Indigenous Studies in the Digital Era:

Structural Topic Modeling and Historical Indigenous Newspapers

Master's Thesis of Eva Matzel (evmatzel@uni-mainz.de); M.A. Digitale Methodik in den Geistes– und Kulturwissenschaften, Universität Mainz; Advisors: Prof. Dr. Scheiding, Dr. Andreas Wagner

"There is still a long way to go…"

Introduction

This motto applies to my research on many levels:

- While the field of periodical studies is well-established among scholars, the field of Indigenous periodical studies has long been neglected and merits more attention still;
- The same is true for the digitization process: while many libraries and archives have started to digitize their collections in the last decades, publications of ethnic minorities and marginalized groups are often forgotten or categorized as not important enough — or even worse: many got lost before they could have been archived;
- Moreover, although the number of digitized newspapers and periodicals has been rising continuously – by now, millions of pages have been made available – the development of user-friendly digital methods and tools to analyze such immense data is still in the beginnings.

Altogether, this is how I came up with the subject of my master's thesis.

Master's Thesis

In bridging Indigenous periodical studies and user-friendly digital research methods, I plan on showing Humanities scholars at my university how to shape future research more efficiently. I want to introduce some of the methods introduced by the Digital Humanities: I am going to analyze the *American Indian Magazine* by using the digital method of Structural Topic Modeling in R.

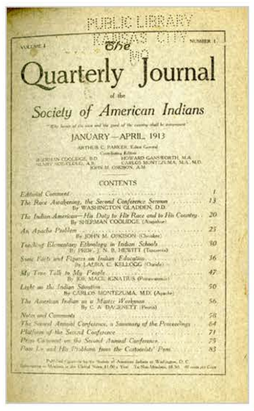
In the end, I want to find out:

- What the key topics of the various issues are;
- Whether the topics change over the course of issues;
- Whether there are relations between the topics of the various issues.

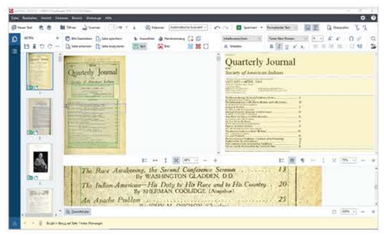
Using my approach, it should be possible to getting these answers without having read the magazine in person.

Optical Character Recognition (OCR)

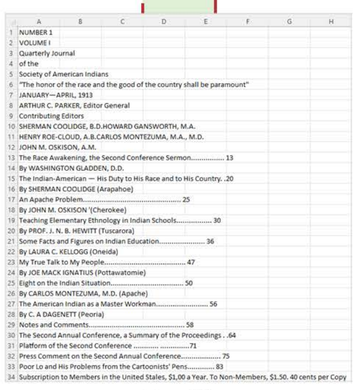
The digitized version of the *American Indian Magazine* can be found on the website of the "American Indian Digital History Project." On there, all issues are available for download as PDF files. Since the pages of the pdf files are scans/ photos, the program R is not able to recognize the text on them. Hence, the pdf files have to be converted into a machine-readable format, for example CSV or Excel files. This process is called Optical Character Recognition.



PDF file



OCR in ABBY FineReader

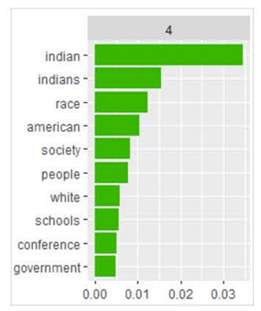


CSV file

```
library(reshape2)
td_beta <- tidy(topicmodel1)

td_beta %>%
  group_by(topic) %>%
  top_n(10) %>% #, beta)
  ungroup() %>%
 #(topic = paste0("Topic ", topic),
  mutate(term = reorder(term, beta)) %>% #_within(term,
  ggplot(aes(term, beta, fill = as.factor(topic))) +
  geom_col(show.legend = FALSE) + #alpha = 0.8,
  facet_wrap(~ topic, scales = "free") +
  coord_flip() #+
```

STM in R (excerpt code)



Visualization

However, there are also some problems and risks when using OCR. For example:

1. There are only a few open-source software available. In addition, those are not user-friendly for researchers without prior IT knowledge. However, an example for a fee-based but user-friendly software is the "ABBY FineReader Pdf 15 OCR Editor." Thankfully, the Johannes Gutenberg-University provides this application to its students for free.
2. The low scan quality of some historical newspaper pages results in a poor OCR output. Thus, it is necessary to manually post-correct the results. Depending on how low the scan quality is, this step can be more or less time-consuming.
3. The OCR software divides the individual articles of one page into separate segments. Unfortunately, it does not recognize if an article continues on the next page or even in another issue. This carries the risk of distorting the results of the subsequent analysis in R.

