zudem sind sprachspezifische Listen leicht zugänglich, indem sie beispielsweise von Natural-Language-Processing-Tools wie MALLET enthalten oder im Internet auffindbar sind. Während Stemming und Lemmatisierung nicht in allen Studien zur Anwendung kommt, gehört die Verwendung von Stopwords-Listen – wie die Fallnormalisierung und Entfernung der Satzzeichen – zum Mindeststandard an sprachlicher Normalisierung.[65]

Da beim Topic Modeling der Text in Tokens unterteilt wird, werden auch semantisch zusammenhängenden Wortpaaren- oder gruppen (z.B. „Heiliger Stuhl“ oder „Vereinigte Staaten von Amerika“) getrennt. Es gibt verschiedene Ansätze zum Umgang mit den sogenannten N-Grammen (N-Grams), wobei beispielsweise MALLET ein Topical-N-Grams-Modell zur algorithmischen Erkennung von zusammenhängenden Begriffen beinhaltet. David Mimno, der technische Betreuer von MALLET, betonte in einem Blog-Beitrag einerseits die Wichtigkeit der Berücksichtigung von N-Grammen für spezifische Fragen, sprach sich andererseits gegen die Verwendung des erwähnten Modells aus. Als pragmatischen und zielführenden Weg beschrieb er eine Verkettung der zusammengehörigen Wörter (z.B. „heilige_stuhl“) im Rahmen

[62] Vgl. Andorfer 2017: Kap. 6; Fechne, Weiss 2017: Kap. 1.2; Hodel, Möbus, Serif 2022: 197.
[63] Vgl. Schöch 2017: Abs. 17-18.
[64] Vgl. Jockers 2013: 131-133.
[65] Vgl. Wallach, Mimno, McCallum 2009: 1; Jockers 2013: 131; Graham, Milligan, Weingart 2015: 86; Lamba, Madhusudhan 2022: 85.

des Preprocessing. Während diese Verbindungen für Menschen gut verständlich sind, interpretiert MALLET sie als singuläre Tokens, womit einzelne Bestandteile nicht durch Stopwords-Listen entfernt werden.[66]

Neben der sprachlichen Vorbearbeitung der Texte bildet die Segmentierung des Korpus in Einzeldokumente ein zentrales Element. Gemäss Tang et al., die den Einfluss unterschiedlicher Parameter bei der Anwendung von LDA untersuchten, spielt die Anzahl der Dokumente eine wichtige Rolle. Während es für 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, benötigen aber eine gewisse Mindestlänge. In der praktischen Anwendung können deshalb entweder kurze Dokumente mit ähnlichen Attributen (z.B. Tweets von einer Person) in ein umfangreicheres Segment zusammengeführt werden oder grössere Dokumente (z.B. Bücher oder Theaterstück) in kleinere Segmente unterteilt werden.[67] 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.[68] So entschied sich Andorfer aufgrund des unterschiedlichen Umfangs der Briefe dafür, längere Dokumente nach 1'000 Wörtern zu trennen.[69] Denselben Wert wählte auch Schöch für die Zergliederung der französischen Theaterstücke.[70] Hodel, Möbus und Serif sprachen sich ebenfalls für eine Unterteilung in kleinere Einheiten aus, wobei sie für die verschiedenen Korpora unterschiedliche Bezugseinheiten (Sätze, Wörter) wählten und jeweils mehrere Werte testeten.[71] Newman und Block wählten dagegen die einzelnen Zeitungsartikel und Werbeblöcke als Segmente, womit sie sich stärker an den vorgegebenen inhaltlich-formalen Strukturen orientierten.[72]

2.1.2. Realisierung

Die Ausgangslage in der vorliegenden Arbeit bestand darin, dass sich die Wetterbeobachtungen nach Standort (Einsiedeln, Pfäffikon, Freudenfels, Fahr) unterteil in vier Excel-Dateien befanden und jede Zeile je einen der insgesamt 5’178 Einträge abbildete. Da die Einträge von

[66] Vgl. Mimno 2015.
[67] Vgl. Tang et al. 2014: 196.
[68] Vgl. Graham, Milligan, Weingart 2015: 117.
[69] Vgl. Andorfer 2017: Kap. 4.2.
[70] Vgl. Schöch 2016: Abs. 16.
[71] Vgl. Hodel, Möbus, Serif 2022: 194, 200.
[72] Vgl. Newman, Block 2006: 757.

Hand transkribiert und mehrfach überprüft wurden, ist das Risiko ist das Risiko von Verzerrungen in Form von Verschreibungen als gering zu erachten.[73] In einem ersten Schritt wurden die vier Tabellen zu einer vereint und dabei Beobachtungen aus der Hand von Reding sowie die nachträglich von Dietrich rekonstruierten Einträge entfernt. Aufgrund der inkonsistenten Orthografie und der sprachlichen Eigenheiten des Autors wurde 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. Für die vorliegende Arbeit wurden zudem Zeichen, die im heutigen deutschen Sprachgebrauch nicht mehr gängig sind, aufgelöst. Dazu gehören Ligaturen im Lateinischen (æ, œ*)* und verschiedene Buchstaben (ë, ÿ , ÿ), bei denen die Diakritika entfernt wurden. Das anlautende „v“ anstelle von „u“ und das Scharf-s („ß“) wurden beibehalten. Die Getrennt- und Zusammenschreibung wurde bei der Transkription möglichst vorlagengetreu abgebildet. Da insbesondere die damalige Verwendung von „zu“ mit Infinitiv von der heutigen abweicht, wurden hier teilweise nachträglich Worttrennungen (z.B. „angefangen zu regnen“ statt „angefangen zuregnen“) vorgenommen.

Weitere Normalisierungen, wie etwa der Gross- und Kleinschreibung und die Entfernung der Satzzeichen, werden in der Java-basierten Version von MALLET vom Tool selbst vorgenommen. Erste Versuche haben gezeigt, dass die Tokenisierung semantisch zusammengehöriger Einheiten zu spezifischen Problemen im vorliegenden Korpus führen kann. So verwendete der Autor beispielsweise im Zusammenhang mit der Temperatur häufig negierende oder qualifizierende Begriffe (z.B. „nit kalt“, „grimmig kalt“, „kein Schnee“, „vill Schnee“), die nur in Kombination mit anderen Wörtern bedeutungstragend sind. Um ihre Funktion und Bedeutung erkennbar zu machen, wurden mit Hilfe von AntConc[74] häufig auftretende Wortpaare- oder gruppen identifiziert und auf Textebene zu N-Grammen zusammengeführt. Trotz der vorgenommenen Auszählungen basierte die Wahl weitgehend auf subjektiven Gesichtspunkten. Dasselbe gilt auch für die Stopwords-Liste, welche im Hinblick auf den Ausschluss von Funktionswörtern erstellt und im Rahmen der unzähligen Testläufe fortwährend modifiziert wurde. Abgesehen von den Durchläufen zur Ermittlung des Optimierungsintervalls[75] wurden für alle Modellierungsprozesse dieselbe Stopwords-Liste verwendet.

[73] Vor allem bei Korpora, die mittels automatischer Texterkennung aufbereitet werden, können dadurch bedingte Fehler zu Verzerrungen des Resultats führen. Vgl. Fechne, Weiss 2017: Kap. 1; Wehrheim 2019: 92.

[74] AntConc wurde von Laurence Anthony entwickelt und eignet sich besonders gut für die Analyse von Wortkonkordanzen. So lassen sich mit AntConc die Wörter auf beiden Seiten eines Stichworts anzeigen, die Auftretenshäufigkeit einzelner Begriffe über Zeiträume hinweg darstellen oder der Kontext eines gesuchten Terminus abbilden. Vgl. Graham, Milligan, Weingart 2015: 79-81. Für den Download vgl. https://www.laurenceanthony.net/software/antconc, 31.08.2022.

[75] Die beiden Stopwords-Listen unterscheiden sich nur marginal. Vgl. dazu Kap. 6.1.

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 witterungsbedingten Jahreszyklus[76] angeordnet. Das heisst, die Daten wurden 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. Die Tagebucheinträge eignen sich aufgrund ihrer chronologischen Anordnung insbesondere auch für diachrone Betrachtungen auf der Zeitachse. Um Hinweise auf Konstanten und Veränderungen über den gesamten Zeitraum zu erhalten, wurden die Daten in einer dritten Art der Segmentierung in Einheiten pro Jahr zusammengeführt. Abgesehen von den Jahren, in denen ein Standortwechsel stattfand, ist in dieser Darstellung auch der ortsbezogene Einfluss ersichtlich, da Aufenthaltsorte und Zeiträume bekannt sind.

2.2. Modellierungsprozess

2.2.1. Forschungspositionen

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[77] – die optimale Anzahl an Topics nicht im Vornherein festsetzen kann. Allerdings ist dies auch für Forschende schwierig, da sich die „optimale“ Anzahl Topics nur iterativ ermitteln lässt.[78] Wesentliche Einflussfaktoren sind dabei Umfang und der Heterogenität des Korpus sowie die erwartete Trennschärfe und Interpretierbarkeit der einzelnen Topics. Letzteres bezieht sich darauf, inwiefern Menschen die ausgegebenen Wortketten nachvollziehen können. Dabei ist zu beachten, dass in

[76] Diese Art der Segmentierung orientiert sich an den Darstellungen von langjährigen Mittelwerten von Monatstemperetaturen, welche die klimatischen Bedingungen für einzelne Ortschaften oder Regionen abbilden. Es wäre ebenfalls möglich gewesen, die Daten zunächst in chronologischer Form modellieren zu lassen und dann aus den ausgegebenen Wahrscheinlichkeiten einzelner Topics den Mittelwert zu berechnen. Tests mit dieser Vorgehensweise haben jedoch gezeigt, dass dabei starke Verzerrungen aufgrund orthografischer oder stilistischer Faktoren möglich sind.

[77] 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 2022: 194-195. Auch Wehrheim stützte sich für die Bestimmung Anzahl Topics auf Berechnungen, wobei er jedoch eine andere Metrik verwendete. Vgl. Wehrheim 2019: 113-114.

[78] Vgl. Hodel 2022: 164.

der Regel nicht alle ausgegebenen Topics verständlich erscheinen. Gemäss Jockers impliziert dies nicht eine Infragestellung des Modells an sich, sondern erfordert eine Auswahl der geeigneten Topics für die jeweilige Analyse.[79] Da sich mit der Veränderung der Anzahl auszugebender Topics auch deren Zusammensetzung wandelt, hat dies einen direkten Einfluss auf die Trennschärfe. Bei einer geringen Anzahl setzen sich die Topics nämlich tendenziell eher aus sehr allgemeinen Begriffen zusammen, im umgekehrten Fall kommen auch eher seltenere Termini in den Wortketten vor.[80] Asmussen und Møller formulierten deshalb als Faustregel, dass eine geringe Anzahl an Topics einem allgemeinen Überblick dient und eine höhere Zahl eine detailliertere Perspektive bietet.[81]

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. Mit jeder zusätzlichen Iteration werden die Wahrscheinlichkeiten präziser berechnet und das Modell somit theoretisch immer besser. Eine höhere Zahl an Iterationen wirkt sich allerdings auch auf die für die Modellierungsprozess erforderliche Zeitdauer aus, weshalb in der Praxis ein Kompromiss zwischen der angestrebten Qualität des Modells und der verfügbaren Zeit erforderlich ist. Gemäss Jockers Beobachtungen verändert sich die Zusammensetzung und Qualität ab einer gewissen Zahl allerdings kaum mehr.[82] Schöch, der die Einflüsse verschiedener Parameter untersuchte, spezifizierte, dass sich ab einer gewissen Anzahl Iterationen weniger Unterschiede in der Zusammensetzung der Topics als vielmehr bei der Gewichtung der einzelnen Wörter innerhalb der Topics ergaben. Somit führt eine höhere Zahl der Wiederholungen dazu, dass sich die Ergebnisse unterschiedlicher Modellierungsprozesse (mit denselben Parametern) weniger voneinander unterscheiden.[83]

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

[79] Vgl. Jockers 2013: 129-130.
[80] Vgl. Jockers 2013: 127-128; Hodel 2022: 164.
[81] Vgl. Asmussen, Møller 2019: Kap. Pre-processing.
[82] Vgl. Jockers, Thalken 2020: 224-225.
[83] Vgl. Schöch 2017: Abs. 14, 20-21.

definiert, wie stark die Abweichung von der standardmässig flachen Wahrscheinlichkeitsverteilung abweicht. Eine sehr starke Optimierung führt zu einer starken Differenz zwischen einigen wenigen Topics, die mit hoher Wahrscheinlichkeit vorkommen, und den Übrigen mit tiefer Auftretenswahrscheinlichkeit. Je nach Erkenntnisinteresse kann eine mehr oder weniger starke Modifikation sinnvoll sein. Mehr ist allerdings nicht zwingend besser, da das Risiko einer zu starken Beeinflussung besteht.[84] So untersuchte Schöch am Beispiel der französischsprachigen Dramen das Zusammenspiel der verschiedenen Parameter und wählte dazu bei 6'000 Iterationen für die Anzahl an Topics sechs Werte zwischen 50 und 100 und für das Optimierungsintervall acht Nummern zwischen 0 und 3'000. Eine Auswertung der 48 Modelle ergab, dass das beste Resultat bei 60 Topics und einem Optimierungsintervall von 300 erzielt wurde.[85]

2.2.2. Realisierung

Da für die Wahl der Parameter, welche den Modellierungsprozess beeinflussen, keine allgemeingültigen Standardwerte bestehen, müssen diese für den jeweiligen Korpus iterativ ermittelt werden. Im Unterschied zu den meisten 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, was mit der allgemein geringen Datenmenge und bis zu einem gewissen Grad wohl auch mit deren inhaltlichen Homogenität begründet werden kann. 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, womit sich die Möglichkeit differenzierterer und weiterführender Analysen ergibt. 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.

Im Weiteren wurde ermittelt, inwiefern sich die Optimierung der Hyperparameter auf das Resultat auswirken. Zu diesem Zweck wurden auf Basis des Datensatzes mit der Segmentierung nach Monaten kumuliert mehrere Modellierungsprozesse mit unterschiedlichen Kombinationen hinsichtlich der Anzahl an Topics, der Anzahl Iterationen und des Optimierungsintervalls

[84] Vgl. Schöch 2016.
[85] Vgl. Schöch 2017: Abs. 20.

durchgeführt. Zuerst wurde der Einfluss des Optimierungsintervalls mit der Standardkonfiguration von 400 Iterationen[86] für 5 und 20 Topics untersucht. Für das Optimierungsintervall wurden die Werte 0, 10, 50 und 100 eingesetzt und die Resultate als Heatmaps visualisiert wurden. Der theoretisch begründete Effekt, dass sich die Manipulation des Optimierungsintervalls auf die Abgrenzbarkeit der Topics auswirkt, liess sich hierbei sowohl beim Modell mit 5 als auch demjenigen mit 20 Topics gut nachvollziehen. Am deutlichsten unterscheiden sich die Modelle ohne Optimierungsintervall von denjenigen mit. Das bedeutet, dass die Entscheidung, ob ein Optimierungsintervall gewählt wird oder nicht, stärkere Konsequenzen hat als die Frage nach dem Wert desselben.

Bei 5 Topics zeigt sich die Tendenz, dass sich die allgemeinen Begriffe in einem Topic bündeln und die übrigen Topics höhere Auftretenswahrscheinlichkeiten in wenigen Monaten abbilden. Bei 20 Topics ist dieser Effekt noch ausgeprägter, da viele Topics erhöhte Auftretenswahrscheinlichkeiten in einzelnen Monaten aufweisen. Umgekehrt bedeutet dies, dass die Zahl der Monate, bei denen die Auftretenswahrscheinlichkeit bei null Prozent liegt, steigt. Während die Trennschärfe vor allem beim Optimierungsintervall 10 und 50 hoch ist, ist sie bei 100 rückläufig. So treten bei 5 Topics am meisten Nullwerte mit dem Optimierungsintervall 50 und bei 20 Topics das Maximum für den Wert 10 auf. Obwohl es weniger Nullwerte gibt, werden auch mit dem Optimierungsintervall 100 noch trennscharfe Topics abgebildet, da bei vielen Monaten Auftretenswahrscheinlichkeiten unter drei Prozent vorkommen.

In einem zweiten Schritt wurde die Zahl der Iterationen vom Standartwert 400 auf 6'000[87] erhöht und dabei der Einfluss des Optimierungsintervalls mit den Werten 0, 10, 50, 100, 500, 1'000 und 2'000 bei 5 und 20 Topics beobachtet. Dabei konnten ähnliche Tendenzen wie bei 400 Iterationen festgestellt werden. So weisen die Topics auf jeden Fall eine höhere Trennschärfe auf, wenn ein Optimierungsintervall gewählt wird. Die meisten Nullwerte zeigten sich sowohl bei 5 als auch bei 20 Topics mit dem Optimierungsintervall 50. Die Topics werden zwar mit höherem Optimierungsintervall wiederum indifferenter, allerdings ist dieser Effekt im Vergleich zur Umsetzung mit 400 Iterationen weniger stark ausgeprägt. Somit scheint die höhere Zahl der Iterationen im vorliegenden Fall erheblich zu einer Stabilisierung beizutragen, womit die Möglichkeit, dass sich wegen der Wahl des Optimierungsintervalls erhebliche Verzerrungen ergeben könnten, geringer ausfällt. Aus diesem Grund wurden alle weiteren Modellie-

[86] Die Darstellungen, welche als Vorlage für die beschriebenen Analysen mit 400 Iterationen verwendet wurden, werden an dieser Stelle nicht abgebildet, weil die übrigen Modellierungsprozesse alle mit 6'000 Iterationen durchgeführt wurden. Für die Modelle mit 400 Iterationen vgl. https://observablehq.com/@lheinzmann/tm_intervalloptimierung_400, 31.08.2022.

[87] Für die Wahl des Werts dienten als Orientierungspunkt die Analysen von Schöch, der ebenfalls mit 6'000 Iterationen arbeitete. Vgl. Schöch 2017: Abs. 19.

rungsprozesse mit 6’000 Iterationen durchgeführt. Da möglichst trennscharfe Topics angestrebt werden, wurde das Optimierungsintervall 50 gewählt. Es handelt sich hierbei um Grundsatzentscheidungen, die im Rahmen der Analyse teilweise kritisch reflektiert werden.

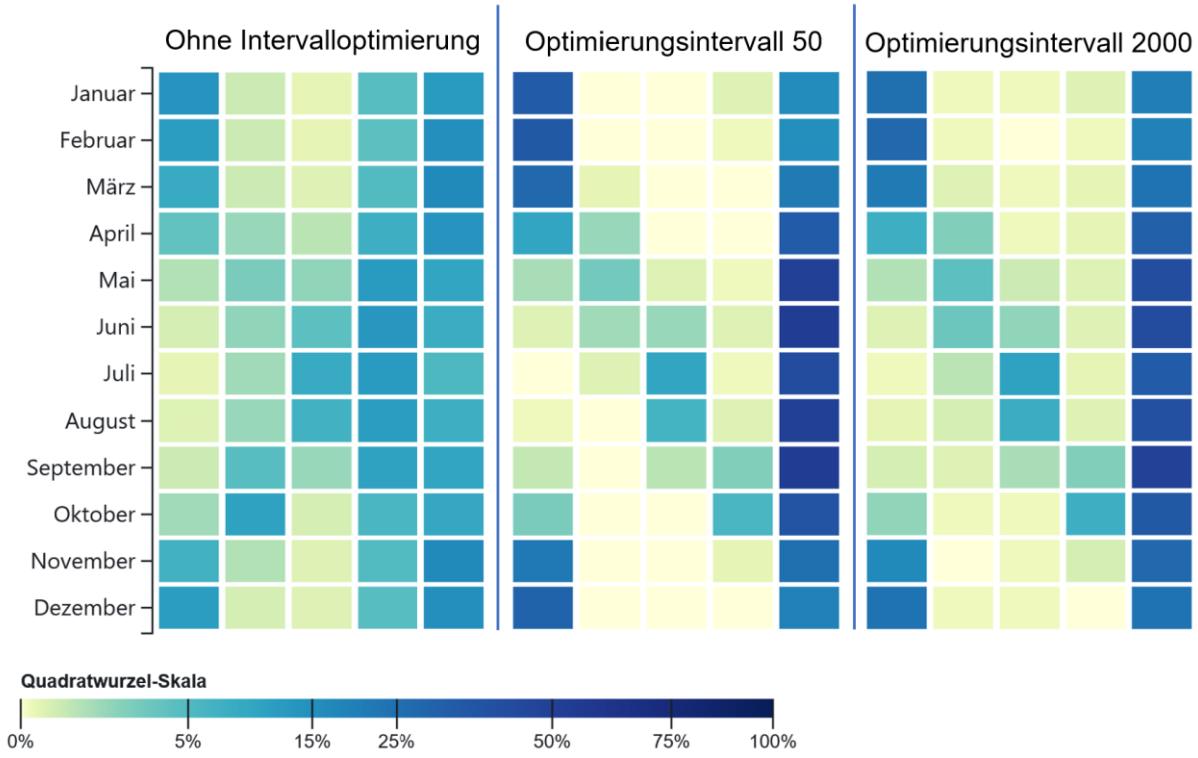

Abb. 1: Intervalloptimierung für 5 Topics bei 6'000 Iterationen*: Die Darstellung repräsentiert die Auftretenswahrscheinlichkeiten von drei Modellen mit jeweils 5 Topics, für die unterschiedliche Optimierungsintervalle verwendet wurden (0, 50, 2’000). Weitere Modelle sind im Observable-Notebook ersichtlich. Vgl. dazu https://observablehq.com/@lheinzmann/tm_intervalloptimierung_6000, 31.08.2022.*

2.3. Postprocessing

2.3.1. Forschungspositionen

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 Befehl die erwünschten Outputs definiert werden können. Ein Outptut enthält die Zusammensetzung der Topics, ein weiterer bildet deren Auftretenswahrscheinlichkeit in den einzelnen Dokumenten ab und ein dritter zeigt die Frequenz der Tokens innerhalb der einzelnen Topics. Im Weiteren kann ein Diagnosedokument im XML-Format erzeugt werden, welches eine Reihe weiterer statistischer Werte zu den Topics und Tokens enthält, die vor allem für stärker mathematisch ausgerichtete Ansätze Möglichkeiten einer eingehenderen Analyse bieten.[88] In dieser Form sind die Ergebnisse – vor allem bei einer

[88] Für weiterführende Informationen zu den im Diagnosedokument abgebildeten Werten vgl. https://mallet.cs.umass.edu/diagnostics.php, 31.08.2022.

