Chronicling America: The Sentiment towards Women in the News

By:

# Austin Steffes

Computer Science

# Hamilton Greene

Computer Science

Prototype:

Live Prototype: <http://sirhamy.com/projects/ChroniclingAmerica/>

GitHub: <https://github.com/SIRHAMY/ChroniclingAmerica>

For:

Assignment 3: News Analysis and Aggregation

CS 4496 – Computational Journalism

Spring 2016 | Georgia Institute of Technology

Data:

Chronicling America: Historic American Newspapers[1]

United States Library of Congress

# Purpose

The goal of our project is to visualize the change in frequency and corresponding sentiment of articles mentioning women over time in a certain context pertaining to an important role women played across time. The data made available for this assignment consisted of newspapers and data from 1836 through 1922, and thus it was hypothesized that there would be a great difference in the attitude towards a role women have lived throughout time going from more negative to more positive as time goes on. In order to study the effects of time on our hypothesized stance on a certain job many women played, we parsed through the data containing the search word “housewife” for the years available (1836-1922) for the Chronicling America database. Then once we obtained the data, we created a Bayes Net classifier that uses training data sentences that are given a positive or negative classification. Through this, we are able to test the overall public attitude towards women and the role of the housewife by running the stories through the Bayes Net Classifier one by one and obtaining the connotation of its reading. We wanted to investigate some interesting ideas of a class of people with less public power and opportunity through most of the time the data is available for. By investigating this, we could then show the changing of public opinion through time as we witness the changing of rights overtime along with these results as well. Overall we wanted to measure the opinion on women; however, in order to simplify the results obtain through the parsing of the database, there was a focus on “housewife” as this gave a much more reasonable amount of results and still delivered the information we desired. Overall, the purpose of this assignment was met.

# Data

The data that we used for this assignment was the queried data from Chronicling America using a search term of housewife. We use this search term in order to control the huge volume of data that would be received if we were viewing information for a search term such as women. The data consists of 1001 entries from of newspaper data entries that consisted of this search term. We queried the Chronicling America database and used a fetcher in order to write the information to a CSV file, where were then able to read all the stories into a list. We could then use this list and we were able to run a text classification module on it in order to classify the overall attitude towards the housewives given the language used. Overall, we have 1001 queried data stories from the Chronicling America database, varying in year of the stories, which were gathered using the search term housewife.

We also fetched a corpus of articles containing over 187K results related to “housewives”. However, the size of it made it difficult to manipulate and increased the length of operations required to classify it. We do have a visualization of it up on the prototype, however the size of the file means that the whole visualization takes awhile to load which is detrimental to the overall user experience. Therefore, while it is available, we decided to stick the results of the smaller, 1001 articles, corpus at the forefront.

## Why this Data?

We believed that the time setting of the available information could allow for some interesting analytics, for marginalized groups of the time. Women were not considered as highly as men overall during the times available, and therefore we wanted to gather and analyze data that would show the changing public opinion over time through newspapers. We believed that since we know that women’s rights were changing, that public opinion on portrayal of women should have been changing as well. Additionally, we wanted to see if there was a difference in the attitude in the writing depending on the time period it took place, and we were able to see some interesting classifications through this idea.

# Methods

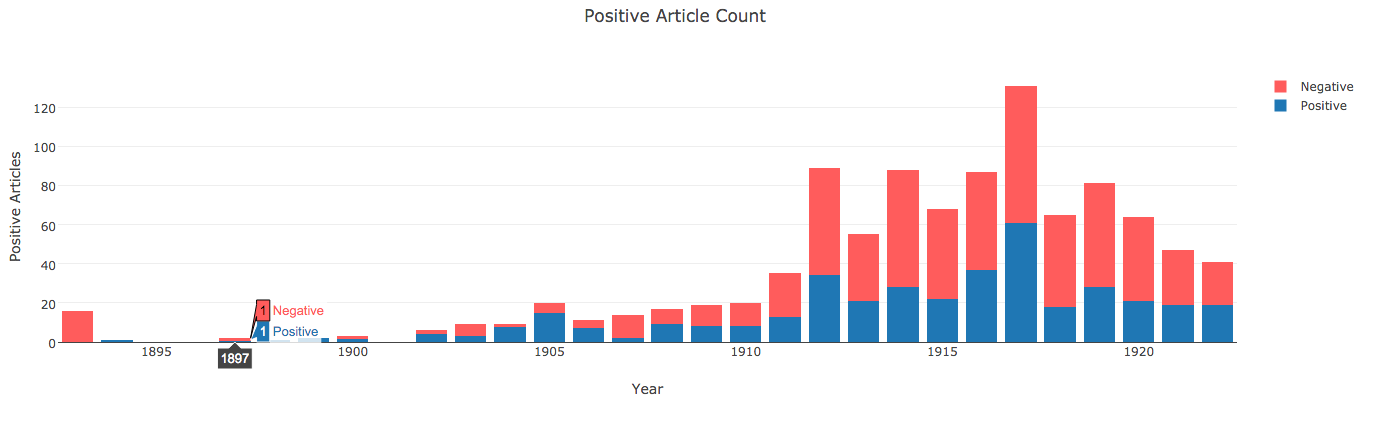
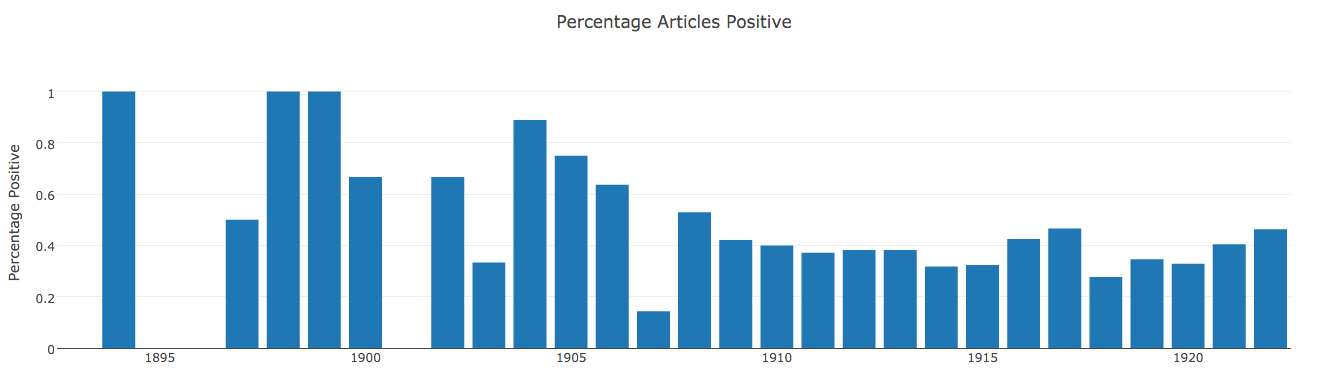
* **Pulled data from API**