Structural Topic Modeling (STM)

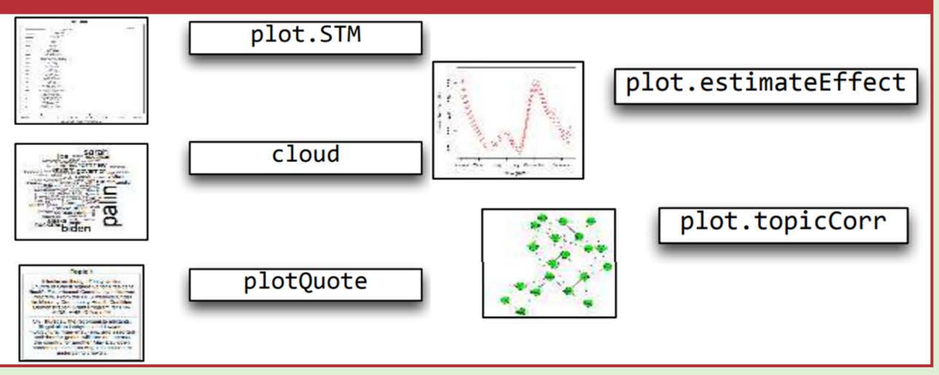
There are numerous digital methods to analyze text data. However, one approach is particularly interesting: the Structural Topic Modeling. It is a topic modeling framework included in R. Its key advantage to other topic modeling frameworks: the ability to integrate document metadata into the topic modeling process. In the end, it enables the researcher to discover topics within documents as well as their relations to the documents' metadata — without having to read them in person.

Apart from the ability to incorporate metadata, there are further advantages to the use of Structural Topic Modeling:

1. R is an open-source environment. Consequently, all integrated packages are free of charge, too. This makes the stm package accessible for any researcher worldwide.
2. Once the researcher has become acquainted with the coding language of R, it facilitates the workflow associated with text analysis. The process becomes less time-consuming and thus much more efficient.
3. The numerous visualization opportunities of the stm package help illustrate the results of the topic modeling. This may improve the researcher's understanding of the results. (Examples of visualization methods: bar graphs, word clouds, correlation models, various graphics, etc.)

Preliminary Outcomes & Outlook

Until now, I have only analyzed one issue of the *American Indian Magazine*: Vol. 1, No. 1. The OCR process itself by the ABBY FineReader only took a few minutes. However, the post-editing of the OCR results took several hours because of the varying quality of the pdf scans. Afterwards, I wrote the script for the Structural Topic Model. By now, I have only written the code for one visualization method: the bar graph. In the upcoming weeks, I want to plot the results as depicted in the examples alongside. Since I have only processed one document so far, the results are not significant, yet. The more documents are fed into the Structural Topic Model, the more significant the results are going to get.



JG|U

GUTENBERG GRADUATE SCHOOL of the Humanities and Social Sciences

OBAMA INSTITUTE

References

"American Indian Digital History Project." aidhp.com. Accessed 20 Sept. 2022.; Hammill, Faye, et al. "Introduction: Magazines and/as Media: Periodical Studies and the Question of Disciplinarity." In: The Journal of Modern Periodical Studies 6.2 (2015): iii-xiii.; Priewe, Marc. "Transnationale Printkultur des 19. Jahrhunderts im digitalen Raum: Die Untersuchung von Zeitungen als Daten." *Handbuch Zeitschriftenforschung*, edited by Oliver Scheiding, and Sabrina Fazil, [transcript], tba.; Riedl, Martin, et al. "Clustering-Based Article Identification in Historical Newspapers." Universität Stuttgart.; Roberts, Margaret E., et al. "stm: R Package for Structural Topic Models." *Journal of Statistical Software*, vol. 1, no. 2, Oct. 2019, DOI: 10.18637/jss.v091.i02. Accessed 21 Sept. 2022; "What is R?" www.r-project.org/about.html. Accessed 21 Sept. 2022.