Daniel James Rodgers

Project Overview

Using MediaWiki I harvested the wikipedia summary of leading left-wing, centre, and right-wing politicians from various countries. Then, I used NLTK to determine how much neutral text these entries contained, and their normalised positivity score. Finally, I compared the average scores (and word count) for the different ideologies. I did this to check, as would be expected all else being equal, whether there was a similarly low amount of non-neutral text in the wikipedia summaries, and a similar positivity score.

Implementation

The 3 objectives of the code was to 1) Harvest the wikipedia summaries of the relevant politicians, 2) Perform a sentiment analysis using NLTK, and 3) Extract and present the information.

For the first step, I manually created a dictionary of the leading left-wing, centre, and right-wing politicians of various countries. Instead of a dictionary, I could have used a list where country[0] was the left-wing politician, country[1] was the centrist, and country [2] was the right. I elected not to do this as a dictionary is more malleable to new information. For instance, if I were to add new keys such as “communist” or “alt-right”, in a dictionary it would not matter if a country had leading politicians in those ideologies. The lists though would all need to be the same length with aligned ideologies for country[0], country[1] etc.

Performing the sentiment analysis using NLTK is a fairly standard procedure, but I had the choice of storing the scores in either a list or a tuple. It does not make a difference when extracting the data as both are accessible by index, but there is a difference in performance when adding to them. Tuples are immutable, meaning that I would not be able to append a new entry like in a list, instead I would have to essentially duplicate the tuple and add the next value to a new tuple object, which is significantly less efficient.

Results

// The results below were taken by running the program at 8PM (EDT) on the 25th March 2020 //

Firstly, the program averages the word-count of the wikipedia summaries, returning:

```
Left-wing politicians:  257
Centrist politicians:   266
Right-wing politicians: 268
```

I did this step mainly to see whether there would be any bias in the sentiment analysis from some of the ideologies having shorter/longer summaries, but the difference is small enough that it would not have a significant impact.

Secondly, the program calculates the proportion of non-neutral text in the summaries, returning:

```
Left-wing politicians:  11.88%
Centrist politicians:    9.54%
Right-wing politicians: 14.14%
```

All else being equal, as wikipedia purports to be factual and not opinion-based, I would have expected the three ideologies to have very similar amounts of non-neutral text, however this is not the case. Moreover, the difference is not directly explained by the word count, as centrist politicians have the lowest proportion of non-neutral text in their wikipedia summaries, but a similar word count to the right-wing politicians. I believe the difference (14.14% is ~1.5x larger than 9.54%) is substantial enough to warrant further investigation.

Finally, the program calculates the normalised positivity score in the summaries, returning:

```
Left-wing politicians:  0.64
Centrist politicians:   0.70
Right-wing politicians: 0.50
```

I was not as surprised with the difference in scores here: only ~10-15% of the text has positive/negative sentiments, thus a small difference in sentiment would have a large effect on the normalised positivity score. However, with more politicians that volatility would be reduced and systematic bias would be visible, for me justifying having the normalised positivity score as an output.

Reflection

I chose this topic because especially during these times of duress, political partisanship has the potential of being further entrenched; as people are taking more of an interest in politics, it is important that the wikipedia summaries of political leaders (often the first step of research for the not politically-inclined), presents factual and not-biased information. I am more than aware that analysing only 14 countries' politicians is not enough to draw any meaningful conclusions, but this code can be scaled very easily: it is possible to see if the scores change over time, and all that is needed to add an additional country is the names of prominent politicians and their ideology. Furthermore, after all of that data has been compiled, it would also be easy to modify the code to compare the sentiments of their wikipedia summaries based on their geographic region or their gender. However, I know that I have some code redundancy, such as writing out the summary/count/neu/compound for all three ideologies, but I just couldn't work out how to write it so that it would also append to the correct tables. I know that I need to work on shortening my code and I am continuing to practice doing so. I hope I will be able to go back in a few weeks and make it shorter and more efficient.