In [7]:
```
import nltk
```

In [8]:
```
nltk.__file__
```

Out[8]:
```
'C:\\Users\\Rehan Shaikh\\AppData\\Local\\Programs\\Python\\Python310\\lib\\site-packages\\nltk\\__init__.py'
```

In [9]:
```
#downloading required collection and packages
nltk.download('popular')
```

```
[nltk_data] Downloading collection 'popular'
[nltk_data]    |
[nltk_data]    | Downloading package cmudict to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\cmudict.zip.
[nltk_data]    | Downloading package gazetteers to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\gazetteers.zip.
[nltk_data]    | Downloading package genesis to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\genesis.zip.
[nltk_data]    | Downloading package gutenberg to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\gutenberg.zip.
[nltk_data]    | Downloading package inaugural to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\inaugural.zip.
[nltk_data]    | Downloading package movie_reviews to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\movie_reviews.zip.
[nltk_data]    | Downloading package names to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\names.zip.
[nltk_data]    | Downloading package shakespeare to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\shakespeare.zip.
[nltk_data]    | Downloading package stopwords to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\stopwords.zip.
[nltk_data]    | Downloading package treebank to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\treebank.zip.
[nltk_data]    | Downloading package twitter_samples to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\twitter_samples.zip.
[nltk_data]    | Downloading package omw to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\omw.zip.
[nltk_data]    | Downloading package omw-1.4 to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\omw-1.4.zip.
[nltk_data]    | Downloading package wordnet to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\wordnet.zip.
[nltk_data]    | Downloading package wordnet2021 to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\wordnet2021.zip.
[nltk_data]    | Downloading package wordnet31 to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\wordnet31.zip.
[nltk_data]    | Downloading package wordnet_ic to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\wordnet_ic.zip.
[nltk_data]    | Downloading package words to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\words.zip.
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     C:\Users\Rehan
```

```
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping chunkers\maxent_ne_chunker.zip.
[nltk_data]    | Downloading package punkt to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping tokenizers\punkt.zip.
[nltk_data]    | Downloading package snowball_data to C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     C:\Users\Rehan
[nltk_data]    |     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping taggers\averaged_perceptron_tagger.zip.
[nltk_data]    |
[nltk_data]  Done downloading collection popular
```

Out[9]:
```
True
```

In [10]:
```
nltk.download('gutenberg')
```

```
[nltk_data] Downloading package gutenberg to C:\Users\Rehan
[nltk_data]     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]   Package gutenberg is already up-to-date!
```

Out[10]:
```
True
```

In [11]:
```
#to check for the names of the books i.e fileids contained in the package gutenber
nltk.corpus.gutenberg.fileids()
```

Out[11]:
```
['austen-emma.txt',
 'austen-persuasion.txt',
 'austen-sense.txt',
 'bible-kjv.txt',
 'blake-poems.txt',
 'bryant-stories.txt',
 'burgess-busterbrown.txt',
 'carroll-alice.txt',
 'chesterton-ball.txt',
 'chesterton-brown.txt',
 'chesterton-thursday.txt',
 'edgeworth-parents.txt',
 'melville-moby_dick.txt',
 'milton-paradise.txt',
 'shakespeare-caesar.txt',
 'shakespeare-hamlet.txt',
 'shakespeare-macbeth.txt',
 'whitman-leaves.txt']
```

In [12]:
```
!cat 'C:\Users\MSCIT\AppData\Roaming\nltk_data\corpora\gutenberg'
```

```
'cat' is not recognized as an internal or external command,
operable program or batch file.
```

In [54]:
```
#to check the raw content of a paticular file in a package
from nltk.corpus import gutenberg
#gutenberg.raw('austen-emma.txt')
```

In [53]:
```
#checking for raw content of two files simultaneously
#gutenberg.raw(['austen-emma.txt','bible-kjv.txt'])
```

In [15]:
```
gutenberg.sents('austen-emma.txt')
```

Out[15]:
```
[['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']'], ['VOLUME', 'I'], ...]
```

In [16]:
```
#number of sentences in a particular fileid
len(nltk.corpus.gutenberg.sents('austen-emma.txt'))
```

Out[16]:
```
7717
```

In [17]:
```
#total number of sentences in the package i.e corpus
len(nltk.corpus.gutenberg.sents())
```

Out[17]:
```
98503
```

In [18]:
```
#total no of words in the corpus
len(nltk.corpus.gutenberg.words())
```

Out[18]:
```
2621613
```

In [19]:
```
#total no of words in a particular fileid in the corpus
len(nltk.corpus.gutenberg.words('austen-emma.txt'))
```

