Text Mining Write-up and Reflection

Anisha Nakagawa

26 February 2015

1 Project Overview

The goal of this project was to measure the objectivity and polarity of news articles from different sources, and to try to detect any resulting bias. The program takes a set of URLs about a given topic, reads in the HTML, and get the article text using the HTML parser from pattern.web. It then runs a sentiment analysis on each article using pattern.en, and graphs each article based on polarity and objectivity using the matplotlib module. The news sources can then be compared to each other based on where they are plotted on the graph.

2 Implementation

The articles to analyze were chosen in sets around a specific event, such as the People's Climate March in September 2014. For each event, news articles were chosen from major news sources such as BBC, NPR, The Economist, and Fox News. These were stored in a dictionary where the keys are the filenames (of the format event_newsSource.txt), and the values are the URLs of the websites. Since the filenames contain the name of the news source, that information can be extracted from the keys as well. Once the text from the URL is saved to a file, that text can be accessed by reading the filenames in the keys of the dictionary. For this project, the URLs were hard-coded into the dictionary because it was important to analyze only the primary article from a news source (not an opinion or blog post). While it would have been possible to automate the process of selecting URLs by performing a google search according to specifications, the desired article did not always immediately show up on the search page. Therefore, the best results were from hard-coding the website url into the values of the dictionary.

The implementation of this project has two main stages. First, it saves the text from news articles. It then reads and analyzes the text using a sentiment analysis, and plots the news source according to sentiment.

To get the article text, the program takes a website url and gets the HTML using the pattern.web module. The HTML is then parsed to get only the text from the article, which includes stripping out the header, footer, and other website-specific tags. While some tags must be stripped out of all the HTML files, each news source has additional tags specific to the company that need to be removed. These website specific tags are saved in a list for each website. After stripping away the extraneous information, the article is converted to a plaintext string and saved in a .txt file. The website article is saved as .txt file so that the HTML only has to be accessed once by the URL.

In order to analyze the articles, the sentiment analysis was run on each article (from the text file). The sentiment analysis from the pattern module returns a measure for how objective a piece of text is, and a measure of how positive (or negative) the text is. The sentiment analysis was run on each article. The data was then plotted on a graph according to the values found for sentiment and polarity. For each current event, the articles were plotted by the sentiment and labelled according with the news source (ie. BBC, Fox).

3 Results

The sentiment analysis from the pattern module returns two values for each article: a measure of how objective the article was, and a measure of how positive the article was. These two values were represented by a point on a two dimensional plane, where the x-coordinate represented the polarity, on a scale from -1 (negative article) to 1 (positive). The y-coordinate represents the subjectivity, where 0 is objective and 1 is subjective. Each news article was plotted as a point in this plane, and labelled by the news source. The graphs for three events - The People's Climate March, The State of the Union Address, and the Keystone Pipeline Veto - are included below.

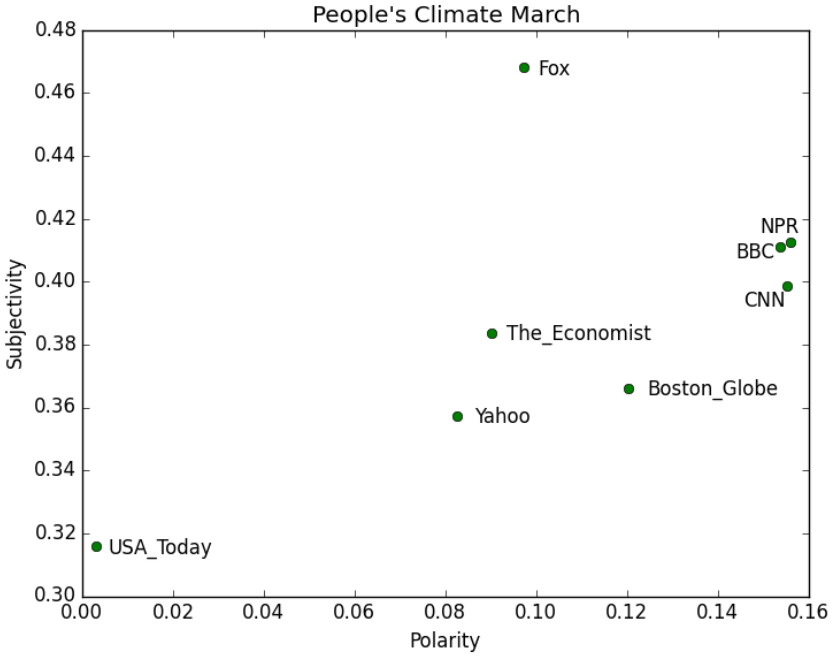

Figure 1: The People's Climate March occurred in September 2014, when large groups of people gathered in New York and other cities to show support for environmental movements.

