NATURAL LANGUAGE PROCESSING TASK 3

In [6]:
```
import nltk
```

In [7]:
```
nltk.download()
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

Out[7]:
```
True
```

BROWN CORPUS

Brown Corpus contains 1 Million words Oldest corpus

In [9]:
```
from nltk.corpus import brown
```

Categories in Brown Corpus

In [10]:
```
brown.categories()
```

Out[10]:
```
['adventure',
 'belles_lettres',
 'editorial',
 'fiction',
 'government',
 'hobbies',
 'humor',
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction']
```

In [11]:
```
brown.words(categories='adventure')    # gives the top words of adventure category
```

Out[11]:
```
['Dan', 'Morgan', 'told', 'himself', 'he', 'would', ...]
```

In [12]:
```
brown.words(categories='adventure')[:100] # gives the first 100 words of adventure category
```

Out[12]:
```
['Dan', 'Morgan', 'told', 'himself', 'he', 'would', ...]
```

In [14]:
```
brown.words(categories='editorial')[:50]  #gives the first 50 words of editorial category
```

Out[14]:
```
['Assembly',
 'session',
 'brought',
 'much',
 'good',
 'The',
 'General',
 'Assembly',
 ',',
 'which',
 'adjourns',
 'today',
 ',',
 'has',
 'performed',
 'in',
 'an',
 'atmosphere',
 'of',
 'crisis',
 'and',
 'struggle',
 'from',
 'the',
 'day',
 'it',
 'convened',
 '.',
 'It',
 'was',
 'faced',
 'immediately',
 'with',
 'a',
 'showdown',
 'on',
 'the',
 'schools',
 ',',
 'an',
 'issue',
 'which',
 'was',
 'met',
 'squarely',
 'in',
 'conjunction',
 'with',
 'the',
 'governor']
```

In [15]:
```
brown.words(categories='hobbies')[:30] # first 30 words of hobbies
```

Out[15]:
```
['Too',
 'often',
 'a',
 'beginning',
 'bodybuilder',
 'has',
 'to',
 'do',
 'his',
 'training',
 'secretly',
 'either',
 'because',
 'his',
 'parents',
 "don't",
 'want',
 'sonny-boy',
 'to',
 '``',
 'lift',
 'all',
 'those',
 'old',
 'barbell',
 'things',
 "''",
 'because',
 '``',
 "you'll"]
```

Inaugural corpus

In [16]:
```
from nltk.corpus import inaugural
```

In [18]:
```
inaugural.fileids()
```

Out[18]:
```
['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 [22]:
```
inaugural.raw('2013-Obama.txt')
y = (inaugural.words(fileids='2013-Obama.txt'))
print(y[:50])
```

```
['Thank', 'you', '.', 'Thank', 'you', 'so', 'much', '.', 'Vice', 'President', 'Biden', ',', 'M
r', '.', 'Chief', 'Justice', ',', 'Members', 'of', 'the', 'United', 'States', 'Congress', ',',
'distinguished', 'guests', ',', 'and', 'fellow', 'citizens', ':', 'Each', 'time', 'we', 'gathe
r', 'to', 'inaugurate', 'a', 'President', 'we', 'bear', 'witness', 'to', 'the', 'enduring', 'str
ength', 'of', 'our', 'Constitution', '.']
```

In [23]:
```
inaugural.raw('1933-Roosevelt.txt')
y = (inaugural.words(fileids='1933-Roosevelt.txt'))
print(y[:15])
```

```
['I', 'am', 'certain', 'that', 'my', 'fellow', 'Americans', 'expect', 'that', 'on', 'my', 'induc
tion', 'into', 'the', 'Presidency']
```

Webtext

In [25]:
```
from nltk.corpus import webtext
webtext.fileids()
webtext.raw('pirates.txt')
x = (webtext.words(fileids="pirates.txt"))
print(x[:30])
```

```
['PIRATES', 'OF', 'THE', 'CARRIBEAN', ':', 'DEAD', 'MAN', "'", 'S', 'CHEST', ',', 'by', 'Ted',
'Elliott', '&', 'Terry', 'Rossio', '[', 'view', 'looking', 'straight', 'down', 'at', 'rolling',
'swells', ',', 'sound', 'of', 'wind', 'and']
```

Books in Book Corpus

In [26]:
```
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 [ ]:
```

```