Josue Ramirez Antonio, Ruhan Xia, Matthew Xue

Fall 2023, DS3500 HW4

2023/11/17

# An extensible reusable library for Natural Language Processing

GitHub link: https://github.khoury.northeastern.edu/ruhan/ds3500_hw4

## Overview

The Civil War was a monumental war that changed the entire landscape of the United States politically, socially, and economically. Slavery was abolished, the Union was preserved, industrialization accelerated, and the list goes on. So, wouldn't it be interesting to analyze historical Civil Letters to give us a better understanding of what it was like to fight in the war from many different perspectives. These letters provide an intimate and personal perspective on the war, offering insights into the struggles, sacrifices, and challenges faced by soldiers, their families, and communities. These letters offer details that may not be found in official records or historical documents, giving a more human and relatable dimension to the war.

## Data Source: University of Washington's collection of Civil War Letters

For our data sources, we chose to utilize the University of Washington's collection of Civil War Letters. We found their database to be abundant in information and easy to navigate. All the letters were presented in high quality, and the transcripts provided for each letter seemed accurate and legible.

From this database, we selected letters from three specific individuals to analyze: Hazard Stevens, James A. Sayles, and Riley M. Hoskinson. Each person came from different backgrounds and offered unique insights into the war.

Hazard Stevens, the son of the first governor of Washington Territory, corresponded with his parents from 1861 to 1865. He shared his opinions on the military and political direction of the war. Fortunately, he survived the war and conveyed the atmosphere when they celebrated General E. Lee's surrender.

James A. Sayles, a young captain from Illinois, wrote letters to his lover, Florence Lee, in 1864. His letters initially portrayed a romantic view of army life, but they progressively revealed increasing homesickness until his death.

Riley M. Hoskinson enlisted in the army alongside his son Stuart and corresponded with his wife, Martha Hoskinson, in 1863. His letters depicted the horrors of field hospitals, the harsh conditions in mountainous regions, and the grim realities of war. He detailed the capture, escape, and survival of himself and his son from the Battle of Chickamauga.

### First visualization: Sankey Diagram

The functions load_all_text, flatten_wordcount_to_dataframe, and plot_sankey were utilized to generate the Sankey Diagram displayed below. Initially, we loaded and combined all the texts from individual text files, organizing them by author. Subsequently, the nested word count data was converted into a dataframe. Finally, utilizing the dataframe, a multi-layered Sankey diagram was created, illustrating the flow from each author to each individual letter and subsequently to each individual word.

The Sankey diagram reveals that Stevens and Sayles shared certain experiences, notably discussing the army extensively in their letters. On the other hand, Hoskinson had distinct experiences, predominantly using words such as "mountain" and "road," reflecting his detailed accounts of the challenging conditions in mountainous regions and battles fought in such terrains.

Visit Exhibit A for the Sankey Diagram.

### Second visualization: WordCloud subplot

The function generate_wordclouds_subplots enables users to create a subplot of WordClouds for each letter, ranging from 2 to 10 (or more, if required), uploaded into the framework. The

WordCloud showcases the most common words found in each letter and titles each WordCloud with the corresponding file title. The generate_wordclouds_subplots function accepts a customizable column parameter that allows users to choose the subplot format of the WordClouds.

We decided to select three letters from Hoskinson, three letters from Sayles, and four letters from Hazard Stevens. The subplot allows us to compare word frequency, key themes, and keywords from each letter. In Hoskinson's letters, frequent topics included mountains, water/canteen, blankets, and guns. These words are relevant as Hoskinson fought in mountainous terrain and recounted his capture, escape, and survival, hence these words are directly related to survival. Sayles' letters, in contrast, were less dark and more informational. Commonly used words were hope, cavalry, army, officer, and time. These words suggest that Sayles likely shared his day-to-day life with his lover and generally seemed upbeat. Hazard Stevens' letters predominantly discussed his experiences in the army, battles fought, and the Confederacy's surrender. Commonly occurring words included Lee, army, troops, soldier, home, church, and Appomattox. These words indicate a breadth of experiences during his military service and a change in mood as the war approached its end.

Visit Exhibit B for the Sankey Diagram.

