```
import numpy as np
import pandas as pd

import nltk #provides a set of diverse algorithm of NLP
from nltk.corpus import stopwords #read a corpus files in variety of
formats

import seaborn as sns

import matplotlib.pyplot as plt

#reading the data

df = pd.read_csv('/home/dara/ass7_dsbd/Resume_Data.csv')
df['Cleaned_Resume'] = ''
df.head()
```

```
       Category                                             Resume  \
0  Data Science  Skills * Programming Languages: Python (pandas...
1  Data Science  Education Details \r\nMay 2013 to May 2017 B.E...
2  Data Science  Areas of Interest Deep Learning, Control Syste...
3  Data Science  Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...
4  Data Science  Education Details \r\n MCA   YMCAUST,  Faridab...

  Cleaned_Resume
0
1
2
3
4
```

Cleaned_Resume is created to keep the clean text.

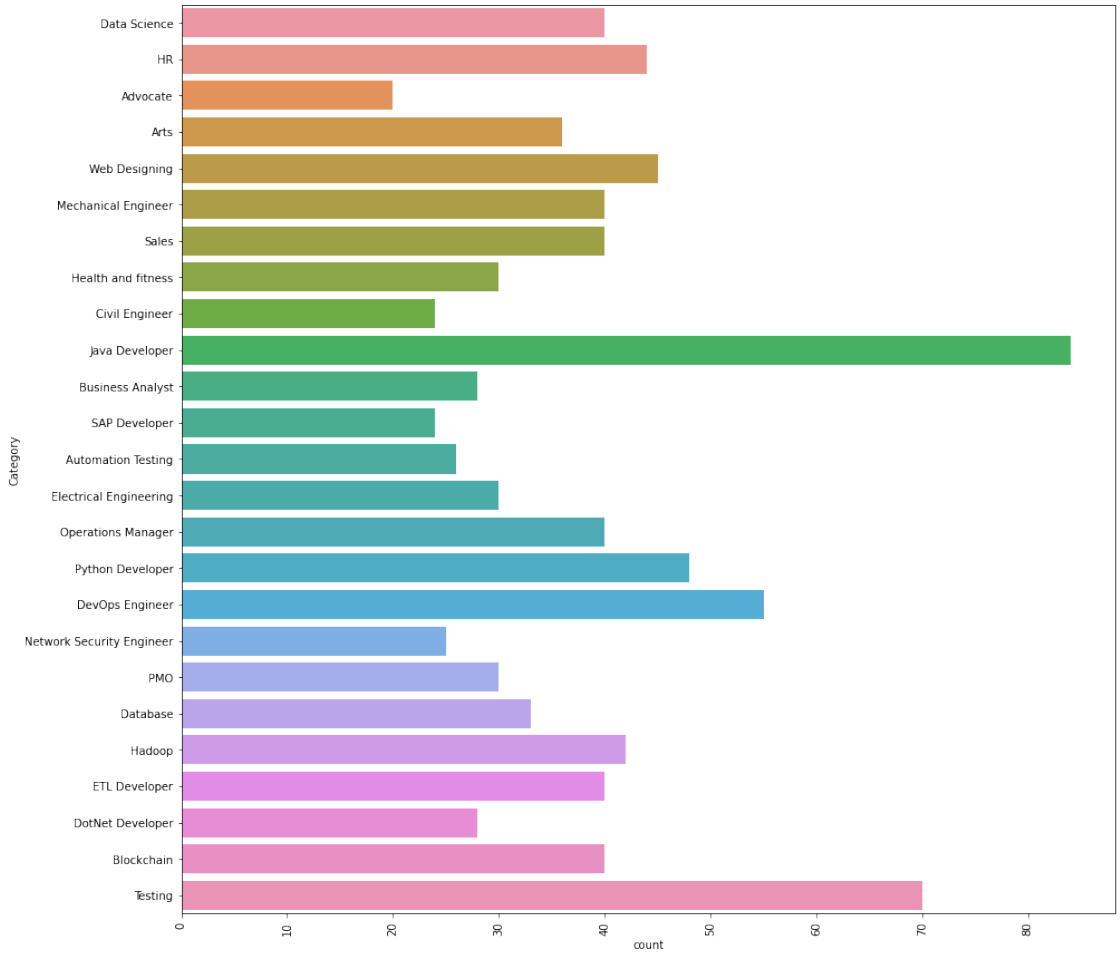
```
print ("Resume Categories")
print (df['Category'].value_counts())
```

```
Resume Categories
Java Developer               84
Testing                      70
DevOps Engineer              55
Python Developer             48
Web Designing                45
HR                           44
Hadoop                       42
Blockchain                   40
ETL Developer                40
Operations Manager           40
Data Science                 40
Sales                        40
Mechanical Engineer          40
Arts                         36
```

```
Database                    33
Electrical Engineering      30
Health and fitness          30
PMO                         30
Business Analyst            28
DotNet Developer            28
Automation Testing          26
Network Security Engineer   25
SAP Developer               24
Civil Engineer              24
Advocate                    20
Name: Category, dtype: int64
```

```
plt.figure(figsize=(15,15))
plt.xticks(rotation=90)
sns.countplot(y="Category", data=df)
```

```
<AxesSubplot:xlabel='count', ylabel='Category'>
```




```
df["Resume"][2]
```

```
'Areas of Interest Deep Learning, Control System Design, Programming
in-Python, Electric Machinery, Web Development, Analytics Technical
Activities q Hindustan Aeronautics Limited, Bangalore - For 4 weeks
under the guidance of Mr. Satish, Senior Engineer in the hangar of
Mirage 2000 fighter aircraft Technical Skills Programming Matlab,
Python and Java, LabView, Python WebFrameWork-Django, Flask, LTSPICE-
intermediate Languages and and MIPOWER-intermediate, Github (GitBash),
Jupyter Notebook, Xampp, MySQL-Basics, Python Software Packages
Interpreters-Anaconda, Python2, Python3, Pycharm, Java IDE-Eclipse
Operating Systems Windows, Ubuntu, Debian-Kali Linux Education Details
\r\nJanuary 2019 B.Tech. Electrical and Electronics Engineering
Manipal Institute of Technology\r\nJanuary 2015    DEEKSHA CENTER\r\
nJanuary 2013    Little Flower Public School\r\nAugust 2000    Manipal
Academy of Higher\r\nDATA SCIENCE \r\n\r\nDATA SCIENCE AND ELECTRICAL
ENTHUSIAST\r\nSkill Details \r\nData Analysis- Exprience - Less than 1
year months\r\nexcel- Exprience - Less than 1 year months\r\nMachine
Learning- Exprience - Less than 1 year months\r\nmathematics-
Exprience - Less than 1 year months\r\nPython- Exprience - Less than 1
year months\r\nMatlab- Exprience - Less than 1 year months\r\
nElectrical Engineering- Exprience - Less than 1 year months\r\nSql-
Exprience - Less than 1 year monthsCompany Details \r\ncompany -
THEMATHCOMPANY\r\ndescription - I am currently working with a Casino
based operator(name not to be disclosed) in Macau.I need to segment
the customers who visit their property based on the value the patrons
bring into the company.Basically prove that the segmentation can be
done in much better way than the current system which they have with
proper numbers to back it up.Henceforth they can implement target
marketing strategy to attract their customers who add value to the
business.'
```

