Name: Muhammad Ghazanfar

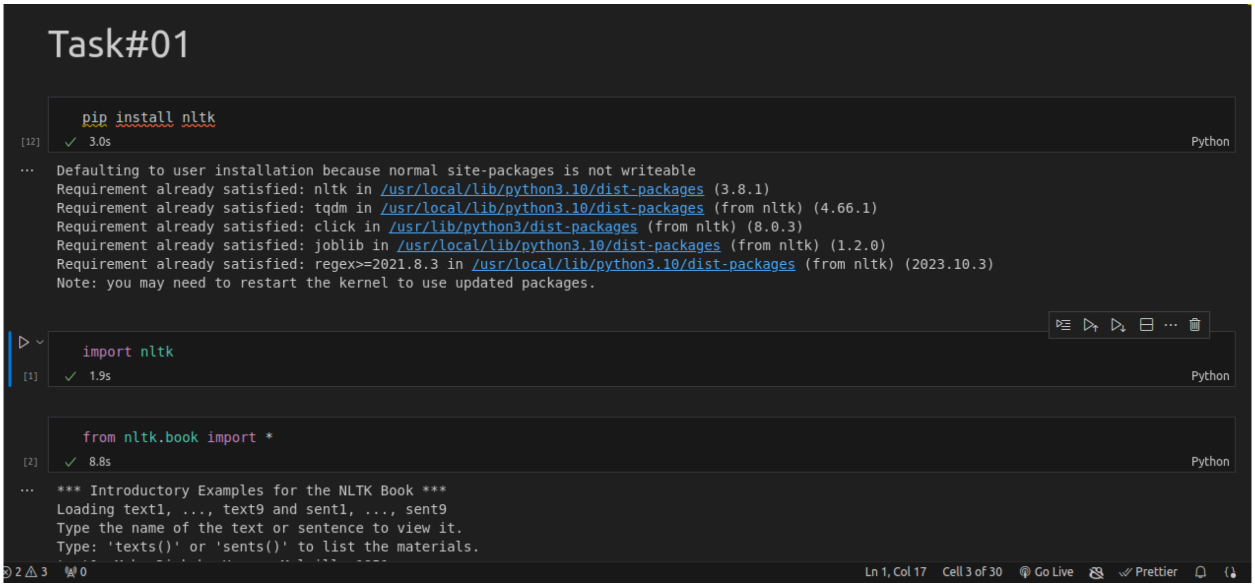
Roll#: 20P-0567

Sec: 7A

Course: NLP

Lab#01

Task 1: Getting Ready



Task#01

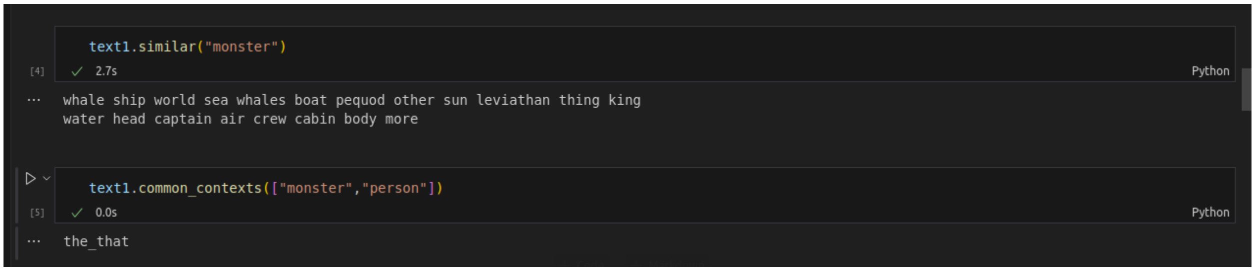
```
pip install nltk
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
Requirement already satisfied: click in /usr/lib/python3/dist-packages (from nltk) (8.0.3)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.10.3)
Note: you may need to restart the kernel to use updated packages.
```

```
import nltk
```

```
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
```

