Koç University
COMP201 – Computer Systems and Programming
Assignment 2 (Due Date: 11.11.2020)

C – Strings
Contact TA: Ahmed Imam Shah
ashah20@ku.edu.tr

Movie Review Sentiment Analysis

Sentiment Analysis is a Big Data problem which seeks to determine the general attitude of a writer given some text they have written. For instance, we would like to have a program that could look at the text "The film was a breath of fresh air" and realize that it was a positive statement while "It made me want to poke out my eye balls" is negative.

The problem that we'll solve in this assignment is about calculating or analyzing sentiments of the critics by their reviews about the movie. You are going to search through a file containing movie reviews from the Rotten Tomatoes website which have both a numeric score as well as text. You'll use this to learn which words are positive and which are negative. The data file looks like this:

```
1 A series of escapades demonstrating the adage that what is good for the goo
4 This quiet , introspective and entertaining independent is worth seeking .
1 Even fans of Ismail Merchant 's work , I suspect , would have a hard time s
3 A positively thrilling combination of ethnography and all the intrigue , be
1 Aggressive self-glorification and a manipulative whitewash .
4 A comedy-drama of nearly epic proportions rooted in a sincere performance b
1 Narratively , Trouble Every Day is a plodding mess .
3 The Importance of Being Earnest , so thick with wit it plays like a reading
1 But it does n't leave you with much .
1 You could hate it for the same reason .
1 There 's little to recommend Snow Dogs , unless one considers cliched dialo
1 Kung Pow is Oedekerk 's realization of his childhood dream to be in a marti
4 The performances are an absolute joy .
3 Fresnadillo has something serious to say about the ways in which extravagar
3 I still like Moonlight Mile , better judgment be damned .
3 A welcome relief from baseball movies that try too hard to be mythic , this
3 a bilingual charmer , just like the woman who inspired it
2 Like a less dizzily gorgeous companion to Mr. Wong 's In the Mood for Love
1 As inept as big-screen remakes of The Avengers and The Wild Wild West .
2 It 's everything you 'd expect -- but nothing more .
```

Note that each review starts with a number 0 through 4 with the following meaning:
- 0 : negative
- 1 : somewhat negative
- 2 : neutral
- 3 : somewhat positive
- 4 : positive

Please read the following sections from your textbook in order to do this assignment easily:
- **Chapter 7: Input and Output** from The C Programming Language by Kernighan and Ritchie
- **C Strings** from Essential C

Task 0: Setting up the environment
1. Accept the assignment from the following link https://classroom.github.com/a/qEj84RHr
2. Create a file main.c using "touch" command.
3. Import libraries such as
   - `#include <stdio.h>`
   - `#include <string.h>`
   - `#include <math.h>`
4. Use editor to edit this file and make a main() function.
5. Write a program to read a text file.

Task 1: Average review based on a word (40 points)
You will ask the user to enter a word, and then you will search every movie review for that word. If you find it, add the score for that review to the word's running score total (i.e., an accumulator variable). You also will need to keep track of how many appearances the word made so that you can report the average score of reviews containing that word back to the user.

Some sample runs of the program might look like this:

```
Enter a word: fantastic
fantastic appears 13.0 times.
The average score for reviews containing the word fantastic is 2.8461538461538463
```

```
Enter a word: horrible
horrible appears 11.0 times.
The average score for reviews containing the word horrible is 0.6363636363636364
```

```
Enter a word: ok
ok appears 462.0 times.
The average score for reviews containing the word ok is 1.9502164502164503
```

(Hint: you can use strstr() functions to find the substring in a given string.)

if (strstr(string, substring) != NULL) checks if the substring is present in the string or not. However, if the word is "impractical" and you search for "practical", this function strstr() will return the index of "(im)practical".

- So it will be a good idea to first split the strings into smaller strings using " " (*space*) as the delimiter. Then you can look for the entire word. You can also use any other approach to solve this issue.

- Use *int* datatype to count the frequency as well as score of the given word.

- Display the average score up to 4 decimal places. i.e. display 1.9502156450 as 1.9502

Task 2: Max occurring word in positive and negative reviews (55 points)

1. For this part of the assignment, you need to open wordList.txt. This file contains some words that are frequently found in the movie reviews. You can open two files simultaneous just create two distinct instances.

2. For each word in wordList file count the number of instances it occurred in Positive reviews and number of instances in occurred in negative reviews. Save these values in 2 arrays (one for positive reviews and negative reviews each).
3. Find maximum of these two arrays to find the most occurring word in positive reviews and most occurring word in negative reviews.
4. Print these two numbers.

Note: Consider 2 – neutral as a negative review.

Oral Assessment
Important Note: We use automated plagiarism detection to compare your assignment submission with others and also the code repositories on GitHub and similar sites. Moreover, we plan to ask randomly selected 10% of students to explain their code verbally after the assignments are graded. And one may lose full credit if he or she fails from this oral part.

How to Read a file
You should follow the following steps to successfully read the file:
1. First, open the text file using the *fopen()* function.
2. Then, use the function *fgets()* to read text from the stream and store it as a string. The newline or EOF character makes the *fgets()* function stop reading so you can check the newline or EOF file character to read the whole line.
3. When you are done reading all the contents of the file, close the text file using the *fclose()* function.

Late Submission Policy
- You may use up to 5 grace days (in total) over the course of the semester for the assignments. That is you can submit your solutions without any penalty if you have free grace days left.
- Any additional unapproved late submission will be punished (1 day late: 20% off, 2 days late: 40% off) and no submission after 2 days will be accepted.

Coding Style Evaluation
We have reserved 5 points for a subjective evaluation of the style of your solutions and your commenting. Your solutions should be as clean and straightforward as possible. Your comments should be informative, but they need not be extensive.

Academic Integrity
All work on assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, but these discussions should be carried out in an abstract way. That is, discussions related to a particular solution to a specific problem (either in actual code or in the pseudocode) will not be tolerated. In short, turning in someone else's work, in whole or in part, as your own will be considered as a violation of academic integrity. Please note that the former condition also holds for the material found on the web as everything on the web has been written by someone else. See Koç University - Student Code of Conduct.

Acknowledgement
This assignment is adapted from Stanford University Course CS1 assignment.