Aparajithan Venkateswaran

TA: Camilla Lambrocco

Recitation: 101

REPORT

Purpose

This assignment requires us to finish an incomplete program that analyzes data (tweets) to determine the opinion (sentiment) conveyed in the data about a particular topic. Apart from learning a very interesting topic (NLP), I think this assignment is aimed at teaching a few other important skills that every beginning software engineer must have.

The first is learning to read up documentation and understand a library someone else has written. Software engineering is the art of standing upon others' shoulders i.e., using code that other people have already written and tested to build something new and radical. This require the ability to understand what somebody else's code. And this project uses many external libraries, most importantly tweepy. While we are not needed to use them, it is essential to understand those parts of the code.

The next skill is closely related to the previous skill – understanding a piece of code upon which what you code depends. Often in industries, engineers will work with only small pieces of a large product. This will depend on parts that are being built by others and it is essential to understand how these work in order to test your piece of software.

Thus, I believe that this assignment's true purpose to teach students to understand other people's code – which is very difficult and essential.

Procedure

Before beginning to code, I made sure I went through all of the code provided to made sure I understood what was happening there. Then, I made a list of the tasks I needed to accomplish in order to get the code working correctly and made sub tasks for each task.

First, I created a function to split a string (tweet) into words by replacing all punctuation marks with a whitespace and removed all whitespaces to create a list of words. Then I filtered all tweets and stored all tweets containing the search query as a Tweet (class) instance.

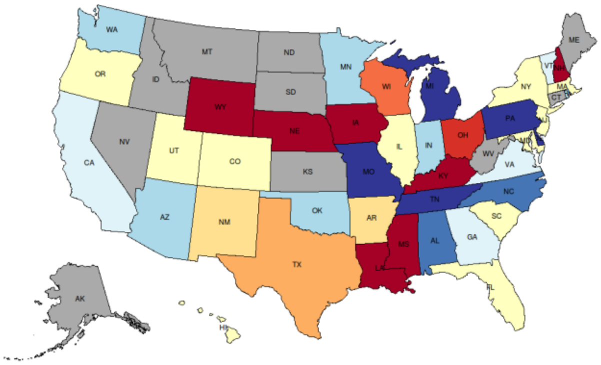
Then I created a list of dictionaries containing the state code as key and a list of sentiment scores as the value. I looped through the tweets and did three things – find the closest state, calculate the sentiment score for each tweet, and append it to the list for the corresponding state. Then I averaged the score and converted into the respective color. After that, I updated the color for each state on the map.

Later, I added functionality to read tweets from multiple files and search for multiple queries at the same time. I also implemented case-insensitivity to broaden search results.

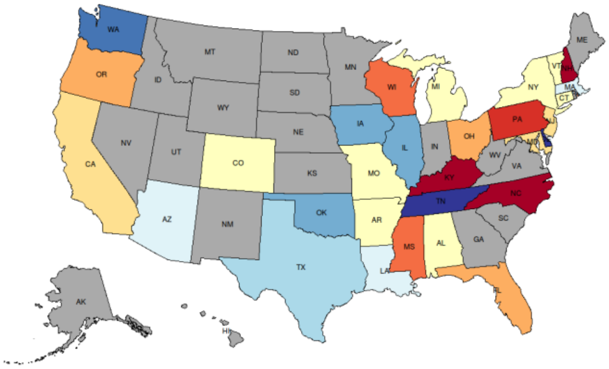
Since the tweets were collected during the first two weeks of November (when the elections were most popular), I decided to analyze the sentiments towards the most popular two Presidential candidates and see how they reflected the actual outcomes of the elections.

Results

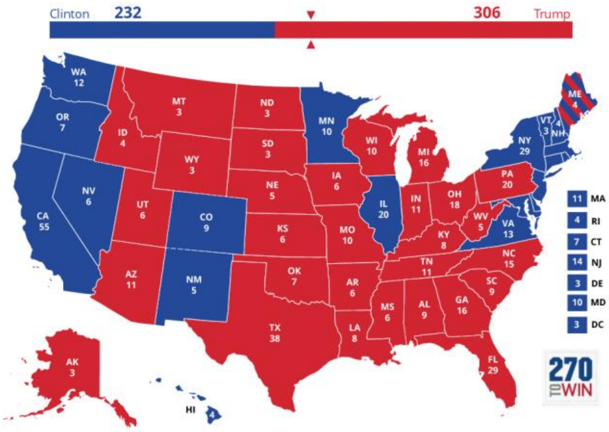
Map for ['Donald', 'Trump']



Map for ['Hillary', 'Clinton']



Actual Results of the Elections



Overall, the sentiments towards both these contenders are conflicting in most states (as must be).

There are a few states like Colorado that seem to have similar feelings towards both the candidates and this seems to reflect in the results as to how close the competition was in these states. For example, Oregon 'likes' Clinton more than Trump and this correctly models the outcome.

Then there are states like California that hates Trump and likes Clinton and states like Texas that likes Trump and hates Clinton. This sentiment is reflected in the results accurately.

There are also states like Wyoming and Nebraska that 'love' Trump and have no opinions about Clinton that seem to have voted for Trump.

While overall these sentiments fairly predicted the results, there are many anomalies. For instance, Arizona hates Trump more than Clinton but, has voted for Trump. There are many reasons for these kind of anomalies – the distance calculation algorithm is flawed; not everybody who voted tweeted about their opinions; sarcasm and satire could have been interpreted literally and; people might have conveyed dishonest opinions to avoid being judged. Perhaps these are the reasons why the actual result of the elections might have come as a shock to many.

Factoring in these anomalies could give a more accurate result but would probably require higher order math. This could definitely prove to be a fun project to work on during the winter break!