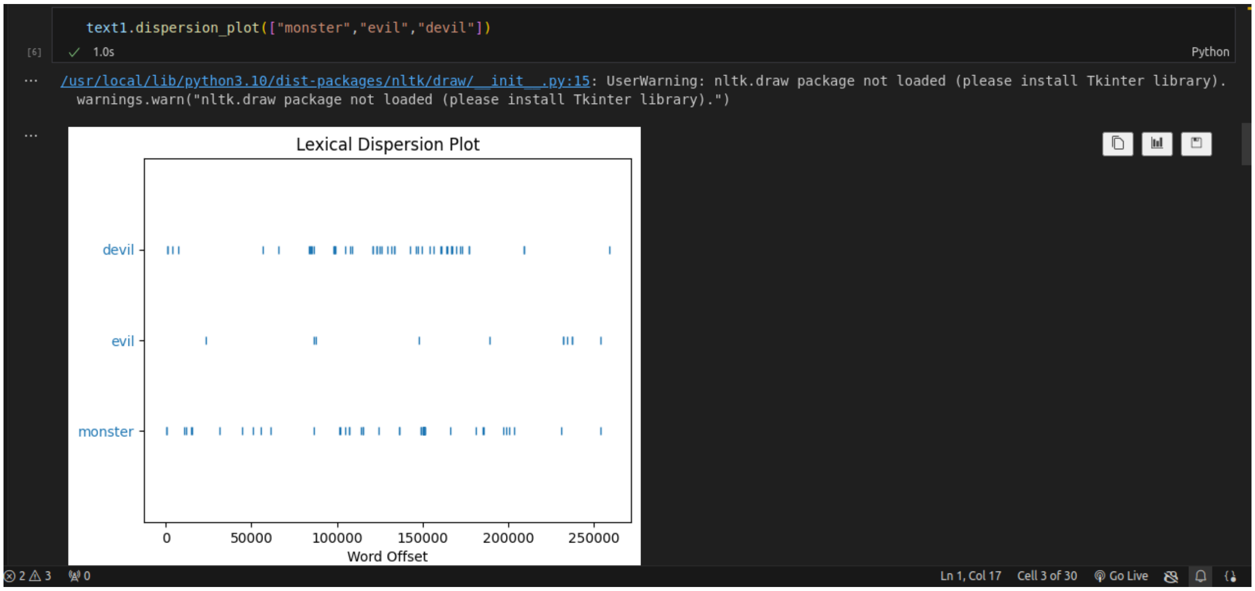


```
text1.similar("monster")
```

```
whale ship world sea whales boat pequod other sun leviathan thing king
water head captain air crew cabin body more
```

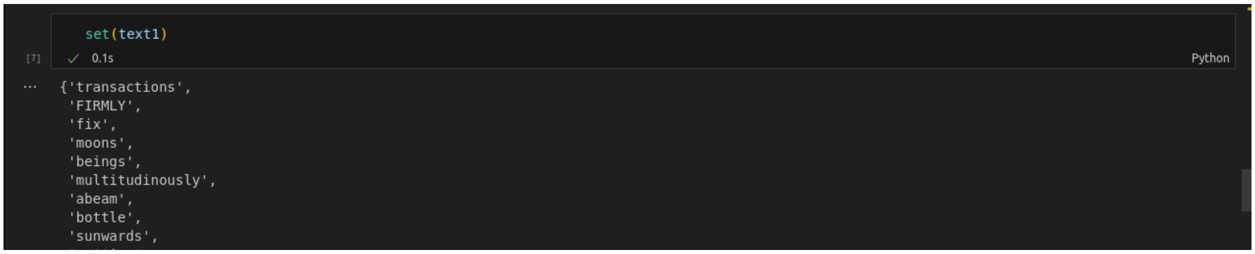
```
text1.common_contexts(["monster","person"])
```

```
the_that
```



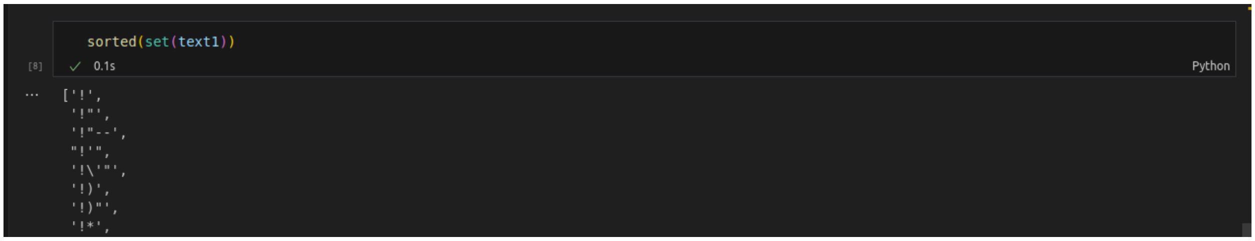
```
text1.dispersion_plot(["monster","evil","devil"])
```

```
/usr/local/lib/python3.10/dist-packages/nltk/draw/__init__.py:15: UserWarning: nltk.draw package not loaded (please install Tkinter library).
  warnings.warn("nltk.draw package not loaded (please install Tkinter library).")
```