Out[19]:
```
192427
```

Brown Corpus

In [20]:
```
#downloading another corpus/package
nltk.download('brown')
```

Out[20]:
```
[nltk_data] Downloading package brown to C:\Users\Rehan
[nltk_data]     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\brown.zip.
True
```

In [21]:
```
from nltk.corpus import brown
```

In [22]:
```
brown.fileids()
```

Out[22]:
```
['ca01',
 'ca02',
 'ca03',
 'ca04',
 'ca05',
 'ca06',
 'ca07',
 'ca08',
 'ca09',
 'ca10',
 'ca11',
 'ca12',
 'ca13',
 'ca14',
 'ca15',
 'ca16',
 'ca17',
 'ca18',
 'ca19',
 'ca20',
 'ca21',
 'ca22',
 'ca23',
 'ca24',
 'ca25',
 'ca26',
 'ca27',
 'ca28',
 'ca29',
 'ca30',
 'ca31',
 'ca32',
 'ca33',
 'ca34',
 'ca35',
 'ca36',
 'ca37',
 'ca38',
 'ca39',
 'ca40',
 'ca41',
 'ca42',
 'ca43',
 'ca44',
 'cb01',
 'cb02',
 'cb03',
 'cb04',
 'cb05',
 'cb06',
 'cb07',
 'cb08',
 'cb09',
 'cb10',
 'cb11',
 'cb12',
 'cb13',
 'cb14',
 'cb15',
```

```
 'cb16',
 'cb17',
 'cb18',
 'cb19',
 'cb20',
 'cb21',
 'cb22',
 'cb23',
 'cb24',
 'cb25',
 'cb26',
 'cb27',
 'cc01',
 'cc02',
 'cc03',
 'cc04',
 'cc05',
 'cc06',
 'cc07',
 'cc08',
 'cc09',
 'cc10',
 'cc11',
 'cc12',
 'cc13',
 'cc14',
 'cc15',
 'cc16',
 'cc17',
 'cd01',
 'cd02',
 'cd03',
 'cd04',
 'cd05',
 'cd06',
 'cd07',
 'cd08',
 'cd09',
 'cd10',
 'cd11',
 'cd12',
 'cd13',
 'cd14',
 'cd15',
 'cd16',
 'cd17',
 'ce01',
 'ce02',
 'ce03',
 'ce04',
 'ce05',
 'ce06',
 'ce07',
 'ce08',
 'ce09',
 'ce10',
 'ce11',
 'ce12',
 'ce13',
```

```
'ce14',
'ce15',
'ce16',
'ce17',
'ce18',
'ce19',
'ce20',
'ce21',
'ce22',
'ce23',
'ce24',
'ce25',
'ce26',
'ce27',
'ce28',
'ce29',
'ce30',
'ce31',
'ce32',
'ce33',
'ce34',
'ce35',
'ce36',
'cf01',
'cf02',
'cf03',
'cf04',
'cf05',
'cf06',
'cf07',
'cf08',
'cf09',
'cf10',
'cf11',
'cf12',
'cf13',
'cf14',
'cf15',
'cf16',
'cf17',
'cf18',
'cf19',
'cf20',
'cf21',
'cf22',
'cf23',
'cf24',
'cf25',
'cf26',
'cf27',
'cf28',
'cf29',
'cf30',
'cf31',
'cf32',
'cf33',
'cf34',
'cf35',
'cf36',
```

```
'cf37',
'cf38',
'cf39',
'cf40',
'cf41',
'cf42',
'cf43',
'cf44',
'cf45',
'cf46',
'cf47',
'cf48',
'cg01',
'cg02',
'cg03',
'cg04',
'cg05',
'cg06',
'cg07',
'cg08',
'cg09',
'cg10',
'cg11',
'cg12',
'cg13',
'cg14',
'cg15',
'cg16',
'cg17',
'cg18',
'cg19',
'cg20',
'cg21',
'cg22',
'cg23',
'cg24',
'cg25',
'cg26',
'cg27',
'cg28',
'cg29',
'cg30',
'cg31',
'cg32',
'cg33',
'cg34',
'cg35',
'cg36',
'cg37',
'cg38',
'cg39',
'cg40',
'cg41',
'cg42',
'cg43',
'cg44',
'cg45',
