Will Work for Data!

Challenges to Text & Data Mining in Cultural Heritage Research

Kristoffer L. Nielbo
University of Southern Denmark, Department of History
knielbo@sdu.dk — knielbo.github.io

Melvin Wevers
KNAW Humanities Cluster, Digital Humanities Group
melvinwevers@gmail.com — melvinwevers.github.io

Abstract

Due to the "data deluge" we are currently witnessing, *Text and Data Mining* (TDM) has become an intricate part of many research areas. Cultural heritage research and the humanities in general are experiencing this change. Concepts such as 'distant reading', 'culture analytics', and 'humanities at scale' are showing up across all fields of the humanities. TDM, however, is a complicated multi-step procedure that presents a number of challenges to researchers who traditionally adhered to the qualitative analysis of a small data set. We identify four interrelated challenges that have impacted our research in different degrees (technical competencies, interdisciplinary respect and understanding, epistemology differences, data access and mobility) and exemplify them through a use case. It turns out that access to data is by far the greatest challenge to TDM in cultural heritage research.

Introduction

The growth in data is accelerating at an unprecedented pace and researchers are forced to find new and innovative ways of managing this data deluge in their respective fields. eScience infrastructures are being made available at all universities, massive data management schemes are being implemented at all levels.

Research in cultural heritage, and the humanities in general, is also confronted with ways to deal with digitized data. This is evinced by an increasing number of papers are dealing with issues related to distant reading, culture analytics, and humanities at scale. These papers typically present research, which utilizes Text and Data Mining (TDM) techniques to find new meaningful patterns in textual cultural heritage data. TDM is, a complicated multi-step procedure - which involves querying databases, sampling data sets, preprocessing and normalization of data, statistical modelling and, finally, validation - before useful knowledge can be derived from the results. This procedure, therefore, requires access to relevant data, understanding of basic mathematical and statistical concepts, and interdisciplinary collaboration. In our own research, we have experienced four interrelated challenges that complicate the application and future development of TDM in cultural heritage research and, more generally, humanities research:

1. Technical competencies
2. Interdisciplinary respect and understanding
3. Epistemology differences
4. Data access and mobility

Just as data scientists and curators have become a scarce and valued commodity outside the humanities, researchers with competencies in programming and statistical analysis have much sought after in cultural heritage research. Institutional and infrastructural initiatives have partly solved this bottleneck in the humanities. However, the dependency on new technical competencies, which have traditionally not been part of the humanities curriculum, have generated scepticism and disputes. At the basis of this conflict, one often finds a lack of interdisciplinary respect and understanding, which can only be overcome through dialogue and successful collaboration ('examplars').

Still, there are researchers who insist that the distinctive nature of the humanities is merely that of hermeneutic examination and idiographic research. These researchers regularly view the quantitative and data-driven approach of TDM as a return to positivist fallacies of the past. Proponents of this perspective tend to ignore the methodological plurality of TDM as well as the intricate relationship between data, method, and analysis in what has also been called data-intensive research. At the same time, experts in the field of TDM can also benefit from the methodological and epistemological insights from the humanities. As the amount of digitized humanities data will certainly increase rapidly in the future, it is, therefore, adamant that researchers from different fields strike interdisciplinary collaborations.

Finally, the last challenge, which to us is by far the greatest, is the lack of access to and mobility of large-scale data collections due to copyright legislations, data regulations, and propensity to construct institutional silos. Since the benefits of TDM depend on having relevant data at the right time and place, we envision that researchers need to communicate closely with data-providers and libraries to get access to a target data set. Even if that means not adhering to institutional guidelines.

In what follows, we will exemplify these four challenges through a use-case that relates to large-scale analysis of Dutch and Danish newspaper articles.

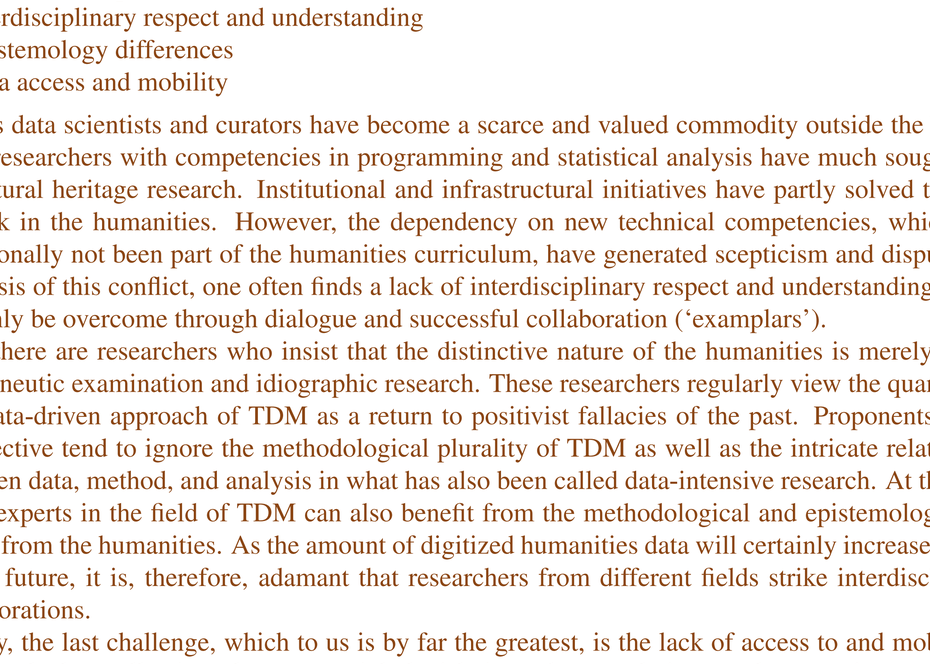

Figure 1: Variation in detrended *Brood* (bread) plotted at biannual intervals in advertisements and the newspapers *De Tijd* and *De Telegraaf* (grey band represents 95% confidence intervals). Lower row: Illustrates the estimation of H-exponent (measure of fractal dynamic) as the slope of the best fit of variance of the residuals, $F(w)$, on multiple time scale (i.e., window size), w on double log scale.