As we can see the text needs a lot of processing. This is not suitable for analyzing

```
#We now have to clean the resume text.
#re--lets you check if a particular string matches a given regular
expression
import re
def cleanResume(resumeText):
    resumeText = re.sub('http\S+\s*', ' ', resumeText)  # remove URLs
    resumeText = re.sub('RT|cc', ' ', resumeText)  # remove RT and cc
    resumeText = re.sub('#\S+', '', resumeText)  # remove hashtags
    resumeText = re.sub('@\S+', '  ', resumeText)  # remove mentions
    resumeText = re.sub('[%s]' % re.escape("""!"#$%&'()*+,-./:;<=>?
@[\]^_`{|}~"""), ' ', resumeText)  # remove punctuations
    resumeText = re.sub(r'[^\x00-\x7f]',r' ', resumeText)
    resumeText = re.sub('\s+', ' ', resumeText)  # remove extra
whitespace
    return resumeText

df['Cleaned_Resume'] = df.Resume.apply(lambda x: cleanResume(x))

df.head()
```

```
       Category                                             Resume  \
0  Data Science  Skills * Programming Languages: Python (pandas...
1  Data Science  Education Details \r\nMay 2013 to May 2017 B.E...
2  Data Science  Areas of Interest Deep Learning, Control Syste...
3  Data Science  Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...
4  Data Science  Education Details \r\n MCA   YMCAUST,  Faridab...

                                        Cleaned_Resume
0  Skills Programming Languages Python pandas num...
1  Education Details May 2013 to May 2017 B E UIT...
2  Areas of Interest Deep Learning Control System...
3  Skills R Python SAP HANA Tableau SAP HANA SQL ...
4  Education Details MCA YMCAUST Faridabad Haryan...
```

Now we see that the text is clean.

```
len(df)
```

```
962
```

```
#getting the entire Cleaned_Resume as single text.

corpus=" "

for i in range(0,962):
    corpus= corpus+ df["Cleaned_Resume"][i]

corpus[1000:2500]
```

```
'review process and run analytics and generate reports Core member of
a team helped in developing automated review platform tool from
scratch for assisting E discovery domain this tool implements
predictive coding and topic modelling by automating reviews resulting
in reduced labor costs and time spent during the lawyers review
Understand the end to end flow of the solution doing research and
development for classification models predictive analysis and mining
of the information present in text data Worked on analyzing the
outputs and precision monitoring for the entire tool TAR assists in
predictive coding topic modelling from the evidence by following EY
standards Developed the classifier models in order to identify red
flags and fraud related issues Tools Technologies Python scikit learn
tfidf word2vec doc2vec cosine similarity Na ve Bayes LDA NMF for topic
modelling Vader and text blob for sentiment analysis Matplot lib
Tableau dashboard for reporting MULTIPLE DATA SCIENCE AND ANALYTIC
PROJECTS USA CLIENTS TEXT ANALYTICS MOTOR VEHICLE CUSTOMER REVIEW DATA
Received customer feedback survey data for past one year Performed
sentiment Positive Negative Neutral and time series analysis on
customer comments across all 4 categories Created heat map of terms by
survey category based on frequency of words Extracted Positive and
Negative words across all the Survey categories and plotted Word cloud
```

```
Created customized tableau dashboards for effective reporting and
visualizations CHAT'
```

As the text has now been cleaned and joined together and is ready for document preprocessing methods.

Tokenization

Tokenization is the process of breaking raw text into small units. Here, we convert the entire text into single words. Tokenization is important because it splits the data into small usable and easy-to-process units. These smaller units of text are called tokens. These tokens can help in understanding the context of the text and also in building the NLP models.

```
#Creating the tokenizer
tokenizer = nltk.tokenize.RegexpTokenizer('\w+')

#Tokenizing the text
tokens = tokenizer.tokenize(corpus)

len(tokens)
```

```
411913
```

```
#now we shall make everything lowercase for uniformity
#to hold the new lower case words

words = []

#Looping through the tokens and make them lower case
for word in tokens:
    words.append(word.lower())
```

Here we have used word tokenization for our analyzing.

