19MAI0017 PARAG PUJARI NATURAL LANGUAGE PROCESSING

LAB1 -----22/05/2020

In [2]:
```
#IMPORT NLTK PACKAGE
import nltk
```

In [3]:
```
#IMPORT BOOK FROM NLTK
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

In [4]:
```
#FROM CORPUS IMPORT BROWN CORPUS AND ACCESS
from nltk.corpus import brown
brown.categories()
```

Out[4]:
```
['adventure',
 'belles_lettres',
 'editorial',
```

```
 'fiction',
 'government',
 'hobbies',
 'humor',
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction']
```

In [5]:
```
brown.words(categories='hobbies')[:100]# LIST OF CATEGORY hobbies IN BR
OWN
```

Out[5]:
```
['Too', 'often', 'a', 'beginning', 'bodybuilder', ...]
```

In [6]:
```
#FROM CORPUS IMPORT INUGURAL CORPUS AND ACCESS
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.words(fileids='2017-Trump.txt')
```

Out[6]:
```
['Chief', 'Justice', 'Roberts', ',', 'President', ...]
```

In [7]:
```
# LIST OF 100 WORDS IN THE ONE OF THE INAUGURAL SPEECHES BY PRESIDENT B
ILL CLINTON
inaugural.words(fileids='1993-clinton.txt')[:100]
```

Out[7]:
```
['My',
 'fellow',
 'citizens',
 ',',
 'today',
 'we',
 'celebrate',
 'the',
 'mystery',
 'of',
```

```
                'American',
                'renewal',
                '.',
                'This',
                'ceremony',
                'is',
                'held',
                'in',
                'the',
                'depth',
                'of',
                'winter',
                ',',
                'but',
                'by',
                'the',
                'words',
                'we',
                'speak',
                'and',
                'the',
                'faces',
                'we',
                'show',
                'the',
                'world',
                ',',
                'we',
                'force',
                'the',
                'spring',
                '.',
                'A',
                'spring',
                'reborn',
                'in',
                'the',
                'world',
                "'",
                's',
```

```
                'oldest',
                'democracy',
                ',',
                'that',
                'brings',
                'forth',
                'the',
                'vision',
                'and',
                'courage',
                'to',
                'reinvent',
                'America',
                '.',
                'When',
                'our',
                'founders',
                'boldly',
                'declared',
                'America',
                "'",
                's',
                'independence',
                'to',
                'the',
                'world',
                ',',
                'and',
                'our',
                'purposes',
                'to',
                'the',
                'Almighty',
                ',',
                'they',
                'knew',
                'that',
                'America',
                ',',
```

```
 'to',
 'endure',
 ',',
 'would',
 'have',
 'to',
 'change',
 '.',
 'Not',
 'change',
 'for']
```

In [8]:
```
#DISPLAY FIELDS OF INAUGURAL CORPUS AND GET WORDS FROM 1829-JACKSON.TXT
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.words(fileids='1829-Jackson.txt')
```

Out[8]:
```
['Fellow', 'citizens', ',', 'about', 'to', 'undertake', ...]
```

In [9]:
```
#DISPLAY FIELDS OF INAUGURAL CORPUS AND GET WORDS FROM 1841-Harrison.tx
t
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.words(fileids='1841-Harrison.txt')
```

Out[9]:
```
['Called', 'from', 'a', 'retirement', 'which', 'I', ...]
```

In [10]:
```
#DISPLAY THE FIELDS OF INAUGURAL
inaugural.fileids()
```

Out[10]:
```
['1789-Washington.txt',
 '1793-Washington.txt',
 '1797-Adams.txt',
 '1801-Jefferson.txt',
 '1805-Jefferson.txt',
 '1809-Madison.txt',
 '1813-Madison.txt',
 '1817-Monroe.txt',
```

```
                '1821-Monroe.txt',
                '1825-Adams.txt',
                '1829-Jackson.txt',
                '1833-Jackson.txt',
                '1837-VanBuren.txt',
                '1841-Harrison.txt',
                '1845-Polk.txt',
                '1849-Taylor.txt',
                '1853-Pierce.txt',
                '1857-Buchanan.txt',
                '1861-Lincoln.txt',
                '1865-Lincoln.txt',
                '1869-Grant.txt',
                '1873-Grant.txt',
                '1877-Hayes.txt',
                '1881-Garfield.txt',
                '1885-Cleveland.txt',
                '1889-Harrison.txt',
                '1893-Cleveland.txt',
                '1897-McKinley.txt',
                '1901-McKinley.txt',
                '1905-Roosevelt.txt',
                '1909-Taft.txt',
                '1913-Wilson.txt',
                '1917-Wilson.txt',
                '1921-Harding.txt',
                '1925-Coolidge.txt',
                '1929-Hoover.txt',
                '1933-Roosevelt.txt',
                '1937-Roosevelt.txt',
                '1941-Roosevelt.txt',
                '1945-Roosevelt.txt',
                '1949-Truman.txt',
                '1953-Eisenhower.txt',
                '1957-Eisenhower.txt',
                '1961-Kennedy.txt',
                '1965-Johnson.txt',
                '1969-Nixon.txt',
                '1973-Nixon.txt',
```

