PATRICK OGUNBUFUNMI
WORD FREQUENCY ANALYSIS IN SOCCER

PROJECT OVERVIEW

My project overview was to analyze the words successful managers use for their interviews, which may or may not have a strong correlation with their success rate, but nevertheless is ‘
An interesting discovery. I used HTML sources to collect my interview data off transcripts posted on the web. Main sources of data were Dailymail.com, Metro, and Mirror.com. For text parsing and processing, I used the Beautiful Soup library and the URL request library to handle source code importing, as well as the use of a dictionary to keep track of word frequencies.

IMPLEMENTATION

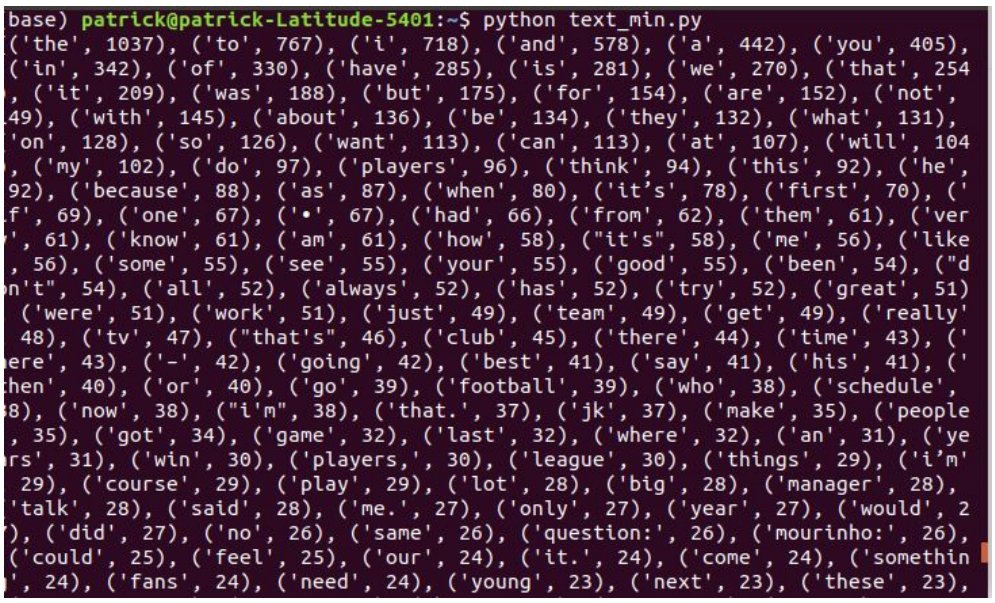
From the Beautiful Soup library as well as the URL request library, I imported the source code to python. From there, I got rid of the unnecessary HTML tags, and therefore got only the paragraph string text, where I managed to get the interview comments. I then compiled all the text into a list, and split them to be a long string.
Using that string, I managed to make a frequency table of the words using a dictionary in my code, and finally sorted them in ascending order, using the sorted function on values.

As for a design decision, a big one for me was to either write the parsed text into a new document, where I could analyze the frequency separately and store the text in a document, or to put them in a list, where I could work with it in python.

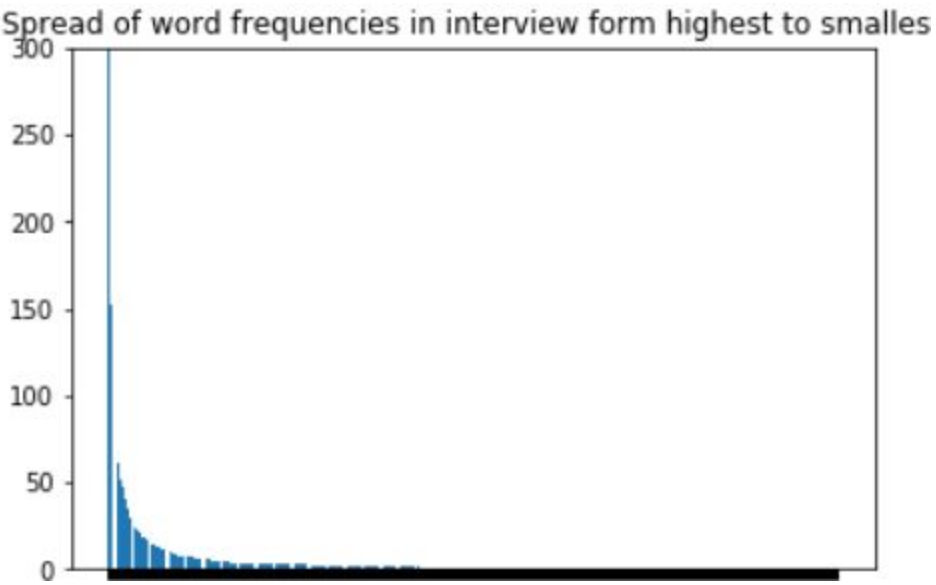
I opted for the list, as it was easier to keep to the python IDE, without using external documents, and it saves less time and data.

RESULT

As for the result, apart from the usual prepositions and conjunctions, the most used “action” words used in soccer pre match interviews by successful managers are “we”,”can”,



“want”,“try”,”work”,”game”,”will”,”do”,”players”,”think”,”win” and “good”. These words imply motivation and try to “hype” players, talking about the game itself, and the words try to stay positive, with “win, going, could”, etc They also talk about the fans by using “fans, they , these,tv”, but don;t forget to allude to the game itself with “game, league, play, schedule”, etc



It looks like these managers seem to hype their teams and use a lot of motivational words, as well as talk about the specifics of the game itself, and finally talk about what their team will attempt to do in the game, focusing on positive actions.

ALIGNMENT

My original idea was to find out what keywords managers commonly use when giving out pre match interviews or speeches, which eventually lead to my success, and I think my program satisfactorily manages to do that .The thing is unfortunately, I ended up with a lot of conjunctions and prepositions, that I should have expected, but don’t provide relevant information to me.

It was interesting to see that managers actually share a lot of words, hence the highest count for some specific words.

As for what sparked my interest, I was very curious to see if there was any link to success on the pitch, as a result of words said by managers. Being a soccer fanatic, things like this naturally came to me, as I had a lot of information from all the sources, and what they have done in the soccer industry. Regardless, I feel I didn’t have as much information as I wanted, as there were only so many websites with interview transcripts available, which was a natural limitation on it’s own.

The tools I used managed to serve its purpose, helping me easily pars text online ,and giving me a clean data source to work with, and I’m confident my program gave an accurate representation of the number of words used in match interview, as it seems to make sense, since the words are very commonly used in interviews and related to soccer.

REFLECTION

As for what went well, I feel like I understand how all of my code works, and it makes logical sense. I also arranged the code so it looks neat and commented and uses docstrings where necessary, as opposed to before. As for what I could improve, I feel like I could've used classes next time for efficient data processing, as well as going more in depth of data analysis to find an even more interesting insight.

As for my unit testing , I rather decided to implement my code stage by stage, and printed the result out and made sure it was what I expected, as there was no way to predict what would come out of the data, to ensure no error was made, and it served me well. I learned conclusively how to text scrape and mine, and going forward, I feel like I can comfortably use packages and modules to parse data online, and do relevant analyses with that data.

As for what I wish I knew, I wish I was more comfortable with the packages, and had more time to implement an even better project. I also wish I tried other data sources such as Twitter, Wiki, but that wasn't the case, because of the time to get the API keys.