Use Case

Digitization of newspapers offers a valuable source of exploration and validation for cultural research. In a series of studies, we explore the dynamics of public discourse by the use of digitized newspapers (Wevers 2017; Wevers & Nielbo submitted). Accessing and working with Dutch and, in part, Danish digitized newspapers has, however, resulted in multiple instances of the above-mentioned four challenges.

Materials and Methods

We have worked on a sample of Dutch newspapers from the 20th century consisting of advertisements (n = 18,564,411), articles from *De Tijd* (n = 3,806,083) and articles from *De Telegraaf* (n = 7,659,137). Our current project will extend the Dutch sample to the 19th century and merge it with a matching sample from Danish newspapers.

In terms of methods, we combine multiple techniques from information retrieval and natural language processing (e.g., concept extraction, word associations, and topic modelling) with techniques for time series analysis (e.g., Granger causality, multi-fractal analysis, change point detection, dynamic time warping) in order to track fundamental time-dependent properties of public discourse.

Results and Future Research

On the Dutch newspaper data set, we showed that during the 20th century, articles, taken as a proxy for public discourse, exhibit a different scaling dynamic than advertisements, which reflect an inherent difference in public and commercial discourse (Figure 1). Furthermore, we have identified commercial product groups that tend to shape newspaper article content and other groups that, together with articles, reflect societal issues. Currently, we are working on a method for detecting the persistence of topical change points in the last 200 years of Danish and Dutch newspapers (Figure 2).

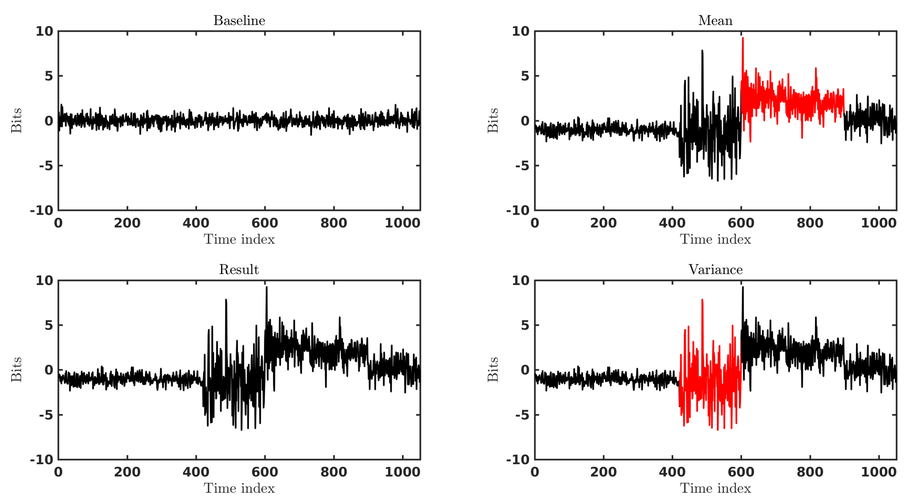

Figure 2: Method for detection of topical change points in temporally ordered documents (simulation data). Each document is represented as a distribution over latent topics, $\theta_{d=1...M}$, and the relative distance between adjacent documents is then plotted, $D_{KL}(\theta_i \mid\mid \theta_{i+1})$. Left column plot a simulation of a no change (upper) vs. change scenario (lower). Right column highlights a region with a change in mean (upper) and change in variance (lower).

Discussion

While lack of technical competencies has been and still is an issue in textual cultural heritage applications of TDM, we have generally found much support from colleagues in mathematics and computer science. This has been made possible by more formalized institutional infrastructures (e.g., Culture Analytics @ UCLA's Institute of Pure and Applied Mathematics), which facilitate collaboration across disciplinary boundaries. Collaboration with new (and strange) bed-fellows does, however, require interdisciplinary respect and understanding from all the involved parties. In our case, we have build this through research network activities (e.g., Calculus of Culture @ AU's Interacting Minds Center), which develop a common language in order to transcend disciplines. Today, many TDM-related workshops and summer schools exists that foster interdisciplinary respect and understanding between researchers very diverse disciplines. Still, we experience both academic indifference and, at occasions, animosity when we present this project to fellow humanities researchers. While the results are interesting, they belong to a positivistic and, it is argued, scientistic discourse that undermine humanities' normative purpose.

But the greatest challenge for us and the humanities at large, is data access and mobility. While other research fields have standards for managing issues related to data protection laws, these issues are new in the humanities. Large data collections (e.g., internet archives, social media) are not de facto available for research due to the lack of a standard procedure for accessing them within a reasonable time frame. Furthermore, textual cultural heritage are often also subject to copyright restrictions, which moves researchers' point of entry back 70-100 years. For newspaper data, copyright has presented multiple complications with respect to Danish data, while the Netherlands have a more flexible research data management policy. There are also socio-cultural hindrances to data access and mobility. Archives and libraries, like other institutions and companies, are protective of their commodities, which can result in data silos. Silos are highly problematic for researchers who want to run their analysis on the most efficient hardware. This is especially problematic when the hardware is HPC clusters that are a varied and scarce resource. Finally, it is our impression that humanities scholars, contrary to what one might think, are somewhat opposed to open science. There is a culture in the humanities that is quite protective of data and not particularly willing to share. This culture is highly problematic for the reliability and future development of the humanities.