Project Overview

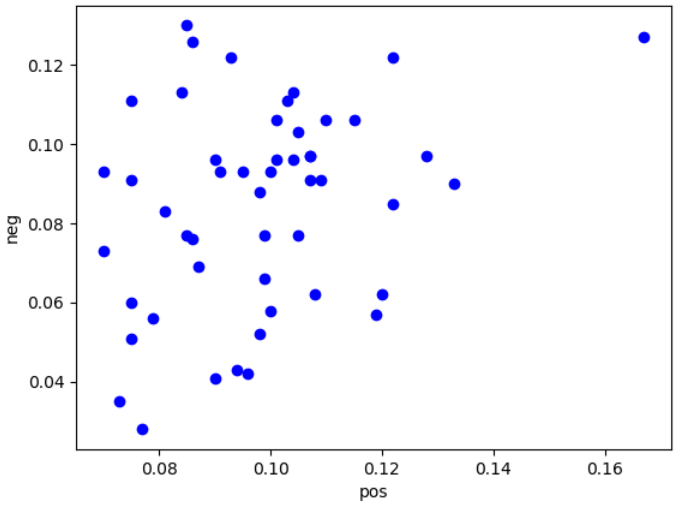
We took the top 50 most significant historical figures according to TIME magazine and put all of their names into a list. Then we got the respective content from the Wikipedia page of all of the figures. We then performed a sentiment analysis of the content on each figures Wikipedia page. This would then allow us to determine the overall sentiment towards the figures that are remembered most. This would give insight into the question of how people are remembered for their legacies.

Implementation

The list of historical figures was inputted manually. The original list from TIME magazine had 100 figures on the list, but for the purposes of runtime, we chose to take only the top 50 figures from the list. At first, we attempted to scrape other websites for a list of more historical figures, but the websites were either structured with really complicated HTML or the request to retrieve the page had been denied.

Once we actually had a list to work with, we used the “wikipedia" library to fetch the content on the Wikipedia pages of all of the figures. Once we had all of the content within a list, we used the Natural Language Toolkit to perform a sentiment analysis. We compared the positive sentiment scores against the negative sentiment scores for each figure.

Results



Above is a plot of the negative sentiment scores over the positive sentiment scores for each of the figures. With the exception of one outlier in the top right corner most of the points fall

towards the left side of the graph. On the other hand, the points are distributed pretty evenly along the vertical axis. This suggests that the negative sentiment had been stronger than the positive sentiment for each of the figures. This makes one begin to wonder whether it is inevitable for your legacy to be criticized when you are a well-remembered figure.

Things that should be considered when evaluating these results are (1) the neutral words had been ignored but could play an important role when evaluating a figures legacy and (2) normally sentiment analysis is used in the context of twitter posts where it is popular for posts to be opinions. In Wikipedia articles, it is more likely that the writer merely wrote out the history associated with the figures. Therefore, rather than an analysis of sentiment, the results of the program could be seen as an analysis of the history associated with the figure.

Reflection

When we approached this assignment, we tried to plan times to meet, but realized that our schedules were very conflicting. Therefore, we decided to take a lot of the planning online over messaging apps and divide the work. At the end of the project we collaborated our combined efforts and revised the finished product. If we could redo this assignment, we would spend time to try and incorporate a lot of the ideas that could not be done within the time frame of the project. We would have added other forms of text analysis that would give more insight into answering our problem (i.e. does a figure with more Wikipedia sub-headings on their page rank higher on the list). In terms of brainstorming and coming up with an app that fit both our interests, we did a great job with coordinating both of our thoughts. The only thing that needed real improvement is managing our time to be able to find a meeting time where we could both be present physically.