Anzeigen als Daten
Dynamisches Tagging und iterative Auswertung eines frühneuzeitlichen Intelligenzblattes

Serif, Ina
ina.serif@unibas.ch
Universität Basel, Switzerland

Reimann, Anna
anna.reimann@unibas.ch
Universität Basel, Switzerland

Engel, Alexander
alexander.engel@unibas.ch
Universität Basel, Switzerland

Proto-Industrialisierung, Industrielle Revolution, Konsumrevolution, Handelsrevolution: Solche Forschungsparadigmen haben das 18. und 19. Jahrhundert zunehmend als eine Zeit des Übergangs zum industriellen Zeitalter gezeichnet, als eine Zeit des institutionellen Wandels, der Intensivierung von Produktion und Konsum, als eine Zeit der verstärkten Arbeitsteilung, letztlich: der Vermarktlichung (Grundlegend Mendels 1972; McKendrick/Brewer/Plumb 1982; Mui/Mui 1989; de Vries 2012; einen aktuellen Forschungsüberblick bieten Blondé/Van Damme 2018). Das SNF-Projekt "Märkte auf Papier – das Basler Avisblatt 1729–1844"[1] nähert sich diesen (und anderen) Themen, indem es ein wöchentlich erschienenes Anzeigenblatt untersucht, das über einen Zeitraum von 116 Jahren einen erheblichen Teil des sozioökonomischen Austauschs in einer Schweizer Grossstadt widerspiegelt, und zwar hinsichtlich Angebot und Nachfrage von Waren, gebrauchten Gütern oder Dienstleistungen, des Wohnungs- und Stellenmarkts, des Geldverleihs und vieler anderer Aspekte (zu Intelligenzzeitungen grundlegend Blome 2006; Tantner 2015).

Das Projekt zielt auf eine digitale Aufbereitung der Quelle ab, die nicht nur in der Bereitstellung digitaler Bilder besteht, sondern die Zeitungsanzeigen in die strukturierte Form einer Datenbank überführt, die als Grundlage für verschiedene historische Studien dient; über unterschiedliche Plattformen und Werkzeuge werden diese auch über das Projektende im Dezember 2022 hinaus zur Verfügung gestellt und analysiert werden.[2]

In einem ersten Schritt hat die Universitätsbibliothek Basel qualitativ hochwertige Digitalisate aller erschienenen Avisblatt-Ausgaben produziert und mitfinanziert. Das technische Rückgrat der daraus resultierenden digitalen Sammlung ist eine von der Data Futures GmbH[3] entwickelte iiif-basierte Annotationsinfrastruktur, genannt Freizo (MongoDB). Sie bietet in Kombination mit dem Mirador-Viewer neben Anzeige- und Speichermöglichkeiten auch Authentifizierungs- und Bearbeitungsmöglichkeiten sowie Import- und Exportfunktionalitäten.

Ausgehend von den einzelnen Anzeigen als Analyseeinheiten haben wir einen eindeutig referenzierbaren Datensatz für jede einzelne Anzeige erstellt: mit Anzeigen-ID, Transkription, Zeitstempel, iiif-Bildfragment und Annotator:in-ID. Für die Layouterkennung wurde dhsegment verwendet, um automatisch Bounding Boxes für die Anzeigen zu erstellen (Ares Oliveira/Seguin/Kaplan 2018). Diese Boxes wurden in Transkribus korrigiert und verfeinert, um dort danach die automatische Texterkennung durchzuführen.[4] Für die Texterkennung haben wir zwei Modelle mit einer Zeichenfehlerrate (CER) im Validierungsset von weniger als 1,7 Prozent trainiert. Das resultierende page-xml wurde dann in die Forschungsumgebung Freizo eingespeist, wo es im Mirador-Viewer angezeigt oder als TSV für die weitere Analyse mit R exportiert werden kann. Durch diesen Prozess haben wir Zugriff auf fast eine Million Datensätze, die aus 6.600 Avisblatt-Ausgaben bzw. 48.000 Seiten extrahiert wurden: der komplette Bestand aller im Avisblatt über die Laufzeit von 116 Jahren veröffentlichten Anzeigen.

Wir nutzen R und Github als unsere Data-Science-Umgebung, um zusätzliche Metadaten zu den Datensätzen hinzuzufügen und spezifische Werkzeuge und Methoden für die Analyse zu entwickeln; diese Skripte und die Daten werden später als öffentliches Repository zur Verfügung gestellt.

