Engineering & Computer Science
Department of Computer Science

CS161 Introduction to Programming and Problem Solving

Due March 9

Exercise 5

This homework is designed to highlight the use of exceptions and simple text files within python. Other skills learned through the course will also be utilized in developing the solution. You are not allowed to use any Python-provided method or function named count() for this assignment, although you may write your own.

Students will write a program that:

1. Accepts as user input the name of a file to process. If the named file does not exist, appropriate error handing will occur and the name of the file will be again requested. This will repeat until either a valid file name is entered or the string "ALL DONE" is entered as the file name. The program will assume that the named file is an ordinary text file and not a pickled file. Use the readline() function, not readlines(), to read the file.
2. The file will be processed ONE LINE AT A TIME:
    a. Certain characters will be removed from the input. Using the string module, the following statement will define which characters to remove:
    (string.punctuation + string.whitespace).replace(' ', '')
    Note that the removal of a character should leave a space in its place and not change the length of the input line. E.g., the string "12.34" would become the string "12 34" not "1234".
    b. Once the specific characters are removed from the input, the remainder of the input line will be split on "word" boundaries where words are separated with the space character (' ').
    c. Each "word" of the processed input will be stored in a dictionary as a key where the value is the number of times that the word occurs in the input. However, if the "word" is an integer, the word will not be stored in the dictionary and instead will be summed so that the total of all integers in the processed file can be displayed.
3. One the file has been processed, the following information will be displayed:
    a. The total of all integers in the file
    b. The 5 most common words in the file
    c. The 5 least common words in the file.

Note that it is very likely that there will be more than 5 words that are the "least common" in the file. In this case, you should print any 5 of those least common words. For example, if there are 7 words with a frequency of '1', then listing any 5 of them will suffice, but only list 5.

Example output:

Enter filename for analysis: HW 5 gettysburg.txt

The total value of the integers: 1882

Top 5 most common words:
that: 13
the: 11
we: 10
here: 8
to: 8

5 least common words:
usa: 1
should: 1
full: 1
consecrate: 1
fathers: 1

>>>