NLP_LAB8_Exploring Part of Speech Tagging on Large Text Files

DINESH KUMAR_K_225229108

In [1]:

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

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

Out[1]:

```
True
```

In [2]:

```
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 [33]:

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

In [34]:

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

Out[34]:

```
11
```

In [35]:

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

Out[35]:

```
349
```

In [36]:
```
top10w=FreqDist(words)
top10w.most_common(10)
```

Out[36]:
```
[('the', 19),
 ('of', 15),
 ('a', 9),
 ('to', 8),
 ('and', 7),
 ('his', 7),
 ('in', 6),
 ('for', 5),
 ('Kurosawa', 4),
 ('he', 4)]
```

In [37]:
```
import nltk
nltk.download('averaged_perceptron_tagger')
```

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

Out[37]:
```
True
```

In [38]:
```
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[38]:
```
15
```

In [39]:
```
top_pos=FreqDist(tagged)
top_pos.most_common(10)
```

Out[39]:
```
[(('Kurosawa', 'NNP'), 4),
 (('film', 'NN'), 3),
 (('Japanese', 'JJ'), 3),
 (('RAN', 'NNP'), 3),
 (('At', 'IN'), 2),
 (('one', 'CD'), 2),
 (('work', 'NN'), 2),
 (('In', 'IN'), 2),
 (('RAN,', 'NNP'), 2),
 (('years', 'NNS'), 2)]
```

In [40]:

```
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)
```

```
98
```

In [41]:

```
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)
```

```
24
```

In [42]:

```
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[42]:

```
44
```

In [43]:

```
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[43]:

```
12
```

In [44]:

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

Out[44]:

```
[(('surely', 'RB'), 1)]
```

In [45]:

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

Out[45]:

```
[(('greatest', 'JJS'), 1)]
```