Name : Arul Kumar ARK

Roll No. : 225229103

Lab : 7

Exploring Part of Speech Tagging on Large Text Files

In [2]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\1mscdsa03\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[2]: True

In [18]:
```
import glob
import nltk
import pandas as pd
from nltk import*
import zipfile
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
```

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

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\1mscdsa03\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[5]: True

No. of Sentences :

In [19]:
```
files="Psycho.txt"
f=open(files,'r')
content=f.read()
f.close()
from nltk.tokenize import sent_tokenize
sentnces=sent_tokenize(content)
len(sentnces)
```

Out[19]: 24

No. Words :

In [20]:
```
word=nltk.tokenize.WhitespaceTokenizer()
words=word.tokenize(content)
len(words)
```

Out[20]: 612

Top 10 Words And Their Counts :

In [8]:
```
top10w=FreqDist(words)
top10w.most_common(10)
```

Out[8]:
```
[('the', 50),
 ('of', 26),
 ('and', 20),
 ('a', 18),
 ('to', 14),
 ('is', 14),
 ('in', 12),
 ('as', 9),
 ('his', 7),
 ('Hitchcock', 5)]
```

In [ ]:
```
import nltk
nltk.download('averaged_perceptron_tagger')
```

Different POS :

In [21]:
```
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)
```

Out[21]: 20

Top 10 POS :

In [22]:
```
top_pos=FreqDist(tagged)
top_pos.most_common(10)
```

Out[22]:
```
[(('Hitchcock', 'NNP'), 5),
 (('The', 'DT'), 5),
 (('Paramount', 'NNP'), 3),
 (('murder', 'NN'), 3),
 (('John', 'NNP'), 3),
 (('June', 'NNP'), 2),
 (('Alfred', 'NNP'), 2),
 (('mystery', 'NN'), 2),
 (('Psycho', 'NNP'), 2),
 (('New', 'NNP'), 2)]
```

No. of Noouns :

In [23]:
```
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)
```

```
161
```

No. of verb :

In [24]:
```
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=='VBG':
        verbs+=1
print(verbs)
```

```
49
```

No. of Adjective :

In [25]:
```
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)
```

Out[25]: 15

No. of Adverb :

In [26]:
```
adj=[]
for i in top_pos.keys():
    (word,pos)=i
    if pos=='JJ' or pos=='JJR'or pos=='JJS':
        adj.append(i)
len(adj)
```

Out[26]: 67

Adverb Frequent :

In [27]:
```
adv=FreqDist(adv)
adv.most_common(1)
```

Out[27]:
```
[(('prior', 'RB'), 1)]
```

Adjective Frequent :

In [28]:
```
adv=FreqDist(adj)
adv.most_common(1)
```

Out[28]:
```
[(('iconic', 'JJ'), 1)]
```