Assignment 2 – Textual Analysis

Design your solution in Pseudo-Code*

- Here is my design (You can have your own)
  1. Read from ‘pos_list.txt’ and get a list of positive words
  2. Read from ‘neg_list.txt’ and get a list of negative words
  3. Setup a Dictionary of positive & negative words for counting
  4. Ask user for the article filename (e.g., Joyful.txt)
  5. Read from the article file and get a list of article words
  6. Find out the number of words in the article
  7. Collect positive & negative word statistics using the Dictionary
  8. Print out the word counts in the Dictionary
  9. Print out the Analysis Report

***Note**: *What’s a Pseudo-Code? Pseudocode is a method of planning which enables the programmer to plan without worrying about syntax. It is a term which is often used in programming and algorithm based fields. It is a methodology that allows the programmer to represent the implementation of an algorithm. ... It has no syntax like any of the programming language and thus can't be compiled or interpreted by the computer.*

Assignment 2 – Textual Analysis

- Here is my design (You can have your own)
  1. Read from ‘pos_list.txt’ and get a list of positive words
  2. Read from ‘neg_list.txt’ and get a list of negative words
  3. Setup a Dictionary of positive & negative words for counting
  4. Ask user for the article filename (e.g., Joyful.txt)
  5. Read from the article file and get a list of article words
  6. Find out the number of words in the article
  7. Collect positive & negative word statistics using the Dictionary
  8. Print out the word counts in the Dictionary
  9. Print out the Analysis Report

Assignment 2 – Textual Analysis

Further Refinement

- We observe that Step1, Step2, & Step5, are very similar.
- Write a function `File2List(filename)` that:
  - Read line by line from a file
  - For each line of words,
  - change all characters to lowercase
  - remove all the punctuations (see the note on next page)
  - remove whitespaces except the blank – ‘ ’
  - split the words within the line by the blank and form a list of separated words
  - Merge all the list of separated words together and the function should return a wordlist of individual words

Assignment 2 – Textual Analysis

A note on the punctuation ***apostrophe (’)*** & word count

- The apostrophe (’) should be replaced by a blank instead of just removing it – replaced by a ‘ ’.
- It is because “Who’s” and “He’s” should be counted as TWO words as of “Who is” and “He is” or “He has”.
- But then I also found out that sometimes the apostrophe is used in abbreviation, like “Int’l Conference” for “International Conference”
- or used to indicate someone’s possession like “Mike’s car” or “Joe’s father”.
- And in such cases, it should be treated as a single word.
- Hence, for a simple program like our Assignment #2, I decided to take an easy way out by removing all the punctuation within the article (i.e., replacing all punctuation by ‘ ’ instead of replacing the apostrophe by a blank.
- Anyway, I have told the TA that both answers will be considered as correct!
- I personally have thanked your classmate in pointing this out because he is interested in getting the assignment right and even with the very fine details of it.

Assignment 2 – Textual Analysis

- Here is my design (You can have your own)
  1. Read from ‘pos_list.txt’ and get a list of positive words
  2. Read from ‘neg_list.txt’ and get a list of negative words
  3. Setup a Dictionary of positive & negative words for counting
  4. Ask user for the article filename (e.g., Joyful.txt)
  5. Read from the article file and get a list of article words
  6. Find out the number of words in the article
  7. Collect positive & negative word statistics using the Dictionary
  8. Print out the word counts in the Dictionary
  9. Print out the Analysis Report

Assignment 2 – Textual Analysis

Further Refinement

- We need to set up a Dictionary to keep track of the statistics for the positive and negative words
- We use the following construct:
  - **`Mydict = {“keyword1: 0”, “keyword2:0”, …, “keywordN:0”}`**
  - where *keyword* can be a positive or a negative word
  - the value associate with the keyword is initialized to zero first
  - this value acts like a counter that will increase by one whenever the keyword is read from the file.
  - Hence, this value will keep track of the number of times this keyword appears in the article.

Assignment 2 – Textual Analysis

Extra Information on positive / negative words and adjectives

- This assignment is a simplified version of sentiment analysis.
- You should do a Google search on “sentiment analysis” if you are interested at this kind of Textual Analysis in Data Science.
- You should also do a Google search on “positive words”, “negative words”, “positive adjectives” and so on.
- I have included some lists of words and adjectives for your reference.
- And for the articles for testing, you can search for sample article databases, or song lyrics, quotes, words from greeting cards, or sympathy cards, …etc.