POS Tagging

POS Tagging is a popular Natural Language Processing process which refers to categorizing word in a text (corpus) in correspondance with a particular part of speech, depending on the definition of the word and it's context.

```
words1 = nltk.word_tokenize(corpus)

print(words1)
```

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.
```

```
Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

```
len(words1)
```

```
411913
```

```
import nltk
nltk.download('averaged_perceptron_tagger')
nltk.pos_tag(words1)
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/dara/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

```
[('Skills', 'NNS'),
 ('Programming', 'VBG'),
 ('Languages', 'NNP'),
 ('Python', 'NNP'),
 ('pandas', 'VBZ'),
 ('numpy', 'JJ'),
 ('scipy', 'JJ'),
 ('scikit', 'NN'),
 ('learn', 'NN'),
 ('matplotlib', 'NN'),
 ('Sql', 'NNP'),
 ('Java', 'NNP'),
 ('JavaScript', 'NNP'),
 ('JQuery', 'NNP'),
 ('Machine', 'NNP'),
 ('learning', 'VBG'),
 ('Regression', 'NNP'),
 ('SVM', 'NNP'),
 ('Na', 'NNP'),
 ('ve', 'FW'),
 ('Bayes', 'NNP'),
 ('KNN', 'NNP'),
 ('Random', 'NNP'),
 ('Forest', 'NNP'),
 ('Decision', 'NNP'),
 ('Trees', 'NNP'),
 ('Boosting', 'NNP'),
 ('techniques', 'NNS'),
 ('Cluster', 'NNP'),
 ('Analysis', 'NNP'),
 ('Word', 'NNP'),
 ('Embedding', 'NNP'),
 ('Sentiment', 'NNP'),
```

```
('Analysis', 'NNP'),
('Natural', 'NNP'),
('Language', 'NNP'),
('processing', 'NN'),
('Dimensionality', 'NNP'),
('reduction', 'NN'),
('Topic', 'NNP'),
('Modelling', 'NNP'),
('LDA', 'NNP'),
('NMF', 'NNP'),
('PCA', 'NNP'),
('Neural', 'NNP'),
('Nets', 'NNP'),
('Database', 'NNP'),
('Visualizations', 'NNP'),
('Mysql', 'NNP'),
('SqlServer', 'NNP'),
('Cassandra', 'NNP'),
('Hbase', 'NNP'),
('ElasticSearch', 'NNP'),
('D3', 'NNP'),
('js', 'NN'),
('DC', 'NNP'),
('js', 'NN'),
('Plotly', 'NNP'),
('kibana', 'NNP'),
('matplotlib', 'NN'),
('ggplot', 'NN'),
('Tableau', 'NNP'),
('Others', 'NNP'),
('Regular', 'NNP'),
('Expression', 'NNP'),
('HTML', 'NNP'),
('CSS', 'NNP'),
('Angular', 'NNP'),
('6', 'CD'),
('Logstash', 'NNP'),
('Kafka', 'NNP'),
('Python', 'NNP'),
('Flask', 'NNP'),
('Git', 'NNP'),
('Docker', 'NNP'),
('computer', 'NN'),
('vision', 'NN'),
('Open', 'NNP'),
('CV', 'NNP'),
('and', 'CC'),
('understanding', 'NN'),
('of', 'IN'),
('Deep', 'NNP'),
```

```
('learning', 'VBG'),
('Education', 'NNP'),
('Details', 'NNP'),
('Data', 'NNP'),
('Science', 'NNP'),
('Assurance', 'NNP'),
('Associate', 'NNP'),
('Data', 'NNP'),
('Science', 'NNP'),
('Assurance', 'NNP'),
('Associate', 'NNP'),
('Ernst', 'NNP'),
('Young', 'NNP'),
('LLP', 'NNP'),
('Skill', 'NNP'),
('Details', 'NNP'),
('JAVASCRIPT', 'NNP'),
('Exprience', 'NNP'),
('24', 'CD'),
('months', 'NNS'),
('jQuery', 'JJ'),
('Exprience', 'NNP'),
('24', 'CD'),
('months', 'NNS'),
('Python', 'NNP'),
('Exprience', 'NNP'),
('24', 'CD'),
('monthsCompany', 'NN'),
('Details', 'NNP'),
('company', 'NN'),
('Ernst', 'NNP'),
('Young', 'NNP'),
('LLP', 'NNP'),
('description', 'NN'),
('Fraud', 'NNP'),
('Investigations', 'NNP'),
('and', 'CC'),
('Dispute', 'NNP'),
('Services', 'NNPS'),
('Assurance', 'NNP'),
('TECHNOLOGY', 'NNP'),
('ASSISTED', 'NNP'),
('REVIEW', 'NNP'),
('TAR', 'NNP'),
('Technology', 'NNP'),
('Assisted', 'NNP'),
('Review', 'NNP'),
('assists', 'VBZ'),
('in', 'IN'),
('a', 'DT'),
```

```
('elerating', 'VBG'),
('the', 'DT'),
('review', 'NN'),
('process', 'NN'),
('and', 'CC'),
('run', 'VB'),
('analytics', 'NNS'),
('and', 'CC'),
('generate', 'VB'),
('reports', 'NNS'),
('Core', 'NNP'),
('member', 'NN'),
('of', 'IN'),
('a', 'DT'),
('team', 'NN'),
('helped', 'VBD'),
('in', 'IN'),
('developing', 'VBG'),
('automated', 'VBN'),
('review', 'NN'),
('platform', 'NN'),
('tool', 'NN'),
('from', 'IN'),
('scratch', 'NN'),
('for', 'IN'),
('assisting', 'VBG'),
('E', 'NNP'),
('discovery', 'NN'),
('domain', 'NN'),
('this', 'DT'),
('tool', 'NN'),
('implements', 'NNS'),
('predictive', 'JJ'),
('coding', 'NN'),
('and', 'CC'),
('topic', 'NN'),
('modelling', 'NN'),
('by', 'IN'),
('automating', 'VBG'),
('reviews', 'NNS'),
('resulting', 'VBG'),
('in', 'IN'),
('reduced', 'JJ'),
('labor', 'NN'),
('costs', 'NNS'),
('and', 'CC'),
('time', 'NN'),
('spent', 'VBN'),
('during', 'IN'),
('the', 'DT'),
```

