Logan Sweet
Software Design Spring 2015
Text Mining and Analysis

Project Overview

In this project, I aimed to evaluate the way that the language of different Wikipedia articles illustrate the sentiment surrounding a group of people. I decided to focus on authors, and looked at groups based on the subjects that they are known for. In doing this I hope to get an idea of the ways the contributors of these pages consciously and subconsciously think of the people they are writing about in order to get an average perception of each genre of author.

Implementation

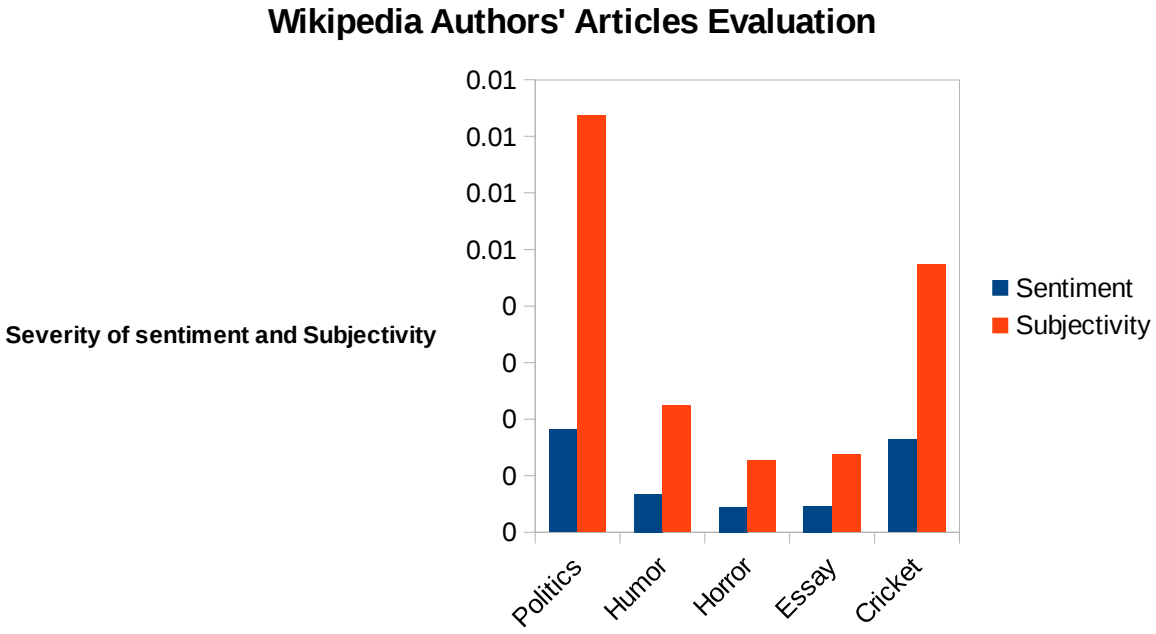
In order to implement this, I decided to break the problem up into three parts:The first part was to retrieve a list of authors' names with their own articles from the lists of authors that Wikipedia provides. To do this, I evaluated the links present in the body of the article. In order to avoid evaluating any links that were included in the article as reference material, I only chose author lists that did not contain this extraneous information. The second part took in the names of authors and evaluated the sentiment and subjectivity present in their corresponding articles. It was able to access the correct articles by searching for the exact title of the article, as inputted by me. The third part took the sentiment and subjectivity values for each article and averaged them over the total number of articles evaluated.

If I had evaluated different sections of the article instead of the while thing I think I could have gotten data that was more aimed at what the Wikipedia contributors thought of the author. For example, some articles have only one sections devoted to the author's life while others spend multiple sections detailing their lives, careers, and public record. This information would have been interesting to evaluate, but since each article had a slightly different setup with varying section titles and numbers of sections I made the decision to evaluate the sentiment of the article as a whole.

Results

I didn't do as many genres as I would have liked since I ran out of time (each genre took around half an hour to run) but still got one significant result out of it. While many of the sentiment and subjectivity results were as I expected in relation to each other, including humor getting a result as higher in both categories, the results for articles on political authors surprised me. These people's articles were higher ranking in both subjectivity, which I would have expected, but also in Sentiment. I expect that this is because the contributors that write articles about political subjects,

including authors, have strong opinions about politics.



Reflection

At the beginning of the project I spent a **lot** of time going in a direction that did not end up working out. I wanted to do a similar sentiment analysis as I did with groups of authors but instead evaluate sentiment around people that were grouped by race. However, the lists I tried to use varied in structure: some were lists while others were charts. Even with the help of a ninja, I wasn't able to get any information from the charts and abandoned the idea after lots of work. In my second iteration I was much more easily able to get working code, but at that point (Wednesday evening) I was very constrained for time and didn't get to do as much as I would have liked with it.