'cg46',
'cg47',
```

```
 'cg48',
 'cg49',
 'cg50',
 'cg51',
 'cg52',
 'cg53',
 'cg54',
 'cg55',
 'cg56',
 'cg57',
 'cg58',
 'cg59',
 'cg60',
 'cg61',
 'cg62',
 'cg63',
 'cg64',
 'cg65',
 'cg66',
 'cg67',
 'cg68',
 'cg69',
 'cg70',
 'cg71',
 'cg72',
 'cg73',
 'cg74',
 'cg75',
 'ch01',
 'ch02',
 'ch03',
 'ch04',
 'ch05',
 'ch06',
 'ch07',
 'ch08',
 'ch09',
 'ch10',
 'ch11',
 'ch12',
 'ch13',
 'ch14',
 'ch15',
 'ch16',
 'ch17',
 'ch18',
 'ch19',
 'ch20',
 'ch21',
 'ch22',
 'ch23',
 'ch24',
 'ch25',
 'ch26',
 'ch27',
 'ch28',
 'ch29',
 'ch30',
 'cj01',
```

```
'cj02',
'cj03',
'cj04',
'cj05',
'cj06',
'cj07',
'cj08',
'cj09',
'cj10',
'cj11',
'cj12',
'cj13',
'cj14',
'cj15',
'cj16',
'cj17',
'cj18',
'cj19',
'cj20',
'cj21',
'cj22',
'cj23',
'cj24',
'cj25',
'cj26',
'cj27',
'cj28',
'cj29',
'cj30',
'cj31',
'cj32',
'cj33',
'cj34',
'cj35',
'cj36',
'cj37',
'cj38',
'cj39',
'cj40',
'cj41',
'cj42',
'cj43',
'cj44',
'cj45',
'cj46',
'cj47',
'cj48',
'cj49',
'cj50',
'cj51',
'cj52',
'cj53',
'cj54',
'cj55',
'cj56',
'cj57',
'cj58',
'cj59',
'cj60',
```

```
 'cj61',
 'cj62',
 'cj63',
 'cj64',
 'cj65',
 'cj66',
 'cj67',
 'cj68',
 'cj69',
 'cj70',
 'cj71',
 'cj72',
 'cj73',
 'cj74',
 'cj75',
 'cj76',
 'cj77',
 'cj78',
 'cj79',
 'cj80',
 'ck01',
 'ck02',
 'ck03',
 'ck04',
 'ck05',
 'ck06',
 'ck07',
 'ck08',
 'ck09',
 'ck10',
 'ck11',
 'ck12',
 'ck13',
 'ck14',
 'ck15',
 'ck16',
 'ck17',
 'ck18',
 'ck19',
 'ck20',
 'ck21',
 'ck22',
 'ck23',
 'ck24',
 'ck25',
 'ck26',
 'ck27',
 'ck28',
 'ck29',
 'cl01',
 'cl02',
 'cl03',
 'cl04',
 'cl05',
 'cl06',
 'cl07',
 'cl08',
 'cl09',
 'cl10',
```

```
 'cl11',
 'cl12',
 'cl13',
 'cl14',
 'cl15',
 'cl16',
 'cl17',
 'cl18',
 'cl19',
 'cl20',
 'cl21',
 'cl22',
 'cl23',
 'cl24',
 'cm01',
 'cm02',
 'cm03',
 'cm04',
 'cm05',
 'cm06',
 'cn01',
 'cn02',
 'cn03',
 'cn04',
 'cn05',
 'cn06',
 'cn07',
 'cn08',
 'cn09',
 'cn10',
 'cn11',
 'cn12',
 'cn13',
 'cn14',
 'cn15',
 'cn16',
 'cn17',
 'cn18',
 'cn19',
 'cn20',
 'cn21',
 'cn22',
 'cn23',
 'cn24',
 'cn25',
 'cn26',
 'cn27',
 'cn28',
 'cn29',
 'cp01',
 'cp02',
 'cp03',
 'cp04',
 'cp05',
 'cp06',
 'cp07',
 'cp08',
 'cp09',
 'cp10',
```

```
 'cp11',
 'cp12',
 'cp13',
 'cp14',
 'cp15',
 'cp16',
 'cp17',
 'cp18',
 'cp19',
 'cp20',
 'cp21',
 'cp22',
 'cp23',
 'cp24',
 'cp25',
 'cp26',
 'cp27',
 'cp28',
 'cp29',
 'cr01',
 'cr02',
 'cr03',
 'cr04',
 'cr05',
 'cr06',
 'cr07',
 'cr08',
 'cr09']
```

