Additional Questions

Student: siyli424 (Siyu Liu)

Student: jinya425 (Jin Yan)

Task1 - A simple text extraction script

Q1: In what way did you "clean up" or divide up the text into words (in the program; the text files should be left unaffected)? This does not have to be perfect in any sense, but it should at least avoid counting "lord", "Lord" and "lord." as different words.

We use regular expression to divide the text into words. In `text_stats.py`, we use `r'[a-zA-Z]+'` as the default exp to find all words and ignore all other non-alphabet values, such as digits. We make it as a parameter of the function, so we can custom the definition about 'what is a word', you can see the usage of this parameter in task 2.

We also convert all the words into lower case to avoid counting "lord", "Lord" and "lord." as different words.

Q2: Which data structures have you used (such as lists, tuples, dictionaries, sets, ...)? Why does that choice make sense? You do not have to do any extensive research on the topics, or try to find exotic modern data structures, but you should reflect on which of the standard data types (or variants thereof) make sense. If you have tried some other solution and updated your code later on, feel free to discuss the effects!

For letter stats, we use numpy array because the `unique` function is powerful and easy to use. We then transform the numpy array back to tuple to maintain a consistent, simple data output format. The final output is a tuple of tuples. Tuple is easy to sort, slice and simple. We also don't have a need to index frequency by word. So we belive tuple is a good choice in this task.

For word stats, the words frequency part is same as letters. In the words chain part, we use use a nested dictionary to store the frequency of each word following another word because dictionary is easy to access values by keys which is the target word we will use in task 2.

Task 2 - Text generator

We add another txt file about the 'The Lord of the Rings' for testing. Please check the `example_output` folder, all parameters, time cunsumption and output are saved there.