Da die Anzeigen sowohl extrem zahlreich als auch in ihrer Art sehr unterschiedlich sind, gehört deren Klassifizierung zu den wichtigsten Metadaten, die hinzugefügt werden müssen: Für Untersuchungen beispielsweise des Basler Wohnungs- oder Büchermarkts, von Aktivitäten im Rahmen der Herbstmesse oder von Auktionen als Transaktionsform wird jeweils nur die relevante Teilmenge an Anzeigen als Forschungsgrundlage benötigt.

Im Avisblatt selbst sind die Anzeigen bereits klassifiziert, wenn auch in allgemeiner, pragmatischer und wechselnder Weise: Die Hauptrubriken, unter denen Anzeigen abgedruckt werden, sind "Zum Verkauf wird angetragen", "Zum Ausleihen wird offeriert", "Zu kaufen begehrt", "Kost, Informationen und Bedienung", "Verlorene und gefundene Sachen" und schliesslich die gut durchmischte Rubrik "Allerhand Nachrichten" – und es gibt weitere, die im Laufe der 116 Jahre neu eingeführt wurden oder wieder verschwunden sind. Wir haben die verschiedenen Hauptrubriken über Texterkennung identifiziert und diese als Metadaten für jede Anzeige als erste Klassifizierung aufgenommen – die aber allein nicht ausreicht. Eine vollständige manuelle Klassifizierung aller Datensätze wäre jedoch allein aufgrund ihrer schieren Anzahl nicht durchführbar – deswegen sind Anzeigenblätter bisher kaum als wirtschafts- und konsumhistorische Massenquelle in entsprechender Breite und Tiefe analysiert worden (eine der wenigen Arbeiten dazu ist Homburg 1991).

Stattdessen haben wir die Strategie des algorithmischen Taggings entwickelt. Mit Hilfe von R haben wir eine Klasse von Funktionen definiert ("Tag-Filter"), die jeweils ein positives und ein negatives Dictionary mit regular expressions enthalten, um Anzeigen zu erfassen, die wir entsprechend taggen wollen (z.B. "Kleidung" oder "Mietangebot"). Jede Funktion kann dabei auch auf Anzeigen beschränkt werden, die nur unter bestimmten Überschriften, also Rubriken, erscheinen. Aufgrund des skriptbasierten Ansatzes von R können wir die Metadaten jederzeit aktualisieren, wenn wir Tagfilter hinzufügen oder ändern, sodass hier ein dynamischer statt eines statischen Tagging-Ansatzes zur Anwendung kommt.

Die Vorteile dieser dynamischen, algorithmischen Verschlagwortung sind enorm. Erstens ist sie skalierbar: Anders als bei einem manuellen Tagging ist die Gesamtzahl der zu klassifizierenden Anzeigen praktisch irrelevant; der einzige Unterschied zwischen der Verschlagwortung von einigen hundert oder einigen hunderttausend Anzeigen sind ein paar Minuten Rechenzeit. Zweitens sind die Ergebnisse, anders als bei den Entscheidungen, die einer manuellen Klassifizierung zugrundeliegen, vollständig reproduzierbar und immer eindeutig nachvollziehbar. Drittens ist der Ansatz extrem flexibel: Anstatt ex ante eine feste Klassifikation der Datensätze zu erstellen, die dann die Analyse vorgibt

und einschränkt (d.h. was gezählt oder zusammengefasst werden kann und was nicht), können die Klassifikationen angepasst und weiterentwickelt werden, wenn sich im Analyseprozess neue Erkenntnisse und Ideen ergeben. Jede:r Forscher:in, der:die mit den Daten arbeitet, kann seine:ihre eigene Klassifizierung in einem überschaubaren Zeitrahmen erstellen.

Obwohl die Vorteile überwiegen, gibt es auch einen Nachteil des algorithmischen Ansatzes: Im Gegensatz zur manuellen Klassifizierung und zur Verwendung der Rubrikeninformation aus der Quelle selbst produziert er zwangsläufig einige Klassifizierungsfehler, sowohl false positives (Anzeigen, die fälschlicherweise als einem Tag zugehörig getaggt sind) als auch false negatives (Anzeigen, die fälschlicherweise nicht als einem Tag zugehörig getaggt sind). Die Optimierung eines Tagfilters zur Vermeidung von false positives führt zu einem übermässig konservativen und vorsichtigen Filter, der die Erkennungsrate absenkt (und mehr false negatives produziert). Die Optimierung zur Vermeidung von false negatives führt wiederum dazu, dass er zu viel einbezieht und mehr Beifang (d.h. false positives) produziert.

Eine Strategie, damit umzugehen, wäre, einzelne fehlende Datensätze manuell ein- und falsch markierte manuell auszuschliessen. Dies vermindert allerdings nicht nur die Reproduzierbarkeit, auch der Zeit- und Arbeitsaufwand steigt mit der Anzahl der zu prüfenden Datensätze. Da dies unter Umständen in Einzelfällen dennoch sinnvoll sein kann, haben wir die Möglichkeit ebenfalls in unseren Tagfilter-Skripten implementiert.