**Third visualization: Scatter plot for sentiment score**

The functions get_sentiment_plot and plot_sentiment are utilized to create a sentiment analysis graph displaying subjectivity (objective, subjective) on the y-axis and polarity (positive, negative, neutral) on the x-axis. This graph facilitates the comparison of different letters and authors, providing insight into their sentiments during the Civil War. The get_sentiment_plot function incorporates a parameter called GroupedAuthor=False, enabling the creation of a figure based on either individual letters or authors. In our case, we generated figures for both individual letters and grouped letters for each author, plotting them together on a single graph using the plot_sentiment function. From the sentiment analysis plot, we can see that based on the letters we chose, Sayles wrote more positive and subjective letters compared to Hoskinson and Hazard.

Hazard and Hoskinson both wrote letters with similar subjectivity and polarity. This may be the case because they both wrote a lot about their experiences in skirmishes and life in the army.

Visit Exhibit C for the Sankey Diagram.

### Code Distribution

Josue Antonio created the sentiment analysis plot, implemented exceptions and custom parsing functions, and made the library as efficient and organized as possible. Ruhan Xia created the Sankey diagram for the text files, took initiative to set up the repository and organized the repository, and set up the skeleton of the library. Matthew Xue created the subplot of word clouds, made adjustments to the library, and took leadership over the report.

### Conclusion

Through our NLP library and app, we gained deeper insights into the Civil War from the perspectives of individuals who directly experienced it. All three authors provided profoundly interesting insights, revealing the remarkable diversity of experiences among just three individuals. The war impacted not only the soldiers who fought but also their loved ones, to whom they wrote letters. Overall, this experience taught us much about text analysis, offering valuable insights into fascinating subjects.

### References

List of letters in Database. ::: Civil War Letters Collection ::: (n.d.).
https://content.lib.washington.edu/civilwarweb/collections.html

## Exhibits:

*Exhibit A*

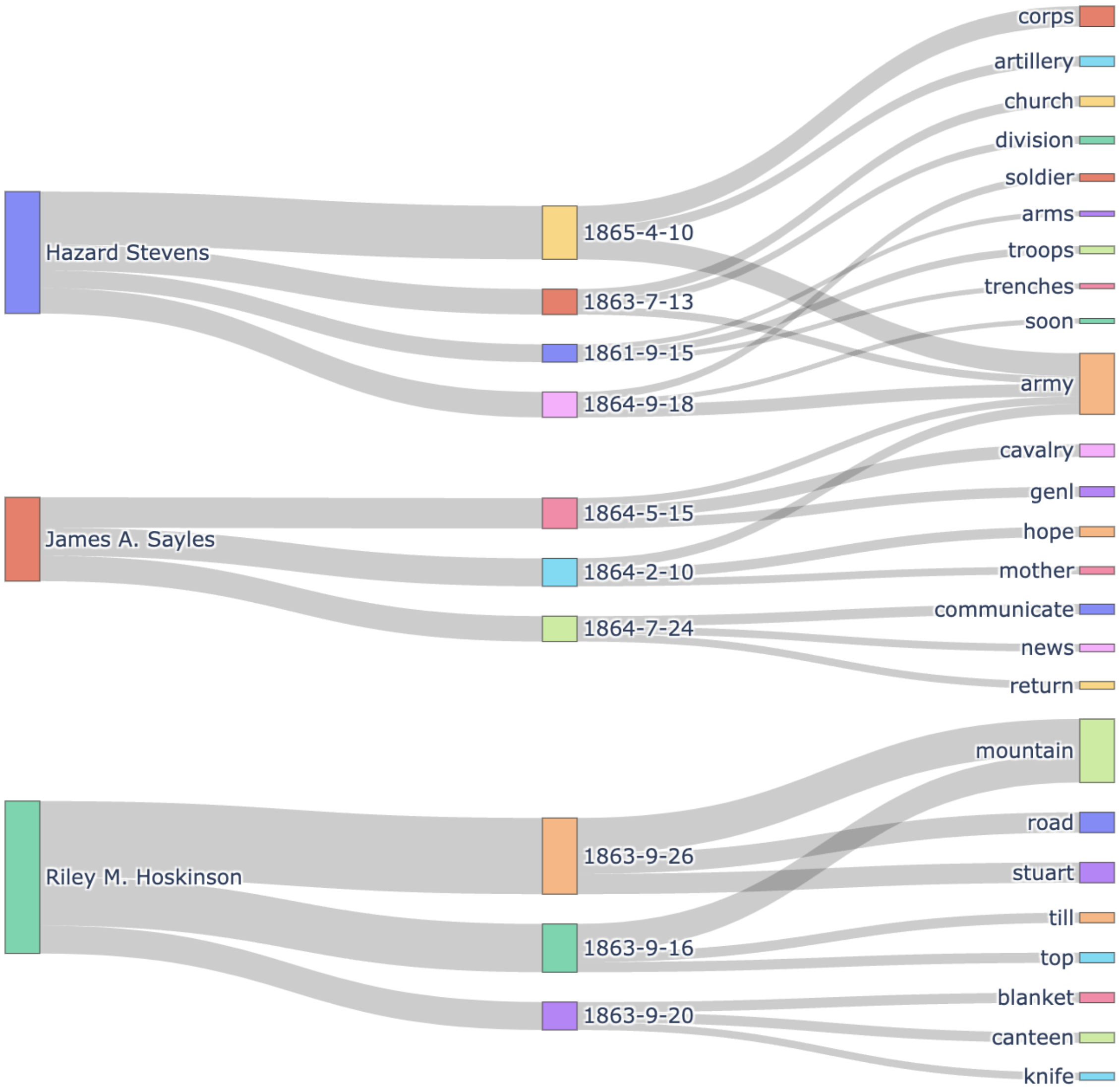
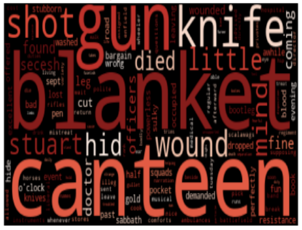
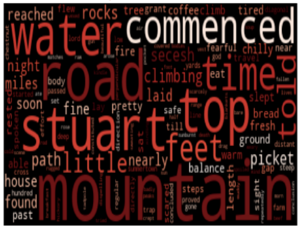