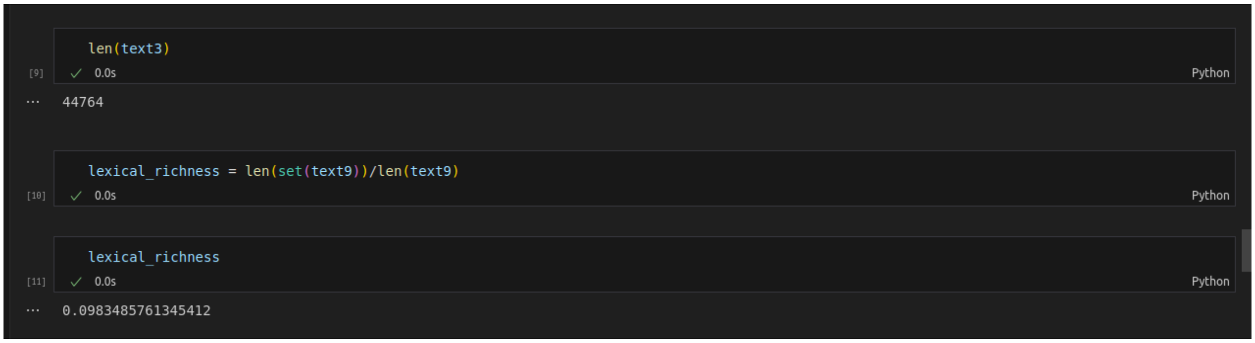
```
set(text1)
```

```
{'transactions',
 'FIRMLY',
 'fix',
 'moons',
 'beings',
 'multitudinously',
 'abeam',
 'bottle',
 'sunwards',
```



```
sorted(set(text1))
```

```
['!',
 '!"',
 '!"--',
 "!'",
 '!\'"',
 '!)',
 '!)"',
 '!*',
```



```
len(text3)
```

```
44764
```

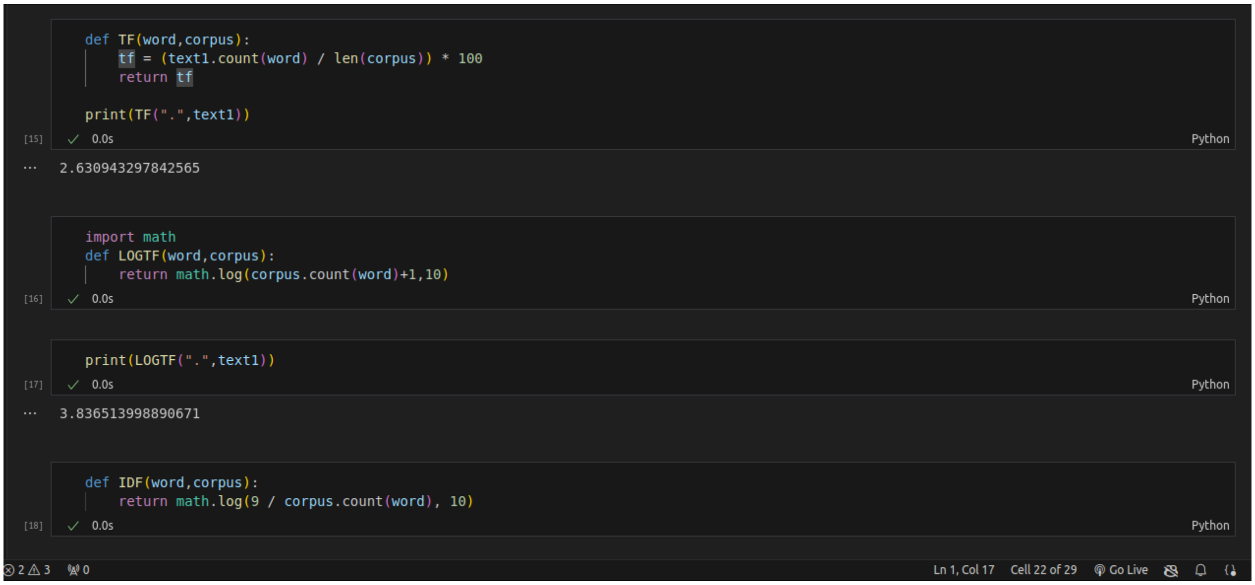
```
lexical_richness = len(set(text9))/len(text9)
```

```
lexical_richness
```

```
0.0983485761345412
```