```
('lawyers', 'NNS'),
('review', 'VBP'),
('Understand', 'IN'),
('the', 'DT'),
('end', 'NN'),
('to', 'TO'),
('end', 'VB'),
('flow', 'NN'),
('of', 'IN'),
('the', 'DT'),
('solution', 'NN'),
('doing', 'VBG'),
('research', 'NN'),
('and', 'CC'),
('development', 'NN'),
('for', 'IN'),
('classification', 'NN'),
('models', 'NNS'),
('predictive', 'JJ'),
('analysis', 'NN'),
('and', 'CC'),
('mining', 'NN'),
('of', 'IN'),
('the', 'DT'),
('information', 'NN'),
('present', 'NN'),
('in', 'IN'),
('text', 'NN'),
('data', 'NNS'),
('Worked', 'VBN'),
('on', 'IN'),
('analyzing', 'VBG'),
('the', 'DT'),
('outputs', 'NNS'),
('and', 'CC'),
('precision', 'NN'),
('monitoring', 'NN'),
('for', 'IN'),
('the', 'DT'),
('entire', 'JJ'),
('tool', 'NN'),
('TAR', 'NNP'),
('assists', 'VBZ'),
('in', 'IN'),
('predictive', 'JJ'),
('coding', 'NN'),
('topic', 'NN'),
('modelling', 'VBG'),
('from', 'IN'),
('the', 'DT'),
```

```
('evidence', 'NN'),
('by', 'IN'),
('following', 'VBG'),
('EY', 'NNP'),
('standards', 'NNS'),
('Developed', 'VBD'),
('the', 'DT'),
('classifier', 'NN'),
('models', 'NNS'),
('in', 'IN'),
('order', 'NN'),
('to', 'TO'),
('identify', 'VB'),
('red', 'JJ'),
('flags', 'NNS'),
('and', 'CC'),
('fraud', 'NN'),
('related', 'JJ'),
('issues', 'NNS'),
('Tools', 'NNP'),
('Technologies', 'NNPS'),
('Python', 'NNP'),
('scikit', 'NN'),
('learn', 'NN'),
('tfidf', 'NN'),
('word2vec', 'NN'),
('doc2vec', 'NN'),
('cosine', 'NN'),
('similarity', 'NN'),
('Na', 'NNP'),
('ve', 'VBZ'),
('Bayes', 'NNP'),
('LDA', 'NNP'),
('NMF', 'NNP'),
('for', 'IN'),
('topic', 'NN'),
('modelling', 'VBG'),
('Vader', 'NNP'),
('and', 'CC'),
('text', 'JJ'),
('blob', 'NN'),
('for', 'IN'),
('sentiment', 'NN'),
('analysis', 'NN'),
('Matplot', 'NNP'),
('lib', 'NN'),
('Tableau', 'NNP'),
('dashboard', 'NN'),
('for', 'IN'),
('reporting', 'VBG'),
```

```
('MULTIPLE', 'NNP'),
('DATA', 'NNP'),
('SCIENCE', 'NNP'),
('AND', 'NNP'),
('ANALYTIC', 'NNP'),
('PROJECTS', 'NNP'),
('USA', 'NNP'),
('CLIENTS', 'NNP'),
('TEXT', 'NNP'),
('ANALYTICS', 'NNP'),
('MOTOR', 'NNP'),
('VEHICLE', 'NNP'),
('CUSTOMER', 'NNP'),
('REVIEW', 'NNP'),
('DATA', 'NNP'),
('Received', 'NNP'),
('customer', 'NN'),
('feedback', 'NN'),
('survey', 'NN'),
('data', 'NNS'),
('for', 'IN'),
('past', 'IN'),
('one', 'CD'),
('year', 'NN'),
('Performed', 'VBD'),
('sentiment', 'JJ'),
('Positive', 'NNP'),
('Negative', 'NNP'),
('Neutral', 'NNP'),
('and', 'CC'),
('time', 'NN'),
('series', 'NN'),
('analysis', 'NN'),
('on', 'IN'),
('customer', 'NN'),
('comments', 'NNS'),
('across', 'IN'),
('all', 'DT'),
('4', 'CD'),
('categories', 'NNS'),
('Created', 'VBN'),
('heat', 'NN'),
('map', 'NN'),
('of', 'IN'),
('terms', 'NNS'),
('by', 'IN'),
('survey', 'NN'),
('category', 'NN'),
('based', 'VBN'),
('on', 'IN'),
```

```
('frequency', 'NN'),
('of', 'IN'),
('words', 'NNS'),
('Extracted', 'NNP'),
('Positive', 'NNP'),
('and', 'CC'),
('Negative', 'NNP'),
('words', 'NNS'),
('across', 'IN'),
('all', 'PDT'),
('the', 'DT'),
('Survey', 'NNP'),
('categories', 'NNS'),
('and', 'CC'),
('plotted', 'VBD'),
('Word', 'NNP'),
('cloud', 'NN'),
('Created', 'NNP'),
('customized', 'VBD'),
('tableau', 'NN'),
('dashboards', 'NNS'),
('for', 'IN'),
('effective', 'JJ'),
('reporting', 'NN'),
('and', 'CC'),
('visualizations', 'NNS'),
('CHATBOT', 'NNP'),
('Developed', 'VBD'),
('a', 'DT'),
('user', 'NN'),
('friendly', 'JJ'),
('chatbot', 'NN'),
('for', 'IN'),
('one', 'CD'),
('of', 'IN'),
('our', 'PRP$'),
('Products', 'NNS'),
('which', 'WDT'),
('handle', 'VBP'),
('simple', 'JJ'),
('questions', 'NNS'),
('about', 'IN'),
('hours', 'NNS'),
('of', 'IN'),
('operation', 'NN'),
('reservation', 'NN'),
('options', 'NNS'),
('and', 'CC'),
('so', 'RB'),
('on', 'IN'),
```