*Exhibit B*

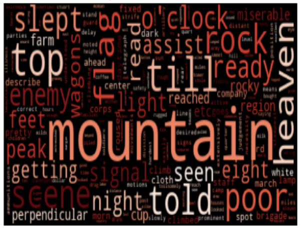
Riley M. Hoskinson 1863-9-20



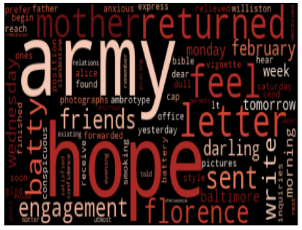
Riley M. Hoskinson 1863-9-26



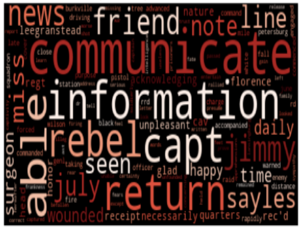
Riley M. Hoskinson 1863-9-16



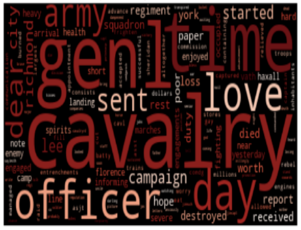
James A. Sayles 1864-2-10



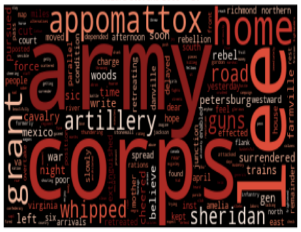
James A. Sayles 1864-7-24



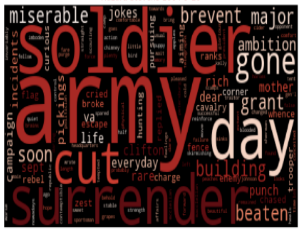
James A. Sayles 1864-5-15



