Caroline Cutter

CS150 Pre-Lab 8

Functions:

- **read_corpus(filename)**

This function is going to take a file that consists of some sort of writing (like a paragraph) and it is going to read the data, make all the letters lowercase, split it into a list of separate words and then remove any sort of extra punctuation that does not effect the actual word itself.

Args: filename: the name of the file that the words will be extracted from

Returns: Returns a list of words from the file

- **count_and_rank(word_list)**

This function is going to take in a list of words and return a tuple of the words and the count of their words from most appearances to least appearances.

Args: word_list: a list of words that are going to be counted and ranked

Returns: a tuple of words and their counts ranked from highest to lowest.

- **count_rank_table(rank_tuple)**

This function is going to take the tuple from count_and_rank and it is going to print out the table desired in the lab so there will be a word and count column in the table

Args: The count_and_rank tuple for information

- **zipfs_law(frequency, rank)**

This function is going to compute zipfs law on a word given the frequency and rank. This function will be implemented when we begin to graph the function

Args: frequency: the frequency of the word
rank: the rank of the word

Returns: the value of zipfs law of a specific word

- **zipfs_graph(filename)**

This function will graph Zipfs law using the data from the previous functions. So, it will use the read_corups function to obtain the list of words and then it will rank and count them using the count_and_rank function. Then, the zipfs law will be calculated using the zipfs_law function, then using matplotlib, we can use the data to graph and label all axis

Args: filename: the file where the information will be extracted

Returns: none, just prints table of top 10 ranked words and shows the graph

Other Aspects:

In this lab, I am going to try to avoid long functions and try to make them as simple as possible. I am going to try to make everything work within each function, but then I am going to revise it to make it more concise. The flow of my program will be based off the flow of the functions described above.

I am going to use the if __name__ = "__main__" function and the sys module to make everything run from command line arguments, but also show the difference between when it is being run vs imported.