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. Gemäss Lamba und Madhusdhan stellt die Visualisierung der Topic-Modeling-Resultate einen Teilbereich der Methode dar, weshalb neben Studien zum Thema auch mehrere Tools und viele Anschauungsbeispiele im Internet existieren. 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.[89]

Für die Frage, ob und welche Visualisierungen genutzt werden sollen, besteht kein allgemeiner Konsens. Fechne und Weiss verzichteten etwa gänzlich auf Visualisierungen, da sie der Gefahr einer vorschnellen Interpretation vorbeugen wollten. Sie verwiesen darauf, dass die Visualisierung ein Hilfsinstrument sei, mit Hilfe dessen die bereits durch Topic Modeling simplifizierten Daten weiter reduziert würden.[90] Im Gegensatz dazu verwendete Schöch ein breites Spektrum an Visualisierungsformen (Balken- und Liniendiagramme, Wordclouds, Heatmaps, Box-Plots, Scatter-Plots, Dendrogramm), um differenzierte Aspekte der Resultate hervorzuheben.[91] Gemäss Andorfer bieten sich für die Darstellung der Rohdaten die beiden Formen Wordcloud und Heatmap an. So können mit Hilfe von Wordclouds alle Begriffe und ihr Gewicht innerhalb eines Topics übersichtlich repräsentiert werden. Heatmaps eignen sich besonders für die Visualisierung der Topic-Dokument-Matrix, womit sich die Gesamtheit der Dokumente gut überblicken lässt.[92] Die Verwendung von Liniendiagrammen bietet sich vor allem für die Darstellung der Auftretenswahrscheinlichkeit von Topics im zeitlichen Längsschnitt. Deshalb ist diese Form überwiegend in den beiden Studien zu den historischen Zeitungen, im Blog zu Martha Ballards Tagebuch und in Wehrheims Aufsatz zum *Journal of Economic History* anzutreffen.[93]

2.3.2. Realisierung

Die mit Hilfe von MALLET erzeugten Rohdaten wurden in der vorliegenden Arbeit in mehreren Schritten weiterverarbeitet. Zuerst erfolgte eine Umwandlung der Textdatei mit den Informationen zur Auftretenswahrscheinlichkeit ins xslx-Format, was eine Vorbearbeitung in Excel ermöglichte. Der Vorteil dieses Schritts besteht darin, dass mit Hilfe der Formatierungsoptionen

[89] Vgl. Lamba, Madhusudhan 2022: 107-108.
[90] Vgl. Fechne, Weiss 2017: Kap. 1.2.
[91] Vgl. Schöch 2017.
[92] Vgl. Andorfer 2017: Kap. 5.2, 6.
[93] Vgl. Newman, Block 2006; Blevins 2010; Nelson 2012; Wehrheim 2019.

schnell eine provisorische Heatmap erzeugt werden kann, welche einen ersten Eindruck vermittelt. Dazu wurden auch die Topics, die MALLET in einer eigenen Datei ausgibt, den jeweiligen Spalten zugeordnet. Um die spätere Analyse und Beschreibung der Resultate zu vereinfachen und bestimmte Muster besser sichtbar zu machen, wurden die Anordnung der Spalten manuell angepasst. Diese Anordnung orientierte sich nicht an statistisch-mathematischen Kriterien, sondern wurde anhand des subjektiven Eindrucks der Auftretenswahrscheinlichkeiten vorgenommen. Schwieriger gestaltete sich die Weiterverarbeitung des Dokuments, welches die Frequenzen der einzelnen Tokens in den Topics abbildet. Um diese Daten in eine Tabellenform zu übertragen, wurde zunächst ein Excel-Makro[94] genutzt. Dabei ergaben sich allerdings Unstimmigkeiten hinsichtlich der Anordnung von Tokens mit gleicher Frequenz sowie Probleme mit der Darstellung von Sonderzeichen (Zeichensatzproblematik), weshalb die Wortfrequenzen dem Diagnosedokument entnommen und in Excel weiterverarbeitet wurden.

Da bezüglich des Modellierungsprozesses entschieden wurde, für jede Art der Segmentierung mehrere Modelle mit unterschiedlicher Anzahl an Topics zu generieren, erwies sich die Frage nach der geeigneten Repräsentation als evident. Trotz vereinzelter kritischer Stimmen[95] in der Forschung wurden Visualisierungen als geeignetes Mittel für die Darstellung der Outputs erachtet, zumal deren grosser Umfang eine Vereinfachung erforderlich machte. Insbesondere die Visualisierungsform „Heatmap“ erwies sich als gewinnbringend, da sie eine kompakte und übersichtliche Darstellung einer grossen Datenmenge erlaubt. Aus diesem Grund wurde sie im Zusammenhang mit den Auftretenswahrscheinlichkeiten der Topics sowohl für die Testläufe im Zusammenhang mit den Konfigurationsparametern für den Modellierungsprozess als auch für die Modelle in allen Arten der Segmentierung konsequent umgesetzt. Damit die Ergebnisse der Heatmaps vergleichbar sind, wurde für die Abbildung des Farbverlaufs immer dieselbe nichtlineare Skala (Quadratwurzel-Skala) verwendet, welche Werte im tieferen Bereich stärker hervorhebt.[96]

---

[94] Das Excel-Makro wurde von David L. Hoover programmiert und ist darauf ausgerichtet, den mit MALLET erzeugten Output tabellarisch darzustellen und zu visualisieren. Vgl. https://wp.nyu.edu/exceltextanalysis/visualize-mallet-topics, 31.08.2022.

[95] Es lässt sich nicht von der Hand weisen, dass Visualisierungen als weiterer Schritt der Datenverarbeitung eine potenzielle Quelle für Fehler und Falschinterpretationen sein können. Allerdings stellt dies eher eine Folge einer unsorgfältigen und unbedachten Arbeitweise dar und charakterisiert nicht den eigentlichen Zweck. Dieser begründet sich darin, grosse Informationsmengen in einer vereinfachten Form zugänglich zu machen. Diese Simplifizierung kann problematisch sein, wenn die zugrundeliegenden Mechanismen zu wenig verstanden und falsche Schlüsse gezogen werden. Allerdings zielen letztlich alle Methoden zur Erkenntnisgewinnung auf eine Vereinfachung ab, womit sich je nach Umsetzung die Möglichkeit unzureichender Resultate ergibt.

[96] Bei der Quadratwurzel-Skala handelt es sich um eine nichtlineare Skala. Während lineare Skalen eine genaue Visualisierung der Daten liefern, entspricht bei nichtlinearen Skalen ein gleichmässiger Abstand zwischen den Werten einem ungleichmässigen Abstand in der Visualisierung. Nichtlineare Skalen sind in der Regel dann sinnvoll, wenn Datensätze Werte mit sehr unterschiedlichen Grössenordnungen vorkommen. Dabei werden grössere Zahlen in einen kleineren Bereich komprimiert, womit die Differenzen in kleineren Wertebereichen trennschärfer abgebildet werden. Im Gegensatz zu der zu

Es wird nicht bestritten, dass sich Wordclouds zur Vermittlung eines allgemeinen Überblicks oder zur vergleichenden Darstellung der Wortfrequenzen einzelner Topics eignen, allerdings erscheint dies im Hinblick auf eine konsequente Umsetzung wenig sinnvoll zu sein, da bei mehreren Modellen eine grosse Menge an Wordclouds generiert wird.[97] Auch wenn es durchaus Möglichkeiten gibt, die Wortfrequenzen mehrerer Topics kompakter und übersichtlicher darzustellen,[98] wurden die Wortfrequenzen hier in anderer Form abgebildet. Für die Visualisierung der Ergebnisse wurde nämlich die webbasierte Plattform Observable genutzt, welche auf JavaScript aufbaut und die im Web weit verbreitete JavaScript-Bibliothek D3.js[99] integriert.[100] Basierend auf anderen bereits veröffentlichten Notebooks wurde Vorlage erstellt, welche sich für jede Art der Segmentierung nutzen liess und bei den einzelnen Modellen mehrere in Bezug stehende Elemente abbildet. An erster Stelle befindet sich eine Liste mit allen Topics und deren Auftretenswahrscheinlichkeit im Gesamtkorpus, dann folgt die Zusammensetzung der einzelnen Topics und die Frequenz der einzelnen Tokens in tabellarischer Form. Beide Elemente können über ein Dropdown-Menü sicht- oder unsichtbar gemacht werden. Unmittelbar darunter befindet sich eine Heatmap, welche die Auftretenswahrscheinlichkeit der Topics für die jeweilige Einheit darstellt sowie eine Visualisierung der verwendeten Skala. Zudem wurden zusätzliche Elemente (Farbwahl, Darstellungsbreite, Sortierung) integriert, welche optional interaktive Veränderungen der Gestalt der Heatmaps ermöglichen.

Insgesamt wurden die Elemente so angeordnet, dass die Informationen der Outputs eines Modells innerhalb eines kompakten Bereichs zugänglich sind und so die in den Heatmaps vermittelten Muster einfach mit den Topics und den Frequenzen der einzelnen Tokens in Beziehung gesetzt werden können. Da sich alle Modelle pro Art der Segmentierung in einem Notebook befinden, lässt sich schnell ein Überblick über die Unterschiede bei einer veränderten Zahl an Topics gewinnen. Im Weiteren wurde in jedem Notebook im Sinne der Transparenz

---

diesem Zweck am häufigsten genutzten logarithmischen Skala erlaubt die Quadratwurzel-Skala das Vorhandensein eines Nullwerts. Vgl. Wilke 2020: 16-20.

[97] Andorfer erwähnte beispielsweise einerseits den inflationären Gebrauch von Wordclouds in den Digital Humanities und erzeugte andererseits im Rahmen seiner Untersuchungen über hundert davon. Entsprechend wies er im Fazit auf das Problem der mangelnden Benutzerfreundlichkeit hin und artikulierte Vorschläge für interaktivere Darstellungsformen. Vgl. Andorfer 2017: Kap. 4.4, 6.

[98] Garry Diz publizierte beispielsweise auf Observable eine Visualisierung der Wortfrequenzen in Form eines Dendrogramms. Vgl. https://observablehq.com/@dizdata/radial-dendrogram-topic-modeling-visualization, 31.08.2022.

[99] Hodel et al. thematisierten die Visualisierungsmöglichkeiten mit der JavaScript-Bibliothek D3.js in einem Aufsatz zum digitalen Zugang zu Informationen in Archiven. Vgl. Hodel et al. 2022: 33-39.

[100] Observable ermöglicht die Erstellung von Notebooks, in denen sowohl Programmcode für die Erstellung von Visualisierungen als auch Auszeichnungssprachen (html, Markdown) für die Gestaltung oder Dokumentation benutzt werden können. Die Notebooks können zudem im Web publiziert und frei zugänglich gemacht werden. Analog zu den Verfahren des Versionsverwaltungsdienstes GitHub können veröffentlichte Notebooks geklont und so für die eigene Arbeit nutzbar gemacht werden. Wie GitHub setzt Observable deshalb auf ein partizipatives Mitwirken der Community. So besteht zwar ein Bezahlmodell, allerdings lassen sich die meisten Funktionen kostenlos nutzen. Aufgrund dieser Ausrichtung kann angenommen werden, dass die publizierten Notebooks noch längere Zeit zugänglich bleiben.

auch die Listen zu den verwendeten Stopwords 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. Diese umfassen auch die zugrundeliegenden Daten, welche bei Bedarf heruntergeladen werden können.[101]

Um bestimmte Tendenzen, welche in den Heatmaps erkennbar wurden, eingehender zu studieren, wurden Begleitanalysen mit Voyant Tools[102] 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 Stopwords-Listen verwendet wie beim Topic Modeling.[103] Die mit Voyant Tools erstellten Balken- und Liniendiagramme wurden soweit möglich direkt in die Arbeit integriert.

[101] 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 momentan noch nicht öffentlich zur Verfügung gestellt.

[102] Voyant Tools ist eine webbasierte Textanalyseplattform, welche im Rahmen des Hermeneuti.ca-Projekts von Séfan Sinclair und Geoffrey Rockwell entwickelt wurde. Dieses zielte darauf ab, die computergestützte Textanalyse für die Geisteswissenschaften zugänglich zu machen. Das Tool kombiniert verschiedene Analysefunktionen und Visualisierungsformen und bietet einen einfachen, schnellen und vielfältigen Zugang zu Texten. Vgl. Graham, Milligan, Weingart 2015: 83-86. Für die Plattform vgl. https://voyant-tools.org, 31.08.2022.

[103] Da Voyant Tools und MALLET unterschiedliche Tokenizer verwenden, können hinsichtlich der Frequenz der Tokens kleinere Unterschiede bestehen. Diese allfälligen Differenzen sind für die Aussagen in der vorliegenden Arbeit allerdings vernachlässigbar. Zu den Unterschieden bei der Verwendung unterschiedlicher Tokenizer vgl. Andorfer 2017: Kap. 2.

3. Analyse

3.1. Segmentierung pro Monat kumuliert

Als Ausgangspunkt für die Analyse wurden die Resultate[104] zur Segmentierung, welche die Daten kumuliert pro Monat gliedert, gewählt. Es lassen sich anhand dieser Art der Datenzusammenstellung bestimmte allgemeine Tendenzen, die auch für die späteren Analysen massgeblich sind, exemplarisch erläutern. Das Augenmerk wird hier nämlich vor allem auf die Implikationen einer variierenden Anzahl an Topics gelegt. Ein wesentliches Element der Analyse bildet dabei 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. 2) 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.[105] 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

[104] Eine übersichtliche und transparente Darstellung der Vielzahl an Resultaten, welche in den Observable-Notebooks erzeugt wurden, erweist sich im vorliegenden Format als schwierig. Es werden hier nur diejenigen Heatmaps abgebildet, deren Topics im Rahmen der Analyse eingehender beschrieben werden. Die dazugehörigen Tabellen zur Zusammensetzung der Topics werden im Anhang aufgeführt. Da in den Heatmaps die prozentualen Werte zu den Auftretenswahrscheinlichkeiten der Topics in den einzelnen Feldern nur via Mouseover im Observable-Notebook ersichtlich sind, werden diese im vorliegenden Text soweit möglich in Klammer aufgeführt. Mit der beschriebenen Vorgehensweise kann garantiert werden, dass die für die Arbeit relevanten Informationen in der vorliegenden Fassung ersichtlich und (langfristig) verankert sind. Dennoch wird empfohlen, die Resultate in den übersichtlicher gestalteten und hier jeweils verlinkten Notebooks zu betrachten.

[105] Die farblichen Unterschiede sind bei Topic 5 in der Heatmap weniger deutlich wahrnehmbar, weil die gewählte nichtlineare Farbskala besonders Differenzen in tieferen Segmenten hervorhebt. Somit erscheint die Auftretenswahrscheinlichkeit trotz der Schwankungsbreite von 40 bis 73 Prozent relativ ausgeglichen.

(„bluest“, „grüene“, „buechen“, „wachßen“) und zu lebensweltlichen Aspekten[106] („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 Erntepraxis („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 Weinernte massgeblich waren.

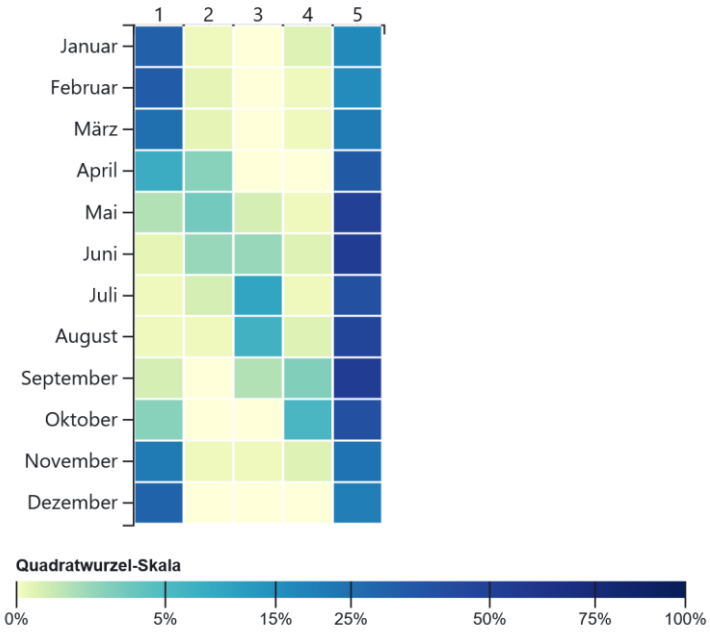

***Abb. 2: Segmentierung pro Monat kumuliert, Modell 5**. Die oben abgebildeten Zahlen stehen für das jeweilige Topic. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.1.1. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, 31.08.2022.*

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,

[106] Insbesondere in Einsiedeln, wo die Mobilität im Winter stark eingeschränkt war, konnten im Frühling wiederum Spaziergänge und Feierlichkeiten im Freien abgehalten werden. So pilgerten beispielsweise jährlich im Frühling Fraktionen aus bestimmten Gemeinden in den sogenannten Kreuzgängen nach Einsiedeln und hielten hier zusammen mit den Konventualen Prozessionen ab. Bei anhaltender Schneebedeckung oder schlechter Witterung fanden diese in der Kirche statt.

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. Das heisst, es bilden sich neben Wortketten mit jahreszeitlicher Ausprägung zunehmend Topics, die vor allem in einem einzelnen Monat eine erhöhte Auftretenswahrscheinlichkeit aufweisen. In Modell 15 (Abb. 3) 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.

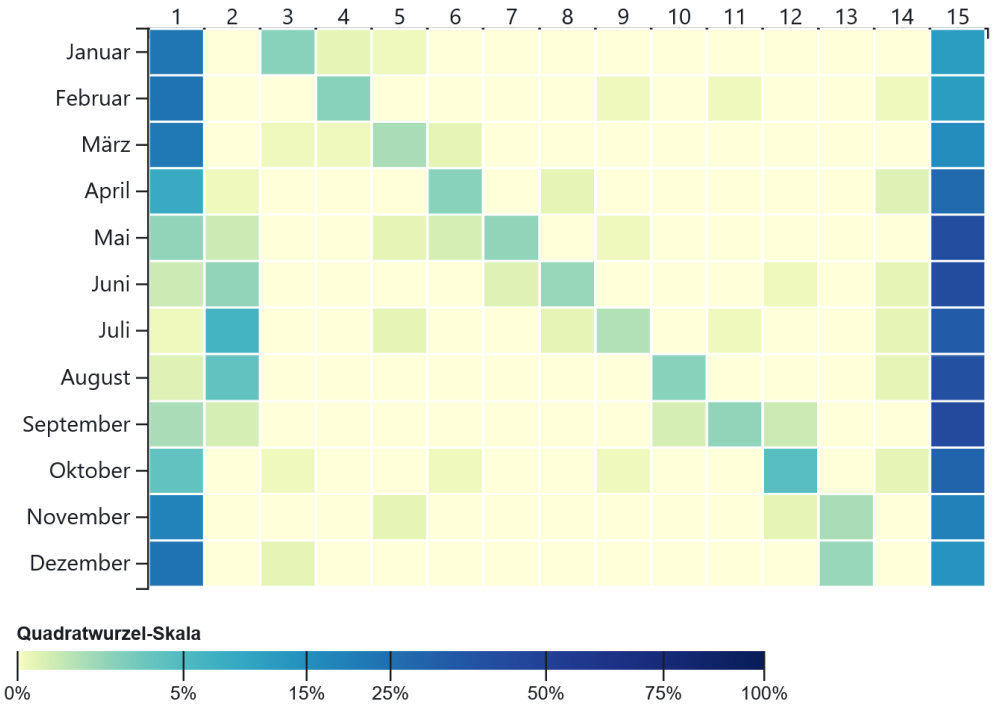

***Abb. 3: Segmentierung pro Monat kumuliert, Modell 15**. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.1.2. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, 31.08.2022.*

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“, „os-

termonntag“). 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 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 Modelle mit steigender Anzahl an Topics allgemein dazu, dass die Auftretenswahrscheinlichkeit einzelner Topics abnimmt. Da bei jedem Modellierungsprozess die Zusammensetzung der Topics neu generiert wird, kommen bei einer grösseren Anzahl an Topics vermehrt Begriffe mit allgemein niedrigerer Frequenz vor. Dies führt generell auch zu Verschiebungen der Auftretenswahrscheinlichkeiten.

