NLP_LAB8_Exploring Part of Speech Tagging on Large Text Files

225229110_HARI PRASATH_S

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 [46]:
```
files="Boyhood.txt"
f=open(files,'r')
content=f.read()
f.close()
```

In [47]:
```
from nltk.tokenize import sent_tokenize
sentences=sent_tokenize(content)
len(sentences)
```

Out[47]: 15

In [48]:
```
word=nltk.tokenize.WhitespaceTokenizer()
words=word.tokenize(content)
len(words)
```

Out[48]: 321

In [49]:
```
top10w=FreqDist(words)
top10w.most_common(10)
```

Out[49]:
```
[('a', 11),
 ('the', 10),
 ('and', 8),
 ('his', 6),
 ('of', 5),
 ('in', 4),
 ('that', 4),
 ('Linklater', 4),
 ('to', 4),
 ('an', 4)]
```

In [50]:
```
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[50]:
```
True
```

In [51]:
```
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[51]:
```
18
```

In [52]:
```
top_pos=FreqDist(tagged)
top_pos.most_common(10)
```

Out[52]:
```
[(('Linklater', 'NNP'), 4),
 (('young', 'JJ'), 3),
 (('Mason', 'NNP'), 3),
 (('He', 'PRP'), 3),
 (('Linklater’s', 'NNP'), 2),
 (('Boyhood,', 'NNP'), 2),
 (('every', 'DT'), 2),
 (('makes', 'VBZ'), 2),
 (('watching', 'VBG'), 2),
 (('end', 'NN'), 2)]
```

In [53]:
```
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)
```

```
92
```

In [54]:
```
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)
```

```
27
```

In [55]:
```
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[55]:
```
33
```

In [56]:
```
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[56]:
```
9
```

In [57]:
```
adv = FreqDist(adv)
adv.most_common(1)
```

Out[57]:
```
[(('ultimately', 'RB'), 1)]
```

In [58]:
```
adv = FreqDist(adj)
adv.most_common(1)
```

Out[58]:
```
[(('new', 'JJ'), 1)]
```