```
('This', 'DT'),
('chat', 'NN'),
('bot', 'VBZ'),
('serves', 'NNS'),
('entire', 'JJ'),
('product', 'NN'),
('related', 'VBN'),
('questions', 'NNS'),
('Giving', 'VBG'),
('overview', 'NN'),
('of', 'IN'),
('tool', 'NN'),
('via', 'IN'),
('QA', 'NNP'),
('platform', 'NN'),
('and', 'CC'),
('also', 'RB'),
('give', 'VB'),
('recommendation', 'NN'),
('responses', 'NNS'),
('so', 'RB'),
('that', 'IN'),
('user', 'JJ'),
('question', 'NN'),
('to', 'TO'),
('build', 'VB'),
('chain', 'NN'),
('of', 'IN'),
('relevant', 'JJ'),
('answer', 'NN'),
('This', 'DT'),
('too', 'RB'),
('has', 'VBZ'),
('intelligence', 'NN'),
('to', 'TO'),
('build', 'VB'),
('the', 'DT'),
('pipeline', 'NN'),
('of', 'IN'),
('questions', 'NNS'),
('as', 'IN'),
('per', 'IN'),
('user', 'NN'),
('requirement', 'NN'),
('and', 'CC'),
('asks', 'VBZ'),
('the', 'DT'),
('relevant', 'NN'),
('recommended', 'VBD'),
('questions', 'NNS'),
```

```
('Tools', 'NNP'),
('Technologies', 'NNPS'),
('Python', 'NNP'),
('Natural', 'NNP'),
('language', 'NN'),
('processing', 'NN'),
('NLTK', 'NNP'),
('spacy', 'NN'),
('topic', 'NN'),
('modelling', 'VBG'),
('Sentiment', 'NNP'),
('analysis', 'NN'),
('Word', 'NNP'),
('Embedding', 'NNP'),
('scikit', 'NN'),
('learn', 'NN'),
('JavaScript', 'NNP'),
('JQuery', 'NNP'),
('SqlServer', 'NNP'),
('INFORMATION', 'NNP'),
('GOVERNANCE', 'NNP'),
('Organizations', 'NNPS'),
('to', 'TO'),
('make', 'VB'),
('informed', 'JJ'),
('decisions', 'NNS'),
('about', 'IN'),
('all', 'DT'),
('of', 'IN'),
('the', 'DT'),
('information', 'NN'),
('they', 'PRP'),
('store', 'VBP'),
('The', 'DT'),
('integrated', 'JJ'),
('Information', 'NNP'),
('Governance', 'NNP'),
('portfolio', 'NN'),
('synthesizes', 'VBZ'),
('intelligence', 'NN'),
('across', 'IN'),
('unstructured', 'JJ'),
('data', 'NNS'),
('sources', 'NNS'),
('and', 'CC'),
('facilitates', 'VBZ'),
('action', 'NN'),
('to', 'TO'),
('ensure', 'VB'),
('organizations', 'NNS'),
```

```
('are', 'VBP'),
('best', 'RB'),
('positioned', 'VBN'),
('to', 'TO'),
('counter', 'VB'),
('information', 'NN'),
('risk', 'NN'),
('Scan', 'NNP'),
('data', 'NNS'),
('from', 'IN'),
('multiple', 'JJ'),
('sources', 'NNS'),
('of', 'IN'),
('formats', 'NNS'),
('and', 'CC'),
('parse', 'JJ'),
('different', 'JJ'),
('file', 'NN'),
('formats', 'NNS'),
('extract', 'VBP'),
('Meta', 'NNP'),
('data', 'NNS'),
('information', 'NN'),
('push', 'NN'),
('results', 'NNS'),
('for', 'IN'),
('indexing', 'VBG'),
('elastic', 'JJ'),
('search', 'NN'),
('and', 'CC'),
('created', 'VBD'),
('customized', 'VBN'),
('interactive', 'JJ'),
('dashboards', 'NNS'),
('using', 'VBG'),
('kibana', 'NN'),
('Preforming', 'NNP'),
('ROT', 'NNP'),
('Analysis', 'NNP'),
('on', 'IN'),
('the', 'DT'),
('data', 'NNS'),
('which', 'WDT'),
('give', 'VBP'),
('information', 'NN'),
('of', 'IN'),
('data', 'NN'),
('which', 'WDT'),
('helps', 'VBZ'),
('identify', 'VB'),
```

```
('content', 'NN'),
('that', 'WDT'),
('is', 'VBZ'),
('either', 'DT'),
('Redundant', 'NNP'),
('Outdated', 'VBD'),
('or', 'CC'),
('Trivial', 'JJ'),
('Preforming', 'NNP'),
('full', 'JJ'),
('text', 'NN'),
('search', 'NN'),
('analysis', 'NN'),
('on', 'IN'),
('elastic', 'JJ'),
('search', 'NN'),
('with', 'IN'),
('predefined', 'VBN'),
('methods', 'NNS'),
('which', 'WDT'),
('can', 'MD'),
('tag', 'VB'),
('as', 'IN'),
('PII', 'NNP'),
('personally', 'RB'),
('identifiable', 'JJ'),
('information', 'NN'),
('social', 'JJ'),
('security', 'NN'),
('numbers', 'NNS'),
('addresses', 'VBZ'),
('names', 'NNS'),
('etc', 'FW'),
('which', 'WDT'),
('frequently', 'RB'),
('targeted', 'VBD'),
('during', 'IN'),
('cyber', 'JJ'),
('attacks', 'NNS'),
('Tools', 'NNP'),
('Technologies', 'NNPS'),
('Python', 'NNP'),
('Flask', 'NNP'),
('Elastic', 'NNP'),
('Search', 'NNP'),
('Kibana', 'NNP'),
('FRAUD', 'NNP'),
('ANALYTIC', 'NNP'),
('PLATFORM', 'NNP'),
('Fraud', 'NNP'),
```

```
('Analytics', 'NNP'),
('and', 'CC'),
('investigative', 'JJ'),
('platform', 'NN'),
('to', 'TO'),
('review', 'VB'),
('all', 'DT'),
('red', 'JJ'),
('flag', 'NN'),
('cases', 'NNS'),
('FAP', 'NNP'),
('is', 'VBZ'),
('a', 'DT'),
('Fraud', 'NNP'),
('Analytics', 'NNPS'),
('and', 'CC'),
('investigative', 'JJ'),
('platform', 'NN'),
('with', 'IN'),
('inbuilt', 'JJ'),
('case', 'NN'),
('manager', 'NN'),
('and', 'CC'),
('suite', 'NN'),
('of', 'IN'),
('Analytics', 'NNS'),
('for', 'IN'),
('various', 'JJ'),
('ERP', 'NNP'),
('systems', 'NNS'),
('It', 'PRP'),
('can', 'MD'),
('be', 'VB'),
('used', 'VBN'),
('by', 'IN'),
('clients', 'NNS'),
('to', 'TO'),
('interrogate', 'VB'),
('their', 'PRP$'),
('A', 'NNP'),
('ounting', 'NN'),
('systems', 'NNS'),
('for', 'IN'),
('identifying', 'VBG'),
('the', 'DT'),
('anomalies', 'NNS'),
('which', 'WDT'),
('can', 'MD'),
('be', 'VB'),
('indicators', 'NNS'),
```