Als guter Weg, um sowohl false negatives als auch false positives zu minimieren, hat sich ein Bottom-up-Ansatz für den Aufbau und die Kombination von Tagfiltern bewährt: Anstatt einen Filter aufzubauen, der eine grosse und allgemeine Gruppe von Anzeigen umfasst, wie z.B. alle Anzeigen, die sich auf materielle Objekte beziehen, oder nur all diejenigen, die Möbel annoncieren, bauen wir viele Filter, die jeweils einen recht engen Anwendungsbereich haben (wie z.B. Einzelfilter für Betten, Schränke, Stühle, Tische usw.) und kombinieren die resultierende Vielzahl von Tags unter Oberbegriffen (wie z.B. “Möbel”). Filter mit einem engen Anwendungsbereich sind in der Regel selektiver, d.h. sie haben gleichzeitig eine niedrige Fehlerquote für false negatives und false positives.

Dieser Bottom-up-Ansatz ist sehr stark auf spezifische Forschungsfragen und -interessen der Beteiligten ausgerichtet und ignoriert alle Anzeigen und potenziellen Gruppierungen von Anzeigen ausserhalb des eigenen Fokus – genau das aber ist beabsichtigt, da jedes Forschungsinteresse durch eine eigene Klassifikation bedient werden kann. Dies unterscheidet sich von einer einzigen, a priori gesetzten und manuellen one-serves-all Klassifikation, die universeller und interoperabler sein und normalerweise die Gesamtheit aller Anzeigen vollständig aufteilen muss, d.h. jede einzelne Anzeige in eine (allgemein gültige, vorgängig festgelegte) Kategorie einordnen muss – und wenn dies nicht der Fall ist, schränkt dies die Nutzbarkeit der Datenbank über ein spezifisches Forschungsprogramm hinaus ein.

In den einzelnen Forschungsprojekten innerhalb des Gesamtprojekts dient das dynamische Tagging verschiedenen methodischen Zwecken und zielt auf unterschiedliche Aspekte:[5] So gibt es (1) eine Reihe von Tagfiltern, die Anzeigen nach ihrem Thema klassifizieren – auf einer allgemeinen Ebene, indem sie (zum Beispiel) den Arbeitsmarkt, den Wohnungsmarkt oder den Tausch von Gegenständen voneinander unterscheiden; oder spezifischer, indem sie verschiedene Arten von Jobs oder verschiedene Kategorien von Dingen klassifizieren. Dann gibt es (2) Tagfilter, die unterschiedliche Arten von Austausch und austauschbezogenen Absichten identifizieren: Sie unterscheiden zwischen dem Verkauf, dem Verleih und der Rückgabe von verlorenen (oder gestohlenen) Gegenständen oder zwischen Angeboten und Gesuchen – einige dieser Informationen sind bereits in den quelleneigenen Überschriften enthalten, allerdings nicht immer zuverlässig bzw. standardisiert. Genauer gesagt erkennen solche Tagfilter verschiedene Formen von Transaktions- oder Marketingpraktiken wie Auktionen, Warenlotterien, Wettbewerbe usw. Sie unterscheiden auch Kleinanzeigen, die auf eine einzelne Transaktion abzielen, von kommerziellen Werbeanzeigen für Unternehmen und Dienstleistungen, die das Ziel hatten, regelmässige zukünftige Geschäfte anzubahnen. Schliesslich gibt es (3) Tagfilter, die bestimmte Umstände prüfen: Ist die Anzeige beispielsweise anonym oder wird der:die Inserent:in genannt, der:die sie ins Avisblatt gestellt hat? Handelt es sich um eine bestimmte Personengruppe – wie z.B. Witwen, die einer Vielzahl von wirtschaftlichen Aktivitäten nachgehen –, oder werden Orte erwähnt (die dann für die Verwendung in GIS-Ansätzen extrahiert werden können)?

Methodisch dient die Identifikation relevanter Teilmengen von Datensätzen als Basis für sehr unterschiedliche Forschungsansätze, die natürlich auch kombiniert werden können und werden: Eine bestimmte Menge relevanter Anzeigen kann einzeln gelesen und hermeneutisch analysiert werden (wenn sie auf einige hundert Datensätze beschränkt ist), sie ist offen für natural language processing oder auch für statistische Analysen.