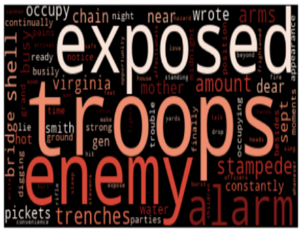
Hazard Stevens 1865-4-10



Hazard Stevens 1864-9-18



Hazard Stevens 1861-9-15



Hazard Stevens 1863-7-13

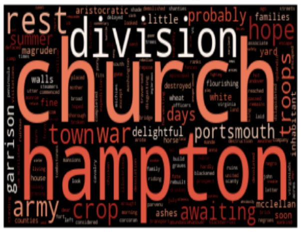
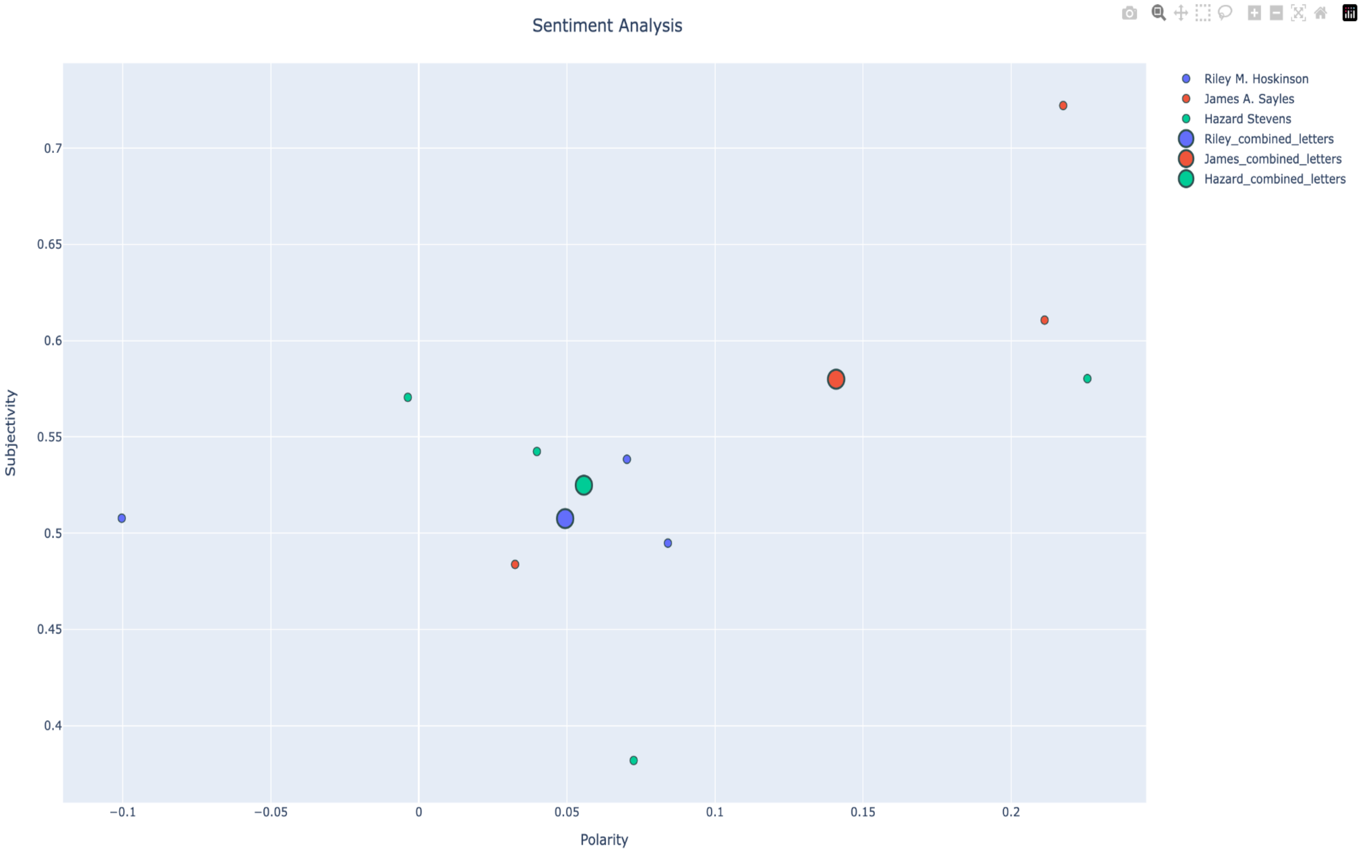


*Exhibit C*


Sentiment Analysis
Subjectivity
0.7
0.65
0.6
0.55
0.5
0.45
0.4
−0.1
−0.05
0
0.05
0.1
0.15
0.2
Polarity
Riley M. Hoskinson
James A. Sayles
Hazard Stevens
Riley_combined_letters
James_combined_letters
Hazard_combined_letters