BERCHMANS KEVIN S

215229107

Natural Language Processing Lab

Lab8. Exploring POS of Large Text Files

In [1]:
```
txt1 = open("12 Angry Men.txt", "r")
txt1 = txt1.read()
print(txt1)
```

```
Lumet's origins as a director of teledrama may well be obvious here in his firs
t film, but there is no denying the suitability of his style - sweaty close-up
s, gritty monochrome 'realism', one-set claustrophobia - to his subject. Script
ed by Reginald Rose from his own teleplay, the story is pretty contrived - duri
ng a murder trial, one man's doubts about the accused's guilt gradually overcom
e the rather less-than-democratic prejudices of the other eleven members of the
jury - but the treatment is tense, lucid, and admirably economical. Fonda, thou
gh typecast as the bastion of liberalism, gives a nicely underplayed performanc
e, while Cobb, Marshall and Begley in particular are highly effective in suppor
t. But what really transforms the piece from a rather talky demonstration that
a man is innocent until proven guilty, is the consistently taut, sweltering atm
osphere, created largely by Boris Kaufman's excellent camerawork. The result, h
owever devoid of action, is a strangely realistic thriller.
```

In [2]:
```
import glob
import nltk
import pandas as pd
from nltk import *
from zipfile import ZipFile
from nltk.corpus import stopwords


import nltk
nltk.download('stopwords')
nltk.download('punkt')
stop_words = set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\1mscdsa07\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\1mscdsa07\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

a. How many sentences in the files??

In [3]:
```
from nltk.tokenize import sent_tokenize
sentences=sent_tokenize(txt1)
len(sentences)
```

Out[3]: 5

b. How many words in the file??

In [4]:
```
from nltk.tokenize import word_tokenize
words_in = nltk.tokenize.WhitespaceTokenizer()
words = words_in.tokenize(txt1)
len(words)
```

Out[4]: 155

c. What are the top 10 words and their counts??

In [5]:
```
top_10 = FreqDist(words)
top_10.most_common(10)
```

Out[5]:
```
[('the', 10),
 ('a', 6),
 ('of', 6),
 ('is', 6),
 ('his', 4),
 ('-', 4),
 ('in', 3),
 ('as', 2),
 ('but', 2),
 ('by', 2)]
```

d. How many different POS tags are represented in this file??

In [6]:
```
nltk.download('averaged_perceptron_tagger')

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)
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\1mscdsa07\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[6]: 18

e. What are the top 10 POS tags and their counts??

In [7]:
```
top_pos = FreqDist(tagged)
top_pos.most_common(10)
```

Out[7]:
```
[(('-', ':'), 4),
 (('rather', 'RB'), 2),
 (("Lumet's", 'NNP'), 1),
 (('origins', 'VBZ'), 1),
 (('director', 'NN'), 1),
 (('teledrama', 'NN'), 1),
 (('may', 'MD'), 1),
 (('well', 'RB'), 1),
 (('obvious', 'VB'), 1),
 (('first', 'JJ'), 1)]
```

f. How many nouns in the file??

In [8]:
```
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)
```

```
41
```

g. How many verbs in the file??

In [9]:
```
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 == 'VBZ
        verbs+=1

print(verbs)
```

```
10
```

h. How many adjectives in the file??

In [10]:
```
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[10]:
```
19
```

i. How many adverbs in the file??

In [11]:
```
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[11]:
```
13
```

j. What is the most frequent adverb??

In [12]:
```
adv = FreqDist(adv)
adv.most_common(1)
```

Out[12]:
```
[(('well', 'RB'), 1)]
```

k. What is the most frequent adjective??

In [13]:
```
adj = FreqDist(adj)
adj.most_common(1)
```

Out[13]:
```
[(('first', 'JJ'), 1)]
```

In [ ]: