Digital Pathways Through Newspaper Advertisements: Workflows from Printed Page to Digital Analysis with the Avisblatt-R-Package

Dickmann, Lars
lars.dickmann@unibas.ch
University of Basel, Switzerland

Reimann, Anna
anna.reimann@unibas.ch
University of Basel, Switzerland

Serif, Ina
ina.serif@unibas.ch
University of Basel, Switzerland

Abstract

This half-day workshop introduces researchers to workflows of and tools developed in the research project ‘Printed Markets’ to analyse historical newspapers,[1] achieved in collaboration between historians, computer and data scientists. Thanks to using mostly open source or easily accessible software and an R package especially developed to enrich and analyse newspaper data, the approach is flexible and can, in theory, be reproduced with and adapted to any similar source and for different research interests.

Introduction

While historical newspapers and other periodical publications have long been neglected for humanities research, recent digitization efforts have begun to mitigate this and allow for better corpus building. The access to these serial sources has been exponentially broadened, but with digital representations still often lacking text recognition or segmentation, let alone further classification of the content to allow for quantitative analysis, in-depth analysis is often difficult without any further processing of the digitized newspapers. Even if, in the best-case scenario, text recognition and segmentation is provided, tools for historical digital analysis are mostly lacking, and general tools are often difficult to adapt for premodern sources which pose a range of problems like noisy OCR, non-standardized spelling and non-modern vocabularies. While new database projects have indeed taken impressive steps to enrich their data and thus facilitate research, these are limited to the newspapers that fit into their data model and rely on a specific interface, which is not adaptable for other research interests and sources.[2]

Historical context

Our research object, the historical source, is not a standard political newspaper, but the *Avisblatt*, an intelligence paper published between 1729 and 1844 in Basel, Switzerland. While our example is an especially well preserved paper, it is by far not the only one of its kind: at the end of the seventeenth century, so-called intelligence newspapers emerged in big European cities like Paris or London, springing up all over the continent during the eighteenth century.[3] They mainly consisted of (classified) advertisements and tried to connect people offering and people seeking something – any kind of thing: things for rent and for sale, lost and found items or animals, second-hand goods, newly invented or well-known medical products, and imported goods like coffee and tea, to name just a few. These new platforms for facilitated communication were soon frequently used by non- or semi-professional sellers, as well as professional suppliers such as craftsmen, traders and shops. Therefore, these intelligencers are, on the one hand, a particularly interesting source for examining the micromechanics of local markets, on the other, helpful for analyzing connections between local, transregional and increasingly global markets of goods in the early modern period – and they are also quite unmanageable for a single (and analog) research endeavor with regards to the sheer mass of data, making collaboration and the use of digital tools a necessary prerequisite for its investigation.

Case study

Our specific case study, the *Avisblatt*, published almost 6000 issues during its 116-year runtime, which consist of around 50 000 pages and contain close to 1 million advertisements. In view of these figures for a *single* intelligence paper, it is perhaps not surprising, that these sources have so far often been neglected or studied only qualitatively with very small-scale interests.[4] The workflows and tools we present during this workshop are a first step to open up the treasure trove of intelligence newspapers to historical research, starting at the digitization of the source and leading to a database of classified ads that can be constantly expanded and enriched. From the start, this approach was also developed with the aim to make it adaptable to other similar publications, allowing for regional as well as trans-national comparisons;[5] the steps that we performed can be done using open source software only, and the data set, the R package and scripts we created are available on GitHub.[6]

At its core, the R package contains different scripts that result in an automated classification of the ads. This is a process that we call “dynamic tagging”, which classifies the single advertisements on basis of the text contained within. Currently, there exist nearly 200 dictionaries for different categories. These dictionaries try to catch all spellings for items that form a category, e.g. “dog”, which could be “poodle”, “boxer”, or “sausage dog”; to broaden the catches, regular expressions can be used. If a term stored in a dictionary, e.g. “poodle” in the dictionary “dogs”, is found in the text of an ad, the tag “dogs” is added to the meta data. This approach makes it possible for individual researchers to focus on their research topic within the advertisements, by filtering thematic subsets, and enriching the data further, contributing to the increasing comprehensiveness of the data set as a whole: If somebody were interested in animals, they could create a subcollection with all the ads tagged with “dogs”, “cats”, etc., and concentrate their analysis on this subcollection. They can also just care for cats, of course. And if somebody else were interested in the appearance of mice in the ads, they could create a new dictionary that runs over the corpus and retags all the ads containing corresponding terms. In the resulting subcollection, they could further enhance the data, looking for places that appear in ads mentioning mice, creating a topography of rodents in Basel.

Workshop content

The workshop will demonstrate how the workflow that we developed for the text-based digital classification of an early modern advertisement newspaper allows for subsequent and dynamic classification of the single ads according to the researcher's interest – of the data set as a whole or of selected subcollections. It teaches how to gather the data from GitHub and how to use different functions from our package for analysis and for visualization of results. It also shows how this approach could be applied to similar publications, using our R package.

While participants will need to install some open source software (see below), prior programming experience is not required.

Target audience

Researchers from all humanities disciplines that are interested in (historical) newspapers. While the newspaper from our project is in German, we will also briefly show an English example (e.g. The publick advertiser, 1650s). The maximum number of participants is 12.

Prerequisites

Installation of R, R Studio, setting up a GitHub account. We recommend installing GitHub Desktop if you are unfamiliar with git and/or using the command line.[7]

Outline

This is a half-day workshop which will cover the following content:

- 30' R and R Studio/trouble shooting
- 30' short overview over project; which steps were performed from analog source to digital data; short Q&A for technical questions
- 30'–60' showcasing of classification/tagging of ads:
  - we explain the meta data model that we built for the ads;
  - we explain the principle of a dictionary, how it is constructed, which elements it is made of;
  - we show how one can filter specific subsets, according to different kind of meta data;
  - we show how one can enhance a subset with other meta data
- 60'-90' group tagging, filtering, enhancing, plotting
  - we build a dictionary together that shall tag all advertisements that contain a good/object chosen by the group;
  - we let the dictionary run over the data set to tag the corresponding ads with the meta data tag;
  - we filter for the created subset, make first visualizations, enhance and or filter the data further

Acknowledgements

The presented package was developed within the project "Printed Markets", SNF grant 182156, with support from Matthias Bannert.

Instructors

Lars Dickmann, lars.dickmann@unibas.ch, is research assistant for Early Modern History at the University of Basel with a focus on Global History.

Anna Reimann, anna.reimann@unibas.ch, is a PhD student in the project "Printed Markets" at the University of Basel and interested in Consumption History and Material Culture.

Ina Serif, ina.serif@unibas.ch, is a PostDoc assistant for Premodern and Digital History at the University of Basel and has a special interest in Book and Print History.

Notes

1. "Printed Markets. The Basel Avisblatt 1729–1844", a project at the Department of History of the University of Basel, financed by the Swiss National Science Foundation, running from 2018 until spring 2023.
2. See for example https://impresso-project.ch/, which enables research through an easily accessible interface, but with restrictions regarding export of publications. The Horizon 2020 project newseye (https://www.newseye.eu/) offers a platform for research as well as code for reuse, but does not seem to be too versatile when it comes to non-standardized languages.
3. See for example Blome 2006; Golob 2012; Lyna / Van Damme 2009; Tantner 2015.
4. See for example Brauner 2019; Fleischmann-Heck 2019; Jones 1996.
5. A comprehensive database with an overview over intelligencers is still missing; however, they not only appeared on the European continent in a high number of issues, but also in Northern America.
6. https://avisblatt.github.io/.
7. For the installation of R (https://www.r-project.org/) and R Studio (https://posit.co/products/open-source/rstudio/), chapter 1.1 in Ismay / Kim (2023) might be useful: https://moderndive.netlify.app/1-getting-started.html#getting-started. You can set up a GitHub account under https://github.com/, and you can install GitHub Desktop via https://desktop.github.com/.

Bibliography

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

Brauner, Christina (2019): “Recommendation und Reklame. Niederrheinische Brandspritzenmacher und Praktiken der Werbung in der Frühen Neuzeit”, in: *Zeitschrift für Historische Forschung* 46: 1–45.

Fleischmann-Heck, Isa (2019): “The ‘Duisburger Intelligenz-Zettel’ as a Source for Textile Research. Supply and Consumption of Silk and Cotton Textiles in Western Prussia in the Second Half of the Eighteenth Century”, in: Siebenhüner, Kim / Jordan, John / Schopf, Gabi (eds.): *Cotton in Context. Manufacturing, Marketing, and Consuming Textiles in the German-Speaking World (1500–1900)*. Wien / Köln / Weimar: Vandenhoeck & Ruprecht 335–55.

Golob, Andreas (2012): “Das Zeitungskomptoir als Informationsdrehscheibe. Michael Hermann Ambros und seine Grazer Anzeigenblätter”, in: Brandstetter, Thomas / Hübel, Thomas / Tantner, Anton (eds.): *Vor Google. Eine Mediengeschichte der Suchmaschine im analogen Zeitalter*. Bielefeld: transcript 109–50.

Ismay, Chester / **Kim, Albert Y.** (2023): *Statistical Inference via Data Science. A ModernDive into R and the Tidyverse*, https://moderndive.com/ [28.04.2023].

Jones, Colin (1996): “The Great Chain of Buying. Medical Advertisement, the Bourgeois Public Sphere, and the Origins of the French Revolution”, in: The American Historical Review 101, 1: 13–40.

Lyna, Dries / **Ilja Van Damme** (2009): “A Strategy of Seduction? The Role of Commercial Advertisements in the Eighteenth-Century Retailing Business of Antwerp”, in: *Business History* 51, 1: 100–121.

Tantner, Anton (2015): *Die ersten Suchmaschinen. Adressbüros, Fragämter, Intelligenz-Comptoirs*. Berlin: Wagenbach.