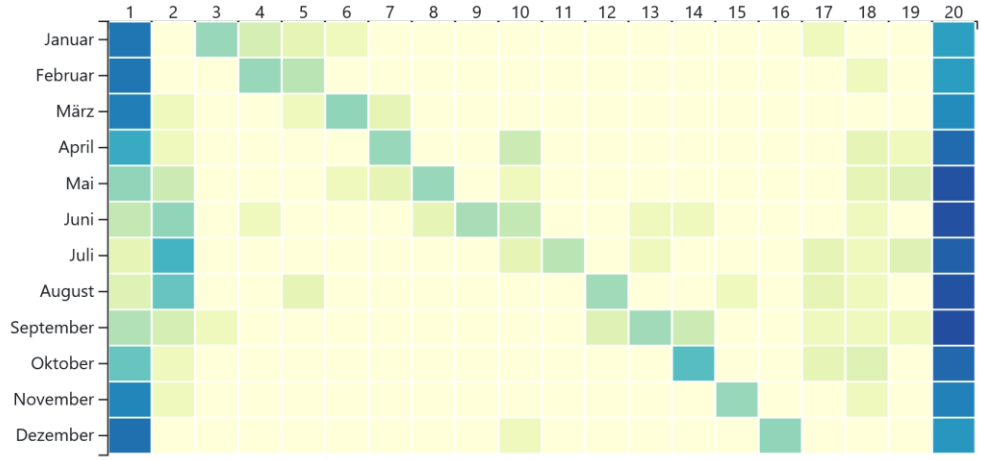

***Abb. 4: Segmentierung pro Monat kumuliert, Modell 20**. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.1.3. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, 31.08.2022.*

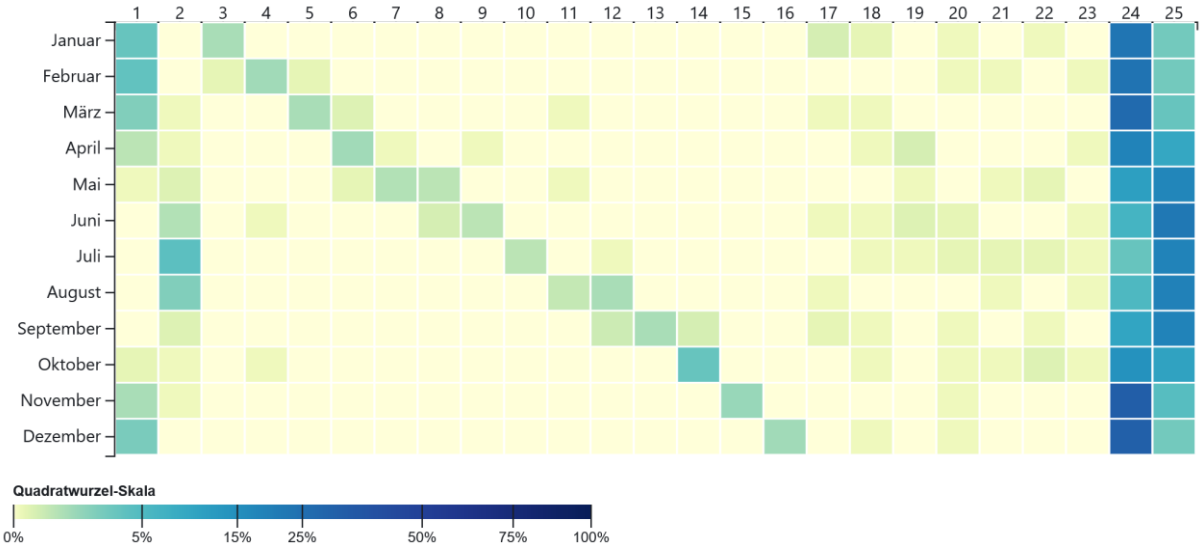

***Abb. 5: Segmentierung pro Monat kumuliert, Modell 25.** Für die Topics und Wortfrequenzen vgl. Kap. 6.4.1.4. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, 31.08.2022.*

In den Heatmaps ist beispielsweise beim jeweils an erster Stelle aufgeführten Topic, welches sich vor allem in den Modellen 5 bis 20 aus ähnlichen Tokens zusammensetzt, erkennbar, dass sich die Farbtiefe bei steigender Anzahl Topics verringert. Ein starkes Ausbleichen zeigt sich insbesondere beim Vergleich der Modelle 20 (Abb. 4) und 25 (Abb. 5). Während in allen Modellen bis und mit 20 ein Topic (in den Heatmaps jeweils am rechten Rand abgebildet) mit durchwegs hoher Auftretenswahrscheinlichkeit, aber geringer Trennschärfe vorkommt, wird bei Modell 25 ein zweites Topic (Topic 24) mit ähnlichen Eigenschaften gebildet, welches vor allem Tokens von Topic 1 und Topic 20 „absorbiert“. So findet sich beispielsweise in Modell 20 der Begriff „luft“ in Topic 1 und Topic 20; beim Modell 25 verschwindet er aus den Entsprechungen und taucht im neugebildeten Topic 24 auf. Im Weiteren wechseln auch die Tokens „kalt“, „kalter“ und „sehr_kalt“ von Topic 1 zu Topic 24. Es verbleiben zwar immer noch viele Begriffe mit direktem oder indirektem Bezug zu Kälte (z.B. „kälte“, „sehr_kalter“, „schnee“,

„schlitten“, „gefrohren“) in Topic 1, allerdings ist deren absolute Zahl sowohl innerhalb des Topics als auch im Gesamtkorpus tiefer, womit die Auftretenswahrscheinlichkeit des Topics insgesamt abnimmt. Im Hinblick auf das Topic mit durchgehend hoher Auftretenswahrscheinlichkeit am äussersten rechten Rand der Heatmaps zeigt sich, dass durch die Absorbierung bestimmter Begriffe eine Fokussierung auf die Sommer- und Übergangsmonate erfolgt und die Trennschärfe so zunimmt.

Neben dem beschriebenen Effekt, dass sich die Zusammensetzungen der Topics verschieben, führt eine höhere Anzahl an Topics auch dazu, dass tendenziell mehr Topics mit einer allgemein 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. 6) 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.

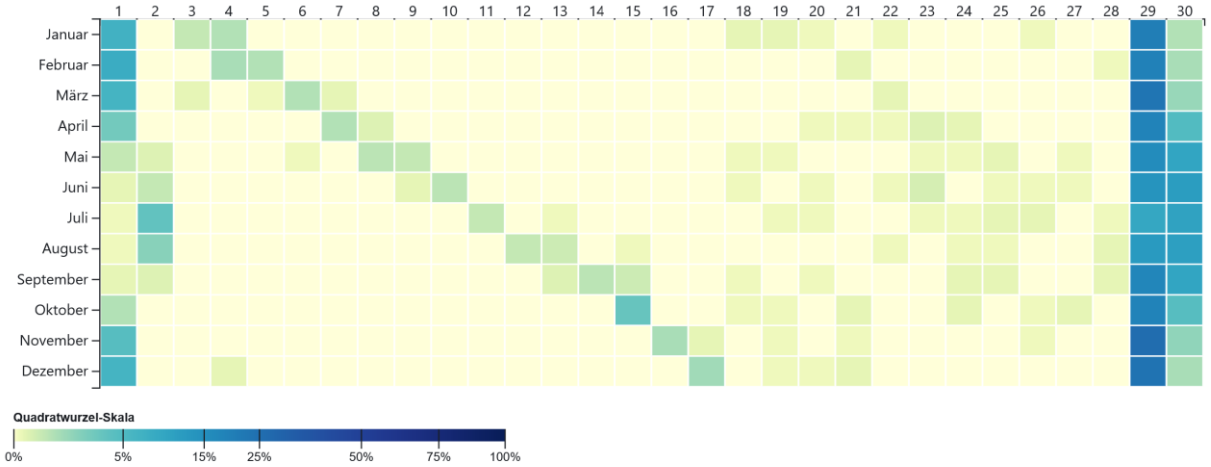

***Abb. 6: Segmentierung pro Monat kumuliert, Modell 30**. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.1.5. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_monate_kumuliert, 31.08.2022.*

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.[107]

Auch bei anderen Modellen fällt auf, dass die erwähnten Topics mit geringer Auftretenswahrscheinlichkeit und Trennschärfe aus Tokens mit tiefer Frequenz bestehen, was am Beispiel der Topics 18 bis 28 in Modell 30 ersichtlich wird. Den höchsten Wert (14) erreichen hier die Begriffe „bluest“ in Topic 24 und „weiden“ in Topic 25. Bei den übrigen Topics liegt das Maximum darunter, wobei der häufigste Begriff in Topic 27 lediglich viermal vorkommt. Bei allen Topics weist der Terminus an zwanzigster Stelle lediglich eine Frequenz zwischen zwei und vier auf. Neben Zweifeln an der statistischen Signifikanz stellen sich auch Fragen bezüglich der Verständlichkeit der Topics. Hierbei zeigt sich, dass beispielsweise die Topics 18 bis 28 in Modell 30 schwierig interpretierbar sind. Topic 22 weist zwar einige Tokens in direktem oder möglichem Bezug zu Wind („windt“, „luften“, „blasste“, „zwerchwind“, „geworffen“, „rüehwiger“) oder Wetter („erwarmet“, „schonte“, „nassem“, „besserte“, „trübe“) auf, allerdings umschreiben diese Begriffe eher allgemeine Zustände des Wetters ohne erkennbare Tendenz. Entsprechend tritt das Topic mit einer Wahrscheinlichkeit von 1 bis 2 Prozent in den Monaten Januar, März, April, Juni, August auf. Am ehesten nutzbar erscheint 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 in Topic 23 von Modell 30 befinden sich in Gesellschaft mit Funktionswörtern wie „här“ und „hinzu“, Adjektiven wie „hochheilig“ oder Substantiven wie „praelat“ und „zulauf“, die das potenzielle Thema inhaltlich scheinbar nicht bereichern. Hierbei offenbart sich unter anderem die Schwierigkeiten bei der Erstellung der Stopwords-Liste. Die Tokens „här“ und „hinzu“ sollten eigentlich nicht berücksichtigt werden, wurden bei den vorgängigen Tests zur Zusammenstellung der Liste aber übersehen. Zudem wurde die Liste iterativ erweitert, womit die Wahrscheinlichkeit, dass zusätzliche unerwünschte Tokens in die Modelle geraten, zunahm. Der Effekt dieser unerwünschten Tokens ist allerdings vernachlässigbar, zumal sie

[107] Die Auftretenswahrscheinlichkeit von Begriffen mit gleicher Frequenz ist über das gesamte Korpus identisch. Bei der Modellierung wird allerdings eine Reihe weiterer Merkmale, beispielsweise zur Bestimmung der Exklusivität eines Tokens innerhalb der einzelnen Topics, berechnet, welche bei der Anordnung von Tokens mit gleicher Frequenz massgeblich sind. Diese sind im Diagnose-Output von MALLET ersichtlich, der für weiterführende Analysen gedacht ist.

nur vereinzelt und tendenziell in Topics, welche für weitere Analysen weniger relevant sind, vorkommen.

Weit evidenter ist die Frage, inwiefern auch Inhaltswörter ausgeschlossen werden sollen. Dass der Begriff „praelat“ (kirchlicher Würdenträger wie Abt oder Bischof) im monastischen Kontext Erwähnung findet, erstaunt nicht und scheint unter dem Hintergrund des Interesses an der Witterung auf den ersten Blick ebenso wenig bereichernd zu sein wie der Terminus „zulauf“. Eine Betrachtung der konkreten Stellen im Quellentext zeigt, dass der Autor den Begriff „zulauf“ immer im Zusammenhang mit der Anzahl der Teilnehmenden (dem „volk“) an den häufig im Frühling stattfindenden kirchlichen Prozessionen verwendete, wobei er das Wetter und die Anwesenheit von Würdenträgern als wesentliche Einflussfaktoren erachtete. Auch dieses Beispiel zeigt, dass ein tieferreichendes Verständnis des Schreibstils des Autors sowie des allgemeinen Kontexts von Vorteil ist und – ausgehend von den Tokens in den Topics – weiterführende Analysen in Form eines Close Readings hilfreich sein können. In Bezug auf die Stopwords kann konstatiert werden, dass mit einer zu aggressiven Ausschlusspraxis die Gefahr besteht, dass ebenjene stilistischen und lebensweltlichen Zusammenhänge aufgelöst werden.

Der Frühlingsbezug in Topic 23 in Modell 30 lässt sich somit nicht nur über die Färbung in der Heatmap, sondern auch eine Dekonstruktion der Bestandteile des Topics nachvollziehen. In der Heatmap zeigt sich ein ähnliches Muster auch bei Topic 19 in Modell 25. Dieses weist teilweise begriffliche Übereinstimmungen („bluest“, „hochheilige“, „zulauf“) auf und beinhaltet noch weitere frühlingsbezogene Termini wie „kirschi“ und „angesäet“. Abgesehen von den drei übereinstimmenden Tokens und dem Frühlingsbezug unterscheiden sich allerdings die übrigen Begriffe, womit es schwierig ist, eine Kontinuität des Topics über mehrere Modelle, wie dies beispielsweise bei Topic 1 möglich ist, auszumachen. Hier dürfte wiederum der bereits erwähnte Umstand, dass Begriffe mit tiefer Frequenz zu einem höheren Grad austauschbar sind, eine tragende Rolle spielen. Auch wenn die Topics mit geringerer Auftretenswahrscheinlichkeit und Trennschärfe weniger stabil und aus inhaltlicher Sicht weniger aussagekräftig oder verständlich sind, bieten sie dennoch einen Gradmesser für die Einschätzung, ab wie vielen Topics eine zu starke Ausdifferenzierung erfolgt, respektive ab welcher Zahl überwiegend wenig wahrscheinliche Topics gebildet werden. Während die Modelle bis 20 Topics überwiegend trennscharfe Auftretenswahrscheinlichkeiten abbilden, werden in den Modellen ab 25 Topics überwiegend Topics mit äusserst geringen Werten kreiert.

Bei der bisherigen Analyse wurde ein starker Fokus auf die Beschreibung der Konsequenzen einer unterschiedlichen Zahl an Topics gelegt. Es zeigte sich, dass bereits mit einer geringen

Anzahl an Topics klar differenzierbare Muster und inhaltlich verständliche Wortketten abgebildet werden. Eine grössere Zahl führt dazu, dass die Topics höhere Wahrscheinlichkeiten in einzelne Segmente (hier Monate) aufweisen. Allerdings geht dieser Prozess damit einher, dass die Auftretenswahrscheinlichkeit der Topics mit einer zunehmenden Menge allgemein abnimmt, da sich die Tokens über mehrere Topics verteilen und somit die Frequenz im Einzelfall reduziert wird. Auch wenn Topics bestehend aus Tokens mit geringer Frequenz bis zu einem gewissen Grad robust und lesbar bleiben können, verlieren sie tendenziell an Klarheit und Trennschärfe. Zudem stellen sich Fragen hinsichtlich der Aussagekraft, weil bereits geringfügige Faktoren einen Einfluss auf die Zusammensetzung haben können. Obwohl bei der Betrachtung der Heatmaps die Anzahl Topics mit generell geringer Auftretenswahrscheinlichkeit als Indikator dafür dienen können, ab welcher Anzahl an Topics kaum aussagekräftige oder relevante Veränderungen mehr stattfinden, 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. Auch wenn die Resultate bei anderen 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 offenbaren sich auch über landwirtschaftliche oder rituelle Praktiken sowie über indirekt mit der Witterung in Bezug stehenden Indikatoren, wie beispielsweise der Vegetationsentwicklung. Je nach Modell prägen diese Bereiche die Topics in unterschiedlichem Masse. Allerdings lässt sich vermuten, 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. Um die Frage, welchen Einfluss die ortsspezifischen Unterschiede bezüglich des Klimas, der Landwirtschaft und dem Alltag haben, eingehender analysieren zu können, wird im Folgenden eine andere Art der Segmentierung verwendet.

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 7 veranschaulicht.

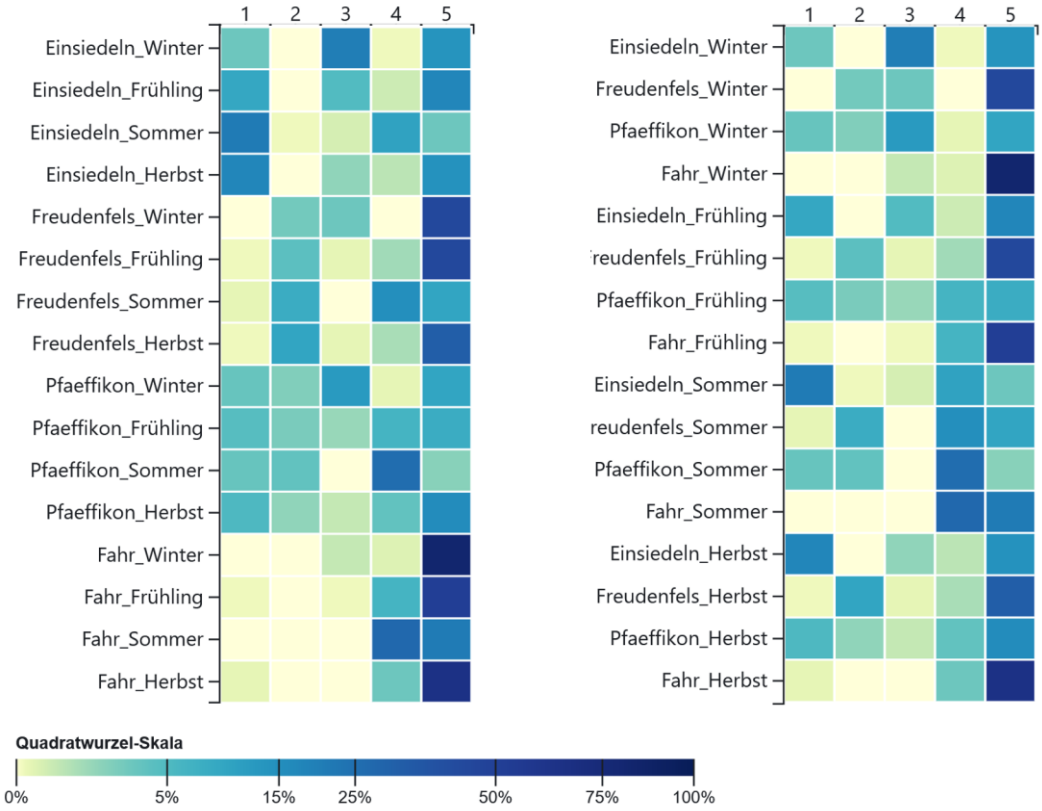

Abb. 7: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 5*. Die linke Heatmap bildet die Sortierung nach Beobachtungsort ab, die rechte Darstellung die Sortierung nach Jahreszeit. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.2.1. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, 31.08.2022.*

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. 8) 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 jahreszeitenspezifische Eigenheiten in den Topics sichtbar zu machen. Der beschriebene Effekt zeigt sich allerdings erst in verstärkter Form ab Modell 15.[108]

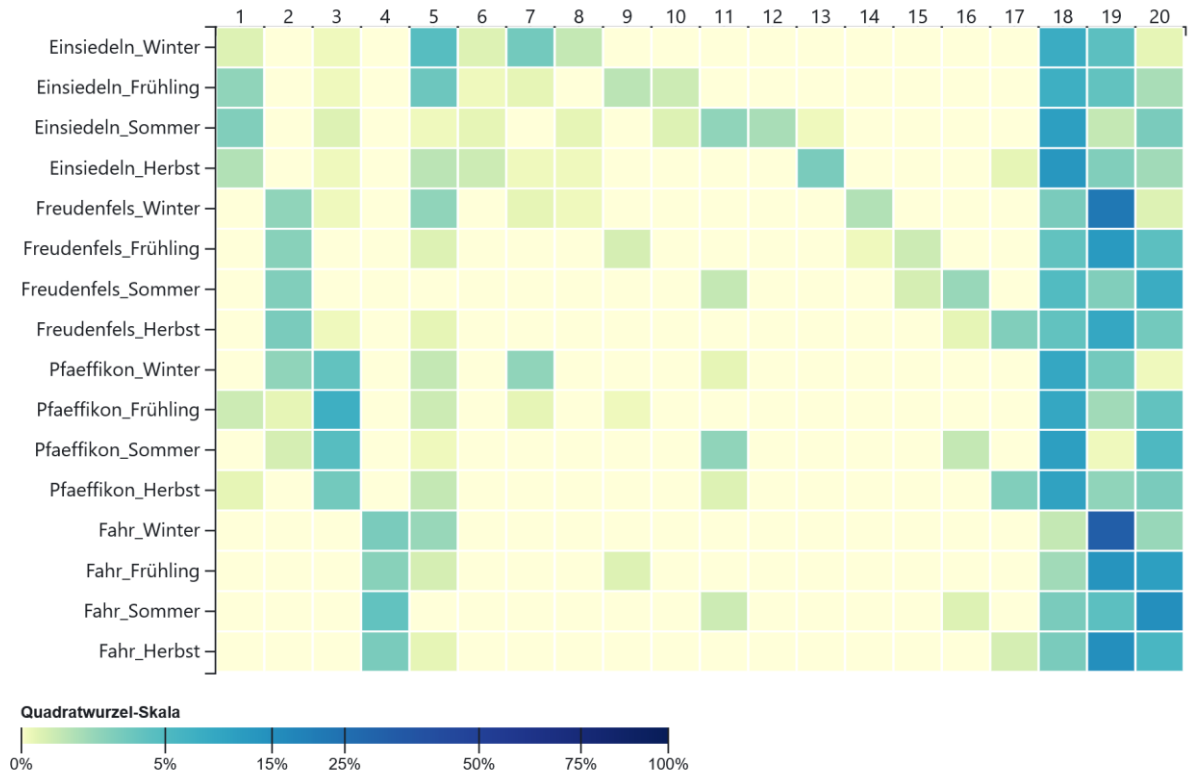

Abb. 8: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 20*. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.2.4. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, 31.08.2022.*

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

[108] 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.