```
('of', 'IN'),
('fraud', 'NN'),
('by', 'IN'),
('running', 'VBG'),
('advanced', 'JJ'),
('analytics', 'NNS'),
('Tools', 'NNP'),
('Technologies', 'NNPS'),
('HTML', 'NNP'),
('JavaScript', 'NNP'),
('SqlServer', 'NNP'),
('JQuery', 'NNP'),
('CSS', 'NNP'),
('Bootstrap', 'NNP'),
('Node', 'NNP'),
('js', 'NN'),
('D3', 'NNP'),
('js', 'NN'),
('DC', 'NNP'),
('jsEducation', 'NN'),
('Details', 'NNP'),
('May', 'NNP'),
('2013', 'CD'),
('to', 'TO'),
('May', 'NNP'),
('2017', 'CD'),
('B', 'NNP'),
('E', 'NNP'),
('UIT', 'NNP'),
('RGPV', 'NNP'),
('Data', 'NNP'),
('Scientist', 'NNP'),
('Data', 'NNP'),
('Scientist', 'NNP'),
('Matelabs', 'NNP'),
('Skill', 'NNP'),
('Details', 'NNP'),
('Python', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Statsmodels', 'NNP'),
('Exprience', 'NNP'),
('12', 'CD'),
('months', 'NNS'),
('AWS', 'NNP'),
('Exprience', 'NNP'),
```

```
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Machine', 'NNP'),
('learning', 'VBG'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Sklearn', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Scipy', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Keras', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('monthsCompany', 'NN'),
('Details', 'NNP'),
('company', 'NN'),
('Matelabs', 'NNP'),
('description', 'NN'),
('ML', 'NNP'),
('Platform', 'NNP'),
('for', 'IN'),
('business', 'NN'),
('professionals', 'NNS'),
('dummies', 'NNS'),
('and', 'CC'),
('enthusiasts', 'VBZ'),
('60', 'CD'),
('A', 'NNP'),
('Koramangala', 'NNP'),
('5th', 'CD'),
```

```
('block', 'NN'),
('Achievements', 'NNS'),
('Tasks', 'NNP'),
('behind', 'IN'),
('sukh', 'JJ'),
('sagar', 'NN'),
('Bengaluru', 'NNP'),
('India', 'NNP'),
('Developed', 'NNP'),
('and', 'CC'),
('deployed', 'JJ'),
('auto', 'NN'),
('preprocessing', 'VBG'),
('steps', 'NNS'),
('of', 'IN'),
('machine', 'NN'),
('learning', 'VBG'),
('mainly', 'RB'),
('missing', 'VBG'),
('value', 'NN'),
('treatment', 'NN'),
('outlier', 'JJR'),
('detection', 'NN'),
('encoding', 'VBG'),
('scaling', 'JJ'),
('feature', 'NN'),
('selection', 'NN'),
('and', 'CC'),
('dimensionality', 'NN'),
('reduction', 'NN'),
('Deployed', 'NNP'),
('automated', 'VBD'),
('classification', 'NN'),
('and', 'CC'),
('regression', 'NN'),
('model', 'NN'),
('linkedin', 'NN'),
('com', 'NN'),
('in', 'IN'),
('aditya', 'NN'),
('rathore', 'NN'),
('b4600b146', 'NN'),
('Reasearch', 'NNP'),
('and', 'CC'),
('deployed', 'VBD'),
('the', 'DT'),
('time', 'NN'),
('series', 'NN'),
('forecasting', 'VBG'),
('model', 'NN'),
```

```
('ARIMA', 'NNP'),
('SARIMAX', 'NNP'),
('Holt', 'NNP'),
('winter', 'NN'),
('and', 'CC'),
('Prophet', 'NNP'),
('Worked', 'VBD'),
('on', 'IN'),
('meta', 'NN'),
('feature', 'NN'),
('extracting', 'VBG'),
('problem', 'NN'),
('github', 'NN'),
('com', 'NN'),
('rathorology', 'NN'),
('Implemented', 'VBD'),
('a', 'DT'),
('state', 'NN'),
('of', 'IN'),
('the', 'DT'),
('art', 'NN'),
('research', 'NN'),
('paper', 'NN'),
('on', 'IN'),
('outlier', 'JJR'),
('detection', 'NN'),
('for', 'IN'),
('mixed', 'JJ'),
('attributes', 'NNS'),
('company', 'NN'),
('Matelabs', 'NNP'),
('description', 'NN'),
('Areas', 'NNP'),
('of', 'IN'),
('Interest', 'NNP'),
('Deep', 'NNP'),
('Learning', 'NNP'),
('Control', 'NNP'),
('System', 'NNP'),
('Design', 'NNP'),
('Programming', 'NNP'),
('in', 'IN'),
('Python', 'NNP'),
('Electric', 'NNP'),
('Machinery', 'NNP'),
('Web', 'NNP'),
('Development', 'NNP'),
('Analytics', 'NNP'),
('Technical', 'NNP'),
('Activities', 'NNP'),
```

