first

May 31, 2020

[2]:
```
import nltk
from nltk.tokenize import sent_tokenize,word_tokenize
file=open('dataadikkuka.txt','r',encoding='UTF-8')
Data=file.read()
sent_tokenz=sent_tokenize(Data)
word_tokenz=[word_tokenize(i) for i in sent_tokenz]
```

[19]:
```
StopWords=['``',"'",'   '  , ' ' , '   ' , '  ' , ' '  , '   ' , '  ' , ' ' ,␣
 ↪'  ' , '  '  , '  '  ,
           '   ' , '   ' , '   ' , '    ' , '    '  , '    '  , '    ' ,␣
 ↪'   '   ,
           ' ' , ' ' ,  ' ' , '  ' , '   '  , '   '  , '  ' , '  ' , '   '  ,␣
 ↪'   '  , '  ' ,
           '   '  , '   '  , '  '  , '  '  , '   ' , '  '  , '   ' , '   '  ,␣
 ↪'  ' , '  ' ,
           ' '  , ' '  , '   ' , '   ' , '  '  , '   ' , '   ' , '  ' ,'"' ,"'",␣
 ↪',' ,'?',"''",
           ':',';','(',')','-',']','['  '   , '  ' , '    ' , '  ' , '   ' ,␣
 ↪'      '  , '   '  ,
           '  '  , '  '  , '  '  , '   ' , '  '  , '  '  , '     ' , '   ' , '   ' ,␣
 ↪'   '  ,
           '  '  , '   ' , '  '  , '   ' , '  '  , '  ','  '  , '   '  , '   '
           , '  ' , '  ' , '    '  , '     ' , '  '  , '    ' , '    ' , '  '␣
 ↪, '  '  ,
           '  '  , '  '  , '  ' , '   '  , '   '  , '   '  , '    '  , '  ' ,␣
 ↪'   '  , '   '  ,
           '     '  , '    ' , '     ' , '     ' , '    ' , '    ' , '    ' ,␣
 ↪'   '  , '  ' ,
           '    '  , '    ' , '   ' , '   ' , '   ' , '    ' , '   ' , '    ' ,␣
 ↪'     '  , '    '  ,
           '    ' , '    ' , '   ' , '      ' , '    ' , '   ' , '    ' , ' ' ,␣
 ↪'  ' , '     ' ,
           '  '  , '    ' , '  '  , '  '  , '  '  , '  '  , '  '  , '  '     ,␣
 ↪'     '  , '  ' ,
           '   ' ,'   '   , '  ' , '    ' , '    ' , '   ' , '  ','  ', '  ',␣
 ↪'  ', '  ',
          ␣
↪'   ','   ','   ','   ','   ','   ',' ','    ',' ','   ','     ','    ',' ','   ','   ',
          ␣
↪'   ','   ','   ','   ','   ','   ','   ','   ','   ','   ',' ','   ','   ','!',"'"]
```

[ ]:

```
Removing Stopwords
```

[21]:

```
NewData=[]
for i in word_tokenz:
    temp=[]
    for j in i:
        if j not in StopWords:
            temp.append(j)
    NewData.append(temp)
```

[ ]:

```
Writing to text file for tagging
```

[24]:

```
filetxt=open('fortagadikkuka.txt','w')
for i in NewData:
    for j in i:
        filetxt.writelines(j)
        filetxt.writelines('\n')
```

[ ]:

```
Store taggeddata from text file to an array
```

[25]:

```
a=open('taggedadikkuka.txt').readlines()
temp=[]
m=[]
for i in a:
    if i=='.\t\t\t/RD_PUNC\n':
        m.append(temp)
        temp=[]
    else:
        temp.append(i)
```

[ ]:

```
Writing the DAta from array to csv file along with its label and sense
```

[28]:

```
import csv
with open('csvadikkuka.csv','w',newline='') as tagfile:
    writer=csv.writer(tagfile)
    writer.writerow(['sentence','ambigous_word','label','sense'])
    for i in m:
        writer.writerow([i,'  ',1,'    '])
```

[ ]: