PROGRAMMING ASSIGNMENT 3: TEXTANALYSIS (60 POINTS)

Text analysis (also known as text mining), is the process of transforming unstructured text into structured data for easy analysis. Text analysis is used as the first step in the natural language processing (NLP) tasks, to examine and prepare the text for further processing.
This assignment requires you to develop your own program to perform simple text analysis tasks.

GETTING STARTED

Reading: For this assignment, you need to have a good command of String methods, lists, loops, as well as function definitions. Review the corresponding Handouts, textbook chapters and programming practice problems.

Programming: Start by creating a new Python project or folder in your environment.

To simplify your task, a starter file **hw3_starter.py** is posted with the assignment as well as files to help you test your program while you are developing it: **tiny.txt, short.txt** and **bentley.txt**. You should also develop your own test case files while testing your code.

The **hw3_starter.py** file includes a function `textFromFile()` that returns the content of a file, the headers for each function that you must define, as well as a header for function **`tester()`**, where you should put your testing code. The **tester()** function content demonstrates how to use function **`textFromFile()`.**

You should develop your code in this file; you can rename it as you wish. Do not change the headers for the functions – parameters are a part of the specification that you must follow precisely for full credit.

Once you create a new project , download the **hw3_starter.py** and **the text files** into it. Run it to see how it works in its current state, review it to understand its components. Then start developing the functions. Once you've developed and tested all your functions, leave a single call to main() on the global level. Keep the tester() code in – it is part of the graded code!

REQUIREMENTS

The program outline below is illustrated with the Sample runs that follow.

Program outline

1. Read from the user the name of the text file. Call function **`textFromFile()`** to obtain the text.
2. Replace all end-of-line characters in the text with a space, and print the text.
3. Output a Text Analysis report, as shown in the sample interactions, including, in this exact order:
    a. Total number of sentences.
    b. Total number of words.
    c. Total number of numeric words.
    d. The numbers specified by the numeric words (in sorted order)
    e. Total number of characters.
    f. Average number of words per sentence.
4. Furthermore, ask the user to enter four parameters (these parameters are used in step 5 for

displaying all or part of the text broken into lines and modified, as shown in the sample runs). User should enter:
   a. a keyword to be capitalized,
   b. number of characters per line,
   c. start and end line numbers, to display the text.
5. Output the modified text split into lines of the specified length (4b above), from start line (4c) to the end line (4c), or to the last line, if the end line number is greater than the number of lines. Modification of the text are as follows:
   a. the keyword (case-sensitive) must be capitalized,
   b. all parenthesized within [] expressions must be removed,
   c. all non-space, non-alpha-numeric characters must be replaced with a #,
   d. each printed line must start with the line number formatted as shown, and end with a vertical bar.

Sample runs illustrate this functionality.

SAMPLE RUNS

In every sample run below user input is show in **green boldface.**

Sample run #1

```
please enter the file name: tiny.txt
Today is March 7, 2024.
******************** Text Analysis Report ********************
                   Sentences: 1
                       Words: 5
            Of them, numbers: 2
                 Numbers are: 7 2024
                  Characters: 23
 Average words per sentence: 5
************************************************************

Which word would you like to highlight? March
Enter number of characters per line:5
start line (press Enter for 1):1
end line (press Enter for the last line):10
  1: Today|
  2:  is M|
  3: ARCH |
  4: 7# 20|
  5: 24#|
************************ End of Report ************************
```

Sample run #2

Note, the modified text does not include parenthesized content and the keyword is not found due to differences in letter-case.

```
please enter the file name: short.txt
Bentley University is a private university in Waltham, Massachusetts, isn't it?
Yes. Founded in 1917 as a school of accounting and finance in Boston's Back Bay
neighborhood, Bentley moved to Waltham in 1968. Bentley awards Bachelor of
Science degrees in 14 business fields and Bachelor of Arts degrees in 11 arts and
```

```
sciences disciplines, [offering 36 minors in arts and science and business
disciplines]... There you go!
******************** Text Analysis Report ********************
                   Sentences: 5
                       Words: 67
            Of them, numbers: 5
                 Numbers are: 11 14 36 1917 1968
                  Characters: 421
 Average words per sentence: 13
************************************************************

Which word would you like to highlight? bentley
Enter number of characters per line:40
start line (press Enter for 1):1
end line (press Enter for the last line):100
  1: Bentley University is a private universi|
  2: ty in Waltham# Massachusetts# isn#t it# |
  3: Yes# Founded in 1917 as a school of acco|
  4: unting and finance in Boston#s Back Bay |
  5: neighborhood# Bentley moved to Waltham i|
  6: n 1968# Bentley awards Bachelor of Scien|
  7: ce degrees in 14 business fields and Bac|
  8: helor of Arts degrees in 11 arts and sci|
  9: ences disciplines# ### There you go#|
************************ End of Report ************************
```