In [23]:

```
#checking for the categories in the corpus
brown.categories()
```

Out[23]:

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

In [24]:

```
#chceking for fileids of a particular category
brown.fileids(['adventure'])
```

Out[24]:
```
['cn01',
 'cn02',
 'cn03',
 'cn04',
 'cn05',
 'cn06',
 'cn07',
 'cn08',
 'cn09',
 'cn10',
 'cn11',
 'cn12',
 'cn13',
 'cn14',
 'cn15',
 'cn16',
 'cn17',
 'cn18',
 'cn19',
 'cn20',
 'cn21',
 'cn22',
 'cn23',
 'cn24',
 'cn25',
 'cn26',
 'cn27',
 'cn28',
 'cn29']
```

In [25]:
```
#finding in which category does a particular fileid belong to
brown.categories(['cn28'])
```

Out[25]:
```
['adventure']
```

In [50]:
```
#checking raw content of fileids simultaneously
#brown.raw(fileids=['cg15','ch15'])
```

In [51]:
```
#brown.raw(categories=['adventure','romance'])
```

In [52]:
```
#brown.raw(fileids=['ca30','cn08'])
```

In [29]:
```
len(brown.sents())
```

Out[29]:
```
57340
```

In [30]:
```
len(brown.words())
```

Out[30]:
```
1161192
```

In [31]:
```
len(brown.words('ca05'))
```

Out[31]:
```
2244
```

In [32]:
```
brown.sents(categories=['adventure'])
```

Out[32]:
```
[['Dan', 'Morgan', 'told', 'himself', 'he', 'would', 'forget', 'Ann', 'Turner',
'.'], ['He', 'was', 'well', 'rid', 'of', 'her', '.'], ...]
```

In [33]:
```
#checking for absolute path of a particular fileid of the corpus
brown.abspath('ca01')
```

Out[33]:
```
FileSystemPathPointer('C:\\Users\\Rehan Shaikh\\AppData\\Roaming\\nltk_data\\corpor
a\\brown\\ca01')
```

In [34]:
```
#checking for the path of the corpus
brown.root
```

Out[34]:
```
FileSystemPathPointer('C:\\Users\\Rehan Shaikh\\AppData\\Roaming\\nltk_data\\corpor
a\\brown')
```

In [35]:
```
brown.encoding('cn05')
```

Out[35]:
```
'ascii'
```

In [36]:
```
brown.readme()
```

Out[36]:
```
'BROWN CORPUS\n\nA Standard Corpus of Present-Day Edited American\nEnglish, for use
with Digital Computers.\n\nby W. N. Francis and H. Kucera (1964)\nDepartment of Lin
guistics, Brown University\nProvidence, Rhode Island, USA\n\nRevised 1971, Revised
and Amplified 1979\n\nhttp://www.hit.uib.no/icame/brown/bcm.html\n\nDistributed wit
h the permission of the copyright holder,\nredistribution permitted.\n'
```

Inaugral Corpus

In [37]:
```
nltk.download('inaugural')
```

Out[37]:
```
[nltk_data] Downloading package inaugural to C:\Users\Rehan
[nltk_data]     Shaikh\AppData\Roaming\nltk_data...
[nltk_data]   Package inaugural is already up-to-date!
True
```

In [38]:
```
from nltk.corpus import inaugural
```

In [39]:
```
inaugural.fileids()
```

Out[39]:
```
['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',
 '2021-Biden.txt']
```

In [40]:
```
inaugural.words()
```

Out[40]:
```
['Fellow', '-', 'Citizens', 'of', 'the', 'Senate', ...]
```

In [41]:
```
inaugural.sents()
```

Out[41]:
```
[['Fellow', '-', 'Citizens', 'of', 'the', 'Senate', 'and', 'of', 'the', 'House', 'o
f', 'Representatives', ':'], ['Among', 'the', 'vicissitudes', 'incident', 'to', 'li
fe', 'no', 'event', 'could', 'have', 'filled', 'me', 'with', 'greater', 'anxieties
', 'than', 'that', 'of', 'which', 'the', 'notification', 'was', 'transmitted', 'by
', 'your', 'order', ',', 'and', 'received', 'on', 'the', '14th', 'day', 'of', 'the
', 'present', 'month', '.'], ...]
```

In [42]:
```
len(inaugural.words())
```

Out[42]:
```
152901
```

In [43]:
```
len(inaugural.sents())
```

Out[43]:
```
5217
```

In [44]:
```
a = inaugural.sents('1789-Washington.txt')
```

In [45]:
```
#chcking for the contents of a particular sentence of a particular fileid
a[0]
```

Out[45]:
```
['Fellow',
 '-',
 'Citizens',
 'of',
 'the',
 'Senate',
 'and',
 'of',
 'the',
 'House',
 'of',
 'Representatives',
 ':']
```

In [46]:
```
len(a[0])
```

Out[46]:
```
13
```

In [47]:
```
max(len(s) for s in a )
```

Out[47]:
```
150
```

In [48]:
```
g = gutenberg.sents('austen-emma.txt')
g[0]
```

Out[48]:
```
['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']']
```

In [49]:
```
max(len(s) for s in g)
```

Out[49]:
```
274
```

In [ ]:
```

```