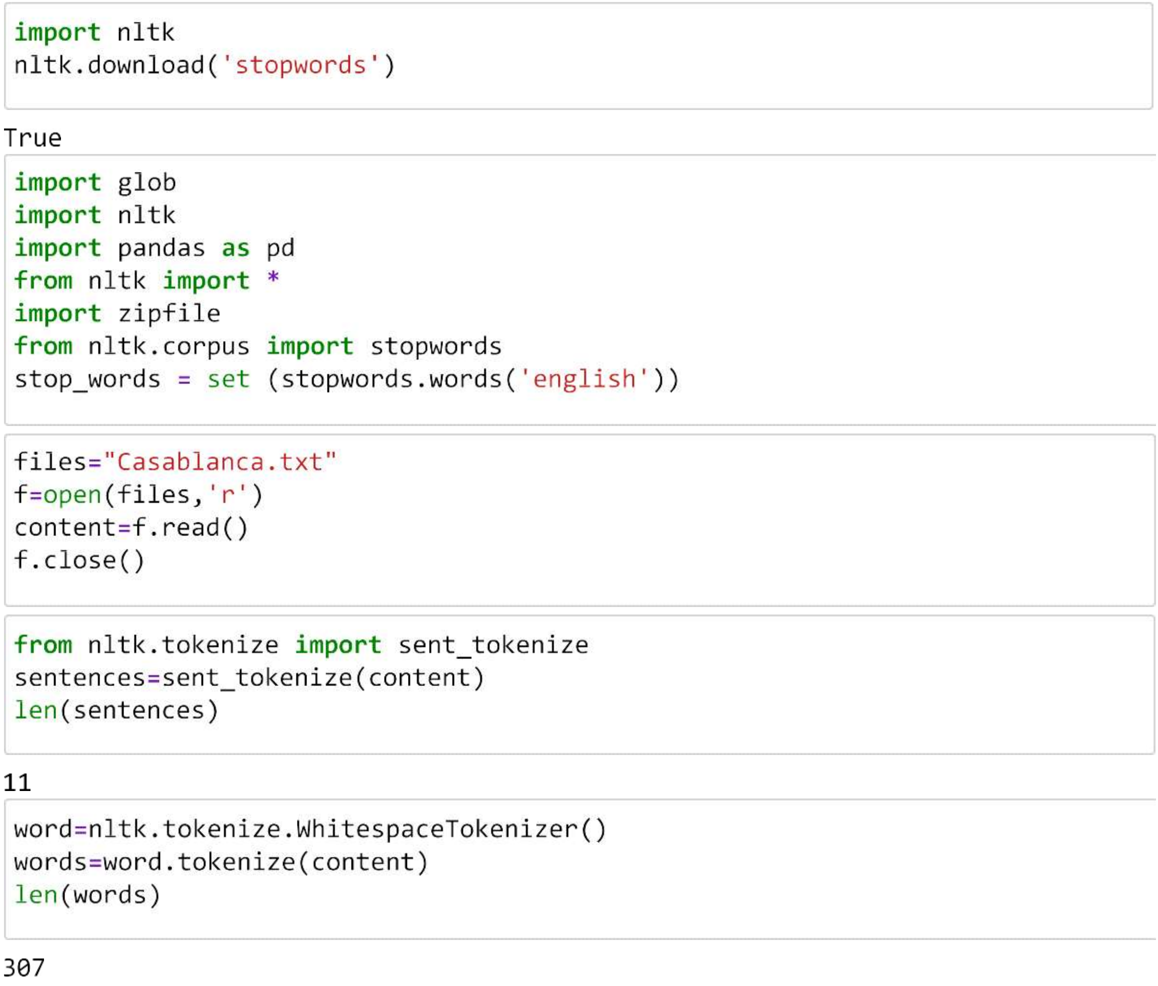
# Part of Speech On Large Text

ANNAPOORNIMA S
225229101



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

```
True
```

```
import glob
import nltk
import pandas as pd
from nltk import *
import zipfile
from nltk.corpus import stopwords
stop_words = set (stopwords.words('english'))
```

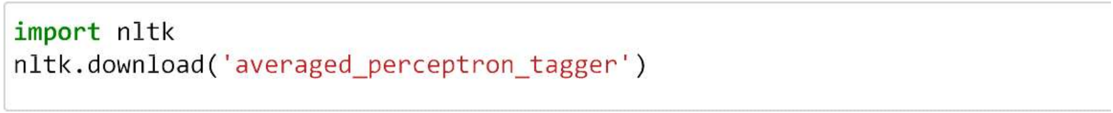
```
files="Casablanca.txt"
f=open(files,'r')
content=f.read()
f.close()
```

```
from nltk.tokenize import sent_tokenize
sentences=sent_tokenize(content)
len(sentences)
```

```
11
```

```
word=nltk.tokenize.WhitespaceTokenizer()
words=word.tokenize(content)
len(words)
```

```
307
```



