Interpreting Historical Data Using Text-Mining Extraction

Andrew Tran, Pavan Govu, Pax Gole, Sophie Horner and Dr. Karen Mazidi

The University of Texas at Dallas

Introduction

The Civil War era was a contentious time period in American history, and even today, researchers are still learning about the war. Our project, Interpreting Historical Data Using Text-Mining Extraction, examined the Civil War era from a different angle - through the eyes of every day people from the South. We sought to learn more about civilians' perspectives during the Civil War era in America by creating programs that could be used as tools to interpret historical data and reach conclusions about the Civil War era.

Corpus

The data used came from 'First Person Narratives' in the 'Documenting the American South' database from the University of North Carolina. The corpus, with a total of 150 texts, is comprised of:

- Letters & Personal Correspondence
- Memoirs
- Biographies and Autobiographies
- Diaries

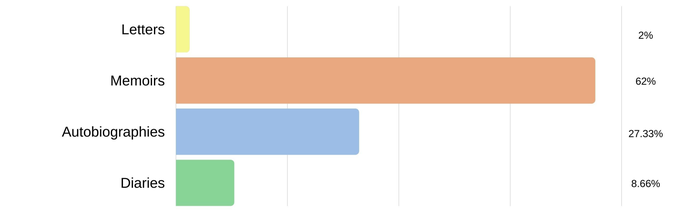

Fig. 1: Various types of texts within the corpus.

The documents were annotated by the research team, as well as our lead and outside volunteers. We created annotations for:

- Polarity
- Topics
- Subjectivity
- Demographics

The curators of the database describe this narrative corpus as diverse in its perspectives, which made it ideal for NLP analysis of the Civil War era.

Extracted Scores

We extracted 4 distinct scores from each document:

- Flesch-Kincaid reading ease
- N-gram frequency
- Topic modeling
- Polarity and subjectivity

Flesch-Kincaid Reading Ease measures how easy it is to read a text. Using the given formula, one can use this score as a metric for comprehensibility.

Topic modeling detects word and phrase patterns within texts, and automatically clusters word groups and similar expressions that best characterize a document.

N-grams measure recurring multi-word sequences in texts. These can pair with topic modeling to better understand and interpret topics that are frequently discussed in a text.

Polarity and subjectivity are sentiment analysis scores. Polarity involves the positivity or negativity of a text, and subjectivity determines whether a text is more opinionated or factual.

Categorical Data Analysis

Categorical data analysis was important for summarizing the overall content of the texts without manual reading. Our analyses used n-gram frequencies and topic modeling to analyze what the authors wrote about, what wording they used, and how topics differed across author demographics and circumstances. We found that:

- The main ideas found in every text were war, politics, and domestic life.
- Women wrote about war less often, and when they did, their texts correlated with the words "hospital" and "news," rather than words such as "cavalry" or "general."
- Texts written by preachers (or others discussing religion) associated war with words such as "sorrow" and "beauty".
- Black authors wrote about religion and education more often than the general population.

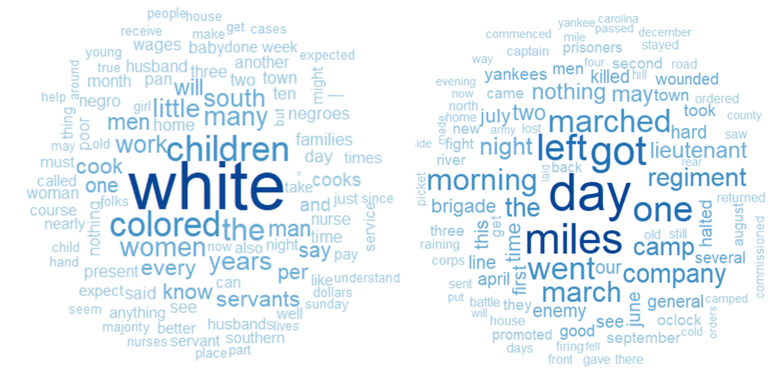

Fig. 2: Word cloud comparison between two different texts from a black, female nurse (left), and a white, male soldier (right).

Numerical Data Analysis

Numerical data analysis was key to understanding an author's emotions and how they were expressed in writing. Our analyses used Flesch-Kincaid reading ease and sentiment analysis scores. We found that:

- Reading ease did not vary significantly by demographics, and the mean score was 60, meaning most texts are easily readable for those with an 8th grade education today.
- Reading ease was often higher than modern texts compared, such as BBC Politics and Entertainment articles.
- Overall, most texts were positive and objective.
- There were visible differences across race, with mixed race authors having the most neutral sentences and white authors having the most negative or positive sentences.

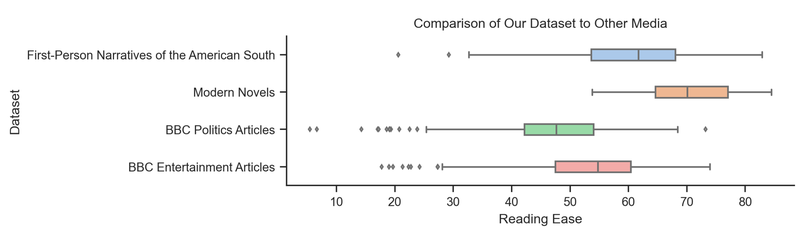

Fig. 3: Comparison of reading ease between the corpus and contemporary texts.

Results

From our analyses, we were able to make general conclusions about those living during the Civil War era:

- By sorting the reading ease scores by demographic, we found that the documents were **consistent** regardless of factors such as race and gender. As seen in Figure 4, black writers, many of whom were former slaves, were writing at **the same** or a similar level as educated white or mixed race writers.
- By combining topic modeling and n-gram frequencies and sorting them by demographic, as mentioned in Categorical Data Analysis, we reached conclusions about subjects certain groups of people were interested in. For example, our programs helped us determine that black authors placed large emphasis on **religion** and **education**, in addition to their own personal experiences as slaves.

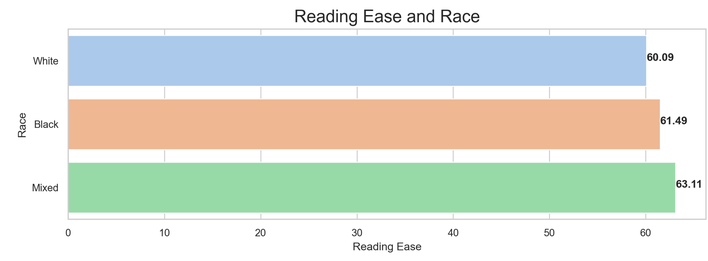

Fig. 4: Reading ease scores of texts written by different races.

Conclusion & Future Work

The programs we created allowed us to establish preliminary conclusions based on the corpus. They also helped us and will help future historians to better understand general perspectives and the context of a document without having to read it first.

For historians reading over 150 documents, our text-mining programs provide a much-needed improvement to annotation speed. Between our team and three other volunteers, annotating *just* 150 documents took the better of three months to complete. Our text-mining methods are capable of annotating the **entire** corpus in an average of **10 minutes**.

In the future, our next course of action would be to implement ways to detect false positives, such as sarcasm, in analyzed documents. Sentences with positive connotations, but negative language skewed our sentiment analysis score. Detection of false positives would help to alleviate the issue and create more accurate analyses.

Acknowledgements and References

Many thanks to the annotators from ACM and to the University of North Carolina Library for their curation of our dataset.

Corpus courtesy of University of North Carolina Library, Documenting the American South, First Person Narratives Collection, docsouth.unc.edu