Do you remember what do we mean by tokenization?

Tokenization is the process of split the words to tokens

Use the following script to install NLTK: pip install nltk

```
In [12]: pip install nltk
```

```
Requirement already satisfied: nltk in d:\users\abadi\anaconda3\lib\site-packag
es (3.4.5)
Requirement already satisfied: six in d:\users\abadi\anaconda3\lib\site-package
s (from nltk) (1.12.0)
Note: you may need to restart the kernel to use updated packages.
```

```
In [11]: import nltk
```

```
In [5]: nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Abadi\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

```
Out[5]: True
```

```
In [8]: from nltk.tokenize import sent_tokenize
```

```
In [11]: text=" Welcome readers. I hope you find it interesting. Please do reply."
```

```
In [12]: sent_tokenize(text)
```

```
Out[12]: [' Welcome readers.', 'I hope you find it interesting.', 'Please do reply.']
```

How many sentence you had? 3

How many sentence will we have if we replace full stop “.” With “,” in text 2

```
In [3]: import nltk
tokenizer = nltk.data.load("tokenizers/punkt/english.pickle")
text="Hello everyone. Hope all are fine and doing well. Hope you find the book i
tokenizer.tokenize(text)
```

```
Out[3]: ['Hello everyone.',
 'Hope all are fine and doing well.',
 'Hope you find the book interesting.']
```

```
In [18]: import nltk
In [19]: import nltk
         tokenizer = nltk.data.load("tokenizers/punkt/english.pickle")
```

```
In [22]: text=" Welcome readers. I hope you find it interesting. Please do reply."
```

```
In [23]: tokenizer.tokenize(text)

Out[23]: [' Welcome readers.', 'I hope you find it interesting.', 'Please do reply.']
```

```
In [24]: Arabic_text="مرحبا بكم. نحن نتعلم اساسيات مبادئ استرجاع المعلومات."
         tokenizer.tokenize(Arabic_text)

Out[24]: ['مرحبا بكم.', 'نحن نتعلم اساسيات مبادئ استرجاع المعلومات.']
```

```
In [36]: text="Welcome readers. I hope you find it interesting. Please do reply. ."
```

```
In [37]: nltk.word_tokenize(text)

Out[37]: ['Welcome',
          'readers',
          '.',
          'I',
          'hope',
          'you',
          'find',
          'it',
          'interesting',
          '.',
          'Please',
          'do',
          'reply',
          '.',
          '.']
```

```
In [39]: nltk.word_tokenize(Arabic)

Out[39]: ['hi.Iam', 'Abdulraheem.bye']
```

Exercise 3: Try to tokenize a given sentence from user into words. Use input() function to enter a text from keyboard.

```
In [40]: Arabic=input("Please write a text")

         Please write a textعبدالرحيم شفيق
```

```
In [41]: nltk.word_tokenize(Arabic)

Out[41]: ['عبدالرحيم', 'شفيق']
```

In [42]:

```
Arabic=input("Please write a text")
```

```
Please write a textI'm a student
```

In [43]:

```
nltk.word_tokenize(Arabic)
```

Out[43]:

```
['I', "'m", 'a', 'student']
```

Exercise 4: Modify the regular expression at step 3 above to find email address

In [44]:

```
from nltk.tokenize import RegexpTokenizer
```

In [54]:

```
tokenizer=RegexpTokenizer("\S+@\S+")
tokenizer.tokenize("Don't hesitate to askquestions or send to me your question t
```

Out[54]:

```
['mohsarem@gmail.com']
```

In [5]:

```
text=[" It is a pleasant evening.","Guests, who came from US arrived at the venue
from nltk.tokenize import word_tokenize
tokenized_docs=[word_tokenize(doc) for doc in text]
print(tokenized_docs)
```

```
[['It', 'is', 'a', 'pleasant', 'evening', '.'], ['Guests', ',', 'who', 'came',
'from', 'US', 'arrived', 'at', 'the', 'venue'], ['Food', 'was', 'tasty', '.']]
```

Exercise 5. What is the role of re.compile(),re.escape() functions?

Type Markdown and LaTeX: $\alpha 2$

Exercise 6. Apply lower () function and upper() function on the sentence below:

In [8]:

```
print(text[0].upper())
print(text[0].lower())
```

```
 IT IS A PLEASANT EVENING.
 it is a pleasant evening.
```

In [58]:

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stops=set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Abadi\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

In [63]:

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stops=set(stopwords.words('english'))
words=["Don't",'hesitate','to','ask','questions']
[word for word in words if word not in stops]
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Abadi\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[63]:

```
["Don't", 'hesitate', 'ask', 'questions']
```

Exercise 7. Tokenize and remove stop words from the sentence below:

In [1]:

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stops=set(stopwords.words('english'))
words=["Don't",'hesitate','to','ask','questions']
[word for word in words if word not in stops]
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Abadi\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[1]:

```
["Don't", 'hesitate', 'ask', 'questions']
```

Exercise 8. Given a text in directory, demonstrate how to use NLTK to treat its content.

In [15]:

```
import nltk
Sentenes = open(r'D:\Users\Abadi\Anaconda3\Sen.txt')
text = Sentenes.read()
text
```

Out[15]:

```
'In computer science, artificial intelligence (AI), sometimes called machine in
telligence, is intelligence demonstrated by machines, in contrast to the natura
l intelligence displayed by humans and animals. Computer science defines AI res
earch as the study of intelligent agents: any device that perceives its environ
ment and takes actions that maximize its chance of successfully achieving its g
oals.'
```

In [21]:

```
import re
from nltk.corpus import stopwords
stops=set(stopwords.words('english'))
words= re.sub("[^\w]"," ", text).split()
[word for word in words if word not in stops]
```

Out[21]:

```
['In',
 'computer',
 'science',
 'artificial',
 'intelligence',
 'AI',
 'sometimes',
 'called',
 'machine',
 'intelligence',
 'intelligence',
 'demonstrated',
 'machines',
 'contrast',
 'natural',
 'intelligence',
 'displayed',
 'humans',
 'animals',
 'Computer',
 'science',
 'defines',
 'AI',
 'research',
 'study',
 'intelligent',
 'agents',
 'device',
 'perceives',
 'environment',
 'takes',
 'actions',
 'maximize',
 'chance',
 'successfully',
 'achieving',
 'goals']
```

In [ ]:

```

```