```
('q', 'NNP'),
('Hindustan', 'NNP'),
('Aeronautics', 'NNP'),
('Limited', 'NNP'),
('Bangalore', 'NNP'),
('For', 'IN'),
('4', 'CD'),
('weeks', 'NNS'),
('under', 'IN'),
('the', 'DT'),
('guidance', 'NN'),
('of', 'IN'),
('Mr', 'NNP'),
('Satish', 'NNP'),
('Senior', 'NNP'),
('Engineer', 'NNP'),
('in', 'IN'),
('the', 'DT'),
('hangar', 'NN'),
('of', 'IN'),
('Mirage', 'NNP'),
('2000', 'CD'),
('fighter', 'NN'),
('aircraft', 'NN'),
('Technical', 'NNP'),
('Skills', 'NNP'),
('Programming', 'NNP'),
('Matlab', 'NNP'),
('Python', 'NNP'),
('and', 'CC'),
('Java', 'NNP'),
('LabView', 'NNP'),
('Python', 'NNP'),
('WebFrameWork', 'NNP'),
('Django', 'NNP'),
('Flask', 'NNP'),
('LTSPICE', 'NNP'),
('intermediate', 'NN'),
('Languages', 'NNP'),
('and', 'CC'),
('and', 'CC'),
('MIPOWER', 'NNP'),
('intermediate', 'VBP'),
('Github', 'NNP'),
('GitBash', 'NNP'),
('Jupyter', 'NNP'),
('Notebook', 'NNP'),
('Xampp', 'NNP'),
('MySQL', 'NNP'),
('Basics', 'NNP'),
```

```
('Python', 'NNP'),
('Software', 'NNP'),
('Packages', 'NNP'),
('Interpreters', 'NNP'),
('Anaconda', 'NNP'),
('Python2', 'NNP'),
('Python3', 'NNP'),
('Pycharm', 'NNP'),
('Java', 'NNP'),
('IDE', 'NNP'),
('Eclipse', 'NNP'),
('Operating', 'VBG'),
('Systems', 'NNP'),
('Windows', 'NNP'),
('Ubuntu', 'NNP'),
('Debian', 'NNP'),
('Kali', 'NNP'),
('Linux', 'NNP'),
('Education', 'NNP'),
('Details', 'NNP'),
('January', 'NNP'),
('2019', 'CD'),
('B', 'NNP'),
('Tech', 'NNP'),
('Electrical', 'NNP'),
('and', 'CC'),
('Electronics', 'NNP'),
('Engineering', 'NNP'),
('Manipal', 'NNP'),
('Institute', 'NNP'),
('of', 'IN'),
('Technology', 'NNP'),
('January', 'NNP'),
('2015', 'CD'),
('DEEKSHA', 'NNP'),
('CENTER', 'NNP'),
('January', 'NNP'),
('2013', 'CD'),
('Little', 'NNP'),
('Flower', 'NNP'),
('Public', 'NNP'),
('School', 'NNP'),
('August', 'NNP'),
('2000', 'CD'),
('Manipal', 'NNP'),
('Academy', 'NNP'),
('of', 'IN'),
('Higher', 'NNP'),
('DATA', 'NNP'),
('SCIENCE', 'NNP'),
```

```
('DATA', 'NNP'),
('SCIENCE', 'NNP'),
('AND', 'NNP'),
('ELECTRICAL', 'NNP'),
('ENTHUSIAST', 'NNP'),
('Skill', 'NNP'),
('Details', 'NNP'),
('Data', 'NNP'),
('Analysis', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('excel', 'VBP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Machine', 'NNP'),
('Learning', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('mathematics', 'NNS'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Python', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
('months', 'NNS'),
('Matlab', 'NNP'),
('Exprience', 'NNP'),
('Less', 'NNP'),
('than', 'IN'),
('1', 'CD'),
('year', 'NN'),
```

```
 ('months', 'NNS'),
 ('Electrical', 'NNP'),
 ('Engineering', 'NNP'),
 ('Exprience', 'NNP'),
 ('Less', 'NNP'),
 ('than', 'IN'),
 ('1', 'CD'),
 ('year', 'NN'),
 ('months', 'NNS'),
 ('Sql', 'NNP'),
 ('Exprience', 'NNP'),
 ('Less', 'NNP'),
 ('than', 'IN'),
 ('1', 'CD'),
 ('year', 'NN'),
 ('monthsCompany', 'NN'),
 ('Details', 'NNP'),
 ...]
```