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", „regenwetter", „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. 9) 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.

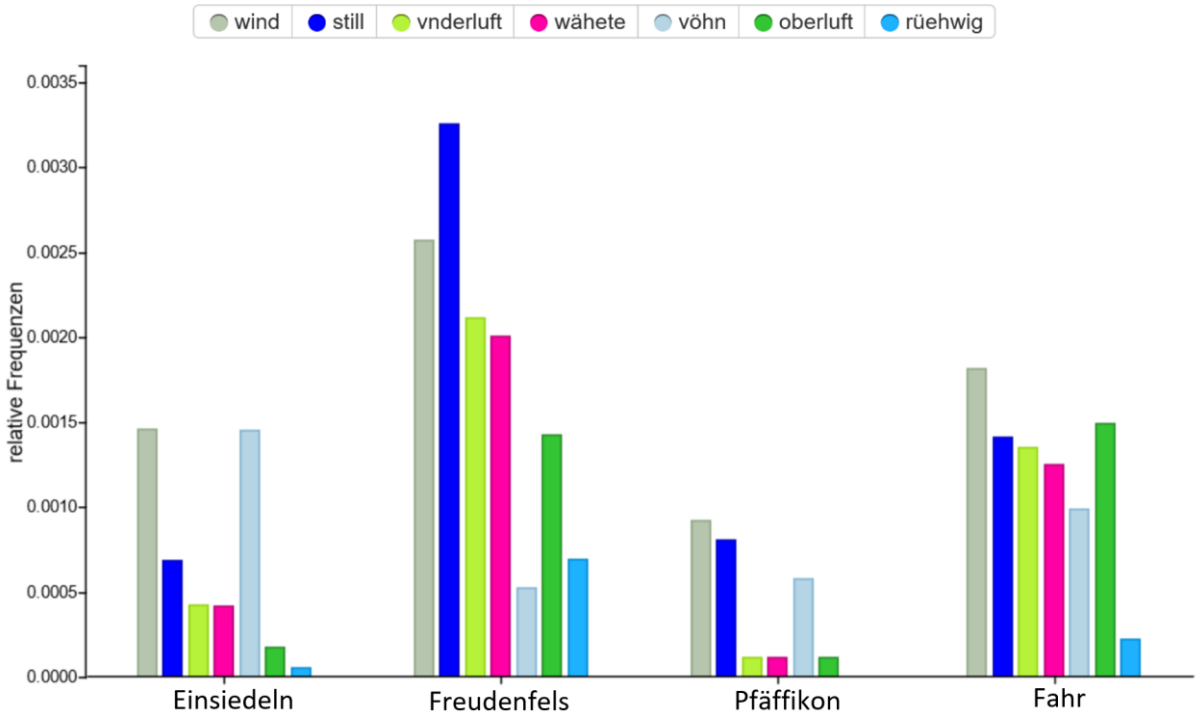

Abb. 9: Relative Frequenzen der Begriffe „wind“, „still“, „vnderluft“, „wähete“, „vöhn“, „oberluft“, und „rüehwig“ pro Beobachtungsstandort*. Die Darstellung wurde mit Voyant Tools erstellt.*

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, son-

dern 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.

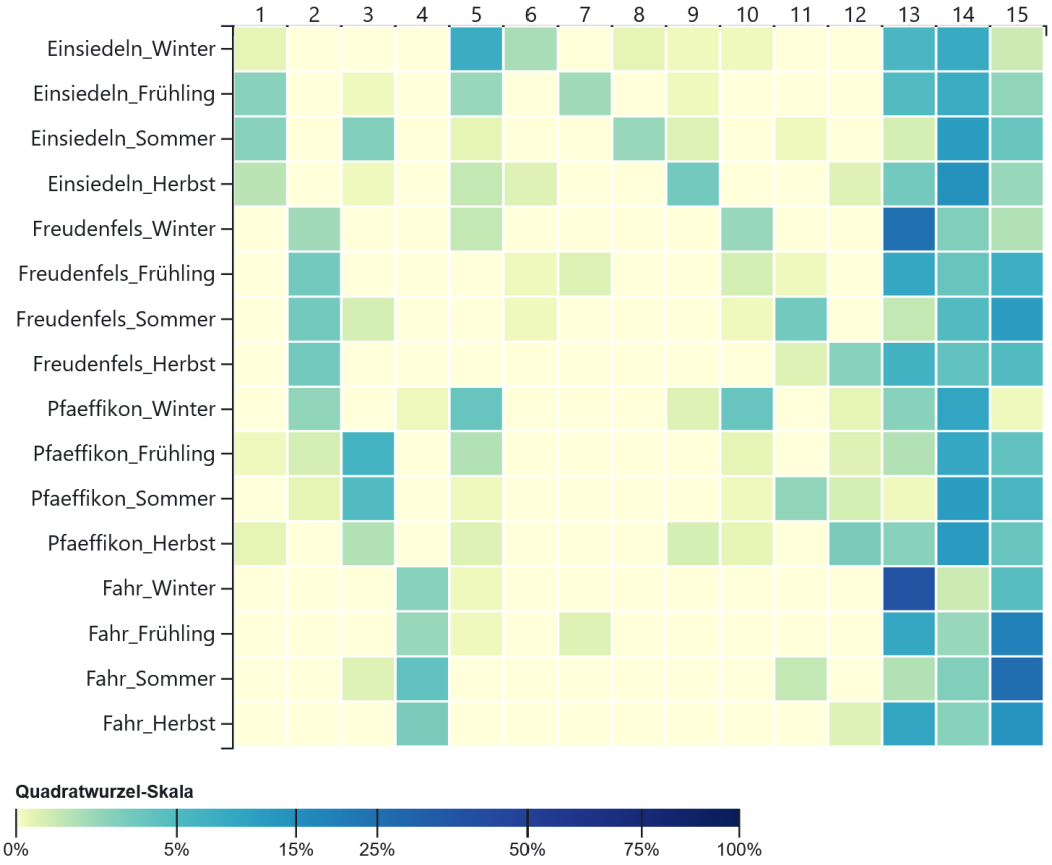

Abb. 10: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 15*. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.2.3. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, 31.08.2022.*

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. 10), 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. 11) 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 Modellierungspro-

zess und individuellen Eigenheiten des Dokumentationsstils des Autors aber nicht als absoluter Gradmesser verstanden werden.[109] Der vergleichsweise 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%, Pfäffikon 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.

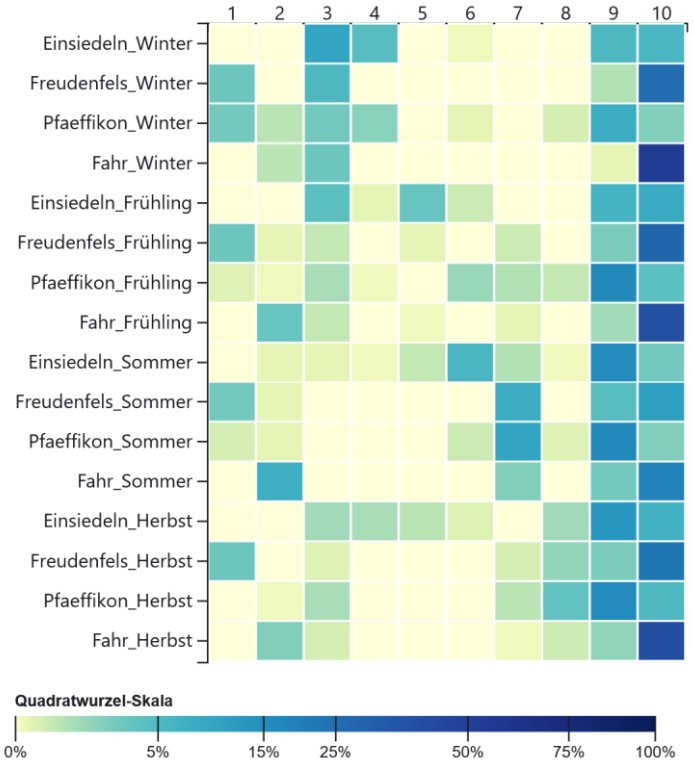

Abb. 11: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 10*. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.2.2. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, 31.08.2022.*

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,

[109] 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.

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[110] 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. 12) 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“,

[110] Diese Darstellungen ähneln zu einem gewissen Grad auch der Grafik, welche Blevins bei der Modellierung von Martha Ballards Tagebuch zum Topic „Cold Weather“ erstellte. Vgl. Blevins 2010.

„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 quasi sukzessive ablösen.

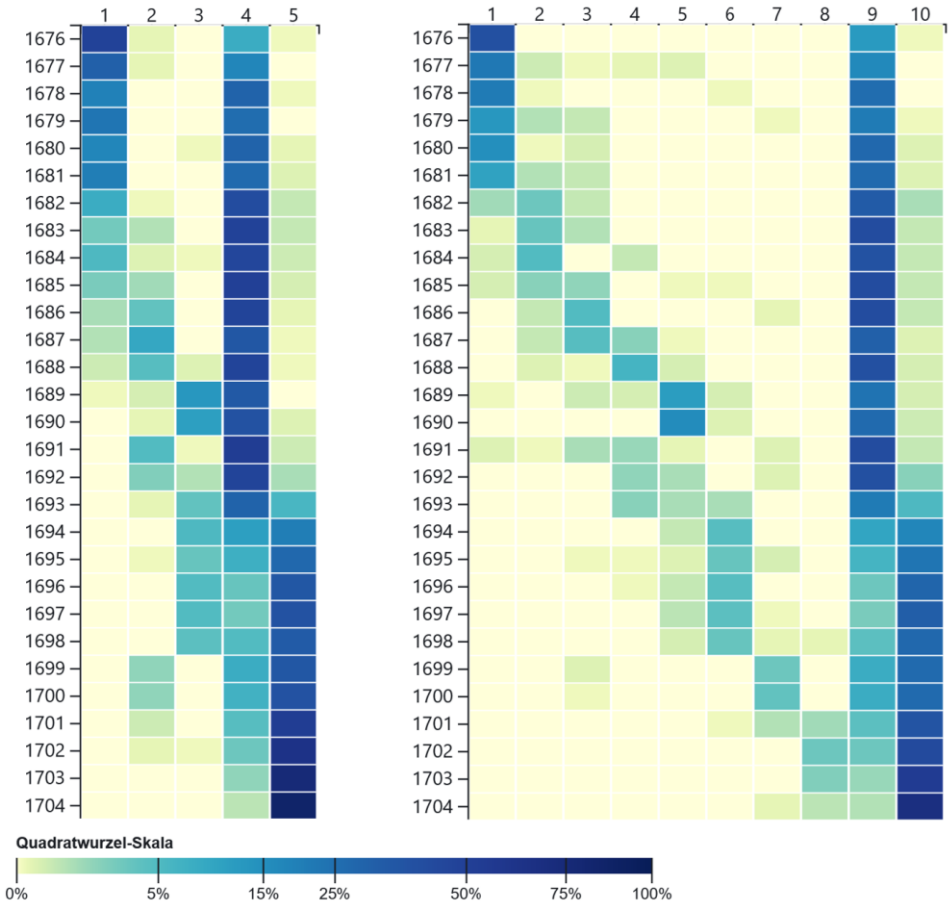

Abb. 12: Segmentierung pro Jahr über den Gesamtzeitraum, Modelle 5 und 10*. Aus Darstellungsgründen wurden die beiden Heatmaps nebeneinander aufgeführt. Für die Topics und Wortfrequenzen vgl. Kap. 6.4.3.1; Kapt. 6.4.3.2. Für das Observable-Notebook vgl. https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, 31.08.2022.*

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[111] die relativen Frequenzen (Abb. 13) ausgewertet wurden. Für den Begriff „Vieh“ benutzte Dietrich im Zeitraum 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.

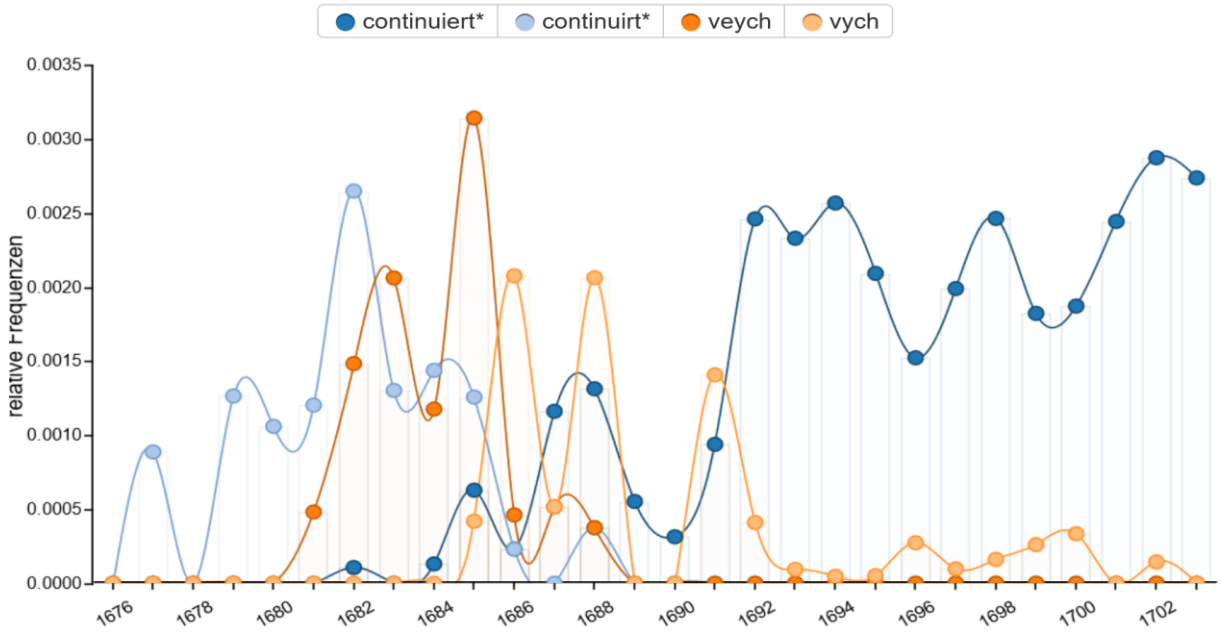

Abb. 13: ***Relative Frequenzen der Begriffe „contiuiert*“, „continuirt*“, „veych“, und „vych“ über den Gesamtzeitraum****. Die Sterne bedeuten, dass alle möglichen Deklinationsformen berücksichtigt wurden. Die Darstellung wurde mit Voyant Tools erstellt.*

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

[111] Aufgrund seines Todes schrieb Dietrich im Jahr 1704 nur bis zum 19. März, womit die Datengrundlage für das letztgenannte Jahr unvollständig ist. Die Werte für das Jahr 1704 wurden in den mit Voyant Tools erstellten Grafiken nicht berücksichtigt.

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 einsiedlischen 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.[112]

Abgesehen von der Erkenntnis, dass mit Hilfe eines starken Optimierungsintervalls die Bildung distinktiver Muster begünstigt wird, ist für weiterführende formale Analysen die Wahl der Stopwords 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.

[112] 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.[113]

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 (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

[113] Vgl. Pfister 1999: 36.

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. 14) 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.[114] 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.[115]

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

[114] Ein Grund hierfür mag sein, dass der atmosphärische Zustand jeweils einfach mit einem Blick nach oben ermittelt werden konnte. Dahingegen beschränkten sich die Möglichkeit des Autors, die Temperatur zu beschreiben, auf seine subjektive Wahrnehmung. Ohne Messungen und statistische Mittel ist es schwierig, einen Tag unter der Berücksichtigung der natürlichen Schwankungen im Tagesablauf und seiner jahreszeitlichen Verortung als kalt, warm oder durchschnittlich zu klassifizieren.

[115] Bei der Transkription fiel im Weiteren auf, dass der Autor mit dem Übergang zur täglichen Berichterstattung häufig mit der Wetterbeobachtung begann und weitere Wetterbeobachtungen zeitlich präziser im Tagesablauf verortete. Eine Analyse mit Voyant Tools ergab, dass die Begriffe „morgen“, „abend“ und „nacht“ ab 1693 eine wesentlich höhere relative Frequenz aufweisen, was tendenziell auch auf „mittag“ zutrifft. Da für alle Modelle dieselbe Stopwords-Liste gewählt und in dieser Liste temporale Begriffe grösstenteils ausgeschlossen wurden, kann der Effekt in den hier erzeugten Topics nicht nachvollzogen werden.

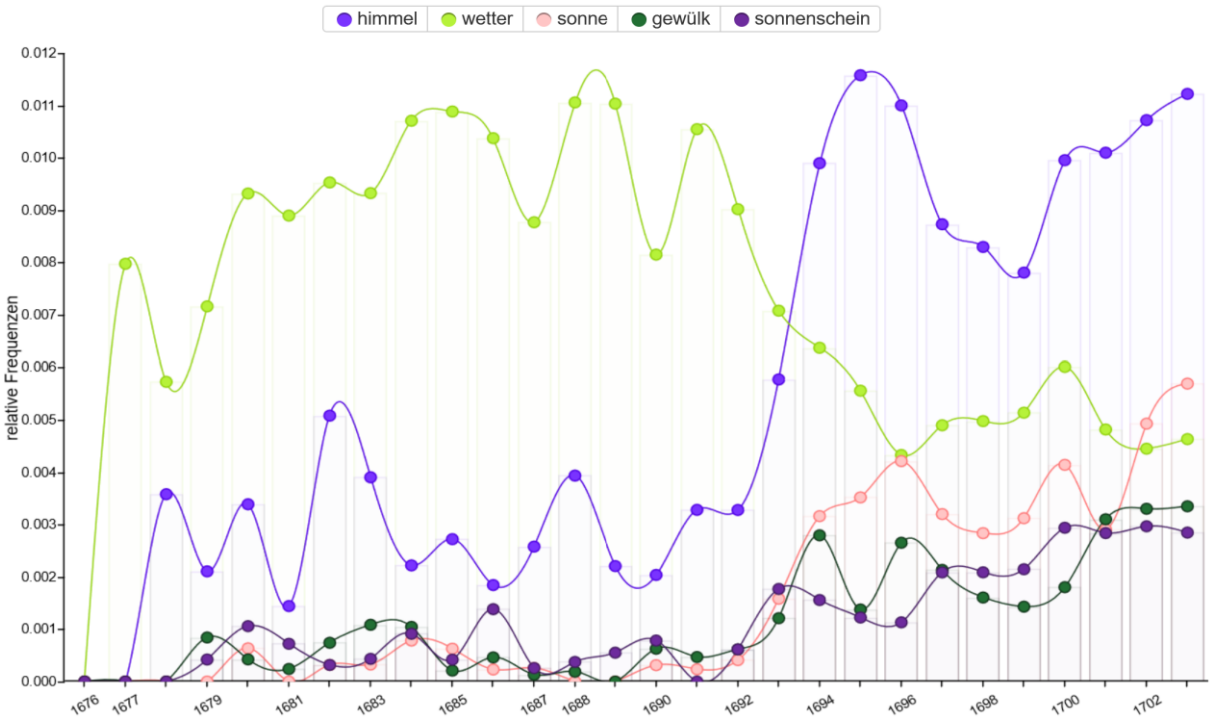

Abb. 14: ***Relative Frequenzen der Begriffe „himmel*“, „wetter*“, „sonne“, „gewülk“ und „sonnenschein“ über den Gesamtzeitraum***. *Die Darstellung wurde mit Voyant Tools erstellt.*

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. 12) 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 für weiterführende Analysen sich ergeben und wie der Modellierungsprozess für eine erfolgreiche Anwendung konfiguriert werden muss. Bezüglich der letztgenannten Fragen kann konstatiert werden, dass für die Anwendung des Ansatzes prinzipiell wenige Einstellungen zwingend erforderlich, im Gegenzug aber mehrere optionale Parameterkonfigurationen für den Modellierungsprozess zu berücksichtigen sind. Zusätzlich kann die Textgrundlage im Rahmen des Preprocessing theoretisch stark modifiziert und die Outputs in unterschiedlicher Art und Weise analysiert und weiterverarbeitet werden. Insgesamt erfordert eine Anwendung von Topic Modeling somit eine Vielzahl an Entscheidungen, deren methodische Begründung und transparente Vermittlung im Hinblick auf wissenschaftliche Ansprüche eine grosse Herausforderung darstellt. Auch wenn die Konsequenzen einzelner Parameterkonfigurationen bis zu einem gewissen Grad mittels weiterführender Untersuchungen eruiert werden können, sind die vielen möglichen Implikationen in ihrer Gesamtheit letztlich schwer zu erfassen, weshalb die Waswäre-wenn-Frage eine ständige Begleiterin beim Topic-Modeling-Prozess ist.