Sample run #3

```
please enter the file name: bentley.txt
Bentley University is a private university in Waltham, Massachusetts. Founded in
1917 as a school of accounting and finance in Boston's Back Bay neighborhood,
Bentley moved to Waltham in 1968. Bentley awards Bachelor of Science degrees in
14 business fields and Bachelor of Arts degrees in 11 arts and sciences
disciplines, offering 36 minors in arts and science and business disciplines.
Bentley University was founded in 1917 as the Bentley School of Accounting and
Finance by Harry C. Bentley, who served as the school's president until 1953. In
1961, the college was accredited to confer four-year Bachelor of Science degrees
under President Thomas Lincoln Morison, who moved the college from its Boylston
Street address in Boston to its current-day location in Waltham, Massachusetts.
Land for this move was purchased from the Lyman Estate in 1962, and the
construction to develop the campus then lasted from 1963 to 1968.[4]  Gregory H.
Adamian, a major driving force in the college's development, became the fourth
president in 1970. Under his guidance, the college became accredited to confer
four-year Bachelor of Arts degrees in 1971 and graduate degrees in 1973. During
this time, the school also changed its name to Bentley College. In 2002, Bentley
College opened up a campus in the Middle Eastern country of Bahrain in
partnership with the Bahrain Institute of Banking and Finance.  The college was
accredited to confer its first doctoral degrees in the fields of business and
accountancy in 2005.[5] A main fixture of the campus, The Bentley Library,
underwent a sweeping renovation in 2006 during which time the school's logo was
changed to showcase the clock tower that sits atop the building.[6] One year
later, Gloria Cordes Larson, a former state and federal government official and
Boston-based lawyer, became the first female president of Bentley College. In
2008, under the leadership of provost Bob Galliers, the school changed its name
to Bentley University after being authorized by the state board of higher
education to do so.[7] Alison Davis-Blake, the former dean of the Carlson School
of Management at the University of Minnesota and of the Ross School of Business
```

```
at the University of Michigan, became Bentley's eighth president in July 2018.
She stepped down in June 2020 and was replaced by Interim President Paul Condrin,
the chair of the board of trustees.[8]  In March 2021, the board unanimously
appointed Dr. E. LaBrent Chrite to serve as Bentley's ninth president.
******************** Text Analysis Report ********************
                  Sentences: 21
                      Words: 402
            Of them, numbers: 19
                 Numbers are: 11 14 36 1917 1917 1953 1961 1962 1963 1968 1970
1971 1973 2002 2006 2008 2018 2020 2021
                  Characters: 2505
 Average words per sentence: 19
************************************************************

Which word would you like to highlight? degree
Enter number of characters per line:70
start line (press Enter for 1):5
end line (press Enter for the last line):15
  5: EGREEs in 11 arts and sciences disciplines# offering 36 minors in arts|
  6:  and science and business disciplines#  Bentley University was founded|
  7:  in 1917 as the Bentley School of Accounting and Finance by Harry C# B|
  8: entley# who served as the school#s president until 1953# In 1961# the |
  9: college was accredited to confer four#year Bachelor of Science DEGREEs|
 10:  under President Thomas Lincoln Morison# who moved the college from it|
 11: s Boylston Street address in Boston to its current#day location in Wal|
 12: tham# Massachusetts# Land for this move was purchased from the Lyman E|
 13: state in 1962# and the construction to develop the campus then lasted |
 14: from 1963 to 1968#  Gregory H# Adamian# a major driving force in the c|
 15: ollege#s development# became the fourth president in 1970# Under his g|
************************ End of Report ************************
```

To implement this functionality, you will need to define and use functions described below.

Start by implementing and testing the following functions. You may also implement and use other functions as needed. Using Python components not introduced in class or course reading will incur a penalty.

1. Function numSentencesAveLen (text) (6 pts)

Define a function **`numSentencesAveLen (text)`** The function must be passed a string of text and must return two pieces of information about it: the number of sentences it contains, and the average number of *words* per *sentence*, rounded to the nearest integer. A word is any sequence of non-blank characters. Treat periods (.), question marks (?), exclamation marks (!), ellipses (...) as sentence terminators.

For example, when passed the following text as a parameter:
"Today is March 7, 2024. Sunny day!" the function should return `(2, 4)`
Test this function by calling it from function **`tester()`** with different parameter strings.

