In [ ]:

```
# This Python 3 environment comes with many helpful analytics libraries inst
# It is defined by the kaggle/python Docker image: https://github.com/kaggle
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will l

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that
# You can also write temporary files to /kaggle/temp/, but they won't be sav
```

In [1]:

```
!pip install nltk
!pip install matplotlib gensim
```

```
Requirement already satisfied: nltk in /opt/conda/lib/python3.10/site-pack
ages (3.2.4)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packa
ges (from nltk) (1.16.0)
WARNING: Error parsing requirements for aiohttp: [Errno 2] No such file or
directory: '/opt/conda/lib/python3.10/site-packages/aiohttp-3.9.1.dist-inf
o/METADATA'
Requirement already satisfied: matplotlib in /opt/conda/lib/python3.10/sit
e-packages (3.7.5)
Requirement already satisfied: gensim in /opt/conda/lib/python3.10/site-pa
ckages (4.3.2)
Requirement already satisfied: contourpy>=1.0.1 in /opt/conda/lib/python3.
10/site-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.10/s
ite-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /opt/conda/lib/python
3.10/site-packages (from matplotlib) (4.47.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/conda/lib/python
3.10/site-packages (from matplotlib) (1.4.5)
Requirement already satisfied: numpy<2,>=1.20 in /opt/conda/lib/python3.1
0/site-packages (from matplotlib) (1.26.4)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.1
0/site-packages (from matplotlib) (21.3)
Requirement already satisfied: pillow>=6.2.0 in /opt/conda/lib/python3.10/
site-packages (from matplotlib) (9.5.0)
Requirement already satisfied: pyparsing>=2.3.1 in /opt/conda/lib/python3.
10/site-packages (from matplotlib) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in /opt/conda/lib/pyth
on3.10/site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: scipy>=1.7.0 in /opt/conda/lib/python3.10/s
ite-packages (from gensim) (1.11.4)
Requirement already satisfied: smart-open>=1.8.1 in /opt/conda/lib/python
3.10/site-packages (from gensim) (6.4.0)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-
packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
WARNING: Error parsing requirements for aiohttp: [Errno 2] No such file or
directory: '/opt/conda/lib/python3.10/site-packages/aiohttp-3.9.1.dist-inf
o/METADATA'
```

In [2]:

```
import nltk
nltk.download('book')
```

```
[nltk_data] Downloading collection 'book'
[nltk_data]    |
[nltk_data]    | Downloading package abc to /usr/share/nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package brown to /usr/share/nltk_data...
[nltk_data]    |   Package brown is already up-to-date!
[nltk_data]    | Downloading package chat80 to /usr/share/nltk_data...
[nltk_data]    |   Package chat80 is already up-to-date!
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     /usr/share/nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     /usr/share/nltk_data...
[nltk_data]    |   Package conll2000 is already up-to-date!
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     /usr/share/nltk_data...
[nltk_data]    |   Package conll2002 is already up-to-date!
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     /usr/share/nltk_data...
```

In [4]:

```
for name in dir(nltk.corpus):
    if name.islower():
        print(name)
```

```
__class__
__delattr__
__dict__
__dir__
__doc__
__eq__
__format__
__ge__
__getattr__
__getattribute__
__gt__
__hash__
__init__
__init_subclass__
__le__
__lt__
__module__
__name__
__ne__
__new__
__reduce__
__reduce_ex__
__repr__
__setattr__
__sizeof__
__str__
__subclasshook__
__weakref__
```

In [5]:

```
corpus=nltk.corpus.brown
print(corpus.paras())
```

```
[[['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday', 'an', 'in
vestigation', 'of', "Atlanta's", 'recent', 'primary', 'election', 'produce
d', '``', 'no', 'evidence', "''", 'that', 'any', 'irregularities', 'took',
'place', '.']], [['The', 'jury', 'further', 'said', 'in', 'term-end', 'pre
sentments', 'that', 'the', 'City', 'Executive', 'Committee', ',', 'which',
'had', 'over-all', 'charge', 'of', 'the', 'election', ',', '``', 'deserve
s', 'the', 'praise', 'and', 'thanks', 'of', 'the', 'City', 'of', 'Atlant
a', "''", 'for', 'the', 'manner', 'in', 'which', 'the', 'election', 'was',
'conducted', '.']], ...]
```