In each of these graphs, it is possible to see how each of the news sources reported the same topic. Some of the articles are more subjective or more positive. When comparing the coverage of multiple events by different news sources, it may be possible to determine a bias from the news sources. For instance, for all of these events, the coverage from Fox News is consistently more subjective than other news sources, and NPR is usually more positive about these issues. However, more data would be needed to form a concrete conclusion about the biases.

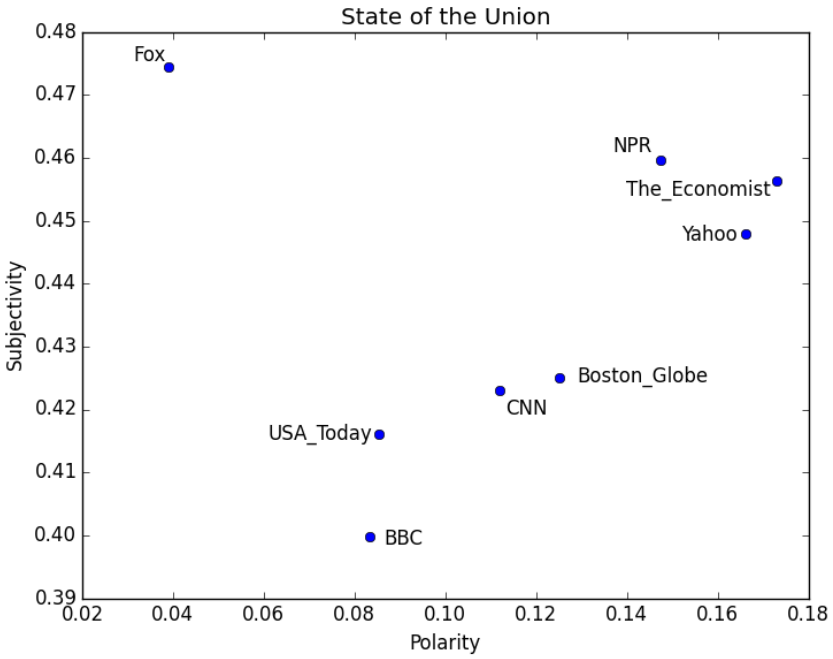

Figure 2: President Obama gave the State of the Union speech in January 2015 about what his administration has done and plans to do in the coming year.

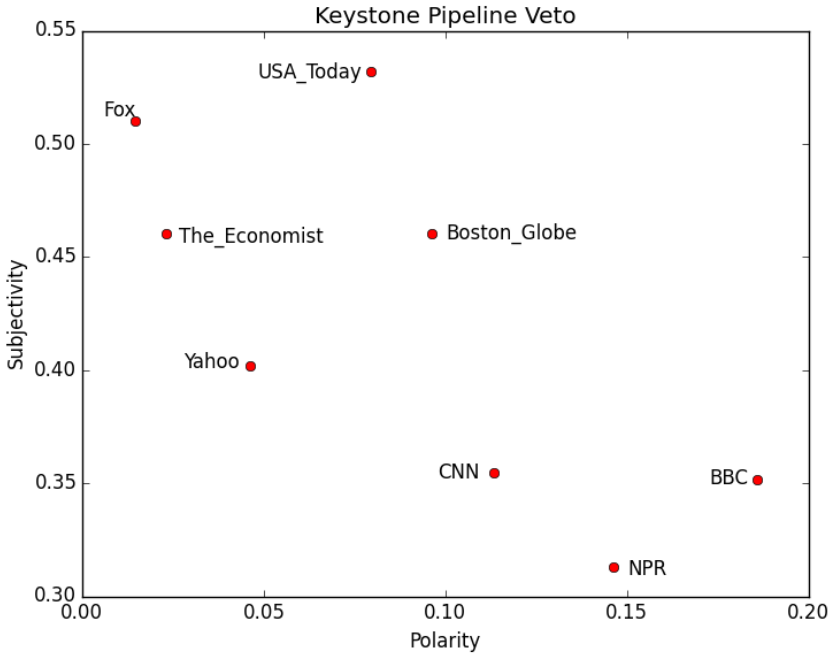

Figure 3: On Tuesday, February 24th 2015, President Obama vetoed a bill that would build the Keystone Pipeline, which would connect oil between Canada and the United States

4 Reflection

The code for this project was broken up into specific functions that were documented clearly, which made it easier to debug and test individual pieces of the code. Furthermore the docstrings included the return type of the function, which made it easier to remember how the functions should fit together. However, in order to get the clearest results, there were a couple places in that were hard-coded, and it would be more elegant if they were automated. In order to get the article text, it was necessary to parse through the HTML and remove specific tags - but those tags were different depending on the news source. This required creating lists of specific tags depending on the article, which was an unexpected hurdle. Furthermore, the URLs of the websites had to be entered directly into the dictionary because the automated search functions were not able to select the best article to analyze. For this project, it was feasible to find the individual URLs by hand, but that would not work on a much larger scale. Overall, the program was efficient at analyzing data that has real applications, and it was interesting to see the results.