2. Function wordAndNumCount(text) (8 pts)

Define a function **`wordAndNumCount(text).`** The function must be passed a string of text and must return two metrics: the total number of words it contains, and the ~~number~~ list of those words that are considered numeric. A word is any sequence of non-blank characters. A number is a word consisting entirely

of digits, or digits followed by a single non-alphanumeric symbol.

The function should return the list of numeric words (as strings) found in the text.

For example, when passed "Today is March 7, 2024. Sunny day!" the function should return ~~(7, 2),~~ `(7, ["7," "2024."])`

because there are 7 words and the two numeric words are "7," and "2024."

Test this function by calling it from function **`tester()`** with different parameter strings.

3. Function processAndHighlight(text, keyword, symbol='#') (14 pts)

Define a function **`processAndHighlight(text, keyword, symbol)`** to work as follows:

1. Replace the occurrences of **`keyword`** in **`text`** with the uppercase version of **`keyword`,**
2. Replace all non-alphanumeric non-space characters with the **`symbol`**.
3. Remove all parenthesized expressions of the form *[some text].*

For example, when passed text "Curiosity [from Latin curiositas, from curiosus 'careful, diligent, curious', akin to cura 'care'] is a quality related to inquisitive thinking such as exploration, investigation, and learning, evident in humans and animals.[2][3] (from a Wikipedia article on Curiosity." and keyword "Curiosity", the function must return

```
"CURIOSITY  is a quality related to inquisitive thinking such as
exploration# investigation# and learning# evident in humans and animals#
#from a Wikipedia article on CURIOSITY#"
```

The easiest approach to implementing parts 2 and 3 is to go through the text one character at a time, constructing the modified text character by character. When a '[' is encountered, record that fact by assigning a value indicating this to a specific variable, e.g. `inparens = True`, and use this variable to decide how to update the modified text. When ']' is encountered, `inparens` should go back to being **`False`**.

Make sure to test your function, by calling it and print out its returned values in **`tester()`**.

4. Function def splitIntoLines(text, lineLen) (6 pts)

Define a function **`splitIntoLines(text, lineLen`**), which will split the text into separate lines of the specified length. The function must be passed text and a positive number representing the length of lines. The function should return a list of strings of the specified length.
For example, when passed "Today is March 7, 2024." `and` 5 the function should return

```
['Today', ' is M', 'arch ', '7, 20', '24.']
```

This can be accomplished by using slicing repeatedly, and appending the slices to a list. Make sure to test your function, by calling it and print out its returned values in **`tester()`**.

5. Function displayLines(linesLst, beginline, endline = -1) (8 pts)

Function **`displayLines(linesLst, beginline, endline = -1)`** must display the lines starting

from the beginline and ending with the endline. The parameters are: a list of lines (strings), line number (1-based) to start with, and to end at. The function should print those lines, appending a line number in front and a vertical bar in the end. The line number must be printed in two positions, right-justified, as shown in the sample run. The function should return nothing.
If the **`endline`** is not specified or if the **`endline`** is greater than the number of lines in the **`linesLst`**, print the lines until the end of **`linesLst:`**

For example, a call
`displayLines(['Today', ' is M', 'arch ', '7, 20', '24.'], 1)` should print

```
 1: Today|
 2:  is M|
 3: arch |
 4: 7, 20|
 5: 24.|
```

`displayLines(['Today', ' is M', 'arch ', '7, 20', '24.'], 2, 4)` should print

```
 2:  is M|
 3: arch |
 4: 7, 20|
```

`displayLines(['Today', ' is M', 'arch ', '7, 20', '24.'], 3, 10)` should print

```
 3: arch |
 4: 7, 20|
 5: 24.|
```

Make sure to test your function, by calling it in **`tester()`**.

6. Function main() (10 pts)

Finally, define and call main() to implement the full functionality described in the Program outline section on page 1. It is your job to figure out how to combine the functions you have already defined. Sample runs below illustrate how the program should run.

OTHER REQUIREMENTS

1. Your program must work correctly on any text file.
2. You DO NOT need to worry about validating incorrect input in this assignment. Your program will be tested only with valid data.
3. Include a **`docstring`** comment at the beginning of your program with the program/file name, your name, date, and a short description of the program.
4. Provide an introductory comment with each function, describing parameters, return value and what the function does.

GRADING

Your program should compile without syntax errors to receive any credit. If a part of your program is

working, you will receive partial credit, **but only if the program compiles without syntax errors**.

Requirement	**Points**
Correct implementation of all functions, and testing code in the tester()	52
Having no code outside of function definitions except for a call to main(), and no global variables	4
Programming style: including docstring, intro comments to functions, descriptive variable names, etc.	4
Total	**60**