CSE101 – Spring 2021
Programming Assignment #6

Due May 13, 2021 by 11:59pm, KST. The assignment is worth 12 points.

Instructions

For each of the following problems, create an error-free Python program.

- Each program should be submitted in a separate Python file that follows a particular naming convention: Submit the answer for problem 1 as "Assign6Answer1.py" and for problem 2 as "Assign6Answer2.py" and so on.
- These programs should execute properly in VS Code using the setup we created in lab.
- At the top of every file add your name and Stony Brook email address in a comment.
- Include all the provided test cases **and write 1 additional test case** of your own for all problems that do not take user input.
- Download these supplemental files – unzip the files and add them to your project directory. The problems will refer to specific files that you will load data from.

Regarding working in pairs:

- You are welcome to work with a partner on the homework assignment, but you **MUST write both your names and email addresses in each file** in a comment. Only one person needs to submit the homework on Blackboard.
- You are only allowed to work together with one other person – larger group submissions or collaborations (beyond high-level discussions of problems, as stated on the syllabus) are not allowed.

Problem 1: (4 points)

Write a function `fourMostCommonCharacters` that will take a string as a parameter and prints on one line the top four most common characters in the string, along with how many times each character occurs. You should print the character that appears the most common character first, followed by the second then third then fourth most common character. If multiple characters occur the same amount, you can choose your own ordering between those characters.

You should make the string uppercase using the `upper()` method, so that it will count upper and lowercase letters together (e.g. "Google" has 2 "G" characters when it is made uppercase). You can assume the string always has at least four distinct characters. Hint: you may find it useful to use a dictionary to keep track of the word count of each letter in the string.

Some example test cases:

```
fourMostCommonCharacters("Hyundaimotorcompany") # prints O: 3 Y: 2 N: 2 A: 2
fourMostCommonCharacters("Mississipi")          # prints I: 4 S: 4 M: 1 P: 1
fourMostCommonCharacters("Daimlerchrysler")     # prints R: 3 L: 2 E: 2 D: 1
```

Problem 2: Iris Dataset (3 points)

Refer to the file `irisdata.txt`. This data set is commonly used in machine learning experiments and consists of 50 samples from each of three species of Iris flower (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Write a Python program to read the data from the text file and print the averages of the sepal length, sepal width, petal length, and pedal width for all three types of Iris flowers.

Expected Output:

```
Class Iris-setosa:
Average sepallength: 5.006
Average sepalwidth: 3.418
Average petallength: 1.464
Average petalwidth: 0.244

Class Iris-versicolor:
Average sepallength: 5.936
Average sepalwidth: 2.77
Average petallength: 4.26
Average petalwidth: 1.326

Class Iris-virginica:
Average sepallength: 6.588
Average sepalwidth: 2.974
Average petallength: 5.552
Average petalwidth: 2.026
```

Problem 3: Text analysis (5 points)

Automated authorship detection is the process of using a computer program to analyze a large collection of texts, one of which has an unknown author, and making guesses about the author of that text. The basic idea is to use different statistics from the text – called ”features” in the machine learning community – to form a linguistic ”signature” for each text. For example, one basic linguistic feature utilizes the number of unique words. Once a signature is obtained, comparisons can be made to known authors and the text of their works.

In this problem, you will write a program to analyze some text files to determine three base-level linguistic features:

1. the average word length,

2. the Type-Token Ratio which is the number of different words divided by the total number of words (this is an indicator of how repetitive the text is), and

3. the Hapax Legomena Ratio which is the number of unique words (words only used once) divided by the total number of words.

We also want to see the most frequent words for the text to see how they differ by text. However, if we don’t do any additional text processing, this list will mainly be commonly used words across all texts such as “the”, “to”, “a”, and “I”. One standard approach to this issue is to keep a list of the common words that you want to exclude and then ignore them.

We will take a simpler approach and just look at the 10 most frequent words that are at least 5 characters long. This will exclude a lot of the common connecting words and be easier to implement.

Write a function called `text_features` that takes a file name as its parameter and output those base-level linguistic features and 10 most frequent words (and the count for how many times each word was used) as follows. With '`file.txt`' as an input file your program should produce the following output. (Actual numbers may vary slightly depending on how you handled your data – I reduced all the strings to lowercase and removed punctuation from any words, similar to the spam filtering problem. I also made sure to ignore any blank strings.)

```
Total Words 33
Average Word Length 4.15
Number of Different Words 25
Number of Words Used Once 20
Type-Token Ratio 0.7576
Hapax Legomena Ratio 0.6061

Most Frequent Words (word, count):
confidence 2, cicero 1, tullius 1, marcus 1, started 1, before 1,
defeated 1, twice 1
```

Assume that words in an input file are separated by the usual white space characters.

You will want to write some helper functions that will be useful to write `text_features`. Some possible helper functions would be:

- `get_average_word_length` which returns the average length of the words contained in the file as a real number (float).
- `get_number_of_different_words` which returns the number of different words (where duplicate words are counted as one word).
- `get_number_of_unique_words` which returns the number of words that appear exactly once (duplicate words are not counted).
- `get_most_frequent_words` which returns the most frequent words

Note that I did not specify what these functions take as their input parameter(s). It depends on how you design your program. There are several possible designs that you could consider. One possibility would be to create a dictionary and pass it to these functions. Another possibility would be to create a list and use it rather than a dictionary. With a list it might be a little easier to write the first function (`get_average_word_length`), it will be a lot easier to write the other functions with a dictionary.

You are provided with a test file, `file.txt`, and two "mystery" files, `mystery1.txt` and `mystery2.txt` to analyze. The mystery files were obtained from Project Gutenberg, a website with text files of free public-domain books. Below is sample output from my solution for the two mystery files.

mystery1.txt Output:

```
Total Words 163401
Average Word Length 4.05
Number of Different Words 10896
Number of Words Used Once 5597
Type-Token Ratio 0.0667
Hapax Legomena Ratio 0.0343
Most Frequent Words (word, count):
there 767, which 656, could 493, would 428, shall 427, helsing 299,
before 277, again 246, seemed 243, about 239,
```

mystery2.txt Output:

```
Total Words 78012
Average Word Length 4.39
Number of Different Words 7693
Number of Words Used Once 3972
Type-Token Ratio 0.0986
Hapax Legomena Ratio 0.0509
Most Frequent Words (word, count):
tuppence 579, tommy 533, there 326, julius 295, would 236, about 213,
don't 189, james 155, think 154, right 142,
```