Assignment 6: Analyzing a Novel

Due Wednesday by 11:59pm　　**Points** 100　　**Submitting** an external tool
Available Oct 26 at 12am - Nov 3 at 11:59pm

Assignment Goals

The goals of this assignment are to:

- Practice with lists and dictionaries
- Work with files
- See how computers can be used in the humanities

Getting Started

Start by downloading the starter code: **A6.zip** ⤓
(https://utah.instructure.com/courses/796373/files/134138663/download?download_frd=1) . This project
has a .py file for code and then several text files that will be used by your working program.

The .py file is broken into several sections:

- Some already implemented functions at top
- Some functions that you need to implement. You should look at the already implemented
  functions for ideas, as well as lecture slides.
- A main function that runs when the .py file is run. It loops over several novels and calls
  functions to gather statistics and report on them.

If you run the starter code, it should run, but produce erroneous results.

Assignment Activities

Implement the required functions. Your results should match the results shown at the bottom of
this document. Some of the required functions call other functions you need to implement - think
about the purpose of each function as you implement it.

Then, when you are satisfied with your code, test your understanding of the overall purpose of
the code by answering the following questions (a short paragraph per question). You should

neatly write up the questions and answers in a word processing document. Title the document with the assignment and add your name and uID. After answering, convert your document to pdf for submission.

Questions

1. The Plato text has the most positive sentiment. Looking at other data presented about this text by the program, why do you think it is estimated by the computer to be so positive?
2. Looking at the "Chunk Sentiment" results for HuckFinn and TaleofTwoCities, what can you guess about the style of the story or plot of each novel? You should not refer to the actual stories, just the numbers of the chunks.
3. Looking at the most common words for TheScarlettLetter, what might you guess a theme of the book might be? (But don't get fooled by the word 'pearl' - that is a name forced to lowercase, so this is not a book about jewelry.)
4. Looking at the most common words for all the books, which one do you think has the most informal writing style? Why?

Submitting

Submit your NovelAnalysis.py file and the pdf with your written work by the due date.

Sample Output

My solution produces the following output:

```
Analyzing: HuckFinn.txt
Getting words from book
    114144  words found
Removing common words
    46130  words remaining
Calculating Sentiment
    Overall: 1.57
    Chunk Sentiment: 0.04 0.56 1.6 1.78 0.13 1.47 2.82 2.38 1.6 1.47 1.95 1.78 2.47 1.82 1.
17 2.04 1.04 0.43 2.56 2.35
Average word length: 5.2
Counting most common words
    Most common words: [('says', 637), ('got', 622), ('said', 481), ('come', 390), ('didnt'
, 342), ('jim', 341), ('dont', 335), ('time', 315), ('right', 297), ('went', 295)]


Analyzing: TheScarletLetter.txt
Getting words from book
    86543  words found
Removing common words
    37849  words remaining
Calculating Sentiment
    Overall: 1.03
    Chunk Sentiment: 2.01 6.39 2.48 1.74 2.17 -1.11 -3.01 1.06 3.49 3.33 -1.95 2.43 -0.37 -
```

```
2.54 -3.12 -0.11 3.54 2.11 -0.21 2.34
Average word length: 6.5
Counting most common words
    Most common words: [('hester', 354), ('little', 234), ('thou', 230), ('pearl', 215), ('
old', 214), ('man', 181), ('child', 177), ('minister', 153), ('life', 153), ('said', 148)]


Analyzing: Plato.txt
Getting words from book
    218836  words found
Removing common words
    82577  words remaining
Calculating Sentiment
    Overall: 5.08
    Chunk Sentiment: 4.82 3.61 7.68 4.36 5.38 2.01 3.95 4.55 5.72 6.95 6.39 2.18 5.98 7.22
5.69 7.12 6.32 4.02 5.64 1.96
Average word length: 6.5
Counting most common words
    Most common words: [('said', 1117), ('state', 684), ('good', 657), ('man', 599), ('true
', 585), ('like', 477), ('yes', 466), ('men', 426), ('say', 420), ('life', 390)]


Analyzing: TaleOfTwoCities.txt
Getting words from book
    135815  words found
Removing common words
    55936  words remaining
Calculating Sentiment
    Overall: 0.14
    Chunk Sentiment: 0.54 0.11 -1.29 -0.5 2.15 0.46 2.04 3.43 -1.79 2.9 3.0 -1.39 -1.22 -2.
5 -1.18 0.11 -0.75 -0.72 -1.07 0.39
Average word length: 6.1
Counting most common words
    Most common words: [('said', 660), ('mr', 620), ('lorry', 336), ('defarge', 280), ('man
', 279), ('little', 265), ('time', 260), ('hand', 247), ('miss', 232), ('know', 229)]
```