NLP_LAB8_Exploring Part of Speech Tagging on Large Text Files

225229117_JOSHUA

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 [3]:

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

In [4]:

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

Out[4]:

```
11
```

In [5]:

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

Out[5]:

```
307
```

In [6]:
```
top10w=FreqDist(words)
top10w.most_common(10)
```

Out[6]:
```
[('a', 13),
 ('to', 12),
 ('of', 11),
 ('the', 8),
 ('and', 7),
 ('its', 6),
 ('that', 4),
 ('for', 4),
 ('you', 4),
 ('in', 3)]
```

In [7]:
```
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[7]:
```
True
```

In [8]:
```
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[8]:
```
18
```

In [9]:
```
top_pos=FreqDist(tagged)
top_pos.most_common(10)
```

Out[9]:
```
[(('one', 'CD'), 2),
 (('Bogart’s', 'NNP'), 2),
 (('Rick’s', 'NNP'), 2),
 (('Michael', 'NNP'), 2),
 (('time', 'NN'), 2),
 (('Casablanca', 'NNP'), 2),
 (('last', 'JJ'), 2),
 (('While', 'IN'), 1),
 (('growing', 'VBG'), 1),
 (('wilds', 'NNS'), 1)]
```

In [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
```

In [11]:

```
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
```

In [12]:

```
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[12]:

```
39
```

In [13]:

```
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[13]:

```
13
```

In [14]:

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

Out[14]:

```
[(('anniversary.', 'RB'), 1)]
```

In [15]:

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

Out[15]:

```
[(('frame-by-frame', 'JJ'), 1)]
```