News Analysis and Aggregation

Brandon Chastain
James Mullenbach
http://jamesmullenbach.org/CS4464Assignment3.html

Concept

We were interested in learning about the changing sentiment of US news toward social progress over time. Our hypothesis was that a more positive sentiment concerning slavery would be seen in southern states, with an overall increase in negative sentiment observed across the country over time. Given access to millions of US news articles published over roughly a century, we found articles concerning the issue of slavery, analyzed the sentiment and number of negations in each, and visualized the average sentiment of each state for each decade on an interactive map of the US. The map has a timeline slider which allows the user to change which decade is displayed. This allows the user to explore the progression of the sentiment that the news media from each state had toward the issue of slavery as time passes.

Methodology

This project was separated into three steps: data collection, data analysis, and data visualization.

Data Collection

We gathered data from Chronicling America, a corpus containing millions of OCR-scanned US news articles from 140,000 publications that were published between 1836 and 1922. This dataset is particularly relevant to the issue of slavery in the US because it contains articles from a period of significance concerning that topic. The issuing of the Emancipation Proclamation, the Civil War, and Reconstruction all happened during this time period.

In order to gather relevant news articles, we wrote a python script which queried the Chronicling America API url `/search/pages/results/?ortext=slavery` overnight in order to download as many pages containing the word “slavery” as possible. After these pages were downloaded, we sorted them by the number of occurrences of the term “slavery” in each one, took the top 20,000 of these, and ended up analyzing around 5,700 articles for sentiment.

Some of the pages were incredibly lengthy, and we could not find an accurate way to pull out only the relevant article from the entire page, which in general would contain many articles. As a solution to this problem, we initially tried to take the subset of the page starting at the first occurrence of the word “slavery”, and ending at the last occurrence of the word. However, this still left most pages very long, and it was infeasible to perform sentiment analysis on them in time. So, we decided to trim each page to contain a maximum of 5,000 characters starting from the first occurrence of the word “slavery.” We believed that this would

approximately keep the topic of slavery intact, as the top 5,000 articles all had at least 27 mentions of the word. These trimmed articles were saved to a CSV file along with their publication title, date, city, and state to be analyzed in the next step.

Data Analysis

We performed sentiment analysis using NLTK's Vader sentiment and intensity analysis tool, developed by researchers at Georgia Tech. Although we recognize that this tool was originally developed for social media text, a medium far removed from 19th century newspapers, the authors found that it generalizes well, and we found it more effective than our attempt to train our own naive Bayes classifier using any of NLTK's sentiment corpora. We measured the success of Vader by testing various articles on both Vader and our own sentiment classifier, and found the results from Vader to be more reasonable, by our own best judgment. Analysis was performed on each of the 5700 articles, using a subset of 5000 characters starting on the first occurrence of "slavery". For each article, we assigned it the compound, positive, and negative sentiment scores that Vader output, and created a new CSV to hold this data. Following the example of the QUOTUS paper, we also searched for occurrences of "not" and the contraction "n't" within the article, and assigned a count of these for each article as well, in order to develop a separate map.

Data Visualization

After analyzing and saving each article's sentiment and negation and aggregating them by state and by decade, we displayed these values for each state on a map of the US by decade using D3, an info visualization tool. We first transformed our data by grouping each result by state and decade. Then, for each state, we assigned a sentiment score by averaging over the positive value minus the negative value of the instances for that state for each decade. The resulting map shows states in shades of red, where a darker shade indicates a more positive sentiment, and a lighter shade is a more negative sentiment. Above the map is a slider which allows the user to change which decade's sentiments are being displayed on the map. The user can drag the slider from left to right in order to see the progression of states' sentiments over time. In the second tab, a similar map was made for the negation count averages. The idea was to explore if there is any correlation between this and sentiment, as there was in the QUOTUS paper (see references). Note that the scaling is different for each decade on the map.

Problems Encountered

The initial idea that we had for this project was to inspect changing sentiments about women's suffrage in the US news. However, we found a low amount of data concerning this issue. Additionally, and we were querying for articles containing multiple words: "women's suffrage," "suffragette," "19th amendment," or "nineteenth amendment," which led to confusion

with the Chronicling America API. We were unsure how to make multi-word queries match both words in sequence. We also observed that some of the resulting articles did not seem to contain any occurrences of our exact queries, so they must have only been partial matches. Switching to only using the single keyword “slavery” allowed us to find many more relevant articles from the API.

Results and Conclusion

As seen in the visualization, there doesn’t appear to be much correlation one can draw from the maps as they change over time, and our hypothesis is not supported. There could be many reasons for this. One important reason is poor coverage; despite performing what we believed was an exhaustive, general search over Chronicling America’s database, many states are not represented at all, even when then-territories such as Hawaii are represented. Across all decades only 31 states/territories are accounted for, and most of those are still only present for a few decades. We do at least see coverage spreading from 1830’s to the 1860’s, leading to a drop in the 1870’s and beyond, so we can confirm that it was a more popular topic in that period, as expected.

Another reason is that, upon further thought, perhaps a simple positive and negative analysis doesn’t work for this case. An abolitionist could write an article on slavery that classifies as having positive sentiment, if he is celebrating the progress made towards abolition. Likewise, an article could be written that expresses outrage over the impending abolition movement, which would have negative sentiment but be pro-slave. The reason this analysis worked for the QUOTUS paper was because they were looking at quotes used by news articles, which are both shorter and can be more indicative of how that article wants to present the POTUS, whereas in this case there is too much possibility for variance of tone across an entire article.

Resources

- Chronicling America API
  (http://chroniclingamerica.loc.gov/about/api/)

- D3.js
  (https://d3js.org/)
  - D3 slider plugin (http://thematicmapping.org/playground/d3/d3.slider/)
  - D3 geomap plugin (https://d3-geomap.github.io/)

- Jupyter Notebook
  (http://jupyter.org/)

- NLTK
  (http://www.nltk.org/)

    - VADER Sentiment Analysis (http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf)
- QUOTUS (http://www.cs.cornell.edu/~cristian/Structure_of_Political_Media_Coverage_files/quoting_patterns.pdf)

- Requests (http://docs.python-requests.org/en/master/)