Eine derartige iterative Analyse kann zu neuen Forschungsfragen führen und neue Forschungsrichtungen aufzeigen. Die Tatsache, dass es dabei mehrere Stränge gibt, mehr als eine Herangehensweise und einen Blickwinkel, um die Quelle zu nutzen, schafft eine zusätzliche Dynamik: Tagfilter und andere Werkzeuge können wiederverwendet und zwischen verschiedenen Forschungsprozessen und Projekten ausgetauscht werden, wodurch die unterschiedlichen Forschungsstränge miteinander verwoben werden. Auch ist eine Weiterverwendung der entwickelten Filter und Analysefunktionen auf weitere Publikationen denkbar – Anzeigenblätter kamen im 18. Jahrhundert in ganz Europa auf und waren gerade auch im deutschsprachigen Raum weit verbreitet; hier finden sich bereits digitalisierte Bestände für viele lokale Intelligenzzeitungen, deren weitere Aufbereitung und anschließende Analyse sich an den für das Avisblatt entwickelten Erschließungsschritten orientieren könnte.

Das Avisblatt-Projekt wird weder von einer spezifischen Forschungsfrage angetrieben, die einen bestimmten und speziell zugeschnittenen Datensatz zur Beantwortung erfordert, noch ist es ein Projekt, das sich auf die Erstellung der Edition einer seriellen Quelle ohne spezifische Forschungsanwendung beschränkt. Letzteres liefe Gefahr, einen Datenfriedhof zu produzieren, ersteres, Single-Use-Datensätze zu produzieren (die später dann zu Datenfriedhöfen werden, nachdem sie verwendet wurden). Indem wir einen Rahmen für dynamisches Tagging und eine skriptbasierte Verarbeitung bereitstellen, fügen wir der digitalisierten Quelle nicht nur nützliche Metadaten hinzu, sondern hoffen, eine gute Grundlage für weitere Forschung und Nachnutzung zu schaffen, mithilfe derer sich neue iterative Analysen entfalten können.

Fußnoten

1. https://avisblatt.ch/.
2. Die Digitalisate sollen in einem nationalen Repositorium bereitgestellt werden; das aktuell noch private Git-Repository, in dem sich die vollständige Datengrundlage und der Code zur Analyse befinden, wird geöffnet.
3. https://www.data-futures.org/.
4. https://readcoop.eu/transkribus.

5. Das Folgende betrifft Tagfilter, die auf die Anzeigen angewendet werden, aber das Konzept besitzt auch eine grundlegendere Anwendung: Der Wortlaut der Überschriften der einzelnen Rubriken ändert sich über die Jahre, sodass ein spezifisches Tagfilter-Set nur für die Klassifizierung der jeweiligen Überschriften erstellt wurde.

Bibliographie

Ares Oliveira, Sofia / Seguin, Benoît / Kaplan, Frédéric (2018): “dhSegment: A generic deep-learning approach for document segmentation”, in: *Frontiers in Handwriting Recognition (ICFHR)*. 16th International Conference 2018: 7–12.

Blome, Astrid (2006): "Vom Adressbüro zum Intelligenzblatt. Ein Beitrag zur Genese der Wissensgesellschaft", in: *Jahrbuch für Kommunikationsgeschichte* 8: 3–29.

Blondé, Bruno / Van Damme, Ilja (2018): “From Consumer Revolution to Mass Market”, in: *The Routledge Companion to the History of Retailing*. New York: Routledge: 31–49.

de Vries, Jan (2012): *The Industrious Revolution. Consumer Behavior and the Household Economy, 1650 to the Present*. Cambridge: Cambridge University Press.

Homburg, Heidrun (1991): "Warenanzeigen und Kundenwerbungen in den ‚Leipziger Zeitungen‘ 1750–1800. Aspekte der inneren Marktbildung und der Kommerzialisierung des Alltaglebens», in: Dietmar Petzina (Hg.): *Zur Geschichte der Ökonomie und Privathaushalte*, Berlin: Duncker & Humblot: 109–131.

McKendrick, Neil / Brewer, John / Plumb, John H. (1982): *The Birth of a Consumer Society. The Commercialization of Eighteenth-Century England*. London: Europa Publ. Ltd.

Mendels, Franklin F. (1972): "Proto-industrialization. The First Phase of the Industrialization Process”, in: *The Journal of Economic History* 32:1: 241–261.

Mui, Hoh-Cheung / Holbrook Mui, Lorna (1989): *Shops and Shopkeeping in Eighteenth-Century England*. Kingston / Montreal / London: McGill Queen's University Press Routledge.

Tantner, Anton (2015): *Die ersten Suchmaschinen. Adressbüros, Fragämter, Intelligenzcomptoirs*. Berlin: Verlag Klaus Wagenbach.