The most technically challenging part of this assignment was effectively pulling the data from Chronicling America using the API provided to us. In order to do so, we passed in multiple URLs into the API settings in order for the fetcher created to pull many needed pieces of information about the articles pertaining to the search term. We could put constrained on the information detailing the years to search; however, it could additionally look at all available information. Through this fetcher, we were able to obtain JSON data that we could search through for using the Bayes Net Classifier to understand the information in much more depth.

* **Trained “classifier”**

One of the most important pieces of this project was using a Bayes Net Classifier that we specifically trained to understand the different attitudes towards women depending on the words in the newspapers. In order to do this, we passed in a training set containing both positive and negative statements about women and housewives in general, in order for the classifier to learn what to look for when testing it. When testing it, we created many statements we understood to be positive or negative and ran the classifier against this to show the correctness of its classifications.

We tried to keep this as simple as possible; through the usage of simple phrases for both positive and negative sentences, we believe it can translate more efficiently to larger sentences and newspaper stories as a whole. The Bayes Net Classifier ran using TextBlob library that helped more efficiently search through the available strings that we have for all the newspaper stories.

* Created Visuals
  + Based on data
  + 
  + A Bar Chart showing the distribution of positive/negative sentiment in articles mentioning housewives over time
  + 
  + Bar Chart showing the percentage of articles containing positive sentiment as a percentage of all the articles that mentioned housewives

## API Usage

## Analytic Techniques; Why?

The analytic technique we used in order to analyze the data we obtained was text classification in order to create a sentiment analysis of a given group of people we searched for (women). Overall, we wanted to see the differences of public opinion towards women as time passed; therefore, we know that text classification and understanding the percentages of positive and negative articles overall for the years gives us a deeper understanding of the theme during that time. We believed that in order to understand opinion on women, that sentiment and text classification was the perfect solution.

# Measuring Success

* **Compared results to what we expected, what we saw in trials**

In beginning of this assignment, we believed that there would be an increase in positive opinion on housewives and women in general as time progressed. We believe we could measure the public opinion of women through the access of “housewife” data. We created a Bayes Net that analyzes strings and classifies them based on the training data provided. In order to assure the Bayes Net classified correctly, we manipulated both the training set and the testing set until it classified, as correctly as we saw it should depending on the data. Overall, we measured success in our ability to effectively obtain a bunch of data from Chronicling America through programming and then additionally our ability to successfully classify large stories of data. Overall, we were able to effectively create a Bayes Net that classified negative and positive sentences (and stories) that we could use effectively on the data we obtained. We had previously believed that there was going to be a strong correlation in negative stories towards women and the earlier it was written. Overall, we saw that there were many negative stories for the classifications as a whole, and that many of them skewed towards the earlier years that were researched.

Success in this assignment is measured in our ability to efficiently gather data using fetchers and then manipulate that data for our use. Not only were we able to gather this data efficiently, we were able to classify it and analyze it depending on the year that it was published. This project is also successful because there is a myriad of ways this it could be extended for further use. This is an important feature of a successful project: setting up a project structure that allows for the continuation of learning processes.

# Work Distribution

Overall, we both worked on every aspect of the assignment together. Both group members communicated and kept track of progress through messaging services and Github. Github was important in keeping a handle on all the different files used for the entire assignment. We believe we were able to accomplish a great amount because of our ability to both work on every aspect of this assignment, offering assistance in any situation that needed it. This allowed for the assignment to be a success overall.

References/Sources

1. Chronicling America: Historic American Newspapers. *National Endowment for the Humanities* through the Library of Congress. Digitized newspapers from 1836 – 1922. URL: <http://chroniclingamerica.loc.gov/>
2. Plotly – Helped with visualizations. <http://plot.ly>
3. D3 – Didn’t actually use it for visualizations, but we did use the CSV functions surprisingly. <https://d3js.org/>
4. Python – Self-explanatory, but we did our data fetching and operations in Python. <https://www.python.org/>
5. Forked base Python APIWrapper from @hugovk’s chroniclingamerica.py. Gave it a HUGE overhaul. Original: <https://github.com/hugovk/chroniclingamerica.py> - Our Fork: <https://github.com/SIRHAMY/ChroniclingAmericaAPIWrapper>
6. We use the Textblob Python Library to help us with classification. It has some nice classifiers built in, which is always useful.
7. For visualization development, used npm’s http-server to provide a quick and dirty local server.