Jonathan Jacobs

I did a sentiment and subjectivity analysis, as well as a word frequency analysis on 9 books obtained through Project Gutenberg. The books were the most popular book of every third year starting in 1900 and ending in 1924. As this spans WWI, I aimed to learn if there were changes of sentiment, subjectivity, and diction in popular culture after the war. My main assumption being that popular, contemporary literature of the 1900's captures the zeitgeist of the time.

Two files contained the code for my text analysis. The first program, BookDownload.py, uses pattern to download the nine books. It saves them as strings in a dictionary with the keys being the names of the texts. That dictionary is then pickled into a file to be utilized by the other program, which does the analysis.

The second program, analyzeTexts.py, first loads the texts obtained in the first program. Then, the Project Gutenburg introduction and licensing are removed, just leaving the novel itself. From there, there are two paths of the program, one being the sentiment, subjectivity analysis. The sentiment, subjectivity analysis splits the book by '?', '!', and '.', which yields individual sentences or phrases. Another option would've been to split the text line by line, but by doing so would inevitably mean that phrases or thoughts would've been separated at odd points, likely yielding odd results from the sentiment analysis. Pattern's sentiment() function is used to analyze each sentence, and the results are averaged to yield the sentiment and subjectivity for the entire book. Although it my seem brutish to only look at the sentiment of the entire novel rather than in parts, having a unified number removes variability due to plot and hopefully provides information on the general feeling of the book. Finally, data from the sentiment analysis is stored in a .txt file to be looked at later.

The word frequency analysis is relatively simple. Rather than splitting the string by new lines, it is written to a file and then the file is read in line by line. There don't appear to be any practical reasons for doing this other than my personal learning how to read and write text files. The word frequency was then done line by line using a dictionary with keys as the words and the value the word frequency.

Results for the sentiment, subjectivity analysis didn't have remarkably meaningful results in relation to WWI. In chronological order:

'toHaveAndHold': (0.04913671617843767, 0.26635317779823553)
'theCallOfTheWild': (0.030089685322022774, 0.3157660215022135)
'theJungle': (0.023910898101177237, 0.30671207286016927)
'thePromiseOfAmericanLife': (0.10337965638888513, 0.3779599520231871)
'theJustandtheUnjust': (0.02454693661501242, 0.25736292955923684)
'theTurmoil': (0.047906985365550346, 0.2643599693501376)
'theMajor': (0.06558307383634716, 0.27680332477802067)
'MainStreet': (0.051719856173551126, 0.2741573945644921)
'PlasticAge': (0.05142029128027152, 0.31085377087564703)

Not surprisingly the novels as whole works are not remarkably positive or negative, and there doesn't appear to be any meaningful trend surrounding WWI. However, in comparison to each other, I hypothesize that the small variations allow me to come to conclusions about how positive or negative the novels are.

Looking at the extremes, The Promise of American Life, and The Just and the Unjust, the results appear to make sense. The Promise of American Life is an idealized book about the future of American politics. As Herbet Crowly put it, the novel “offers a manifesto of Progressive beliefs.” The Just and the Unjust, on the other hand, is a fire and brimstone novel, certainly not an overwhelmingly positive book.

An interesting point to note about the subjectivity of all of the novels is that they are all very similar and not very subjective. I would argue this speak to the general way in which novels are narrated. I wouldn't expect plot narration to be subjective, and all of the novels except The Promise of American Life are fiction. The Promise of American Life is political commentary, which interestingly and expectedly is more subjective than anything else.

Similar to the sentiment analysis, the word frequency analysis didn't have significance in relation to time. It did, however, give an interesting peek into the content of the novels. For example, in The Call of the Wild, words like trail (used 41 times), dog (used 112 times), river, (used 17 times), saw (used 40 times) project (used 87 times), life (used 63 times) and men (used 74 times) give a sense that the novel is a rugged outdoor tale involving dogs. Unsurprisingly, that is exactly what it is.

From a process point of view I think that things went smoothly. Particularly, doing iterative coding where I continually added complexity to simple functionality helped me not encounter too many awful bugs. The area that needs improvement the most is unit testing my code. I found functionality like splitting up the large strings into lines incredibly hard to unit test. I wasn't sure they worked properly until I had to rely on them for other purposes. I do think my project was appropriately scoped for the time frame we had. I had enough time to develop a thorough understanding of all of the concepts I was working with. However, if I could've known more about a certain subject before beginning, that subject would be dealing with files. I think my code could have been much more elegant if I had a better understanding from the beginning.