Die erwähnte Unsicherheit ist zu einem gewissen Grad auf eine falsche Erwartungshaltung zurückzuführen. Ursprünglich als Methode im Bereich des Information Retrieval konzipiert, besteht das Ziel – wie bei anderen Ansätzen dieser praxisorientierten Subdisziplin – von Topic Modeling darin, eine möglichst gute Annäherung zu erreichen. Auch wenn bei einer erfolgreichen Anwendung grosse Textmengen weitgehend thematisch erschlossen werden können, basieren diese Resultate auf einem Modell, dass immer eine vereinfachte Form der Wirklichkeit widerspiegelt und wesentlich von den zugrundeliegenden Annahmen und Konfigurationen abhängt. Dies ist kein Argument gegen die in vielen Disziplinen verbreitete Arbeit mit Modellen, sondern für eine sorgfältige Berücksichtigung der zugrundeliegenden Annahmen. Die zielführende Anwendung von Topic Modeling bedingt somit einerseits eine Auseinandersetzung mit dem Einfluss zugrundeliegender Modellierungsoptionen und andererseits eine Interpretation der generierten Outputs. Letztere bilden nämlich keine direkten oder impliziten Wahrheiten ab, sondern bestehen lediglich aus Rohmaterial, dessen Sinn und Nutzen sich erst durch eine weiterführende Auseinandersetzung erschliesst.

Um diesem doppelten Anspruch gerecht zu werden, wurde in der vorliegenden Arbeit ein Ansatz gewählt, der beide Bereiche zusammenführt. So wurde beispielsweise 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 von 5 bis 30 Topics ausgegeben. Dadurch konnten die Veränderungen bei einer unterschiedlichen Zahl an Topics

direkt verglichen werden. Aus arbeitsorganisatorischer Sicht erwiesen sich in diesem Zusammenhang die individualisierbaren Notebooks in Observable als nützliches Instrument für die vergleichende Darstellung der Vielzahl an generierten Modellen. 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 eignen.

Für die vorliegende Arbeit liess sich aufgrund des letztgenannten Punkts schliessen, dass es weniger eine Idealzahl als vielmehr einen Idealbereich für die Bestimmung der Anzahl Topics gibt. Inwiefern diese Erkenntnis auch für andere Forschungsarbeiten hilfreich ist, lässt sich hier nicht abschätzen. Der Ansatz steht aber nicht im Widerspruch zur mathematischen Ermittlung einer optimalen Anzahl an Topics, zumal diese für die Bestimmung des Bereichs nützlich sein kann. Aus epistemologischer Sicht ist der Umkehrschluss, dass es kein Modell mit einer idealen Anzahl an Topics geben kann, evident. Auch wenn bestimmte Topics in mehreren Modellen hinsichtlich ihrer Zusammensetzung und Auftretenswahrscheinlichkeiten Ähnlichkeiten aufweisen können, sind sie dennoch nie identisch. Entsprechend können die abgebildeten Werte nicht als absolute Gradmesser, sondern immer nur als Tendenz verstanden werden. Dies unterstreicht wiederum den explorativen Charakter des Ansatzes.

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 eher 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 diese Elemente auch diejenigen Faktoren, die die Resultate bei anderen Arten der Segmentierungen beeinflussen können. So konnte beispielsweise anhand der Segmentierung pro Ortschaft und Jahreszeit gezeigt werden, dass ortsspezifische Eigenheiten einen wesentlichen Faktor darstellen können. 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 ergeben, konnte exemplarisch anhand des Windbezuges in Freudenfels illustriert werden. Dieser offenbarte sich nämlich sowohl bei der ortsbezogenen synchronen Analyse als auch in leicht abweichender Form bei der diachronen Untersuchung über den Gesamtzeitraum.

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 bereist bekannt waren und so viele Begriffe einfacher eingeordnet werden konnten. 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 Ansätze, um Thesen, die ausgehend von den Topic-Modeling-Resultaten formuliert wurden, weiterzuverfolgen. Durch diese Vorgehensweise konnten unter anderem die Veränderungen im Hinblick auf die Orthografie und den Stil des Autors präziser charakterisiert werden.

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, die im Rahmen der vorliegenden Arbeit nur exemplarisch analysiert und beschrieben werden konnten. Dennoch offenbarte sich die breite Palette an Möglichkeiten, die insbesondere im Bereich sprachlicher und stilistischer Phänomene im Zusammenhang mit der Schreibpraxis des Autors vielversprechend ist. 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 extrem kalten oder warmen

Monaten oder Jahreszeiten 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 einem 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.

Allgemein würden weiterführende Studien mit ähnlichem Fokus dazu beitragen, die Eignung der hier erzielten Resultate zu bewerten und konkretere Anwendungszwecke zu umreissen. Als Grundlage würden sich die umfangreichen Daten der Plattform Euro-Climhist eignen. Hiermit könnten in ähnlicher Art und Weise wie in der vorliegenden Arbeit mit Hilfe von Topic Modeling bestimmte Aspekte gezielt untersucht und den bisherigen Resultaten gegenübergestellt werden. Dabei wäre möglich, dass sich auch Tendenzen zeigen, die stärker Rückschlüsse zur Art der Erfassung als zu den effektiven Ergebnissen zulassen und somit zum Zweck der Gewährleistung der Datenqualität genutzt werden könnten. Ebenso eignen sich die Daten aufgrund ihrer inhaltlichen Homogenität für Untersuchungen zu sprachlichen oder anderen Phänomenen. 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. Da im Rahmen der zunehmenden Digitalisierung und der Verfügbarkeit zuverlässigerer Verfahren der automatischen Texterkennung vermehrt vormoderne Quellen in digitaler Form zugänglich sind, wären die Resultate nicht nur für digitale Editionsprojekte, sondern auch für Archive und Spezialbibliotheken interessant. Abgesehen davon können weitere Anwendungsbeispiele in den Geisteswissenschaften generell dabei helfen, digitale Methoden stärker in den einzelnen Disziplinen zu verankern. Im Weiteren kann es auch für Bibliotheken in Zukunft relevant sein, im Rahmen ihrer Tätigkeiten zur Forschungsunterstützung Ansätze wie Topic Modeling zu berücksichtigen, fördern und vermitteln.

5. Bibliografie

5.1. Forschungsliteratur

- Andorfer, Peter (2017): 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, https://zfdg.de/2017_002, 31.08.2022.
- Asmussen, Claus Boye (2019); Møller, Charles: Smart Literature Review. A Practical Topic Modelling Approach to Exploratory Literature Review. In: Journal of Big Data 6/93, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0255-7, 31.08.2022.
- Blei, David M. (2012): Probabilistic Topic Models. In: Communications of the ACM 55/4: 77-84, https://dl.acm.org/doi/10.1145/2133806.2133826, 31.08.2022.
- Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. (2003): Latent Dirichlet Allocation. In: Journal of Machine Learning Research 3: 993-1022, https://dl.acm.org/doi/10.5555/944919.944937, 31.08.2022.
- Blevins, Cameron (2010): Topic Modeling Martha Ballard's Diary, http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary, 31.08.2022.
- Brett, Megan R. (2012): Topic Modeling. A Basic Introduction. In: Journal of Digital Humanities 2/1, http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett, 31.08.2022.
- Carbou, Guillaume (2017): Analyser les textes à l'ère des humanités numériques. In: Les Cahiers du Numérique 13/3-4: 91-114, https://www.cairn.info/revue-les-cahiers-du-numerique-2017-3-page-91.htm, 31.08.2022.
- Chappelier, Jean-Cédric: Modèles génératifs à base de thèmes pour l'accès à l'information textuelle. In: Gaussier, Eric; Yvon, Francois (Hg.): Modèles statistiques pour l'accès à l'information textuelle. Cachan 2011: 169-221.
- Cohen, Daniel J.; Troyano, Joan Fragaszy (2012): Pacing Scholarly Conversations. In: Journal of Digital Humanities 2/1, http://journalofdigitalhumanities.org/2-1/pacing-scholarly-conversations, 31.08.2022.
- Crain, Steven P.; Zhou, Ke; Yang, Shuang-Hong; Zha, Hongyuan: Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In: Aggarwal, Charu C.; Zhai, ChengXiang (Hg.): Mining Text Data. New York 2012: 129-161. (=Crain et al. 2012)

- Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K.; Harshman, Richard: Indexing by Latent Semantic Analysis. In: Journal of the American Society of Information 41: 391-407, https://dl.acm.org/doi/10.1145/57167.57214, 31.08.2022.
- Falk, Ingrid; Bernhard, Delphine; Gérard, Christophe (2014): De la quenelle culinaire à la quenelle politique: identification de changements sémantiques à l'aide des Topic Models. In: Proceedings of TALN 2: 525-530, https://hal.inria.fr/hal-00998868, 31.08.2022.
- Fechne, Martin; Weiss, Andreas (2017): Einsatz von Topic Modeling in den Geschichtswissenschaften. Wissensbestände des 19. Jahrhunderts. In: Zeitschrift für digitale Geisteswissenschaften 2, https://zfdg.de/2017_005, 31.08.2022.
- Graham, Shawn; Milligan, Ian (2012): Review of MALLET, produced by Andrew Kachites McCallum. In: Journal of Digital Humanities 2/1, http://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan-and-shawn-graham, 31.08.2022.
- Graham, Shawn; Milligan, Ian; Weingart, Scott: Exploring Big Historical Data. The Historian's Macroscope. London 2015.
- Graham, Shawn; Weingart, Scott; Milligan, Ian (2012): Getting Started with Topic Modeling and MALLET, https://doi.org/10.46430/phen0017, 31.08.2022.
- Henggeler, Rudolf: Professbuch der Fürstl. Benediktinerabtei U. L. Frau zu Einsiedeln. Festgabe zum tausendjährigen Bestand des Klosters (Monasticon-Benedictinum Helvetiae 3). Einsiedeln 1934.
- 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 (Digital Humanities Research 2). Bielefeld 2022: 157-177, https://doi.org/10.48350/169050, 31.08.2022.
- Hodel, Tobias; Gasser, Sonja; Schneider, Christa; Schoch, David (2022): Zugang zu Informationen in digitalen Sammlungen: Fokus Archive. In: Informationswissenschaft – Theorie, Methode und Praxis 7/1: 27-91, https://doi.org/10.48350/170882, 31.08.2022. (=Hodel et al. 2022)
- 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 (Digitale Kultur 1). Hagen 2022: 181-205, https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001838, 31.08.2022.
- Hofmann, Thomas (1999): Probabilistic Latent Semantic Indexing. In: Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in

Information Retrieval. Berkeley: 50-57, https://dx.doi.org/10.1145/312624.312649, 31.08.2022.

- Jockers, Matthew L.: Macroanalysis. Digital Methods & Literary History. Urbana 2013.
- Jockers, Matthew L.; Thalken, Rosamond: Text Analysis with R. For Students of Literature. Cham ^{2}2020.
- Lamba, Manika; Madhusudhan, Margam: Text Mining for Information Professionals. An Uncharted Territoriy. Cham 2022.
- Mauelshagen, Franz: Klimageschichte der Neuzeit: 1500-1900. Darmstadt 2010.
- Meeks, Elijah; Weingart, Scott B. (2012): The Digital Humanities Contribution to Topic Modeling. In: Journal of Digital Humanities 2/1, http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling, 31.08.2022.
- Mimno, David (2015): Using Phrases in Mallet Topic Models, http://www.mimno.org/articles/phrases, 31.08.2022.
- Nelson, Robert K. (2012): Richmond Daily Dispatch, 1869-1865 and Mining the Dispatch. In: Journal of American History 99/1: 386-388, https://doi.org/10.1093/jahist/jas157, 31.08.2022.
- Newman, David J.; Block, Sharon (2006): Probabilistic Topic Decomposition of an Eighteenth Century American Newspaper. In Journal of the American Society for Information Science and Technology 57/6: 753-767, https://dl.acm.org/doi/10.5555/1124169.1124187, 31.08.2022.
- Pfister, Christian: Wetternachhersage. 500 Jahre Klimavariationen und Naturkatastrophen (1496-1995). Bern 1999.
- Rhody, Lisa Marie: Topic Modeling and Figurative Language. In: Journal of Digital Humanities 2/1, http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody, 31.08.2022.
- Schmitt, Benjamin M. (2012): Words Alone: Dismantling Topic Models in the Humanities. In: Journal of Digital Humanities 2/1, http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt, 31.08.2022.
- Schöch, Christof (2016): Topic Modeling with MALLET. Hyperparameter Optimization. In: The Dragonfly's Gaze. Computational Analysis of Literary Texts, https://dragonfly.hypotheses.org/1051, 31.08.2022.
- Schöch, Christof (2017): Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama. In: Digital Humanities Quarterly 11/2, http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html, 31.08.2022.

- Steyvers, Mark; Griffiths, Tom: Probabilistic Topic Models. In: Landauer, Thomas K.; McNamara, Danielle S.; Dennis, Simon; Kintsch, Walter (Hg.): Handbook of Latent Semantic Analysis. London 2007: 427-448.
- Tang, Jian; Meng, Zhaosi; Nguyen, Xuanlong; Mei, Qiaozhu; Zhang, Ming (2014): Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In: Proceedings of the 31st International Conference on Machine Learning 32/1: 190-198, https://proceedings.mlr.press/v32/tang14.html, 31.08.2022. (=Tang et al. 2014)
- Templeton, Clay (2011): Topic Modeling in the Humanities: An Overview, https://mith.umd.edu/news/topic-modeling-in-the-humanities-an-overview, 31.08.2022.
- Underwood, Ted (2012): Topic Modeling Made Just Simple Enough, https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough, 31.08.2022.
- Viehhauser, Gabriel: Mittelalterliche Texte als Modellierungsaufgabe. In: Fischer, Martin (Hg.): Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen. Akten der Tagung Bamberg, 08.-10. November 2018 (Bamberger interdisziplinäre Mittelalterstudien 15). Bamberg 2020: 15-50.
- Wallach, Hanna; Mimno, David; McCallum, Andrew: Rethinking LDA (2009). Why Priors Matter. In: Advances in Neural Information Processing Systems 22, https://papers.nips.cc/paper/2009/hash/0d0871f0806eae32d30983b62252da50-Abstract.html, 31.08.2022.
- Weingart, Scott (2012): Topic Modeling for Humanists. A Guided Tour, http://www.scottbot.net/HIAL/index.html@p=221.html, 31.08.2022.
- Wilke, Claus O.: Datenvisualisierung – Grundlagen und Praxis: Wie sie aussagekräftige Diagramme und Grafiken gestalten. Heidelberg 2020.

5.2. Observable-Notebooks

- Intervalloptimierung bei 400 Iterationen, https://observablehq.com/@lheinzmann/tm_intervalloptimierung_400, 31.08.2022.
- Intervalloptimierung bei 6’000 Iterationen, https://observablehq.com/@lheinzmann/tm_intervalloptimierung_6000, 31.08.2022.
- Segmentierung pro Monat kumuliert, https://observablehq.com/@lheinzmann/tm_monate_kumuliert, 31.08.2022.
- Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, https://observablehq.com/@lheinzmann/tm_orte_jahreszeiten_kumuliert, 31.08.2022.
- Segmentierung pro Jahr über den Gesamtzeitraum, https://observablehq.com/@lheinzmann/tm_jahre_gesamtzeitraum, 31.08.2022.

6. Anhang

6.1. Abbildungsverzeichnis

Abb. 1: Intervalloptimierung für 5 Topics bei 6'000 Iterationen. 28
Abb. 2: Segmentierung pro Monat kumuliert, Modell 5. 34
Abb. 3: Segmentierung pro Monat kumuliert, Modell 15. 35
Abb. 4: Segmentierung pro Monat kumuliert, Modell 20. 37
Abb. 5: Segmentierung pro Monat kumuliert, Modell 25. 37
Abb. 6: Segmentierung pro Monat kumuliert, Modell 30. 38
Abb. 7: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 5....42
Abb. 8: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 20....43
Abb. 9: Relative Frequenzen der Begriffe „wind“, „still“, „vnderluft“, „wähete“, „vöhn“, „oberluft“, und „rüehwig“ pro Beobachtungsstandort. 45
Abb. 10: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 15....46
Abb. 11: Segmentierung pro Beobachtungsort und Jahreszeit kumuliert, Modell 10....48
Abb. 12: Segmentierung pro Jahr über den Gesamtzeitraum, Modelle 5 und 10. 50
Abb. 13: Relative Frequenzen der Begriffe „contiuiert*“, „continuirt*“, „veych“, und „vych“ über den Gesamtzeitraum. 51
Abb. 14: Relative Frequenzen der Begriffe „himmel*“, „wetter*“, „sonne“, „gewülk“ und „sonnenschein“ über den Gesamtzeitraum. 55

6.2. Stopwords-Liste

Abgesehen von den Durchläufen zur Ermittlung des Optimierungsintervalls[116] wurden für alle Modellierungsprozesse die vorliegende Stopwords-Liste verwendet.