In [6]:

```
print(corpus.sents())
```

```
[['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday', 'an', 'inv
estigation', 'of', "Atlanta's", 'recent', 'primary', 'election', 'produce
d', '``', 'no', 'evidence', "''", 'that', 'any', 'irregularities', 'took',
'place', '.'], ['The', 'jury', 'further', 'said', 'in', 'term-end', 'prese
ntments', 'that', 'the', 'City', 'Executive', 'Committee', ',', 'which',
'had', 'over-all', 'charge', 'of', 'the', 'election', ',', '``', 'deserve
s', 'the', 'praise', 'and', 'thanks', 'of', 'the', 'City', 'of', 'Atlant
a', "''", 'for', 'the', 'manner', 'in', 'which', 'the', 'election', 'was',
'conducted', '.'], ...]
```

In [8]:

```
counts=nltk.FreqDist(corpus.words())
vocab=len(counts.keys())
words=sum(counts.values())
lexicaldiversity=float(words)/float(vocab)
```

In [9]:

```
lexicaldiversity
```

Out[9]:

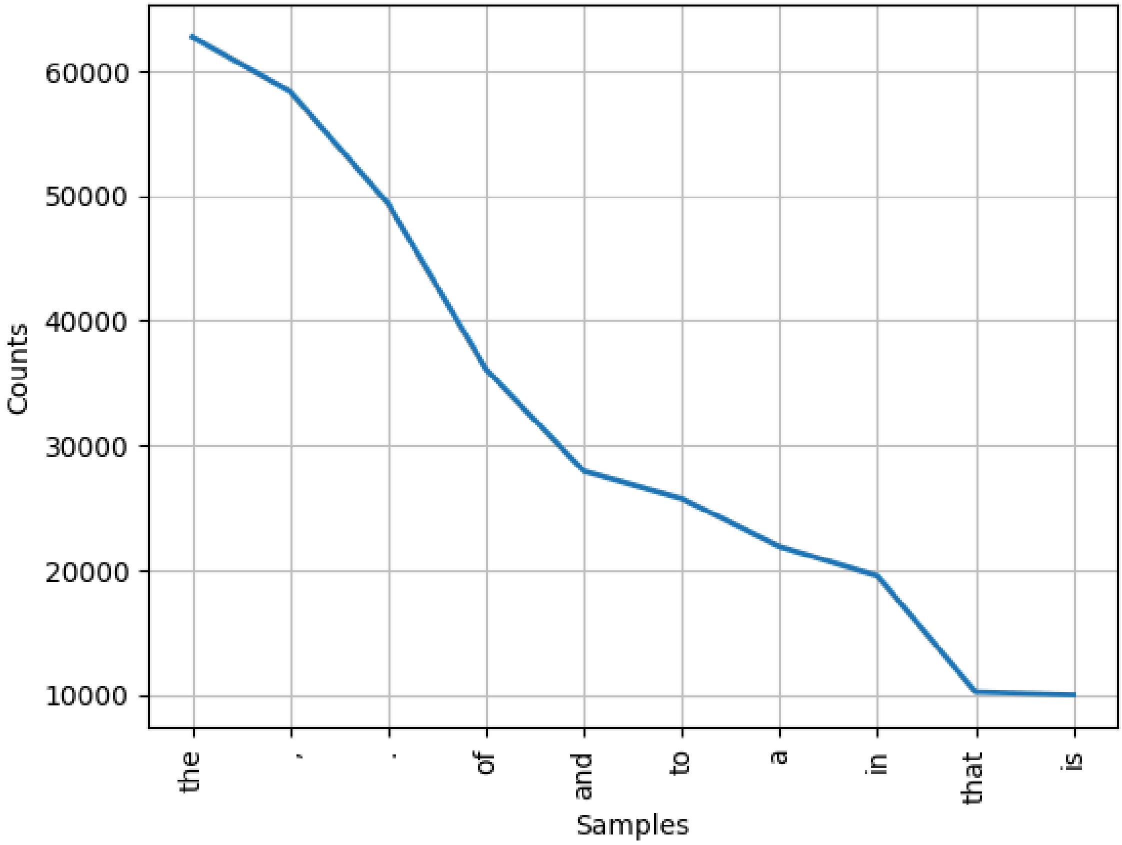
```
20.714487039977165
```

In [10]:

```
print(counts.most_common(10))
print(counts.max())
print(counts.hapaxes()[0:9])
print(counts.freq('a')*100)
```

```
[('the', 62713), (',', 58334), ('.', 49346), ('of', 36080), ('and', 2791
5), ('to', 25732), ('a', 21881), ('in', 19536), ('that', 10237), ('is', 10
011)]
the
['term-end', 'presentments', 'September-October', 'Durwood', 'Pye', 'Mayor
-nominate', 'Merger', 're-set', 'disable']
1.884356764428277
```

```
In [12]: counts.plot(10,cumulative=False)
```



```
In [16]: text= u'Hello world from NIT Rourkela to New York. However lets see.'
         sent=nltk.sent_tokenize(text)
         print(sent)
```

```
['Hello world from NIT Rourkela to New York.', 'However lets see.']
```

```
In [17]: for s in sent:
             print(list(nltk.word_tokenize(s)))
```

```
['Hello', 'world', 'from', 'NIT', 'Rourkela', 'to', 'New', 'York', '.']
['However', 'lets', 'see', '.']
```

```
In [18]: for s in sent:
             print(list(nltk.pos_tag(nltk.word_tokenize(s))))
```

```
[('Hello', 'NNP'), ('world', 'NN'), ('from', 'IN'), ('NIT', 'NNP'), ('Rour
kela', 'NNP'), ('to', 'TO'), ('New', 'NNP'), ('York', 'NNP'), ('.', '.')]
[('However', 'RB'), ('lets', 'NNS'), ('see', 'VBP'), ('.', '.')]
```

```
In [19]: from nltk.stem.porter import PorterStemmer
         s=list(nltk.word_tokenize('flying is very informative for sleeping'))
         port=PorterStemmer()
         out=[port.stem(t) for t in s]
         print(out)
```

```
['fli', 'is', 'veri', 'inform', 'for', 'sleep']
```

In [20]:
```
!python3 -m nltk.downloader wordnet
```

```
/opt/conda/lib/python3.10/runpy.py:126: RuntimeWarning: 'nltk.downloader'
found in sys.modules after import of package 'nltk', but prior to executio