Corpus	Text1	Text2	Text3	Text4	Text5	Text6	Text7	Text8	Text9
Corpus Name	Moby Dick by Herman Melville 1851	Sense and Sensibility by Jane Austen 1811	The Book of Genesis	Inaugural Address Corpus	Chat Corpus	Monty Python and the Holy Grail	Wall Street Journal	Personals Corpus	The Man Who Was Thursday by G . K . Chesterton 1908
Corpus length	260819	141576	44764	152901	45010	16967	100676	4867	69213
Unique Words	19317	6833	2789	10025	6066	2166	12408	1108	6807
Lexical Richness	0.07406285585022564	0.0482638300276883 1	0.0623045304262353 7	0.0655653004231496 2	0.13477005109975562	0.1276595744680851	0.12324685128531129	0.22765564002465585	0.0983485761345412

Task 2: Term Frequency



```
def TF(word,corpus):
    tf = (text1.count(word) / len(corpus)) * 100
    return tf

print(TF(".",text1))
```

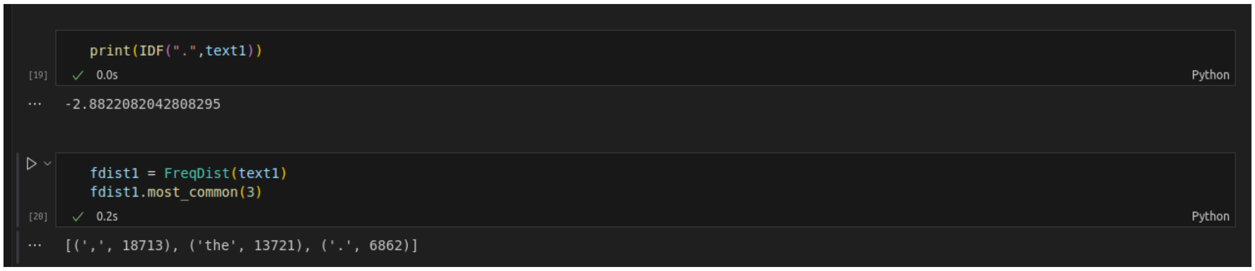
```
2.630943297842565
```

```
import math
def LOGTF(word,corpus):
    return math.log(corpus.count(word)+1,10)
```

```
print(LOGTF(".",text1))
```

```
3.836513998890671
```

```
def IDF(word,corpus):
    return math.log(9 / corpus.count(word), 10)
```