Zahlen: 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680 1681 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1692 1693 1694 1695 1696 1697 1698 1699 1700 1701 1702 1703 1704 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Monatsnamen/Latein: ianuarius ianuarii februarii februarius martius martii merz merzen april aprilen aprilis aprili maius maii meyen iunius iunii iulius iulii augustus augusti september

[116] Die für die Ermittlung des Optimierungsintervalls verwendete Stopwords-Liste unterscheidet sich lediglich darin, dass nachfolgende Tokens darin noch nicht enthalten waren: infra nostri denen dennen derer deren einen allem anders seinen selbige hervor wesßwegen wohin worauf worzu habe hetten sollen nechst sondere sonderem soderen sonders ste stem.

septembris october octobris november novembris december decembris admodum apostolorum baptistae christi corporis dato dominica infra nostri pater patris patri patre patrem patres patrum patribus post reverendus reverendi reverendo reverendum reverendorum reverendis reverendos sanctissima sanctissimae sanctissimus sanctissimum trinitatis

Artikel: das daß dem demme den denen dennen der derer deren des dessen deß desßen die dis dise disem disen diser dises diß diße dißem dißen dißer dißere dißes ein eine einem einen einer eines ihnen

Pronomen: alle allem allen alles aller ander anders andere andres anderen andern anderst dieselbe einer er es eß etwas ich ihr ihre ihren ihro ihme jenige kein keine keines mann mehr mein meine meinem meines meiner mich mir nichts sein seinen selbige sich sie solche solchen vns vnser vnsere vnserem vnseren vnß vnßer vnßere vnßerem vnßeren was welche welchem welchen welches welcher wer wir

Adverben: abend abends abendt abendts allda allenthalben allgemach allhero alligklich also alß anoch annoch anzu ausß auß bald beneben da dahär dahin dann dannen darauf darauß darby darnach darüber darvon darzu derentwegen deretwegen dergestalten dermasßen desßwegen doch dort dorten endtlich erst fort gester gestert har her hernach hervor heüt hiemit hier hiermit hierüber hin hinauß hinein hinunder hinüber hinwider hochster jedoch lang mahl männigklich mithin mittag morgen nacher nacht nit nur öefters sehr so sonsten stund theils vast vberal vorhero wenig weniges wesßwegen wider widerumb wohin wohl wol worab worauf worvon worzu zusamen zwar

Verben: ankommen bin bleibte erfolget führen gangen gab geben gebracht gefallen gefahren gegeben gehabt gehalten gehen gemacht gemachte gewesen geweßen gieng giengen hab habe haben hadt hat hatt hatte hatten hatts hette hetten ist kam komen kommen konnen können könnte könnten könte lasßen liesß liesße mal machen machte machten möchte mögen musß müssen müßen müesßen müssten nachgends sehen sein sehn seye seyn seyn seynd seynd sindt sollen solte stuend sye synd sÿnd syndt thate thut vermeinte war ward ware waren were weren werde werden wollen wollte wolte worden

Partikel: aniezo auch eben gar gleich gleiches iez immer noch oder schon

Präpositionen: ab bei bey bis biß by durch für gegen halber halben hindurch im in mit nechst nach ohne vber vmb vnder vom von vor wegen weilen

Adjektive: dermahlen fest fürstlich gantz gantze ganz ganze ganzen gantzen gebliben gnug grosß grosße grosßen grosßem gut guten halb halbe heütige heütigen hochfürstlich hochwürden in meiste meisten meistes mornderige nechsten nur sankt schier sonder sondere sonderem sonderen sonders sonderlich stark starke starkem vbriger vill wenig weniges zimlich zimlichem zimmlichen zimmlicher zimmlich zimmlichem zimmlichen zimmlicher

Konjunktionen/Kontraktionen: aber als am an auf desto gleichwohl masßen obschon vnd vnnd von vom wenn wann wie zu zum zur

Abkürzungen: dz etc ss ste sten stem

Substantive: anfang christi dag decan end gnaden herr herrn herren leüt tag tags tagen theil vhren vhr zeit

6.3. N-Gramme

Für alle Modellierungsprozesse wurden die nachfolgend aufgeführten N-Gramme verwendet. Der Stern bedeutet, dass alle Deklinationsformen des entsprechenden Begriffs berücksichtigt wurden.

nit_wenig_schaden nit_wenig_frisch* nit_wenig_kalt* nit_wenig_warm* nit_kalt* nit_warm* nit_sonders_kalt* nit_sonders_warm* nit_sonderbar_kalt* nit_sonderbar_warm* nit_gar_kalt* nit_gar_warm* sehr_kalt* sehr_warm* sehr_heisß* sehr_frisch* erschröklich_heisß* erschröklich_kalt* grimmig_heisß* grimmig_kalt* grimmig_warm* bruetig_warm* bruetig_heisß* zimmlich_kalt* zimmlich_frisch* zimmlich_warm* zimmlich_milt* kein_reegen kein_schnee kein_reifen kein_heüw kein_nebel kein_tropfen kein_sonnenschein manns_gedenken

6.4. Topics und Wortfrequenzen

6.4.1. Segmentierung pro Monat kumuliert

6.4.1.1. 5 Topics

1	schnee (1023) - himmel (471) - nebel (319) - luft (265) - kalt (255) - kälte (222) - sehr_kalt (209) - sonne (191) - kalter (190) - wind (173) - nit_kalt (173) - hell (163) - heller (163) - fanden (152) - still (147) - milt (139) - vöhn (130) - fieng (128) - milter (126) - hellen (121)
2	vngewitter (23) - tauw (23) - procession (23) - spazieren (20) - kirchen (20) - bluest (17) - oberwind (17) - mangel (15) - zug (14) - iedermann (14) - buechen (13) - tundern (12) - grüene (12) - maria (12) - strengen (12) - pauli (11) - schwarz (11) - hörten (11) - neüw (10) - wachßen (10)
3	heüw (159) - tach (142) - schaden (100) - frucht (67) - korn (53) - vngewitter (50) - hagel (49) - zehendten (47) - ligende (41) - ernd (40) - plizgen (39) - veld (38) - schneiden (37) - plazreegen (37) - embd (36) - heisßer (32) - hiz (31) - abgelofen (31) - ernezt (29) - roggen (29)

4	trauben (119) - reifen (91) - wein (67) - wimmlen (64) - nit_kalt (49) - wimmet (47) - herpst (40) - reeben (37) - pfefiken (26) - eymer (21) - kalt (20) - vechtnauw (18) - reif (17) - iedermann (17) - gelten (16) - kälte (16) - weins (15) - eimer (14) - herpsten (14) - reifens (14)
5	wetter (1817) - himmel (1767) - reegen (1084) - sonne (593) - gewülk (513) - luft (486) - warm (408) - nebel (407) - schön (394) - starker (393) - sonnenschein (384) - schöner (379) - wind (376) - still (374) - hell (363) - angefangen (358) - fieng (341) - gott (321) - starken (319) - scheinte (313)

6.4.1.2. 15 Topics

1	schnee (1056) - himmel (627) - luft (412) - nebel (383) - kalt (310) - kalter (262) - sonne (244) - kälte (239) - sehr_kalt (224) - hell (208) - still (198) - wind (197) - nit_kalt (196) - fanden (175) - heller (172) - vöhn (158) - milter (140) - milt (139) - sehr_kalter (139) - continuierte (135)
2	heüw (184) - tach (139) - schaden (107) - hagel (71) - frucht (69) - vngewitter (63) - korn (58) - reegenwetter (53) - zürrich (45) - hiz (42) - plizgen (41) - ligende (40) - zehendten (38) - ernd (37) - limmet (37) - erschröklich (36) - schneiden (34) - tundern (33) - plazreegen (32) - heisßer (29)
3	meinradi (16) - schwär (9) - darinn (8) - batzen (7) - hochheiligen (7) - byßwind (7) - rosßen (7) - knecht (7) - grosser (7) - kälterer (6) - zwechtenen (6) - milterung (6) - ochßen (6) - antonii (6) - haufen (6) - verwähet (5) - heütiger (5) - pik (5) - fuosß (5) - erstlich (5)
4	schlittlin (13) - schwär (9) - faßnacht (8) - obervogt (7) - folgende (7) - tachtraufen (6) - jeger (6) - brechen (5) - scholasticae (4) - duhr (4) - marmel (4) - vahr (4) - verwähete (4) - getragen (4) - thäte (4) - erzitteret (4) - matthaei (3) - nit_warmer (3) - septentrionalischer (3) - macht (3)
5	benedicti (10) - laetare (9) - behenkt (6) - zwerchluft (6) - hintringen (6) - heyligen (5) - gebett (5) - zimmlich_milt (5) - abgeritten (5) - vnwandelbahr (5) - zwerch (5) - bewegt (5) - merz (4) - zufriden (4) - förchten (4) - deki (4) - ligende (4) - vberall (4) - bauren (4) - verlegt (4)
6	oster (21) - spazieren (15) - grüene (12) - albis (9) - jörgen (9) - georgii (9) - bluest (9) - grüen (9) - weiden (8) - heylige (8) - gruenen (8) - reifen (8) - wässerig (8) - schnees (8) - korn (7) - grüenen (7) - ostern (6) - kirschi (6) - hochheilige (6) - ostermonntag (6)
7	laub (16) - graß (16) - zug (14) - buechen (11) - tauw (11) - frischen (10) - saagen (9) - hiz (9) - gütigste (9) - ascensionis (7) - vngewitter (7) - abendtröthe (7) - noth (7) - kein_reifen (6) - abgeenderet (6) - pfingsttag (5) - hagelsteinen (5) - pfingstzinstag (5) - pfingstmonntag (5) - sennten (5)
8	pauli (11) - petri (10) - maurer (10) - regnender (7) - zürrich (7) - bluest (6) - weitem (6) - verlehnung (5) - pliz (5) - bitten (5) - segnen (5) - aderläsße (5) - abgeloffen (5) - gasßen (5) - nit_wenig_warm (5) - iagte (5) - vnversehener (5) - sonne (5) - ioannis (4) - rüehig (4)
9	eingebracht (12) - schlosßwiß (8) - apostel (8) - jacobi (7) - edlem (7) - reitete (7) - arbeiten (7) - annae (6) - beicht (6) - erscheinen (6) - geschochet (5) - hew (5) - sollten (5) - angestellt (5) - zehendtheüw (4) - anna (4) - verlehnt (4) - bemühet (4) - ordinari (4) - jährliche (4)
10	embd (36) - garben (25) - haber (18) - juchert (10) - embdt (9) - zelglin (9) - laurentii (8) - assumptionis (8) - gewahret (8) - sigenthal (8) - küeler (8) - rütti (7) - bartholomaei (6) - weyningen (6) - spazierte (6) - nebels (6) - mariae (6) - rieth (5) - deiparae (5) - probsten (5)
11	sehr_warmer (10) - schlosßwiß (9) - haber (9) - engelweyhung (8) - mauritii (8) - winter (8) - abfallen (7) - fahrten (7) - angelasßen (6) - saath (5) - wölklin (5) - geschütz (5) - brachten (5) - abgeenderet (5) - gottlieben (5) - zarth (5) - reegenwetter (5) - räben (4) - ableßen (4) - verbrennt (4)
12	trauben (96) - reifen (75) - wimmlen (64) - wein (60) - wimmet (44) - reeben (37) - nit_kalt (29) - herpst (28) - eymer (19) - reif (15) - herpsten (14) - gelten (14) - reifens (14) - vechtnauw (13) - eimer (12) - priester (12) - rütti (11) - weins (11) - torkel (10) - quantitet (10)
13	wyenacht (8) - vohn (8) - stephani (7) - leidenlicher (7) - andreae (6) - eimer (6) - feücht (6) - stillem (6) - continuiert (6) - vollen (6) - abbtysßin (5) - genebleter (5) - vnsichtbar (5) - adventi (4) - conceptionis (4) - immaculatae (4) - adventus (4) - monachorum (4) - catharinae (4) - lauf (4)

14	ohrt (6) - nunn (6) - versorgt (5) - verbleibten (5) - schmuzigen (5) - schlaf (5) - geschüz (4) - vnnutz (4) - luegeten (4) - wurzel (4) - bevorstehende (4) - mehrste (4) - worüber (4) - hoche (4) - vngfahr (4) - dietland (3) - allerhöchsten (3) - beflisßen (3) - aufgehebt (3) - tröpflete (3)
15	wetter (1835) - himmel (1613) - reegen (1099) - sonne (535) - gewülk (511) - schön (397) - starker (395) - warm (393) - schöner (385) - sonnenschein (370) - wind (358) - fieng (344) - nebel (342) - luft (339) - still (322) - hell (318) - starken (317) - gott (303) - ernstlich (300) - scheinte (290)

6.4.1.3. 20 Topics

1	schnee (1056) - himmel (621) - luft (379) - nebel (378) - kalt (309) - kalter (259) - kälte (242) - sonne (225) - hell (209) - sehr_kalt (209) - wind (195) - nit_kalt (191) - still (188) - heller (170) - fanden (167) - vöhn (160) - continuierte (160) - milter (146) - milt (137) - strasßen (136)
2	heüw (205) - tach (157) - schaden (150) - vngewitter (69) - korn (65) - hagel (65) - frucht (61) - zehendten (52) - zürrich (50) - wasßer (49) - hiz (48) - ernezt (47) - ernd (47) - gott (46) - ligende (43) - plazreegen (43) - reegenwetter (42) - limmet (41) - schneiden (37) - plizgen (36)
3	meinradi (10) - allezeit (8) - grosser (7) - byßwind (6) - gaden (6) - wasßers (6) - gestanden (6) - hochheiligen (5) - fuosß (5) - darinn (5) - abgefahren (5) - antonii (5) - pik (4) - fendenbach (4) - alte (4) - grimmiger (4) - bruk (4) - epiphanium (3) - glorwürdigen (3) - ehrlicher (3)
4	schlitten (27) - schlittlin (10) - rosß9) - fasßacht (8) - schwär (8) - zinstag (5) - nebelgewülk (5) - obervogt (5) - erzitteret (5) - kirchen (5) - scholasticae (4) - wohnen (4) - schwarzen (4) - frauen (4) - erdbeben (4) - oberstad (4) - keiner (4) - verwähete (4) - abgefahren (4) - wägen (4)
5	thurm (6) - fortzukommen (5) - stürmiger (4) - stüklin (4) - septentrionalischer (4) - obervogt (4) - drey (4) - ofene (4) - reißen (4) - stösßen (4) - abwechßlen (3) - nit_warmer (3) - burgermeister (3) - anzufangen (3) - hornung (3) - höflich (3) - krachen (3) - aussgebrochen (3) - vnthier (3) - invitiert (3)
6	reebstok (11) - laetare (9) - benedicti (7) - zwerchluft (7) - wald (7) - gebett (5) - alten (5) - dike (5) - zwerch (5) - bekommen (5) - mertz (4) - herdern (4) - rühewige (4) - erforderet (4) - ertruknet (4) - gerold (4) - brausende (4) - doctor (4) - anfänglich (4) - beißwind (4)
7	oster (17) - ernezt (10) - jörgen (10) - albis (9) - grüen (9) - georgii (8) - grüenen (8) - schwarz (7) - ostern (7) - grüene (7) - gruenen (7) - ostermonntag (6) - jährliche (6) - reebstok (6) - erdrich (6) - carrfreytag (5) - heylige (5) - volks (5) - feria (5) - iagte (5)
8	graß (21) - laub (14) - buechen (12) - tauw (11) - saagen (9) - frischen (8) - strengen (8) - ascensionis (7) - vngewitters (7) - zug (7) - abendtröthe (7) - pfingstmonntag (6) - enthalten (6) - pfingsten (6) - kein_reifen (6) - spazieren (6) - heiligen (6) - halden (5) - himmeltauw (5) - wachßen (5)
9	petri (10) - sanctorum (8) - pauli (8) - nit_wenig_warm (5) - dike (5) - vnversehener (5) - volk (5) - strenge (5) - ioannis (4) - außgang (4) - fronleichnambs (4) - alpen (4) - bitten (4) - höhenen (4) - spruz (4) - geschosßen (4) - vertriben (4) - ylends (3) - gemähet (3) - heisße (3)
10	bluest (11) - maurer (10) - hinzu (9) - dritten (9) - wähen (9) - hochheilige (7) - betten (7) - meynung (7) - geschaden (6) - anderem (6) - kein_tropfen (6) - wohnen (6) - starker (6) - währender (5) - geleßen (5) - gestrahlet (5) - aderläßße (5) - kirschi (5) - zulauf (5) - bygewohnet (5)
11	nachlasß (9) - jacobi (7) - zehendtgarben (6) - zornig (5) - geschochet (5) - erlösst (5) - hew (5) - abgeritten (5) - grund (5) - furren (4) - zehendtheüw (4) - außgegosßen (4) - verderbt (4) - feyrtag (4) - augenblik (4) - gemeint (3) - runß (3) - tractiert (3) - rüedi (3) - occasion (3)
12	embd (35) - haber (15) - schnitter (14) - juchert (10) - laurentii (8) - assumptionis (8) - rösch (8) - embdt (8) - bartholomaei (6) - deiparae (5) - guetem (5) - zerstreüwt (5) - spazierte (5) - gestillet (5) - kornschnitt (4) - schneideten (4) - vychpresten (4) - rieth (4) - augsten (4) - brachten (4)
13	schlosßwiß (9) - engelweyhung (8) - mauritii (8) - benno (6) - scheür (6) - angelasßen (6) - sommer (6) - regen (6) - saath (5) - abfallen (5) - saur (5) - daran (5) - geschütz (4) - ennet (4) - continuiren (4) - edel (4) - mehrentheils (4) - säen (4) - weyn (4) - hofnung (4)
14	trauben (95) - reifen (67) - wimmlen (64) - wein (62) - wimmet (46) - herpst (37) - reeben (35) - nit_kalt (34) - eymer (19) - gelten (14) - herpsten (13) - vechtnauw (13) - reifens (12) - reif (12) - rütti (12) - weins (12) - priester (12) - eimer (11) - jahrzeit (9) - pfefiken (9)

15	eimer (7) - andreae (6) - sehr_kalter (6) - martini (5) - monachorum (4) - catharinae (4) - rüeben (4) - steinbruch (4) - aufgezogen (4) - finger (4) - omnium (3) - fierling (3) - kabis (3) - baumnusß (3) - fueder (3) - wuhr (3) - kreuzer (3) - vermischet (3) - verwahren (3) - luftete (3)
16	wyenacht (8) - stephani (7) - thomae (6) - adventi (4) - conceptionis (4) - immaculatae (4) - vmbgeben (4) - evangelistae (3) - vnbeflekten (3) - kugel (3) - zürrichbiet (3) - fewr (3) - dingen (3) - berge (3) - stillem (3) - tach (3) - heiligen (3) - kleines (3) - erwöhlen (2) - gaudete (2)
17	führten (11) - schreiben (9) - keinen (7) - sonne (7) - einsidlen (6) - bemüheten (5) - gottshaus (5) - gästen (5) - giesste (4) - gewüsst (4) - vertröstete (4) - gesagt (4) - handel (4) - schlaf (4) - hoche (4) - cappell (4) - verwichen (4) - bestanden (3) - geblizget (3) - kurzen (3)
18	mehrers (11) - fortgefahren (8) - geschüz (5) - byren (5) - früestüklin (5) - heüwmangel (5) - mannigklich (5) - güeteren (5) - trostlich (5) - hoof (5) - gehinderet (5) - fueder (4) - sehr_warme (4) - prima (4) - edleste (4) - kreftigen (4) - schuldigkeiten (4) - dero (4) - cantzleren (4) - eins (4)
19	weiden (11) - apostel (8) - schwarzem (6) - tonderen (6) - langem (6) - maur (5) - rosenkranz (5) - zeitlich (5) - bremgarten (4) - kreuzer (4) - regnenden (4) - barmherzigkeit (4) - gewahet (4) - starkh (4) - lieblichem (4) - beziehen (4) - nebelin (4) - bach (4) - geleütet (4) - einführung (3)
20	wetter (1835) - himmel (1618) - reegen (1103) - sonne (551) - gewülk (513) - warm (402) - fieng (397) - schön (396) - starker (387) - schöner (385) - luft (372) - wind (359) - sonnenschein (356) - nebel (344) - still (333) - starken (316) - hell (314) - angefangen (299) - scheinte (283) - ernstlich (282)

6.4.1.4. 25 Topics

1	schnee (666) - kälte (166) - strasßen (105) - milt (101) - schneelin (85) - schnees (82) - nit_sonders_kalt (67) - milter (66) - schlitten (64) - gefrohren (45) - schneyen (45) - schmilzen (43) - geworfen (43) - ochßen (39) - mennweeg (34) - erschröklich_kalt (29) - kelte (28) - brünnen (26) - schmilzte (25) - kälter (25)
2	heüw (187) - tach (156) - schaden (98) - frucht (70) - vngewitter (63) - korn (63) - hagel (62) - zehendten (50) - hiz (46) - ligende (45) - ernd (44) - zürrich (41) - geschehen (41) - limmet (39) - brüel (33) - schneiden (30) - plizgen (30) - tundern (29) - ernezt (29) - bringen (29)
3	meinradi (13) - schlittlin (7) - see (7) - beschwärlich (6) - wasßers (6) - fuhr (6) - schneefezlin (5) - vnrüehwig (5) - saum (5) - möglichkeit (5) - nebelgewülk (4) - nit_sonderbar_kalt (4) - pik (4) - gedult (4) - stunden (4) - fahrte (4) - gnueg (4) - abgefahren (4) - weeg (4) - herein (4)
4	jeger (9) - faßnacht (8) - selbsten (6) - duhr (5) - mehrers (5) - frauen (5) - erzitteret (5) - vberschosßen (4) - scholasticae (4) - sezten (4) - marmel (4) - müsße (4) - verfrohren (4) - wägen (4) - nezen (3) - oefters (3) - geschneyet (3) - matthaei (3) - aufgerichtet (3) - nidere (3)
5	laetare (9) - gewülk (9) - zwerchluft (7) - benedicti (6) - tief (6) - schlittlin (5) - hoch (5) - bedekt (5) - mertz (4) - gebett (4) - frater (4) - geflosßen (4) - zartem (4) - gefahr (4) - brugg (4) - vsßerhalb (3) - deki (3) - angegerifen (3) - kazenstrik (3) - verwandten (3)
6	oster (20) - schnee (12) - hoch (10) - jörgen (9) - grüene (9) - spazieren (9) - georgii (8) - albis (7) - ostern (7) - gruenen (7) - grüenen (7) - ertrochnet (7) - vnlustiger (7) - legen (7) - wässerig (7) - ostermonntag (6) - palmarum (6) - mammeren (6) - etwaß (6) - schwarz (5)
7	laub (16) - buechen (13) - grüen (9) - versicheret (6) - wald (6) - horn (5) - meyens (5) - weiden (5) - versehen (5) - himmeltauw (4) - platzreegen (4) - horte (4) - schikten (4) - disponieren (4) - baümen (4) - getragen (4) - beliebt (4) - glarus (4) - scheinte (4) - reinauw (3)
8	vngewitter (10) - zug (10) - saagen (8) - pfingsten (8) - ascensionis (7) - halden (7) - tauw (7) - pfingstmonntag (6) - tundern (6) - eingehalten (6) - pfingsttag (5) - sennten (5) - morgenröthe (5) - gesungen (5) - weiter (5) - octavam (5) - eingegangen (5) - erdrich (5) - legen (5) - gewendt (5)
9	pauli (11) - petri (10) - heüwet (6) - maurer (6) - schwarze (6) - strenge (6) - nit_wenig_warm (5) - vnversehener (5) - ioannis (4) - außgang (4) - fronleichnambs (4) - alpen (4) - gesprüzt (4) - enthalten (4) - strahl (4) - segnen (4) - glauben (4) - ylends (3) - rüehig (3) - grath (3)

