Progress Report

The project of our group focuses on reproducing the results of the paper “A cross-collection mixture model for comparative text mining” by Zhai and his co-workers.

In the paper, Zhai used his model to examine two sets of news data, one on the Iraq War and the other on the Afghanistan war. Specifically, the Iraq war news excerpts were a combination of 30 articles from the CNN, BBC websites over 2013-2014. The Afghanistan war data consists of 26 news articles downloaded from the CNN and BBC websites for one year starting from Nov. 2001. To examine the robustness of the model, we decided to use different news data for our project. The Covid-19 pandemic is an event that we believe is both appropriate and comparable. With the help of our work and experiences from MP2, we extracted 35 articles from CNN, BBC, and HUFFPOST websites to form our dataset.

We have also begun to input the modeling formulas into our scripts. The formulas involve a large number of variables being implemented, which caused us some problems. We also found some of the formulas hard to understand, but we believe we can fix this problem by looking through the literature more.

Our next steps would be to finish the modeling and run the dataset. Plotting and analyzing the data may involve the usage of R Markdown. After analyzing the data, we will check if our results correspond to the conclusions from the paper.

To view our work, please take a look at the "Progress Report" folder.