
Program One Report**Definition of paragraph, sentence, and word**

1. Paragraph: Paragraph are separated by “\n” or lines that begin with indentation.
2. Sentence: Sentences are strings begin with a space or a line beginning, ended with period, exclamation, question mark, and followed by optional parenthesis or quotation mark.
3. Word: Words are strings begin with a space or a line beginning, and followed by non-space alphabet(s).
 - a) Contractions such as “he’s”, “he’ll”, “he’d” are counted as two words.
 - b) Punctuations are not counted as words.
 - c) Words connected by hyphen such as “beats-per-minute” are counted as one word because compound word is usually considered as one in parsing.
 - d) Abbreviations such as “B.M.P.” is counted as one word.

Program Explanation**Convert File into a Single String**

I find when searching for the patterns in the file, the search will stop at every “\n”. Thus I remove all “\n” from the text and join the text into one single string.

Abbreviation

Avoid Counting Sentence-Middle Punctuations as a Sentence. According to the definition of sentence, the system will originally consider “ Dr. Karageorghis” as two sentences: “ Dr.” and “ Karageorghis” The same is true for sentence-middle floats (4.5), abbreviations (B.M.P), or websites (Fitnessmagazine.com). In other words, if there are periods within a sentence, that sentence would be recognized as two separate ones. Thus, I counted how many such words there are in the text and abstracted the number of quantity from the sentence number counted originally.

Cases Where Sentences End with Abbreviation. However, there might be cases where sentences do end with abbreviations (such as *One of the most important elements, Dr. Karageorghis found, is a song's tempo, which should be between 120 and 140 beats-per-minute, or B.P.M.*) If I abstract this instance from the original sentence number, I would miss that sentence. Thus, I search for the abbreviations which are followed by a space and a capital letter. However, in this sample text, the sentence which is ended with “B.P.M.” has two spaces after the period. I deleted the extra space in the original file and the results were turned out to be 21 sentences. If the extra space was not deleted, the result of sentence number would be 20. This program is going to assume the text submitted for process is composed following standard writing conventions.

Ignore Sentence-Final Parenthesis and Quotation Mark

I do not make special notification for the cases where the sentence ended with period/exclamation mark/question mark plus parenthesis or quotation mark (such as *For a high-intensity workout like a hard run, he suggested Glenn Frey's "The Heat Is On."*) Although the search will not capture the final parenthesis, it will still stop at the periods immediately before the parenthesis. Thus, such sentences will still be correctly counted.

Make the Program Accessible by Command Prompt.

The command prompt is accessed through “sys” module. By using “sys.argv”

method, the input in the command line could be imported in the program. Thus I use “sys.argv” to open the file. Once the file name is typed in the command prompt, the program will know which file to access.

Limitations of This Program

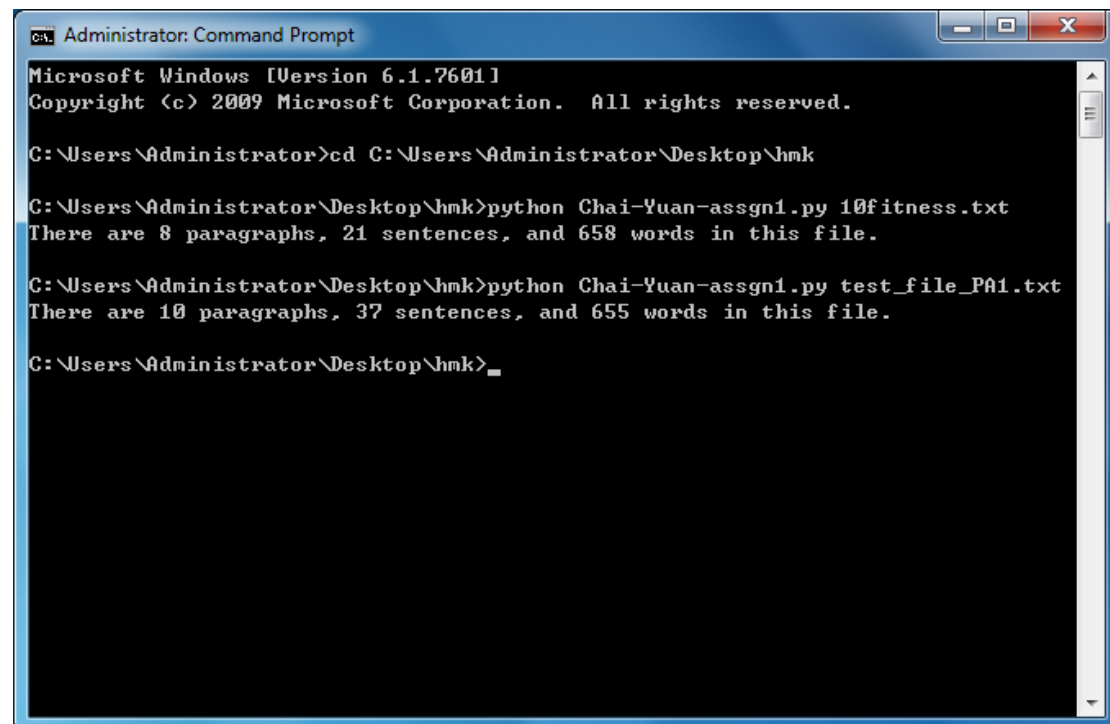
Sentence-Final Addressment

In terms of counting sentences, the program could not distinguish the sentence if it is ended by addressment because whether it is at the end of the sentence, the addressment would be followed by a space and a capital letter. By no means can the program recognize whether “Dr.” is at the end or is in the middle of a sentence. Thus, sentences such as “Please address him as Dr.” would be abstracted from the sentence count.

Not for Segmentation

This program is unable to do sentence or word segmentation while counting the number of the sentences. When counting sentences, as aforementioned, the program will first count all strings ended with periods, exclamation marks, and question marks as sentences and then abstract abbreviations, addressments, and digital numbers from the original number. As a result, the *findall* list would be a list containing all such sentence fragments. In addition, the program will ignore the sentence final parenthesis and quotation marks. When counting words, the program will find all strings separated by space. As a result the sentence-final words would be segmented with the punctuation attached. However, as the objective of this project is not segmenting but counting, I think the program is proper enough.

Result:



```
Administrator: Command Prompt
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>cd C:\Users\Administrator\Desktop\hmk

C:\Users\Administrator\Desktop\hmk>python Chai-Yuan-assgn1.py 10fitness.txt
There are 8 paragraphs, 21 sentences, and 658 words in this file.

C:\Users\Administrator\Desktop\hmk>python Chai-Yuan-assgn1.py test_file_P01.txt
There are 10 paragraphs, 37 sentences, and 655 words in this file.

C:\Users\Administrator\Desktop\hmk>
```