Number of Words:

Assumptions:

- Words are separated by spaces.
- All special characters such as “ ” ‘’, ! @ # $ % etc are not counted as words.
- Words such as self.com, go-go(without space in between) are taken as single word.
- *https://wordcounter.net/ is taken as reference tool for counting number of words.*

Steps:

- Read the file.
- Split it using space.
- Prepared a list of special character.
- If the word is not in special list, add it to count of number of words.

Number of paragraphs:

Assumptions:

- Both development Set and test set don’t have any blank line at the end. If there is, remove it from the end.
- Paragraphs are separated by a single line.

Steps:

- Read the entire file in a string.
- Using regex “\n\n”, split it.
- Count number of splits.

Number of Sentences:

Assumptions:

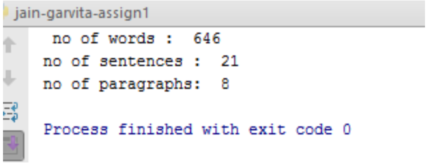
- Sentences ends with “.” or “ !” or “?”
- Sentences begin with an uppercase character.
- Dr|Mr|Ms|Mrs are followed by a dot but not considered as a sentence.
- “.” followed by “)” which is further followed by space and uppercase character is taken as a sentence.
- “.” followed by single or double quotes is taken as a sentence.

Steps:

- All new lines are replaced by spaces.
- index of “.” or “!” or “?” is found.
- Characters occurring before and after the characters mentioned above are analyzed to get a count of number of sentences.

Output:

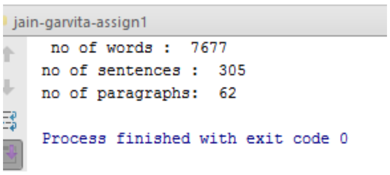
Development Article:



```
 no of words :  646
no of sentences :  21
no of paragraphs:  8

Process finished with exit code 0
```

Test Article:



```
 no of words :  7677
no of sentences :  305
no of paragraphs:  62

Process finished with exit code 0
```

How the system should be evaluated?

It should be evaluated according to the assumptions described above. Some of the examples are:

- *“(We don’t know why.)*” is taken as a sentence.
- “*Fitnessmagazine.com*” is taken as a single word.
- “*"Mr. Brightside" by the Killer (150 B.P.M.), and "Dancing Queen" by Abba.”* is a single sentence with “*M.)*” not considered as end of sentence and “*Mr.*” handled.