```
from zipfile import zipfile.
import glob.
import nltk.
import pandas as pd.
from nltk. import.
from nltk. Corpus import stopwords.
stop_words = set (stopwords. words ('english'))
nltk. download ('averaged_perceptron_tagger')
```

1) files = [file for file in glob. glob('imaries/*')]
for file in files[:]:
with open (files[-3], 'r', encoding = 'cp1252') as f:
cont = f. read()
print(count)

2) (a) from nltk. tokenize import sent_tokenize.
st = sent_tokenize (cont)
len(st)

(b) from nltk. tokenize. import word_tokenize.
tokenizer = nltk. tokenize. WhitespaceTokenizer()
tok = tokenizer. tokenize (count)
len(tok)

Natural Language Processing Lab
Lab8. Exploring POS of Large Text Files

EXERCISE-1

1. Open any movie file from your **movies** sub directory.

2. Tokenize your movie file and print the following

   a. How many sentences in the file?
   b. How many words in the file?
   c. What are the top 10 words and their counts?
   d. How many different POS tags are represented in this file?
   e. What are the top 10 POS tags and their counts?
   f. How many nouns in the file?
   g. How many verbs in the file?
   h. How many adjectives in the file?
   i. How many adverbs in the file?
   j. What is the most frequent adverb?
   k. What is the most frequent adjective?

c) fokfd = freqDist(tok)
fokfd.most_common(10)

d) tag = []
tem = []
tok = [w for w in fok if not w in stop_words
taggged = nltk.pos_tag(tok)
for i in fagged :
(word, pos) = i.
tag.append(pos)

for i in tag:
if i not in tem:
tem.append(i)

len(tem

e)
```
fokpos = FreqDist(tagged)
fokpos.most_common(10)
```

f)
```
n=0
for p in fokpos.keys():
    (word, pos) = p
    if pos == 'NN' or pos == 'NNS' or pos == 'NNP'
       or pos == 'NNPS':
        n+1
print(n)
```

How many verbs in the file?
```
V=0
for i in fokpos.keys()
    (word, pos) = p
    if pos == 'VB' or pos == 'VBD' or pos == 'VBP'
       or pos == 'VBZ':
        V += 1
print(V)
```

h. How many adjectives in the file?
```
adj = []
for i in fokpos.keys():
    (word, pos) = p
    if pos == 'JJR' or pos == 'JJS':
        adj.append p
len(adj)
```

NOTES

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

(j)
```
adv = .freqDist(adv)
adv.most_common(1)
```

(k)
```
.adj = Freq.Dist(adj)
adj.most_common(1).
```