Assignment 3

This assignment deals with using `textblob` and other open-source libraries to perform NLP-based analysis on documents using Python. **All parts should use the same three documents (as outlined in Part 1 below). In addition to your .ipynb and/or .py files, you must submit the three documents in .txt format, as well as a report document in .txt/.pdf format that answers various questions below. It is not necessary to submit the .csv file for Part 3, since we will be executing your code. *Just make sure it works correctly!***

Part 1:
Select and download three texts of your choosing that represent different media or writing formats (for example, you could choose i. a novel, movie script, and play script or ii. a short story, poem, and novel, etc.) **Make sure you briefly descibe your documents and explain the difference between them in a paragraph.**

Part 2:
(a) Compute word counts for each of your documents after excluding English stop words (and optionally, performing lemmatization and/or other preprocessing that you would like to employ).
(b) Create and display a bar plot for each document that include word counts for the 25 most frequent words (after the above processing).
(c) Create and display a word cloud for each document (using a mask image of your choice) that includes only the 100 most frequent words. Note that you'll likely want to use the approach outlined in Session 25 that utilizes the `fitwords` method, since you will want data consistent with those for part (b).
(d) Do you see any notable difference between the documents wrt (b) and/or (c) above? Try to explain why or why not, and whether or not these results are expected.

Part 3:
(a) Using your approach from **Part 2**, compute the 25 most *cumulatively commmon* words across the three documents, along with the *cumulative counts*. Remember that a given word can appear in 2 or even all 3 documents.
Ex: if the word "spider" appears 10 times in document 1, 6 times in document 2, and 5 times in document 3, its cumulative count will be 21.
(b) Create a CSV file named **MCW.csv** with the following specifications: i. The csv file should use the standard delimiter (,)
ii. The first row in the file should be a column header row denoted by the string "Word,Count"
iii. The next 25 rows should be populated with the pairs of the 25 most cumulatively common words and counts, in descending order by count.
iv. One final row should added of the form "Sum,(totalcount)" where (totalcount) represents the sum of the top 25 cumulative counts.
A sample csv file is included to give you an idea of what to generate in practice.

Part 4:
(a) Use **Textatistic** to compute the *average* of the Flesch–Kincaid, Gunning Fog, SMOG, and Dale–Chall scores for each document.
(b) Are there noticeable differences among your documents's readability scores, and would you expect these differences (or lack of differences, if there are none) to be present among documents were you judging their readability manually?

Part 5:
(a) Use spaCy to compute the pairwise similarity between your documents (i.e. doc. 1 to doc. 2, doc. 1 to doc. 3, doc. 2 to doc. 3).
(b) Do any of these similarity scores seem higher or lower than you would expect? Explain your response.