CS410P

LAB 16L

Purpose

The purpose of this assignment is to give you practice with dictionary objects

Problem

There is a data file called Gettysburg.txt available in the public folder for this lab which contains the text of the Gettysburg address. Our goal is to create and use a dictionary object as a way to figure out the word frequency in the Gettysburg address.

Details

- We want to read the contents of the data file into a string or a list
- We want to treat uppercase and lowercase alphabets the same i.e. while figuring out how many times a specific word occurs in the Gettysburg address it should not matter whether it's in upper case or lowercase.
- We also want to strip out any punctuation or special characters from the words, for example if word has a "," or ":" or any other non-alphabetic character in it, we want to remove them before we use the word in the dictionary
- After making the word case insensitive, and stripping out any character that's not an alphabet, we want to store every unique word in the dictionary as a key (you can make the initial value of each key zero or one based on how you write your program)
- Every time you encounter an occurrence of the same word you need to increment the value for that particular key in the dictionary by one
- After you go through all the words in the Gettysburg address your dictionary object will contain a key for every unique word and the value for a key is the number of times that word occurs in total
- Print all the words in the Gettysburg address that appears at least 3 times
- Sort the way the most frequently occurring words are printed (i.e. all words that occur more than 3 times),

Input

The input data file consists text which is the Gettysburg address

Output

The output should indicate how many total words are in the address, how many unique words there are, and what are the words that occur more than 3 times (including their frequency of occurrence). Sample output is shown below.

```
The Gettysburg Address contains 268 words.
The Gettysburg Address contains 139 different words.

The most common words and their frequencies are:
    that: 13
    the: 11
    we: 10
    here: 8
    to: 8
    a: 7
    and: 6
    it: 5
    have: 5
    for: 5
    nation: 5
    of: 5
    in: 4
    this: 4
    dedicated: 4
    dead: 3
    so: 3
    people: 3
    who: 3
    are: 3
    shall: 3
    they: 3
    is: 3
    us: 3
    cannot: 3
    great: 3
```

Grade Key

A	Comments (including Name, brief description about program)	**5**
B	Reading input data correctly	**10**
C	Words are made case insensitive	**5**
D	Words are stripped of punctuations	**8**
E	Unique words stored correctly as keys in a dictionary	**22**
F	Dictionary value for each key indicates frequency of occurrence	**20**
G	Output accuracy (total words: 5 points, total unique words: 5 points, words occurring 3 or more times is accurate 10 points)	**20**
H	Frequently occurring words sorted in descending order	**10**