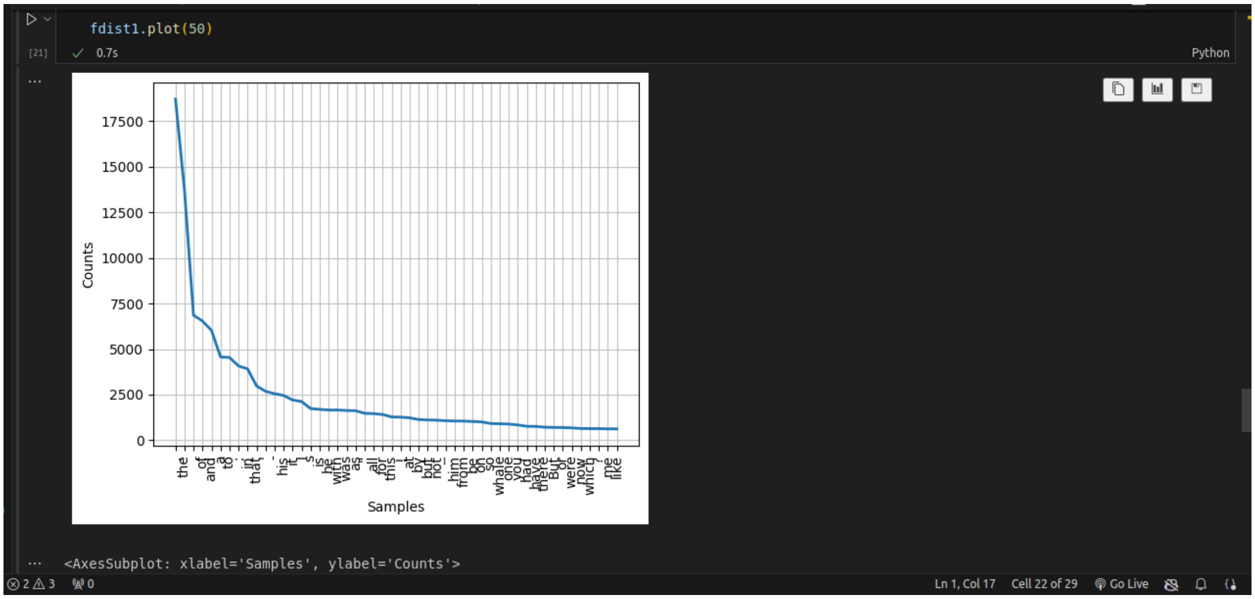


```
print(IDF(".",text1))
```

```
-2.8822082042808295
```

```
fdist1 = FreqDist(text1)
fdist1.most_common(3)
```

```
[(',', 18713), ('the', 13721), ('.', 6862)]
```

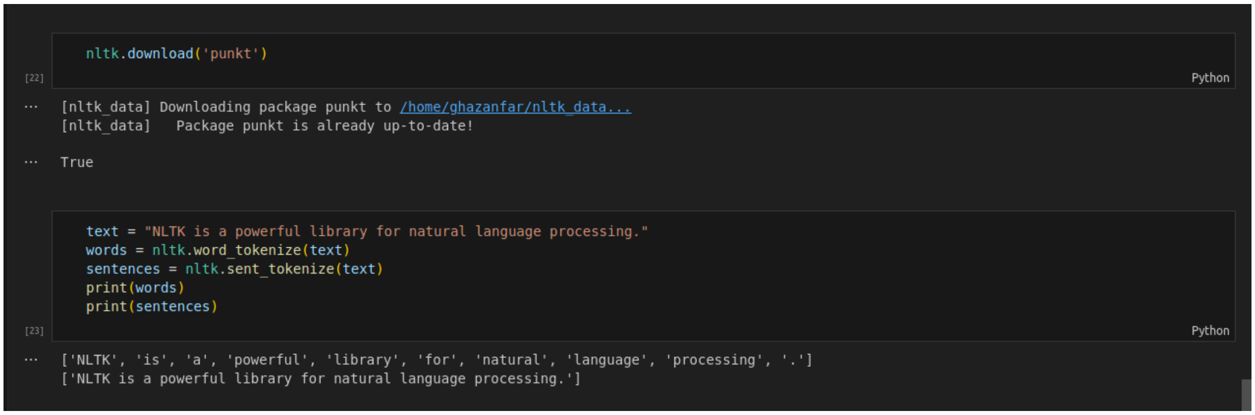


```
fdist1.plot(50)
```

```
<AxesSubplot: xlabel='Samples', ylabel='Counts'>
```

Tokens	TF()	LOGTF()	IDF()
monster	0.018786974875296663	1.6989700043360185	-0.7359535705891886
evil	0.004217484155678842	1.0791812460476247	-0.08715017571890013
devil	0.01955379017632918	1.716003343634799	-0.7533276666586114
the	5.260736372733581	4.137417414990392	-3.1831432548946452
Common word 1 (,)	7.174707364110744	4.272166625140787	-3.3179009081517252
Common word 2 (.)	2.630943297842565	3.836513998890671	-2.8822082042808295
Common word 3 (the)	5.260736372733581	4.137417414990392	-3.1831432548946452

