NLP Assignment 2

- *Shreyash Arya (2015097)*

General Assumption: Regex has been used for all the tasks; Outputs mentioned in each task are tested on the provided 'Development Set.txt'

Task 1: Number of paragraphs, sentences, and words contained in the article.

Number of paragraphs:

The paragraph will end with dot(.), exclamation(!) or question mark(?) followed by zero or more closing braces/quotation tags of any type followed by 1 or more slash n's (\n) or EOF(\Z) character.

Regex: **r'[.!?]+[\]})>\'\"]*([\n]+|\Z)'**

Output: **9**

* It can also be considered that the paragraph should have 2 or more '\n' at the end for which this regex can be used: **r'[.!?]+[\]})>\'\"]*\n[\n]+'** and output will be: **8**

Number of sentences:

Sentences are considered to follow two patterns joined by a pipe(|). The first pattern can be for the sentences which can be found in between the paragraphs and the second pattern sentence can be the end of a paragraph followed by a newline character or EOF.
Also, salutation can be problematic which considering the sentence boundaries and hence they are stripped from the text before passing through the regex; interjection words are considered as one-word sentences are also possible like 'Stop!', 'Oh!' etc. and hence a word with an exclamation mark would be considered as a sentence.

Regex: **r'[.!?]+[\])}>\'\"]*(\n|\Z)|[.!?]+[\])}>\'\"]*[\s]+[\"\'{\[(<]*[A-Z0-9]'**

(dot(.), exclamation(!) or question mark(?) one or more times followed by zero or more closing braces/quotation tags followed by one or more spaces followed by one or more opening braces/quotation tags followed by a capital word or number**) or**(|) **(**dot(.), exclamation(!) or question mark(?) one or more times followed by zero or more closing braces/quotation tags followed by one or more newline characters**)**

Output: **21**

Number of words:

The words are considered to be followed or preceded by any number of characters without space.

Regex: **r'.*[A-Za-z0-9].*'** (matching the word format)**; r'\s+'** (Used for splitting the terms)
The regex follows the pattern of (zero or more number of any character (placeholder '.' is used here)) followed by (capital letter or small letter or digits) followed by (zero or more number of any character).

Output: **646**

* The output is checked with the inbuilt 'wc' command in Linux which gives the output as 651 because it considers '-' as a word given in the Development Set.txt

Task 2: Given a word as input, number of sentences starting with the word.

Sentence end is checked using the regex for detecting the sentence boundary followed by the word that is given as input; word input is case sensitive- 'The' and 'the' are considered different words.

Regex: **r'[.!?]+[>)\'}\]\"]*(\s|\n|\Z)+[\'\"{\[\(<]*'+word+r'[^A-Za-z0-9]'**

* word should not be followed by another alphabet or digit hence r'[^A-Za-z0-9] is used.

Task 3: Given a word as input, number of sentences ending with the word.

Given input word followed by the pattern used to match sentence boundary is used in this case; word input is case sensitive - 'The' and 'the' are considered different words.

Regex: **r'[^A-Za-z0-9]'+word+r'([.!?]+[\])}>\'\"]*(\n|\Z)|[.!?]+[\])}>\'\"]*[\s]+[\"\'{\[(<]*[A-Z0-9])'**

* word should not be preceded by another alphabet or digit hence r'[^A-Za-z0-9] is used.

Task 4: Given a word as input, count of that word in the input file.

The word can be at the start or the end and it shouldn't be followed or preceded by any alphabet or digit; the input word is not case sensitive - 'The' and 'the' is considered as same word.

Regex:
r'(^|[^a-zA-Z0-9])['+word[0].lower()+word.upper()+r']'+word[1:len(word)]+r'([^a-zA-Z0-9]|$)'

This regex checks the word at the start of the sentence using the carrot character(^) and at the end of the sentence using the dollar character($). The word in between is case-insensitive and also it shouldn't be followed or preceded by any word or digit. Though characters can be attached to the word.

References:

https://stackoverflow.com/questions/4947561/can-i-combine-2-regex-with-a-logic-or
https://stackoverflow.com/questions/13332268/python-subprocess-command-with-pipe
https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
https://stackoverflow.com/questions/10677020/real-word-count-in-nltk
https://blogs.transparent.com/english/one-word-sentences-in-english/