n of 'nltk.downloader'; this may result in unpredictable behaviour
  warn(RuntimeWarning(msg))
[nltk_data] Downloading package wordnet to /usr/share/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
Archive:  /usr/share/nltk_data/corpora/wordnet.zip
   creating: /usr/share/nltk_data/corpora/wordnet/wordnet/
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/lexnames
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/data.verb
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/index.adv
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/adv.exc
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/index.verb
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/cntlist.rev
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/data.adj
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/index.adj
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/LICENSE
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/citation.bib
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/noun.exc
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/verb.exc
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/README
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/index.sense
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/data.noun
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/data.adv
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/index.noun
  inflating: /usr/share/nltk_data/corpora/wordnet/wordnet/adj.exc
```

In [22]:
```
!unzip /usr/share/nltk_data/corpora/wordnet.zip -d /usr/share/nltk_data/corp
```

```
Archive:  /usr/share/nltk_data/corpora/wordnet.zip
  inflating: /usr/share/nltk_data/corpora/wordnet/lexnames
  inflating: /usr/share/nltk_data/corpora/wordnet/data.verb
  inflating: /usr/share/nltk_data/corpora/wordnet/index.adv
  inflating: /usr/share/nltk_data/corpora/wordnet/adv.exc
  inflating: /usr/share/nltk_data/corpora/wordnet/index.verb
  inflating: /usr/share/nltk_data/corpora/wordnet/cntlist.rev
  inflating: /usr/share/nltk_data/corpora/wordnet/data.adj
  inflating: /usr/share/nltk_data/corpora/wordnet/index.adj
  inflating: /usr/share/nltk_data/corpora/wordnet/LICENSE
  inflating: /usr/share/nltk_data/corpora/wordnet/citation.bib
  inflating: /usr/share/nltk_data/corpora/wordnet/noun.exc
  inflating: /usr/share/nltk_data/corpora/wordnet/verb.exc
  inflating: /usr/share/nltk_data/corpora/wordnet/README
  inflating: /usr/share/nltk_data/corpora/wordnet/index.sense
  inflating: /usr/share/nltk_data/corpora/wordnet/data.noun
  inflating: /usr/share/nltk_data/corpora/wordnet/data.adv
  inflating: /usr/share/nltk_data/corpora/wordnet/index.noun
  inflating: /usr/share/nltk_data/corpora/wordnet/adj.exc
```