10	roggen (15) - sanctae (11) - heüwet (9) - annae (8) - zehendtgarben (7) - jacobi (7) - velder (6) - abgeworfen (6) - zornig (5) - grichten (5) - erlösst (5) - hew (5) - baden (5) - abgeritten (5) - feyrtag (5) - roken (4) - besichtiget (4) - steinerberg (4) - verlehnt (4) - wachß (4)
11	laurentii (8) - bartholomaei (6) - rütti (6) - zimmlich_warmer (5) - vychpresten (4) - augsten (4) - aufnemmen (4) - yberg (4) - weyd (4) - gewährt (4) - folgte (4) - verhofet (3) - augustini (3) - kornernd (3) - einsamblen (3) - zeitig (3) - streiteten (3) - vendenbacher (3) - erlittenem (3) - guetem (3)
12	embd (36) - haber (35) - garben (28) - schnitter (19) - knecht (12) - rösch (9) - meister (9) - embdt (8) - ieder (8) - geschlagen (8) - lenger (8) - assumptionis (7) - schneiden (7) - stiller (7) - comet (7) - brugg (7) - probsten (6) - benno (6) - einführen (6) - wuhr (6)
13	engelweyhung (8) - mauritii (8) - schlosßwiß (7) - sommer (7) - saath (5) - herpsts (5) - fahrten (5) - reegenwetter (5) - geschütz (4) - continuiren (4) - mehrentheils (4) - angelasßen (4) - zugeschinen (3) - betrüebt (3) - krämer (3) - schütz (3) - nebell (3) - embdet (3) - saur (3) - befürderet (3)
14	trauben (84) - reifen (65) - wimmlen (64) - wein (62) - wimmet (46) - reeben (36) - herpst (33) - eymer (18) - warm (17) - reifens (15) - herpsten (14) - vechtnauw (13) - gelten (12) - reif (12) - rütti (12) - fertig (12) - weins (12) - priester (12) - eimer (10) - trinken (10)
15	eimer (8) - andreae (6) - diken (6) - vohn (5) - gulden (5) - monachorum (4) - catharinae (4) - martini (4) - weisßen (4) - weegweiser (3) - praesentations (3) - omnium (3) - fierling (3) - rüeben (3) - steinbruch (3) - kreuzer (3) - verwahren (3) - heyligen (3) - erhellete (3) - gebesßeret (3)
16	wyenacht (8) - stephani (7) - seiten (5) - adventi (4) - immaculatae (4) - feücht (4) - heiligen (4) - dorf (4) - evangelistae (3) - conceptionis (3) - vnbeflekten (3) - kugel (3) - zürrichbiet (3) - fewr (3) - dingen (3) - aufgehört (3) - vile (3) - vmbgeben (3) - gewulkte (2) - gaudete (2)
17	gaden (7) - stigen (6) - collation (5) - gesauset (5) - hierauf (5) - eröfnet (4) - limmet (4) - celebrierten (4) - vermehrt (4) - stürmen (4) - gesorget (4) - grüenenberg (4) - karrer (4) - langem (4) - schnees (4) - pferdten (4) - flachß (3) - passiert (3) - gehanget (3) - genötiget (3)
18	grossen (6) - beglüke (6) - ehrlicher (5) - porten (5) - mehrtheil (5) - hinzu (5) - brugg (5) - kreftigen (4) - außgestreüwet (4) - nit_gar_warm (4) - wahete (4) - mittwoch (4) - annemlich (4) - eingeladen (4) - mittwochen (4) - weitem (4) - geredt (3) - nemmlich (3) - redlich (3) - gemerkt (3)
19	bluest (14) - hochheilige (11) - aderläsße (9) - regnender (8) - kirschi (8) - gasßen (7) - neüw (6) - bygewohnet (6) - ehrlich (5) - benedict (5) - vorkommen (5) - staub (5) - gehangen (4) - schwarzes (4) - zulauf (4) - vollendet (4) - verursachen (4) - angetroffen (4) - vorherigen (4) - angesäet (4)
20	garten (7) - lauf (5) - tunderte (5) - sizte (5) - herrschaft (5) - sanctorum (5) - vnversehens (5) - herabgefallen (4) - sanften (4) - wahete (4) - hierinn (4) - gemeint (3) - reife (3) - reiche (3) - giesßen (3) - hofentlich (3) - erfrischete (3) - vnkösten (3) - ordnung (3) - tunderklapf (3)
21	schafrath (5) - gewachßen (5) - vermischet (4) - verzehrt (4) - beichtstuhl (4) - bestellt (4) - karrer (4) - verlehnt (3) - bridenwyß (3) - schreken (3) - gemeinen (3) - trochen (3) - ziegel (3) - zurüsten (3) - verehren (2) - bikel (2) - deken (2) - weyniger (2) - würrenloß (2) - abgelupft (2)
22	gestosßen (9) - fürderling (7) - gemein (7) - erfahrten (7) - schwarze (6) - fasß (5) - meilen (5) - hagelsteinen (5) - erachtete (5) - fratres (5) - verwunderet (5) - monntag (5) - laggeyen (4) - schwarzem (4) - anbefohlen (4) - leider (4) - sezte (4) - langer (4) - verreißen (3) - molest (3)
23	täglich (12) - zeitung (10) - besßerte (7) - fratres (6) - kein_sonnenschein (5) - zimliche (5) - weinreben (4) - freüwete (4) - wohlen (4) - violentia (4) - priorin (4) - burgermeister (4) - klein (4) - augen (4) - abgeleßen (3) - gewüsst (3) - gewülkh (3) - canzler (3) - hiesigem (3) - reegen-lechten (3)
24	himmel (1445) - luft (750) - nebel (730) - wetter (542) - hell (511) - sonne (484) - wind (455) - still (413) - schnee (378) - kalt (371) - heller (328) - fieng (301) - wähete (301) - sehr_kalt (301) - kalter (297) - fanden (291) - vöhn (280) - continuierte (268) - sonnenschein (260) - beständig (258)
25	wetter (1291) - reegen (1103) - himmel (794) - gewülk (499) - starker (394) - gott (322) - starken (318) - ernstlich (304) - sonne (301) - schöner (299) - schön (295) - warm (259) - geregnet (253) - trüeb (228) - warmer (225) - regnete (217) - sonnenschein (216) - edler (209) - regnen (209) - angefangen (176)

6.4.1.5. 30 Topics

1	schnee (1044) - kälte (239) - kalt (239) - nit_kalt (176) - luft (175) - kalter (162) - hellen (144) - milt (138) - strasßen (132) - sehr_kalt (123) - milter (109) - nebel (106) - see (102) - sehr_kalter (99) - gefrohren (97) - heller (96) - fanden (95) - nit_sonders_kalt (95) - schnees (87) - schneelin (84)
2	heüw (157) - tach (118) - schaden (98) - frucht (70) - vngewitter (63) - korn (57) - zehendten (48) - ligende (42) - ernd (42) - plizgen (36) - limmet (36) - schneiden (35) - stuk (35) - zürrich (35) - bergen (34) - heisßer (31) - geschehen (30) - tunderwetter (27) - hiz (26) - roggen (25)
3	grosser (9) - beißwind (5) - grimmiger (5) - vblen (4) - einzig (4) - mennweg (4) - lieben (4) - neben (4) - verwandten (3) - angestossen (3) - glorwürdigen (3) - pfausete (3) - grüenenberg (3) - heimkommen (3) - wälder (3) - ellen (3) - heüfiger (3) - wenigem (3) - rahthauß (3) - catharren (3)
4	schlitten (39) - ochßen (20) - zwechtenen (16) - knecht (16) - schwär (13) - vngangbar (13) - schlittlin (12) - tächern (11) - strenger (10) - brünnen (8) - gebraucht (8) - milterte (7) - ezel (7) - jeger (7) - seedorf (6) - herunder (6) - antonii (6) - abgangen (6) - fahren (6) - duhr (5)
5	faßacht (8) - frauen (5) - scholasticae (4) - picht (4) - zinstag (4) - horgenberg (4) - kuchimeister (4) - vahr (4) - liesßen (4) - matthaei (3) - septentrionalischer (3) - strich (3) - hornung (3) - erdbidem (3) - vberwähet (3) - senften (3) - purificationis (3) - eingebracht (3) - geendeter (3) - schwösterhauß (3)
6	reebstok (12) - laetare (9) - mertz (4) - frater (4) - zimmlich_milt (4) - continuieren (3) - thomae (3) - dünnen (3) - rühewige (3) - benedicti (3) - andächtig (3) - achten (3) - zuger (3) - zimmlich_frischer (3) - zwerchluft (3) - gestigen (3) - menni (3) - starken (3) - vbrigens (2) - verdienen (2)
7	oster (18) - albis (8) - ostern (7) - vngestümmen (7) - kirchen (7) - palmarum (6) - grüenen (6) - carrfreytag (5) - ostermonntag (5) - kirschi (4) - osterzinstag (4) - feüchtigkeit (4) - frische (4) - jährliche (4) - wurzel (4) - scharpfem (4) - ostertag (4) - tächern (4) - byß (4) - sabbatho (3)
8	laub (17) - buechen (13) - jörgen (10) - grüene (10) - georgii (9) - grüen (9) - besser (8) - beziehen (8) - matten (8) - kein_reifen (7) - ersorgen (6) - aufgeschoben (5) - ertruknet (5) - mangel (5) - schikten (4) - wachßen (4) - meyens (4) - bluest (4) - feüwr (4) - gottes (4)
9	graß (18) - saagen (9) - vngewitter (8) - zug (8) - ascensionis (7) - pfingsten (7) - gewähet (6) - processionem (5) - enthalten (5) - sennten (5) - jammer (5) - frauen (5) - abendtröthe (5) - außgang (4) - pfingsttag (4) - meinen (4) - zelg (4) - schliesßen (4) - himmeltauw (4) - glöklin (4)
10	pauli (11) - petri (9) - ferne (8) - sanctorum (8) - höhenen (5) - nit_wenig_warm (5) - vnversehener (5) - ioannis (4) - fronleichnambs (4) - veychpresten (4) - alpen (4) - sennen (4) - bitten (4) - segnen (4) - geschosßen (4) - ylends (3) - grath (3) - wekgeführt (3) - pfingstmonntag (3) - oberland (3)
11	nachlasß (9) - annae (8) - eingebracht (8) - ohrt (8) - jacobi (7) - erscheinen (6) - geschochet (5) - erlösst (5) - heüws (5) - benedicti (5) - inzwüschen (5) - pro (5) - abgeritten (5) - ligen (5) - abgeworfen (5) - roken (4) - lauf (4) - besichtiget (4) - steinerberg (4) - meilen (4)
12	laurentii (8) - ettlich (7) - bartholomaei (6) - wuhr (6) - spazierte (6) - syhlthal (6) - kornschnitt (4) - vychpresten (4) - rieth (4) - gewahret (4) - ehe (4) - gasß (4) - baumnusß (3) - deiparae (3) - geschadet (3) - aussgebrochen (3) - genebleter (3) - wenden (3) - stehende (3) - möglichkeit (3)
13	embd (35) - garben (22) - schnitter (12) - rösch (9) - assumptionis (8) - embdt (8) - benno (7) - blaßen (7) - scheüren (6) - drey (6) - wasßer (6) - geschütz (5) - embdet (5) - verschikt (5) - emds (4) - einige (4) - embden (4) - schneideten (4) - tonders (4) - farb (4)
14	mauritii (8) - engelweyhung (7) - saath (5) - abfallen (5) - sehr_warmer (5) - kalte (5) - hiervon (4) - säen (4) - natae (3) - zugeschinen (3) - betrüebt (3) - krämer (3) - schütz (3) - räben (3) - scheür (3) - saur (3) - vberlaufen (3) - sausete (3) - schiff (3) - gestrichen (3)

15	trauben (84) - reifen (80) - wein (66) - wimmlen (64) - nit_kalt (46) - wimmet (45) - reeben (40) - herpst (35) - warm (20) - eymer (19) - gelten (15) - reif (15) - vechtnauw (13) - fertig (13) - weins (13) - hell (13) - herpsten (12) - priester (12) - reifens (11) - rütti (11)
16	eimer (7) - andreae (6) - martini (5) - monachorum (4) - catharinae (4) - rüeben (4) - vohn (4) - kreuzer (4) - gulden (4) - weegweiser (3) - praesentationis (3) - omnium (3) - fierling (3) - verwahren (3) - schuchs (3) - heyligen (3) - erdtreich (3) - limmet (3) - frau (3) - wolt (2)
17	wyenacht (8) - stephani (7) - thomae (6) - sonderliches (6) - gewähret (6) - geneblet (5) - adventi (4) - conceptionis (4) - immaculatae (4) - vohn (4) - vnversehens (4) - evangelistae (3) - vnbeflekten (3) - zürrichbiet (3) - fewr (3) - adventus (3) - adventii (3) - meteorum (3) - baron (3) - gintelhart (3)
18	ruhen (7) - weyn (7) - schlafen (5) - pfarrer (5) - sonderer (4) - alsgemach (4) - verdeütete (4) - neblechter (4) - preyß (4) - warfe (4) - trinken (4) - rukkehr (4) - eingetretten (4) - rothe (3) - streich (3) - ernetzt (3) - spühren (3) - genötiget (3) - vermehrt (3) - gesagtem (3)
19	einsidlen (4) - gabe (4) - allerlei (4) - zeichen (4) - vorherige (4) - raht (4) - erfahren (4) - gintelhart (3) - kehrte (3) - mutter (3) - preiß (3) - verspühren (3) - sambt (3) - nebelecht (3) - kurtz (3) - aderläßer (3) - geschah (3) - vernemmen (3) - fuhr (3) - spahren (2)
20	hinab (7) - heüßer (6) - pliz (5) - geredt (5) - mettin (4) - licentia (4) - gewüßßen (4) - versus (3) - cum (3) - gedankt (3) - antworten (3) - hinderlaßßen (3) - denniken (3) - gewohnheit (3) - vnlustige (3) - verschmelzen (3) - ittendorf (3) - gelofen (3) - angefangene (3) - schlipfe (3)
21	iezunder (9) - galli (5) - streichete (5) - verlangt (5) - aufgezogen (4) - schmuzigen (4) - gemeiner (4) - constanz (4) - baaß (4) - margstahl (4) - sprüzlin (4) - schön (4) - saum (4) - harter (4) - hohe (4) - stürmig (3) - näßße (3) - dünnen (3) - spater (3) - verhofte (3)
22	windt (5) - luften (5) - liecht (5) - prediger (4) - erwarmet (4) - zugefüegt (4) - arbeiteten (3) - schonte (3) - geworffen (3) - weyd (3) - freytag (3) - fuesß (3) - billich (3) - rüehwiger (3) - käli (3) - blasste (3) - zwerchwind (3) - nassem (2) - besserte (2) - trübe (2)
23	bluest (11) - früchten (9) - hochheilige (9) - iagte (9) - hinzu (9) - här (8) - kein_tropfen (6) - gestrahlet (6) - reichenauw (5) - geschwind (4) - watten (4) - march. (4) - zulauf (4) - vorherigen (4) - praelat (4) - lustiger (4) - legen (4) - sand (4) - gearbeitet (3) - erfrischete (3)
24	weiden (14) - vorkommen (9) - mehreren (8) - früestüklin (7) - zierlich (7) - blizgen (6) - einsmahls (6) - nassen (5) - gemeinlich (5) - processio (5) - vorauß (5) - graß (5) - sterkeren (4) - mantel (4) - wettingen (4) - aussgesehen (4) - gewohnnheit (4) - expediert (4) - verhofte (4) - angelangt (4)
25	creüz (14) - nam (9) - limmat (5) - plizgens (4) - ersezt (4) - ennet (4) - befürderet (4) - bedörfen (4) - gebettet (4) - entsinnen (4) - gewässer (4) - erndtag (3) - tropfeter (3) - aussgebliben (3) - steig (3) - hagelsteinen (3) - yberg (3) - beständige (3) - verhoften (3) - walderen (3)
26	eschentz (5) - landtvogt (5) - gloken (5) - gedanken (5) - kurzem (4) - schein (4) - brauch (4) - zehendt (3) - reiche (3) - anna (3) - ofener (3) - molset (3) - streichten (3) - ohnzweifenlich (3) - abgeschüttet (3) - sonniger (3) - fortgang (3) - entschlosßen (3) - weegs (3) - bach (3)
27	bringend (4) - oster (4) - herberg (3) - eodem (3) - raucher (3) - ellenden (3) - außgeben (3) - angangen (3) - jüngere (2) - jagte (2) - absönderlich (2) - saumbte (2) - subpriore (2) - nammlich (2) - vorbereitung (2) - ende (2) - pfingstzinstag (2) - natürlicher (2) - gefrieren (2) - vmbgern (2)
28	erfahren (5) - apostoli (4) - sezten (4) - prognosticieren (4) - zwüschen (4) - vnderlegt (3) - vorgesagt (3) - gegenwertig (3) - daselbst (3) - mehren (3) - nidergehauwen (3) - vnheyl (3) - dunst (3) - aufgestellt (3) - stiesße (3) - nominis (3) - sigentahl (3) - erfolgen (3) - fortgangen (3) - meister (3)
29	himmel (2239) - wetter (954) - sonne (784) - nebel (623) - luft (573) - wind (538) - still (520) - hell (511) - sonnenschein (476) - fieng (469) - angefangen (356) - vnderluft (355) - wähete (337) - beständig (308) - warm (261) - oberluft (254) - erdtrich (244) - heller (234) - continuierte (227) - continuiert (224)
30	reegen (1101) - wetter (875) - gewülk (510) - gott (322) - starker (263) - geregnet (255) - ernstlich (234) - regnete (219) - schöner (214) - edler (210) - schön (206) - regnen (199) - starken (189) - vbel (171) - procession (167) - trüeb (165) - reifen (148) - warm (131) - scheinte (130) - reegenwetter (126)

6.4.2. Segmentierung pro Beobachtungsort und Jahreszeit

6.4.2.1. 5 Topics

1	himmel (1967) - luft (751) - wetter (746) - nebel (729) - sonne (710) - wind (537) - reegen (488) - fieng (465) - hell (461) - sonnenschein (431) - warm (318) - heller (315) - still (309) - kalt (308) - beständig (304) - kalter (295) - schnee (289) - gewülk (270) - sehr_kalt (269) - continuierte (257)
2	reegen (515) - wetter (479) - himmel (274) - gewülk (243) - tach (174) - starker (172) - schaden (150) - heüw (139) - wasßer (127) - schöner (127) - geregnet (127) - ernstlich (116) - gott (113) - edler (100) - hagel (97) - folgte (91) - warmer (83) - regnen (82) - regnete (80) - sonne (75)
3	schnee (767) - kälte (165) - schnees (74) - strasßen (70) - vöhn (67) - kalt (66) - schlitten (64) - pfefiken (62) - schneelin (57) - kelte (49) - winter (44) - sehr_kalt (44) - sehr_kalter (43) - strasß (41) - ochßen (41) - hellem (41) - wein (38) - warme (38) - glüklich (37) - wasßer (36)
4	wetter (600) - procession (159) - heüw (123) - gott (89) - heiligen (84) - reegen (83) - ohrten (72) - statthalter (71) - warm (71) - starken (70) - graß (67) - veych (67) - angefangen (66) - vöhn (66) - monat (66) - hell (64) - pfefiken (63) - halten (61) - brüel (61) - schaden (60)
5	still (192) - vnderluft (121) - hauß (113) - wähete (93) - haber (70) - schön (68) - veld (55) - eschenz (55) - regnete (54) - rüehwig (51) - fanden (51) - ernstlich (51) - pferdt (50) - reeben (49) - schif (47) - treüwete (45) - statthalter (45) - gott (45) - starken (43) - knecht (41)

