Text tokenization

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
text='Bruce banner is a scientist. He is an Avenger'
from nltk .tokenize import word_tokenize,sent_tokenize
sentence=sent_tokenize(text)
print(sentence)
```

```
['Bruce banner is a scientist.', 'He is an Avenger']
```

```
words=[word_tokenize(sent) for sent in sentence]
print(words)
```

```
[['Bruce', 'banner', 'is', 'a', 'scientist', '.'], ['He', 'is', 'an', 'Avenger']]
```

Removing stopwords

```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
from nltk.corpus import stopwords
```

```
from string import punctuation
customStopWords=set(stopwords.words('english')+list(punctuation))


output=[i for i in word_tokenize(text) if i not in customStopWords]
print(output)
```

```
['Bruce', 'banner', 'scientist', 'He', 'Avenger']
```

Identify the bigrams(pair words which occur frequently)

```
from nltk.collocations import *
bigram_measures=nltk.collocations.BigramAssocMeasures()
finder=BigramCollocationFinder.from_words(output)
sorted(finder.ngram_fd.items())
```

```
[(('Bruce', 'banner'), 1),
 (('He', 'Avenger'), 1),
 (('banner', 'scientist'), 1),
 (('scientist', 'He'), 1)]
```

Stemming

```
text1='mary Closed on clothing night when she was in the mood to close'
from nltk.stem.lancaster import LancasterStemmer
st=LancasterStemmer()
stemmedWords=[st.stem(word) for word in word_tokenize(text1)]
print(stemmedWords)
```

```
['mary', 'clos', 'on', 'cloth', 'night', 'when', 'she', 'was', 'in', 'the', 'mood', 'to', 'clos']
```

Part of speech(POS)

```
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
True
```

```
nltk.pos_tag(word_tokenize(text1))
```

```
[('mary', 'NN'),
 ('Closed', 'VBD'),
 ('on', 'IN'),
 ('clothing', 'NN'),
 ('night', 'NN'),
 ('when', 'WRB'),
 ('she', 'PRP'),
 ('was', 'VBD'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('mood', 'NN'),
 ('to', 'TO'),
 ('close', 'VB')]
```