**CSCE100 Introduction to Informatics**

**Fall 2023**

**Programming Assignment 2: Processing Text**

(Assigned: September 14, 2023  Due: September 21, 2023)

## Objectives

1. To familiarize with writing and running Python programs and the Python environment
2. To practice computational thinking skills to develop the solution approach
3. To familiarize with the use of loops (e.g., the for and while loops)
4. To familiarize with data structures, particularly arrays/lists
5. To familiarize with file input/output in Python
6. To be exposed to the use of built-in functions
7. To be exposed to the use of built-in modules or packages (e.g., import math)
8. To familiarize with the use of online documentations on Python

## Relevance to Informatics or Data Science

1. A user interface program for collecting data
2. A program for pre-processing or processing data
3. A solution for parsing texts and computing certain properties automatically

## Problem

You are given an input file that consists of lines of texts. In the starter code, it has the code that reads in the input file and stores each line of text as an element in an array. It also has a rudimentary processing that counts the number of characters in each line and sums them to find the total number of characters in the file. This is done using a for loop. Note also that a line of texts can consist of multiple sentences.

Now, given the above starter code, extend the program that prompts the user to enter a search keyword and a search option continuously until the user enters "-1". When the user enters a search keyword, the program will prompt the user to choose whether they want to search within a song's title, a song's lyrics, or both. Depending on this choice, the program then prints out the song in which this key is found, along with the number of occurrences the key is found within that song's title, or lyrics, or both. This process repeats until the user enters "-1".

Please see the example output below to help you envision the specification for this program.

## Additional Requirements

- The program extension part is required to display an explanation of the program (e.g., its expected range of input values) in the beginning before prompting the user for a search key. (5 points)
- The program is required to use an array to store the lines such that each line is an element of the array. (5 points)
- The program extension part is required to use at least one loop structure. (10 points)

- The program extension part should display proper error messages when an invalid input is entered. (5 points)
- You must document your program (see https://devguide.python.org/documenting/).
    - Name, Date, Affiliation, a description of the program, what inputs does it need, what outputs does it generate (5 points)
    - Inline comments in the program (5 points)

**Bonus** (10 points)

Once you have a working solution, extend the program such that it counts the appearance of a search key *regardless* of whether it is uppercase or lowercase. For example, given the current data file, if the user wanted to search "wind" in both the lyrics and song titles, the program would output 2 songs. However, if the user entered "Wind" instead, the program would not output any songs because there are no uppercase versions of this word in the songs' titles and lyrics.

Update your program accordingly to make sure the whole program still works properly to meet all requirements.

Example Session Runs (red texts provided by user)

```
Welcome to the Text Analysis program!
This program finds the songs that a search key is found in using the input file "songs.txt".

Please enter the search key (or -1 to exit):
challenge
Please select where you would like to search (1, 2, or 3):
1) Within a song's title
2) Within a song's lyrics
3) Within both the title and lyrics
1
There are no song titles containing this search key.

Please enter the search key (or -1 to exit):
time
Please select where you would like to search (1, 2, or 3):
1) Within a song's title
```

2) Within a song's lyrics

3) Within both the title and lyrics

3

Number of times the search key appears in a song's title and lyrics:

Bohemian Rhapsody: 3

Party In The USA: 3

Island In The Sun: 2

Out Of Time: 17

Ain't It Fun: 1

Suga Suga: 1

Please enter the search key (or -1 to exit):

love

Please select where you would like to search (1, 2, or 3):

1) Within a song's title

2) Within a song's lyrics

3) Within both the title and lyrics

5

Sorry, this input is invalid.

Please enter the search key (or -1 to exit):

I said

Please select where you would like to search (1, 2, or 3):

1) Within a song's title

2) Within a song's lyrics

3) Within both the title and lyrics

2

Number of times this search key appears in a song's lyrics:

Wonderwall: 5

The Less I Know The Better: 1

Please enter the search key (or -1 to exit):

-1

Thank you for using the Text Analysis Program. Goodbye!

## Handin

1. The submission deadline is 11:00 AM September 21, 2023 (Thursday). Late submissions will not be accepted or graded.
2. Using the Assignment link on the Canvas course website, you are required to submit a screen capture(s) of your "testing session(s)" using your program. (10 points)
3. Using the Assignment link on the Canvas course website, you are required to submit all program files. (10 points)

## Think About

Now, think about what if we want to build a system that computes statistics for thousands of text files. How would we input the files? What would be some common challenges or issues with that approach? By the same token, what if we want to generate different types of statistics, for different subsets of sentences, and thus will produce many different tables? How should we store the tables of sentences? (Hint: Think about Big Data, Scalability, and Reliability, and how they relate to Informatics or Data Science.) Note also that looking for words that start with an uppercase letter is a very valid application in today's text processing, with the purpose of finding place names, pronouns, etc., to automatically metatag documents or texts.