Homework Problem

•Extract a text from corpus •Tokensizethe words •POS_tagging? •Determine the stems in those words

In [1]:

```python
import nltk
from nltk.corpus import webtext
from nltk.stem import LancasterStemmer

content = webtext.raw(webtext.fileids()[0])[:] # Fetching the Firefox.txt webtext
print("length of corpus text : ",len(content))

words = nltk.word_tokenize(content)
print("\nContent: \n",words[:20]) # Webtext need to be word tokenize

tagged = nltk.pos_tag(words)
print("\n POS :\n",tagged[:20]) # POS_Tagging of words

stemmerLan = LancasterStemmer()
print("\nStemming :")
for word in words[:30]:
    stemm = stemmerLan.stem(word) #finding Stem for each words
    print(stemm)
```

```
length of corpus text :  564601

Content: 
 ['Cookie', 'Manager', ':', '``', 'Do', "n't", 'allow', 'sites', 'that', 'set', 'removed',
'cookies', 'to', 'set', 'future', 'cookies', "''", 'should', 'stay', 'checked']

 POS :
 [('Cookie', 'NNP'), ('Manager', 'NNP'), (':', ':'), ('``', '``'), ('Do', 'VBP'), ("n't", 'RB'), (
'allow', 'VB'), ('sites', 'NNS'), ('that', 'WDT'), ('set', 'VBP'), ('removed', 'VBN'), ('cookies',
'NNS'), ('to', 'TO'), ('set', 'VB'), ('future', 'JJ'), ('cookies', 'NNS'), ("''", "''"),
('should', 'MD'), ('stay', 'VB'), ('checked', 'VBD')]

Stemming :
cooky
man
:
``
do
n't
allow
sit
that
set
remov
cooky
to
set
fut
cooky
''
should
stay
check
when
in
ful
screen
mod
press
ctrl-n
should
op
a
```

In [9]:

```
webtext.fileids()[0]
```

Out[9]:

```
'firefox.txt'
```

In [10]:

```
content = webtext.raw(webtext.fileids()[0])[:] # Fetching the Firefox.txt webtext
```

In [11]:

```
len(content)
```

Out[11]:

```
564601
```