6.4.2.2. 10 Topics

1	still (167) - hauß (111) - vnderluft (104) - wähete (69) - rüehwig (66) - fanden (66) - tiefe (66) - schön (61) - eschenz (53) - haber (48) - reeben (47) - wind (41) - pferdt (38) - nit_sonders_kalt (37) - vnderwind (35) - cell (33) - freüwdenfels (33) - vnrüehwig (33) - see (32) - clingenzell (31)
2	zürrich (63) - folgte (52) - limmet (47) - finster (41) - sonnenscheiniger (40) - küeler (39) - sprüzlin (37) - finstere (36) - sonnenscheinig (36) - zierlicher (34) - gelegen (33) - gewulket (32) - klosterfrauwen (27) - enderung (26) - sehr_heisßer (25) - tröpflen (25) - andermahl (25) - morgenröthe (24) - frau (24) - trüeb (23)
3	schnee (1019) - kälte (204) - kalt (149) - vöhn (144) - hell (129) - strasßen (123) - sehr_kalt (113) - schnees (88) - wind (88) - schneelin (85) - sehr_kalter (66) - milt (64) - milter (61) - strasß (58) - heller (58) - kalter (55) - hellem (53) - schneyen (47) - nit_sonders_kalt (47) - winter (47)
4	pfefiken (46) - ochßen (41) - wein (35) - mennweeg (34) - schlitten (32) - monat (30) - brunnen (29) - continuirt (28) - kälte (24) - veych (22) - gottshauß (21) - recreation (19) - knecht (17) - kelte (16) - strenger (15) - eingefrohren (14) - brauchen (14) - priester (13) - rosß (13) - mennen (12)
5	procession (69) - graß (52) - erdtrich (29) - heüw (28) - veych (27) - frisch (22) - hoch (21) - halten (20) - spazieren (19) - matten (18) - continuirt (18) - kloster (17) - zug (15) - bauren (15) - mangel (15) - weiden (14) - armen (13) - außgebliben (13) - oster (13) - mitte (13)
6	heüw (120) - brüel (61) - procession (55) - schaden (47) - tunderwetter (31) - tach (31) - syhlthal (27) - festo (24) - regenwetter (22) - ligende (21) - plazreegen (19) - gottshauß (19) - ernezt (18) - stehende (18) - veych (18) - heüwen (17) - gewohnnte (17) - complet (14) - erlaubt (14) - gloken (14)
7	reegen (229) - tach (140) - gewülk (117) - heüw (115) - schöner (93) - regnete (73) - frucht (71) - starker (71) - schaden (69) - vngewitter (66) - ernstlich (66) - korn (61) - zehendten (51) - treüwete (49) - gott (49) - ernd (47) - scheinte (44) - ernezt (39) - hagel (38) - garben (37)
8	trauben (97) - reifen (67) - wein (65) - wimmlen (64) - wimmet (46) - herpst (34) - eymer (19) - eimer (18) - gelten (15) - reifens (14) - regnete (14) - herpsten (12) - sehr_kalt (11) - torkel (10) - pferdt (10) - ertragen (9) - engelweyhung (9) - vnreif (8) - fechtnauw (8) - hofte (8)
9	wetter (1354) - reegen (485) - gott (237) - warm (197) - wasßer (193) - ohrten (156) - angefangen (153) - vbel (139) - himmel (139) - starken (136) - ettliche (124) - reegenwetter (111) -

	statthalter (110) - geregnet (109) - allein (109) - geschehen (105) - starker (101) - schaden (97) - gesehen (94) - weniger (92)
10	himmel (2070) - sonne (766) - nebel (730) - luft (730) - sonnenschein (476) - wetter (476) - fieng (469) - wind (426) - reegen (389) - gewülk (385) - still (355) - hell (300) - beständig (296) - scheinte (280) - schöner (275) - continuierte (272) - wähete (268) - heller (266) - oberluft (254) - vnderluft (249)

6.4.2.3. 15 Topics

1	procession (139) - heüw (93) - graß (68) - volk (43) - veych (37) - gottshauß (35) - heiligen (32) - tunderwetter (31) - lieben (31) - brüel (30) - regenwetter (29) - vesper (28) - matten (27) - bauren (26) - beste (24) - kloster (23) - capell (21) - feld (21) - gewohnnte (20) - gloken (20)
2	still (205) - vnderluft (146) - wähete (114) - hauß (103) - oberluft (93) - fanden (93) - haber (67) - rüehwig (65) - eschenz (52) - schön (48) - veld (47) - see (46) - vnderwind (43) - voller (43) - wind (40) - schif (39) - sonnenberg (34) - reeben (34) - cell (33) - tiefe (33)
3	heüw (121) - schaden (91) - tach (58) - hagel (39) - reegenwetter (31) - brüel (27) - plazreegen (24) - angelofen (24) - embd (23) - ernezt (23) - heüwen (22) - ligende (22) - erschröklicher (22) - zugefüegt (21) - bach (18) - stehende (17) - hagels (16) - schif (16) - gestrahlet (16) - mariae (16)
4	zürrich (52) - folgte (50) - limmet (47) - gelegen (39) - festo (37) - sprüzlin (35) - sonnenscheiniger (33) - zierlicher (30) - diker (30) - angefangen (29) - klosterfrauwen (28) - sonnenscheinig (28) - sehr_heisßer (23) - zart (22) - gewähret (21) - oberwind (20) - bezogener (20) - sonn (20) - finstere (19) - vesper (19)
5	schnee (596) - kälte (138) - schnees (73) - schlitten (63) - pfefiken (58) - strasßen (55) - kelte (48) - vöhn (43) - ochßen (41) - frischen (34) - schneelin (34) - mennweeg (34) - monats (32) - vngelegenheit (30) - brunnen (28) - weeg (26) - wind (26) - leidenlich (25) - winter (25) - hellem (25)
6	jeger (10) - menschen (7) - früe (7) - müsße (6) - catharren (6) - wenden (6) - veich (6) - seedorf (5) - aufwart (5) - derselbe (5) - sonntag (5) - jesu (5) - zureden (5) - viler (5) - brauchen (5) - herein (5) - angenemb (4) - durchtrungen (4) - besuechen (4) - vorherigem (4)
7	oster (19) - laub (14) - grüen (14) - buechen (13) - grüene (11) - rauch (11) - ertrochnet (10) - albis (9) - laetare (9) - jörgen (9) - georgii (9) - schneyen (9) - neüwen (9) - erfreüwt (8) - korn (8) - wordurch (8) - heylige (8) - gewülket (7) - zimmlich_milt (7) - schneyete (7)
8	syhlthal (14) - veychpresten (10) - syhltal (10) - angestellt (10) - grund (10) - embdt (9) - aussgegosßen (9) - ergosßen (9) - annae (8) - höhe (8) - abgelofen (8) - fenster (8) - dorf (8) - brütten (7) - bitten (7) - gutes (7) - erschröklich (7) - yberg (7) - comet (7) - sturmwind (7)
9	sehr_kalt (38) - continuirt (27) - solemnitet (26) - reifen (25) - wimmet (18) - gefolget (17) - recreation (16) - sehr_warm (15) - veych (15) - manchem (14) - obwohlen (14) - priester (13) - pfefiken (13) - verhoft (12) - herpst (12) - denn (11) - jahrzeit (10) - guter (10) - recreationem (9) - compagnia (9)
10	see (53) - zwechtenen (14) - erschröklich_kalt (12) - zuruk (12) - schlittlin (11) - eynsidlen (11) - strassen (10) - byßwind (9) - verdekt (9) - obervogt (8) - wohnen (8) - oft (8) - angesicht (7) - wahrheit (7) - geschlagen (7) - selber (7) - fortzukommen (7) - sehr_kalte (7) - schwär (7) - brieflin (6)
11	tach (84) - frucht (70) - korn (55) - heüw (49) - garben (38) - schneiden (37) - ernd (37) - zehendten (32) - vngewitter (28) - scheür (26) - sak (25) - gersten (23) - eingeführt (21) - gedanket (21) - schönen (20) - schnitt (19) - tondern (19) - geschnitten (18) - aker (18) - frücht (17)
12	trauben (87) - wimmlen (64) - reifen (48) - wein (33) - wimmet (28) - reeben (23) - eymer (20) - herpst (19) - eimer (18) - gelten (14) - embd (12) - reifens (12) - laub (11) - obs (10) - truken (10) - torkel (9) - reif (9) - vnreif (8) - fechtnauw (8) - stad (8)
13	himmel (1008) - luft (511) - nebel (476) - schnee (460) - kalt (374) - sonne (311) - wind (299) - kalter (278) - sehr_kalt (277) - hell (242) - heller (228) - nit_kalt (217) - beständig (204) - fanden

	(197) - sonnenschein (197) - fieng (193) - still (192) - hellen (188) - continuierte (186) - wähete (166)
14	wetter (1407) - reegen (371) - angefangen (267) - hell (264) - warm (244) - starken (180) - himmel (176) - wasßer (171) - gott (161) - ohrten (157) - ettliche (148) - schöner (140) - monat (134) - vöhn (130) - wein (128) - starker (124) - vbel (122) - festo (113) - statthalter (112) - gesehen (112)
15	himmel (1054) - reegen (732) - gewülk (508) - sonne (471) - wetter (427) - sonnenschein (284) - fieng (275) - starker (264) - nebel (250) - ernstlich (247) - schöner (242) - luft (239) - regnete (211) - scheinte (209) - edler (207) - trüeb (199) - warmer (187) - geregnet (167) - warm (166) - schön (162)

6.4.2.4. 20 Topics

1	procession (129) - heüw (88) - veych (63) - graß (61) - vesper (48) - boden (45) - brüel (42) - volk (41) - gottshauß (39) - continuirt (34) - halten (33) - tunderwetter (32) - regenwetter (29) - lieben (28) - matten (26) - frischer (26) - kloster (24) - feür (23) - continuirte (22) - feld (22)
2	still (205) - vnderluft (109) - wähete (92) - hauß (83) - schön (75) - fanden (74) - tiefe (70) - rüehwig (63) - haber (60) - eschenz (54) - oberluft (51) - reeben (43) - veld (38) - frühe (38) - cell (33) - treüwete (33) - sonnenberg (32) - namen (32) - clingenzell (31) - reitete (31)
3	schif (26) - bach (20) - see (18) - brüel (17) - reegenwetter (14) - erschröklicher (12) - strasß (11) - starkes (11) - eynsidlen (10) - caspar (9) - beste (9) - weyer (8) - maur (8) - wahrheit (8) - verhoften (8) - getunderet (8) - gewalt (8) - vnnder (7) - spaziert (7) - vfnauw (7)
4	gelegen (53) - zürrich (52) - limmet (47) - folgte (47) - sonnenscheiniger (46) - festo (45) - trüeber (34) - milter (33) - zierlicher (32) - finster (28) - sonnenscheinig (25) - sehr_heisßer (24) - sprüzlin (24) - vesper (24) - bezogener (23) - gewähret (22) - sancti (21) - tröpflen (20) - klosterfrauwen (20) - gewulket (19)
5	schnee (922) - kälte (141) - vöhn (114) - strasßen (96) - kalt (95) - schneelin (86) - schnees (85) - sehr_kalt (71) - winter (64) - warme (59) - hellem (55) - milt (52) - milter (49) - sonnen-scheiniger (48) - frisch (48) - frischen (47) - schneyen (46) - leidenlich (45) - schmilzen (45) - kelte (43)
6	pfeffiken (11) - iezunder (9) - gaden (9) - metti (9) - schönem (8) - leib (8) - begleitung (7) - vohn (7) - täcberen (7) - kalten (7) - catharren (7) - haupt (7) - sehr_kaltem (7) - fratrum (6) - aufs (6) - weyn (6) - gachlingen (6) - allmächtige (6) - gottshaußes (6) - gefrörne (6)
7	kälte (51) - schlitten (47) - ochßen (41) - pfefiken (38) - mennweeg (33) - weeg (24) - brunnen (24) - geführt (20) - getragen (19) - seümer (15) - ezel (14) - vngangbar (14) - eingefrohren (14) - noth (13) - gottshauß (12) - fuhr (12) - grimmiger (11) - knecht (11) - rosßen (10) - zugericht (10)
8	milter (13) - meinradi (9) - antonii (7) - fahrten (6) - fortgefahren (6) - jahren (6) - saum (6) - steken (6) - böß (6) - harten (5) - comet (5) - hergangen (4) - gehört (4) - tiefenen (4) - zufuhr (4) - nutzen (4) - schlaf (3) - verdekte (3) - stangen (3) - güsel (3)
9	oster (19) - laub (17) - grüen (14) - graß (14) - spazieren (14) - buechen (13) - jörgen (10) - warmen (10) - albis (9) - grüene (9) - laetare (9) - georgii (9) - ertrochnet (9) - reifen (8) - rauch (8) - ostern (7) - ertruknet (7) - wald (7) - ostermonntag (6) - palmarum (6)
10	sonnenscheinig (10) - capell (9) - kurz (8) - hochheilige (6) - exponiert (5) - gewülket (5) - gebettet (5) - feria (5) - sennten (5) - heüriges (5) - pfingsten (5) - ezel (5) - wunderlich (5) - verenderet (4) - verstekt (4) - frater (4) - teüre (4) - verschoben (4) - gesezt (4) - knye (4)
11	heüw (175) - tach (146) - schaden (89) - hagel (36) - ligende (35) - zehendten (33) - plazreegen (31) - abgelofen (22) - heüwen (20) - embd (20) - zugefüegt (20) - nebel (19) - angelofen (19) - ernezt (18) - stehende (18) - erschröklich (18) - getrofen (16) - heüwet (15) - eintragen (14) - erbärmlich (14)
12	syhlthal (13) - nidergelegt (11) - veychpresten (10) - embdt (9) - syhltal (9) - dorf (9) - gloken (8) - syhl (7) - streich (7) - annae (7) - strahl (7) - zugericht (7) - brütten (6) - bitten (6) - gärten (6) - weyd (6) - außgebrochen (6) - angestellt (6) - yberg (6) - grund (6)

13	vöhn (29) - sehr_kalt (23) - reifen (22) - pfefiken (22) - wimmet (21) - solemnitet (17) - herpst (16) - priester (12) - recreation (11) - continuirt (10) - jahrzeit (8) - engelweyhung (8) - sehr_warm (7) - gebesßeret (7) - trinken (7) - mauritii (6) - hofte (6) - recreationem (6) - angelasßen (6) - mehrentheils (6)
14	zwechtenen (15) - schlittlin (9) - obervogt (7) - tahl (6) - ziehete (6) - pik (5) - byßwind (5) - möglich (5) - zuruk (5) - oberstad (4) - stekenborn (4) - rühewig (4) - bygewohnet (4) - beunrüehwiget (4) - legte (4) - wägen (4) - zimmeren (4) - vngelegenheit (4) - anfänglich (4) - geführt (4)
15	iagte (11) - zelg (10) - sigenthal (10) - dietlandus (10) - zimmlich_warm (9) - spazierte (8) - mammeren (7) - oberwind (7) - gütigste (7) - reegenlechten (6) - neüw (6) - hellete (6) - klar (6) - ernezte (5) - luftig (5) - plizgete (5) - korn (5) - caspar (5) - gottsdienst (5) - erscheinten (4)
16	frucht (41) - garben (37) - korn (32) - schneiden (31) - scheür (26) - gersten (25) - sak (24) - gedanket (23) - eingeführt (22) - ernd (22) - schnitter (20) - schnitt (18) - geschnitten (15) - zehendten (15) - darumb (15) - einzu (14) - dank (13) - juchert (12) - karrer (12) - hofnung (12)
17	trauben (77) - wimmlen (64) - reifen (54) - wein (35) - wimmet (24) - nebel (23) - reeben (22) - eymer (19) - tach (19) - eimer (18) - herpst (16) - herpsten (13) - gelten (13) - regnete (11) - truken (10) - embd (10) - laub (10) - reifens (10) - fertig (9) - erstlich (9)
18	wetter (1542) - reegen (346) - himmel (287) - warm (269) - hell (259) - wasßer (255) - angefangen (210) - gott (196) - starken (178) - vbel (176) - ohrten (151) - wein (139) - ettliche (136) - monat (135) - schön (132) - festo (128) - schöner (123) - gesehen (118) - vneracht (117) - allein (115)
19	himmel (1371) - luft (702) - nebel (679) - wind (462) - sonne (442) - fieng (336) - sonnenschein (310) - kalter (293) - kalt (278) - hell (266) - continuierte (263) - heller (257) - wähete (245) - still (240) - hellen (239) - nit_kalt (225) - fanden (219) - sehr_kalt (219) - beständig (206) - erdtrich (197)
20	reegen (750) - himmel (581) - gewülk (424) - sonne (343) - wetter (293) - schöner (253) - starker (224) - ernstlich (211) - edler (206) - regnete (181) - trüeb (173) - warmer (168) - geregnet (152) - scheinte (145) - reifen (143) - starken (141) - fieng (133) - regnen (132) - gott (127) - folgte (114)

6.4.3. Segementierung pro Jahr über den Gesamtzeitraum

6.4.3.1. 5 Topics

1	veych (67) - continuirt (45) - regen (29) - continuirte (24) - regenwetter (23) - soll (22) - sagen (21) - solemnitet (20) - see (20) - allezeit (19) - früe (18) - manns_gedenken (18) - monat (17) - geführt (17) - immerdar (16) - comet (16) - obwohlen (15) - tunderwetter (15) - destoweniger (12) - dorf (12)
2	procession (52) - vych (42) - volk (35) - kloster (25) - continuiert (22) - gloken (22) - syhlthal (22) - pfefiken (21) - wald (20) - heimkommen (18) - feür (17) - dardurch (14) - mehrentheil (14) - gottshauß (14) - vesper (14) - geringer (13) - sonntag (12) - hoch (12) - vil (12) - monntag (11)
3	still (177) - see (133) - hauß (109) - schön (87) - haber (72) - wähete (63) - tiefe (63) - reeben (62) - luft (57) - veld (56) - eschenz (55) - rüehwig (52) - schif (48) - wimmlen (44) - bedektem (38) - garben (37) - freüwdenfels (37) - namen (36) - reebstok (34) - reitete (33)
4	wetter (1367) - schnee (626) - reegen (479) - himmel (410) - gott (276) - heüw (246) - wind (230) - luft (210) - kälte (209) - hell (208) - starken (204) - schaden (202) - angefangen (196) - wasßer (183) - starker (177) - warm (163) - erdtrich (157) - vbel (149) - kalt (143) - monat (139)
5	himmel (1831) - sonne (779) - nebel (689) - reegen (620) - luft (484) - wetter (468) - gewülk (448) - fieng (435) - schnee (428) - sonnenschein (386) - vnderluft (354) - schöner (343) - still (341) - hell (318) - heller (308) - wind (307) - fanden (286) - kalter (277) - wähete (274) - scheinte (272)

6.4.3.2. 10 Topics

1	regenwetter (22) - regen (18) - monat (18) - comet (16) - manns_gedenken (13) - allezeit (13) - continuirte (12) - veich (11) - guter (11) - anietzo (10) - gutem (10) - feld (10) - kelte (10) - weyn (9) - lüften (9) - grimmig_kalt (9) - sommer (9) - immerdar (9) - jeger (8) - soll (8)
2	veych (67) - continuirt (44) - mennweeg (18) - ligen (13) - blasste (12) - grimmiger (12) - continuirte (12) - weyden (11) - fuhren (11) - fratres (10) - krankheiten (10) - saagen (9) - mänigklich (9) - veychpresten (9) - vngfahr (9) - schmilzen (8) - hiz (8) - embdt (8) - zürricher (8) - eingefrohren (8)
3	syhlthal (22) - feür (13) - lachen (12) - persohnen (11) - hohe (11) - heim (10) - kuchimeister (10) - gaden (9) - brauch (9) - verhoft (9) - engelweyhung (8) - aufs (8) - gebesßeret (8) - gienge (8) - einte (8) - fürsten (8) - aderen (7) - predigen (7) - althar (7) - metti (7)
4	continuiert (27) - hagel (19) - mehrentheil (11) - hinab (10) - wimmet (10) - gachlingen (9) - weyer (8) - fechtnauw (8) - rothen (8) - wenigist (8) - spath (8) - mancher (8) - wald (8) - maur (7) - mehrentheils (7) - vnlustigem (7) - brütten (7) - gäst (7) - gedulden (7) - gehaußet (7)
5	hauß (80) - pferdt (53) - schif (47) - reeben (36) - freüwdenfels (32) - sonnenberg (26) - frucht (25) - wimmlen (23) - zehendten (23) - namen (22) - erschröklich (22) - reitete (21) - geritten (21) - kamen (20) - cell (19) - fahrten (17) - celebriert (16) - gedanket (16) - ernd (16) - reiteten (15)
6	still (242) - schön (97) - fanden (94) - wähete (89) - see (88) - luft (86) - rüehwig (67) - vnderluft (66) - tiefe (65) - haber (63) - veld (59) - eschenz (48) - himel (45) - hell (40) - zimmlich_kalt (39) - bedektem (38) - oberluft (38) - korn (37) - reeben (37) - kalte (37)
7	schnee (77) - wald (39) - sonnenscheiniger (34) - milter (26) - schneelin (25) - kelte (21) - vesper (19) - procession (18) - frisches (17) - außgegosßen (16) - edler (15) - capell (14) - nit_wenig_kalt (14) - monntag (13) - pfefiken (13) - kloster (12) - halten (12) - angefangen (12) - volk (12) - vorgefallen (11)
8	sonnenscheiniger (67) - zürrich (50) - limmet (47) - zierlicher (33) - bezogener (28) - sehr_heisßer (26) - klosterfrauwen (26) - gewulket (18) - gewulkter (17) - küeler (16) - finster (16) - finstere (16) - genebleter (15) - aussgegosßen (15) - sprüzlin (14) - vesper (14) - ettlich (13) - tunder (12) - baden (12) - weyningen (11)
9	wetter (1365) - schnee (633) - reegen (503) - himmel (327) - heüw (260) - gott (255) - kälte (234) - starken (221) - schaden (214) - luft (211) - wind (209) - hell (207) - angefangen (200) - wasßer (186) - starker (185) - warm (177) - erdtrich (165) - ohrten (157) - procession (149) - schön (139)
10	himmel (1914) - sonne (782) - nebel (729) - reegen (600) - gewülk (474) - fieng (470) - wetter (466) - luft (453) - sonnenschein (447) - wind (345) - schnee (343) - schöner (334) - heller (321) - scheinte (293) - vnderluft (289) - hell (278) - still (277) - ernstlich (276) - kalter (270) - continuierte (265)