Will Thorbecke
Feb. 26, 2016

Text Mining (Mini Project 3)

Project Overview:

I used Wikipedia to analyze book summaries of National Book Award winning fictional novels. I used Beautiful Soup and an html syntax stripper from Stack Overflow to pull and clean up the summaries. Then I used Indico to analyze sentiment. I was hoping to observe some kind of trend in sentiment over time.

Implementation:

The most important function in my program was the wiki_summary function. This took a book title and using a few filters found an appropriate Wikipedia url. Then, using the Beautiful Soup api, wiki_sumamry found the location of book summary from within the book’s Wikipedia page. wiki_summary saved this summary (html syntax and all) as a string and sent it to the stripHTMLTags function which I found on Stack Overflow. This function helped strip my summary of any html syntax. The last major component of my program was analyzing my filtered summary’s sentiment and finally graphing my results.

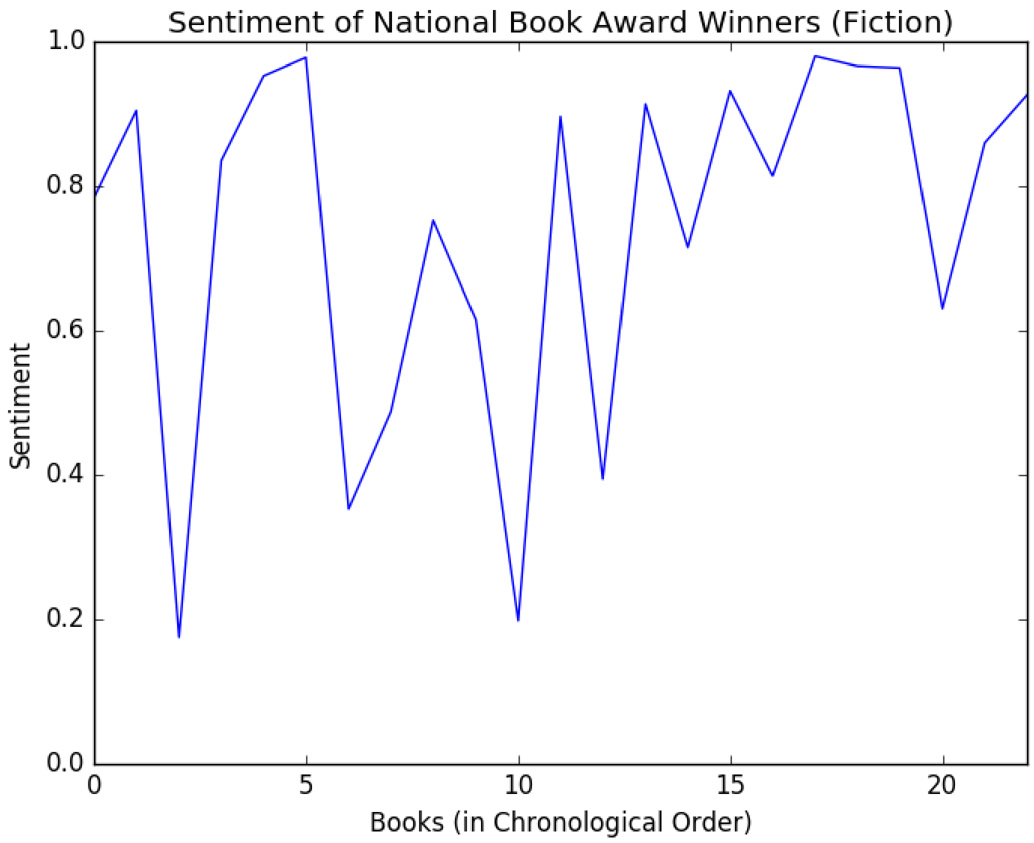
In terms of design decisions, one thing I could’ve done differently was use a search function that used Wikipedia’s search query to find the appropriate book instead of writing a complex algorithm to figure out the Wikipedia url for a specific book. One algorithm in my wiki_summary function that I enjoyed completing was an if statement that checked whether a specific Wikipedia contained certain html IDs such as “Plot” or “Synopsis”, if the page didn’t have these tags then generally that meant “_(novel)” had to be added to the url and the correct page could now be found.

Results:

I wasn’t sure what to expect during this project, but I somehow came up with the idea of analyzing the sentiment of National Book Award winners in fiction using Wikipedia summaries of the book. I came across many roadblocks near the end of the project that I wasn’t able to tackle. One problem was that more than a handful of the books didn’t have summaries or even Wikipedia pages. Furthermore some of these books were anthologies which I also decided to throw out. In the end I had 23 books to analyze from the past 50+ years worth of winners.

I graphed the results in chronological order for the 23 books I was able to analyze. If I were to unscientifically extrapolate information from this data I might say that there is a positive trend in sentiment starting with the 10th book analyzed, and prior to that the sentiment has a much wider range.

Below is a graph of my results:



Reflection:
In terms of what went well, I was able to pull a decent number of book summaries from Wikipedia and analyze their sentiment. In terms of what I could improve, I could've tried to pull from more sources to get more summaries to analyze. Getting all the summaries would've allowed for me to plot years instead of the location of my the book in a chronological list. I would've prefered this because then it could be possible to interpret the sentiment in comparison to the time period. In terms of my plan for unit testing, my scope was never large enough. I could try 20 different books before I ran across an error in one of my functions, thus unit testing was a hard thing to implement in this project.