In [23]:

```
import nltk
from nltk.stem.wordnet import WordNetLemmatizer
lemmatizer=WordNetLemmatizer()
out=[lemmatizer.lemmatize(t) for t in s]
print(out)
```

```
['flying', 'is', 'very', 'informative', 'for', 'sleeping']
```

In [28]:

```
import sklearn
from sklearn.feature_extraction.text import TfidfVectorizer
tf_idf_model=TfidfVectorizer()
corpus=['data science is one of the most important fields in science','this
       'data scientists analyze data']
wordset=set()
for doc in corpus:
    words=doc.split(' ')
    wordset=wordset.union(set(words))
print('Number of words:',len(wordset))
```

```
Number of words: 15
```

In [29]:

```
tf_idf_vector=tf_idf_model.fit_transform(corpus)
print(tf_idf_vector.shape)
```

```
(3, 15)
```

In [30]:

```
print(tf_idf_vector.toarray())
```

```
[[0.         0.         0.         0.18848062 0.31912543 0.31912543
  0.31912543 0.24270312 0.31912543 0.31912543 0.31912543 0.48540623
  0.         0.24270312 0.        ]
 [0.         0.44350256 0.44350256 0.26193976 0.         0.
  0.         0.33729513 0.         0.         0.         0.33729513
  0.         0.33729513 0.44350256]
 [0.54270061 0.         0.         0.64105545 0.         0.
  0.         0.         0.         0.         0.         0.
  0.54270061 0.         0.        ]]
```

In [31]:

```
import gensim
from gensim.scripts.glove2word2vec import glove2word2vec
glove_input_file='/kaggle/input/glove6b50dtxt/glove.6B.50d.txt'
word2vec_out_file='/kaggle/working/glove6b50dtxt.word2vec'
```

In [32]:

```
glove2word2vec(glove_input_file,word2vec_out_file)
```

```
/tmp/ipykernel_33/379660289.py:1: DeprecationWarning: Call to deprecated `
glove2word2vec` (KeyedVectors.load_word2vec_format(.., binary=False, no_he
ader=True) loads GLoVE text vectors.).
  glove2word2vec(glove_input_file,word2vec_out_file)
```

Out[32]:

```
(400000, 50)
```

In [33]:

```
from gensim.models import KeyedVectors
model=KeyedVectors.load_word2vec_format(word2vec_out_file,binary=False)
```

In [34]:

```
result=model.most_similar(positive=['woman','king'], negative=['man'])
print(result)
```

```
[('queen', 0.8523604273796082), ('throne', 0.7664334177970886), ('prince',
0.7592144012451172), ('daughter', 0.7473883628845215), ('elizabeth', 0.746
0219860076904), ('princess', 0.7424570322036743), ('kingdom', 0.7337412238
121033), ('monarch', 0.721449077129364), ('eldest', 0.7184861898422241),
('widow', 0.7099431157112122)]
```

In [35]:

```
result=model.most_similar(positive=['woman','doctor'], negative=['man'])
print(result)
```

```
[('nurse', 0.8404642939567566), ('child', 0.7663259506225586), ('pregnan
t', 0.7570130228996277), ('mother', 0.7517457604408264), ('patient', 0.751
6663074493408), ('physician', 0.7507280707359314), ('dentist', 0.736034393
3105469), ('therapist', 0.7342537045478821), ('parents', 0.728634595870971
7), ('surgeon', 0.7165213227272034)]
```

In [ ]:

```

```