Task 3: Tokenization & POS



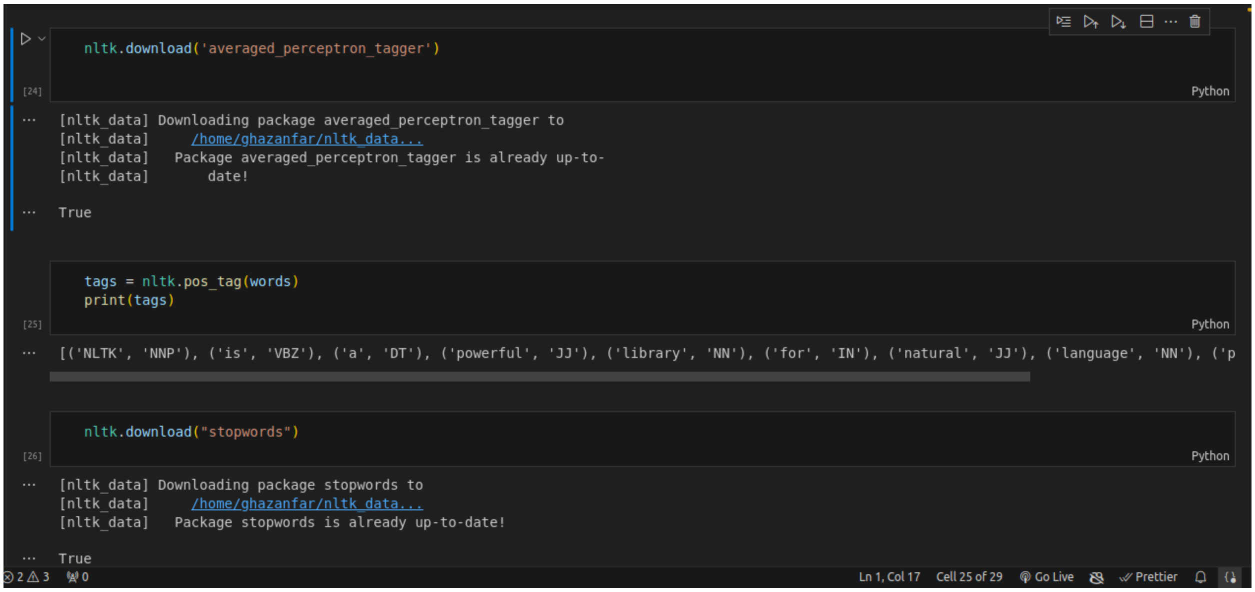
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /home/ghazanfar/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
True
```

```
text = "NLTK is a powerful library for natural language processing."
words = nltk.word_tokenize(text)
sentences = nltk.sent_tokenize(text)
print(words)
print(sentences)
```

```
['NLTK', 'is', 'a', 'powerful', 'library', 'for', 'natural', 'language', 'processing', '.']
['NLTK is a powerful library for natural language processing.']
```



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

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/ghazanfar/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

```
True
```

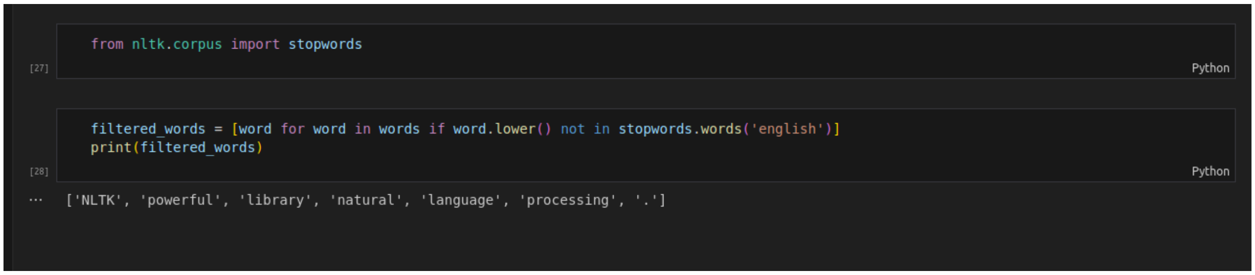
```
tags = nltk.pos_tag(words)
print(tags)
```

```
[('NLTK', 'NNP'), ('is', 'VBZ'), ('a', 'DT'), ('powerful', 'JJ'), ('library', 'NN'), ('for', 'IN'), ('natural', 'JJ'), ('language', 'NN'), ('p
```

```
nltk.download("stopwords")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/ghazanfar/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
True
```



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

```
filtered_words = [word for word in words if word.lower() not in stopwords.words('english')]
print(filtered_words)
```

```
['NLTK', 'powerful', 'library', 'natural', 'language', 'processing', '.']
```