```
 '1977-Carter.txt',
 '1981-Reagan.txt',
 '1985-Reagan.txt',
 '1989-Bush.txt',
 '1993-Clinton.txt',
 '1997-Clinton.txt',
 '2001-Bush.txt',
 '2005-Bush.txt',
 '2009-Obama.txt',
 '2013-Obama.txt',
 '2017-Trump.txt']
```

In [11]:
```
#IMPORT WEBTEXT AND DISPLAY THE DATA
from nltk.corpus import webtext
webtext.fileids()
for fileid in webtext.fileids():
    print(fileid,webtext.raw(fileid)[:50])
```

```
firefox.txt Cookie Manager: "Don't allow sites that set remove
grail.txt SCENE 1: [wind] [clop clop clop]
KING ARTHUR: Who
overheard.txt White guy: So, do you have any plans for this even
pirates.txt PIRATES OF THE CARRIBEAN: DEAD MAN'S CHEST, by Ted
singles.txt 25 SEXY MALE, seeks attrac older single lady, for
wine.txt Lovely delicate, fragrant Rhone wine. Polished lea
```

In [12]:
```
#FREQUENCY DISTRIBUTION
text1="The basis for the work is melvilles 1841 whaling voyage aboard t
he acushnet"
fd=nltk.FreqDist(text1.split())
```

In [13]:
```
fd
```

Out[13]:
```
FreqDist({'the': 2, 'The': 1, 'basis': 1, 'for': 1, 'work': 1, 'is': 1,
'melvilles': 1, '1841': 1, 'whaling': 1, 'voyage': 1, ...})
```

In [14]:
```
#CONDITIONAL FREQUENCY DISTRIBUTION
from nltk.probability import ConditionalFreqDist
```

```
cfd=ConditionalFreqDist((len(word),word) for word in text1.split())
cfd[3]
```

Out[14]:
```
FreqDist({'the': 2, 'The': 1, 'for': 1})
```

In [15]:
```
cfd[6]
```

Out[15]:
```
FreqDist({'voyage': 1, 'aboard': 1})
```

HOME WORK1--------22/05/2020

In [16]:
```
from nltk.tokenize import sent_tokenize
#converting list to single string str1
text1=inaugural.words(fileids='2017-Trump.txt')
str1=" ".join(text1)
str1[:500]
```

Out[16]:
```
'Chief Justice Roberts , President Carter , President Clinton , Preside
nt Bush , President Obama , fellow Americans , and people of the world
: Thank you . We , the citizens of America , are now joined in a great
national effort to rebuild our country and restore its promise for all
of our people . Together , we will determine the course of America and
the world for many , many years to come . We will face challenges , we
will confront hardships , but we will get the job done . Every 4 years
, we'
```

In [17]:
```
# frequency distribution of words in a text
text='"Chief Justice Roberts , President Carter , President Clinton , P
resident Bush , President Obama , fellow Americans , and people of the
 world : Thank you . We , the citizens of America , are now joined in a
 great national effort to rebuild our country and restore its promise f
or all of our people . "'
fd=nltk.FreqDist(text.split())
fd
```

Out[17]:
```
FreqDist({',': 8, 'President': 4, 'of': 3, 'and': 2, 'people': 2, 'th
e': 2, '.': 2, 'our': 2, '"Chief': 1, 'Justice': 1, ...})
```

In [18]:
```
from nltk.probability import ConditionalFreqDist
cfd=ConditionalFreqDist((len(word),word) for word in text.split())
#list of conditons
cfd.conditions()
```

Out[18]:
```
[6, 7, 1, 9, 4, 5, 3, 2, 8]
```

In [19]:
```
cfd[8]
```

Out[19]:
```
FreqDist({'citizens': 1, 'national': 1})
```

In [20]:
```
inauguraldata = inaugural.words(fileids = '2017-Trump.txt')
cdf = ConditionalFreqDist((len(word), word) for word in speech)
cdf[4]
```

```
---------------------------------------------------------------------
----
NameError                                 Traceback (most recent call l
ast)
<ipython-input-20-f14740b1322c> in <module>
      1 inauguraldata = inaugural.words(fileids = '2017-Trump.txt')
----> 2 cdf = ConditionalFreqDist((len(word), word) for word in speech)
      3 cdf[4]

NameError: name 'speech' is not defined
```

In [21]:
```
data = []
for words in inauguraldata:
    if len(words)>4:
        data.append(words)
```

In [22]:
```
# FINDING OUT THE FREQUENCY DISTRIBUTIONOF THE WORDS
fd = nltk.FreqDist(data)
fd
```

Out[22]:
```
FreqDist({'America': 20, 'American': 11, 'people': 10, 'their': 10, 'co
untry': 9, 'again': 9, 'world': 6, 'great': 6, 'Nation': 6, 'while': 6,
...})
```

In [23]:
```
sorted_fd = sorted(fd.items(), key = lambda x:x[1])
```

In [24]:
```
word = list(sorted_fd)[len(sorted_fd)-1]
```

In [25]:
```
print("Most frequently Used used word are :",word)
```

```
Most frequently Used used word are : ('America', 20)
```

In [ ]:
```

```