Name: Sofyan Mahmoud

Sec.: 1

ID: 25

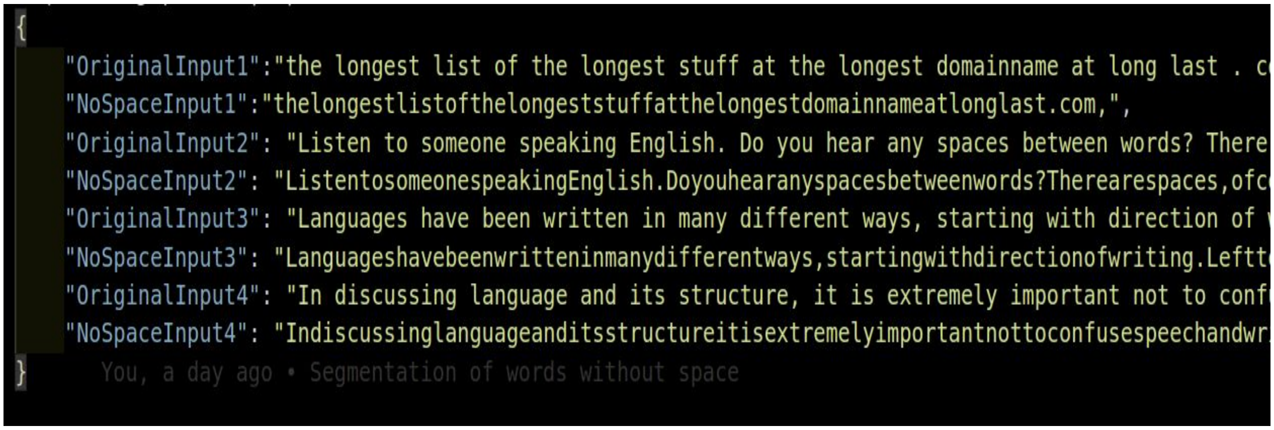
Problem No.: 1

Description of the code files

- **Segmentation.py**: The main file which has the python code
- **Input.json**: The file which take the input
- **Words.csv**: The dataset

How to run

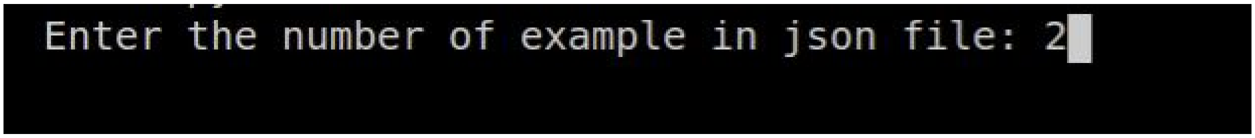
- Install python3
- In the (input.json) file, you will put your input, for each input you will have two keys
  - OriginalInput<number of input> : The original input or expected output
    If there is no original input -expected output- you must write the key and leave its value empty
  - NoSpaceInput<number of the input> : the input with no space, that you want to do the segmentation on it
  - Example of the output:



```
{
    "OriginalInput1":"the longest list of the longest stuff at the longest domainname at long last . c
    "NoSpaceInput1":"thelongestlistofthelongeststuffatthelongestdomainnameatlonglast.com,",
    "OriginalInput2": "Listen to someone speaking English. Do you hear any spaces between words? There
    "NoSpaceInput2": "ListentosomeonespeakingEnglish.Doyouhearanyspacesbetweenwords?Therearespaces,ofc
    "OriginalInput3": "Languages have been written in many different ways, starting with direction of
    "NoSpaceInput3": "Languageshavebeenwritteninmanydifferentways,startingwithdirectionofwriting.Leftt
    "OriginalInput4": "In discussing language and its structure, it is extremely important not to conf
    "NoSpaceInput4": "Indiscussinglanguageanditsstructureitisextremelyimportantnottoconfusespeechandwr
}      You, a day ago • Segmentation of words without space
```

- Run the file (Segmentation.py) using this command
  python3 segmentation.py

- It will ask you about the number of input you want to select which you have assigned in the json file as we mentioned above



```
Enter the number of example in json file: 2
```

Write any textual answers to the problem:

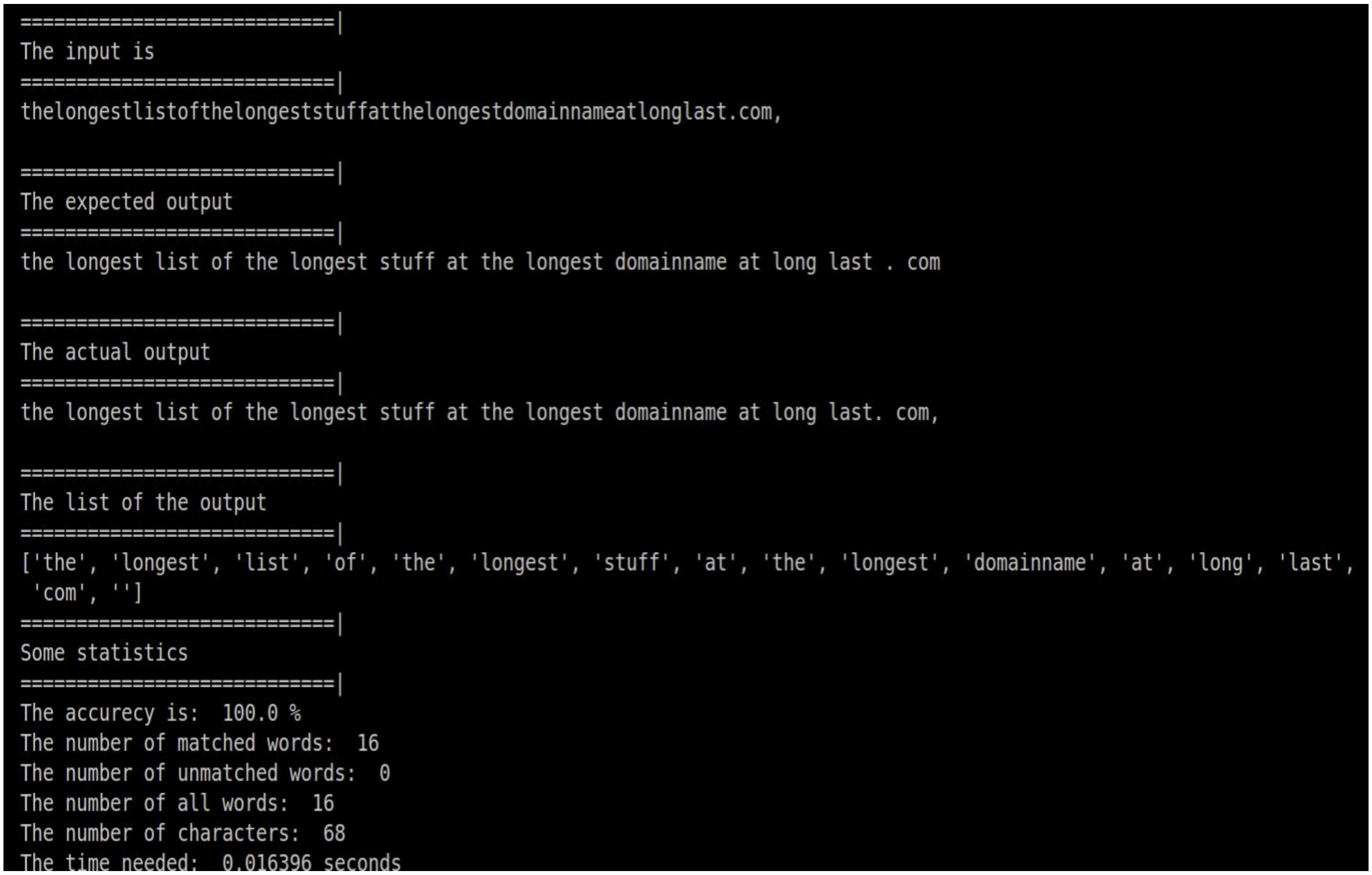
It does segmentation for words, if there are numbers, it may cause an error

How I wrote the code:

The solution depends on dynamic programming aim to do segmentation and its criteria to do right segmentation is to get the minimum number of words with highest frequency and I get this frequencies from the dataset which consists of (1/3) Million Most Frequent English Words on the Web and their frequency

Results and Conclusion

It will show the original input, the expected output, the actual output, a list of the actual output and some statistics about the error, the number of words and characters, etc..



```
============================|
The input is
============================|
thelongestlistofthelongeststuffatthelongestdomainnameatlonglast.com,

============================|
The expected output
============================|
the longest list of the longest stuff at the longest domainname at long last . com

============================|
The actual output
============================|
the longest list of the longest stuff at the longest domainname at long last. com,

============================|
The list of the output
============================|
['the', 'longest', 'list', 'of', 'the', 'longest', 'stuff', 'at', 'the', 'longest', 'domainname', 'at', 'long', 'last',
 'com', '']
============================|
Some statistics
============================|
The accurecy is:  100.0 %
The number of matched words:  16
The number of unmatched words:  0
The number of all words:  16
The number of characters:  68
The time needed:  0.016396 seconds
```

Any Assumption

Assume that the maximum length of the input is 800 characters