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

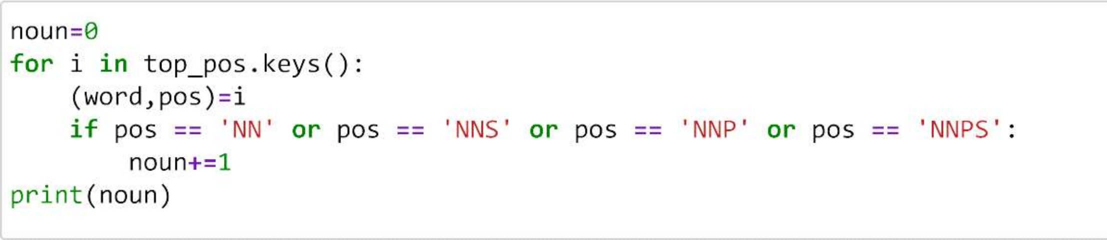
```
True
```

```
tag = []
d_tags = []
words = [w for w in words if not w in stop_words]
tagged = nltk.pos_tag(words)
for i in tagged:
    (word,pos)=i
    tag.append(pos)
for j in tag:
    if j not in d_tags:
        d_tags.append(j)
len(d_tags)
```

```
18
```

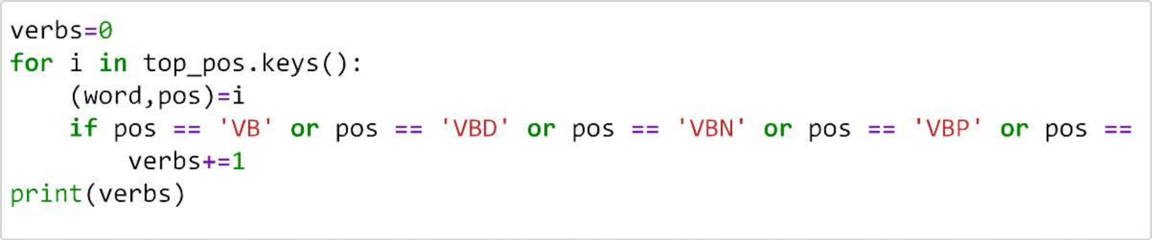
```
top_pos=FreqDist(tagged)
top_pos.most_common(10)
```

```
top10w=FreqDist(words)
top10w.most_common(10)
```



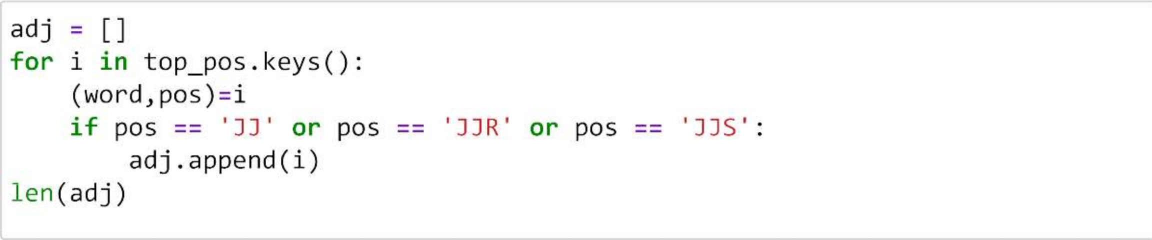
```
noun=0
for i in top_pos.keys():
    (word,pos)=i
    if pos == 'NN' or pos == 'NNS' or pos == 'NNP' or pos == 'NNPS':
        noun+=1
print(noun)
```

```
101
```



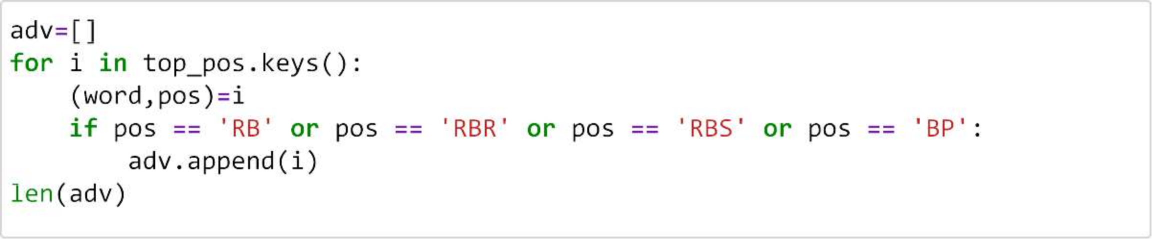
```
verbs=0
for i in top_pos.keys():
    (word,pos)=i
    if pos == 'VB' or pos == 'VBD' or pos == 'VBN' or pos == 'VBP' or pos ==
        verbs+=1
print(verbs)
```

```
20
```

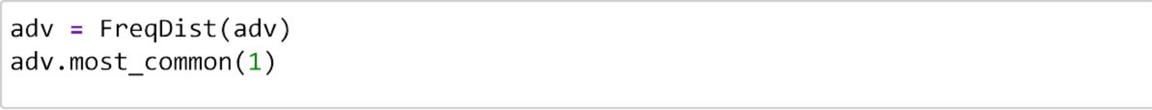


```
adj = []
for i in top_pos.keys():
    (word,pos)=i
    if pos == 'JJ' or pos == 'JJR' or pos == 'JJS':
        adj.append(i)
len(adj)
```

```
39
```



```
adv=[]
for i in top_pos.keys():
    (word,pos)=i
    if pos == 'RB' or pos == 'RBR' or pos == 'RBS' or pos == 'BP':
        adv.append(i)
len(adv)
```



```
adv = FreqDist(adv)
adv.most_common(1)
```

```
'anniversary.
```

```
adv = FreqDist(adj)
adv.most_common(1)
```