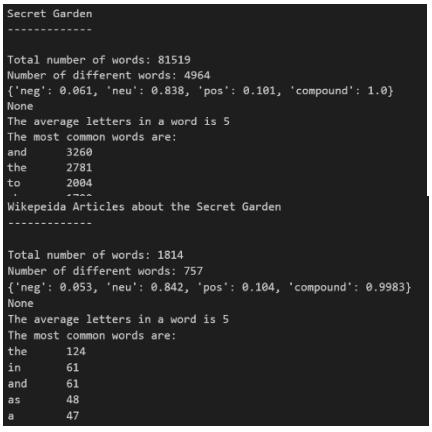
Assignment 3

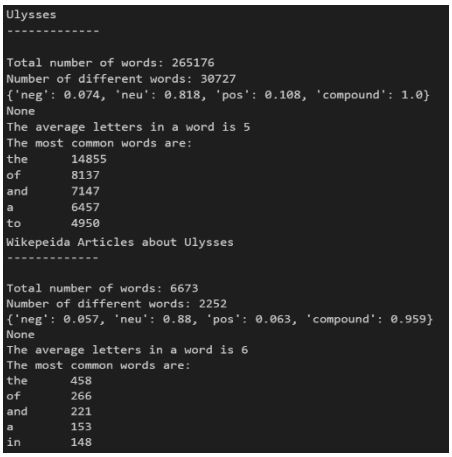
As part of the assignment, I decided to analyze four pieces of text: 2 books and 2 Wikipedia pages. The two books were sourced from Project Guttenberg. The first book was the Secret Garden, written by Burnett, and noted as a classic of English children's literature. The next book was Ulysses, which has been heralded as the most difficult novel to read. As a point of comparison, 2 Wikipedia pages (the novels' respective pages) were analyzed to assess the level of writing. In order to process these bodies of writing, I printed the contents and saved them into text files.

The points of comparison between the four texts were built into seven functions. The first two functions work together to process the text file. The functions read the text file and extracts the punctuations / whitespaces. As for the Gutenberg texts, one of the functions skips the standard header just before the text. To better analyze the text, the code compared total words / unique words, cataloged word frequency in a dictionary, selected random words, printed word count and calculated sentiment. After writing the functions, I called them for each respective text.

One major decision that I had to make related to calculating character count. The purpose of creating the character count function was to later determine the average letters in a single word. I had to decide between using the lens function vs. a for-loop with a count value. The lens function would require the entire text to be placed into a string. In addition to requiring that done, the output of that function would not be an integer value. An integer value is needed as the value will be divided. The path that I decided to take was a for-loop that adds to the count value for every character in the text file.



The most important information gained from the text analysis code were the total number of words/ unique words, the sentiment analysis, the average word length and the most common words. The results of both the secret garden and its respective Wikipedia page are included on the side. The novel has exponentially more words than the Wikipedia page, due to the nature of the text. However, the Secret Garden repeats select words several times, seen through the 6:100 unique word to total words ratio compared to the 42:100 ratio for the Wikipedia page. In both the Secret Garden and the Wikipedia page, the top five most common words represent a majority of the body of work. The Secret Garden's five most common words make up 14% of the entire novel whereas the Wikipedia page's five most common words make up 18% of the entire novel. In both pieces of writing, the average character count per word is 5. Lastly, it is interesting to note that the sentiment analysis for the Secret Garden is strikingly similar to that of its Wikipedia page, where both are slightly skewed to positive.



Next, I reflected the same analysis for Ulysses and its Wikipedia page. Similar to the earlier analysis, the novel has exponentially more words than the Wikipedia page. Ulysses is a stronger and more difficult body of work seen through the 12:100 unique word to total words ratio compared to the 6:100 ratio of the Secret Garden. The ratio for the Wikipedia page is 34:100. Similarly, to the initial analysis, in both Ulysses and the Wikipedia page, the top five most common words represent a majority of the body of work. Ulysses' five most common words make up 16% of the entire novel whereas the Wikipedia page's five most common words make up 19% of the entire novel. In Ulysses, the average character count per word is 5, which seems to be a general average. However, the Wikipedia page has an average word length of 6. Lastly, it is interesting to note that the sentiment analysis for Ulysses is strikingly similar to that of the Secret Garden and its Wikipedia page, where it's slightly skewed to positive. The Wikipedia page has a generally neutral sentiment with a similar negative and positive concentration.

To conclude, I was pleased by the performance of my project. The research helped me conclude that Ulysses was the most sophisticated novel due to the high total words count as well as the high unique word to total words ratio. In addition, the most common words were similar in all bodies of work. I would have ideally developed more variables to better compare the texts. The variables currently used allow me to make a general comparison of the texts. Moving forward, I would research points of comparison that are used by experts in regards to text analysis. I would reflect expert used research in the variables in my code.