```
import nltk
nltk.download('tagsets')
nltk.help.brown_tagset()
```

```
(: opening parenthesis
    (
): closing parenthesis
    )
*: negator
    not n't
,: comma
    ,
--: dash
    --
.: sentence terminator
    . ? ; ! :
:: colon
    :
ABL: determiner/pronoun, pre-qualifier
    quite such rather
ABN: determiner/pronoun, pre-quantifier
    all half many nary
ABX: determiner/pronoun, double conjunction or pre-quantifier
    both
AP: determiner/pronoun, post-determiner
    many other next more last former little several enough most least
only
    very few fewer past same Last latter less single plenty 'nough
lesser
    certain various manye next-to-last particular final previous
present
    nuf
```

AP$: determiner/pronoun, post-determiner, genitive
    other's
AP+AP: determiner/pronoun, post-determiner, hyphenated pair
    many-much
AT: article
    the an no a every th' ever' ye
BE: verb 'to be', infinitive or imperative
    be
BED: verb 'to be', past tense, 2nd person singular or all persons plural
    were
BED*: verb 'to be', past tense, 2nd person singular or all persons plural, negated
    weren't
BEDZ: verb 'to be', past tense, 1st and 3rd person singular
    was
BEDZ*: verb 'to be', past tense, 1st and 3rd person singular, negated
    wasn't
BEG: verb 'to be', present participle or gerund
    being
BEM: verb 'to be', present tense, 1st person singular
    am
BEM*: verb 'to be', present tense, 1st person singular, negated
    ain't
BEN: verb 'to be', past participle
    been
BER: verb 'to be', present tense, 2nd person singular or all persons plural
    are art
BER*: verb 'to be', present tense, 2nd person singular or all persons plural, negated
    aren't ain't
BEZ: verb 'to be', present tense, 3rd person singular
    is
BEZ*: verb 'to be', present tense, 3rd person singular, negated
    isn't ain't
CC: conjunction, coordinating
    and or but plus & either neither nor yet 'n' and/or minus an'
CD: numeral, cardinal
    two one 1 four 2 1913 71 74 637 1937 8 five three million 87-31 29-5
    seven 1,119 fifty-three 7.5 billion hundred 125,000 1,700 60 100 six
    ...
CD$: numeral, cardinal, genitive
    1960's 1961's .404's
CS: conjunction, subordinating
    that as after whether before while like because if since for than altho
    until so unless though providing once lest s'posin' till whereas

    whereupon supposing tho' albeit then so's 'fore
DO: verb 'to do', uninflected present tense, infinitive or imperative
    do dost
DO*: verb 'to do', uninflected present tense or imperative, negated
    don't
DO+PPSS: verb 'to do', past or present tense + pronoun, personal, nominative, not 3rd person singular
    d'you
DOD: verb 'to do', past tense
    did done
DOD*: verb 'to do', past tense, negated
    didn't
DOZ: verb 'to do', present tense, 3rd person singular
    does
DOZ*: verb 'to do', present tense, 3rd person singular, negated
    doesn't don't
DT: determiner/pronoun, singular
    this each another that 'nother
DT$: determiner/pronoun, singular, genitive
    another's
DT+BEZ: determiner/pronoun + verb 'to be', present tense, 3rd person singular
    that's
DT+MD: determiner/pronoun + modal auxillary
    that'll this'll
DTI: determiner/pronoun, singular or plural
    any some
DTS: determiner/pronoun, plural
    these those them
DTS+BEZ: pronoun, plural + verb 'to be', present tense, 3rd person singular
    them's
DTX: determiner, pronoun or double conjunction
    neither either one
EX: existential there
    there
EX+BEZ: existential there + verb 'to be', present tense, 3rd person singular
    there's
EX+HVD: existential there + verb 'to have', past tense
    there'd
EX+HVZ: existential there + verb 'to have', present tense, 3rd person singular
    there's
EX+MD: existential there + modal auxillary
    there'll there'd
FW-*: foreign word: negator
    pas non ne
FW-AT: foreign word: article
    la le el un die der ein keine eine das las les Il

FW-AT+NN: foreign word: article + noun, singular, common
    l'orchestre l'identite l'arcade l'ange l'assistance l'activite
    L'Universite l'independance L'Union L'Unita l'osservatore
FW-AT+NP: foreign word: article + noun, singular, proper
    L'Astree L'Imperiale
FW-BE: foreign word: verb 'to be', infinitive or imperative
    sit
FW-BER: foreign word: verb 'to be', present tense, 2nd person singular or all persons plural
    sind sunt etes
FW-BEZ: foreign word: verb 'to be', present tense, 3rd person singular
    ist est
FW-CC: foreign word: conjunction, coordinating
    et ma mais und aber och nec y
FW-CD: foreign word: numeral, cardinal
    une cinq deux sieben unam zwei
FW-CS: foreign word: conjunction, subordinating
    bevor quam ma
FW-DT: foreign word: determiner/pronoun, singular
    hoc
FW-DT+BEZ: foreign word: determiner + verb 'to be', present tense, 3rd person singular
    c'est
FW-DTS: foreign word: determiner/pronoun, plural
    haec
FW-HV: foreign word: verb 'to have', present tense, not 3rd person singular
    habe
FW-IN: foreign word: preposition
    ad de en a par con dans ex von auf super post sine sur sub avec per
    inter sans pour pendant in di
FW-IN+AT: foreign word: preposition + article
    della des du aux zur d'un del dell'
FW-IN+NN: foreign word: preposition + noun, singular, common
    d'etat d'hotel d'argent d'identite d'art
FW-IN+NP: foreign word: preposition + noun, singular, proper
    d'Yquem d'Eiffel
FW-JJ: foreign word: adjective
    avant Espagnol sinfonica Siciliana Philharmonique grand publique haute
    noire bouffe Douce meme humaine bel serieuses royaux anticus presto
    Sovietskaya Bayerische comique schwarzen ...
FW-JJR: foreign word: adjective, comparative
    fortiori
FW-JJT: foreign word: adjective, superlative
    optimo
FW-NN: foreign word: noun, singular, common
    ballet esprit ersatz mano chatte goutte sang Fledermaus oud def

kolkhoz
    roi troika canto boite blutwurst carne muzyka bonheur monde piece force
    ...
FW-NN$: foreign word: noun, singular, common, genitive
    corporis intellectus arte's dei aeternitatis senioritatis curiae
    patronne's chambre's
FW-NNS: foreign word: noun, plural, common
    al culpas vopos boites haflis kolkhozes augen tyrannis alpha-beta-
    gammas metis banditos rata phis negociants crus Einsatzkommandos
    kamikaze wohaws sabinas zorrillas palazzi engages coureurs corroborees
    yori Ubermenschen ...
FW-NP: foreign word: noun, singular, proper
    Karshilama Dieu Rundfunk Afrique Espanol Afrika Spagna Gott Carthago
    deus
FW-NPS: foreign word: noun, plural, proper
    Svenskarna Atlantes Dieux
FW-NR: foreign word: noun, singular, adverbial
    heute morgen aujourd'hui hoy
FW-OD: foreign word: numeral, ordinal
    18e 17e quintus
FW-PN: foreign word: pronoun, nominal
    hoc
FW-PP$: foreign word: determiner, possessive
    mea mon deras vos
FW-PPL: foreign word: pronoun, singular, reflexive
    se
FW-PPL+VBZ: foreign word: pronoun, singular, reflexive + verb, present tense, 3rd person singular
    s'excuse s'accuse
FW-PPO: pronoun, personal, accusative
    lui me moi mi
FW-PPO+IN: foreign word: pronoun, personal, accusative + preposition
    mecum tecum
FW-PPS: foreign word: pronoun, personal, nominative, 3rd person singular
    il
FW-PPSS: foreign word: pronoun, personal, nominative, not 3rd person singular
    ich vous sie je
FW-PPSS+HV: foreign word: pronoun, personal, nominative, not 3rd person singular + verb 'to have', present tense, not 3rd person singular
    j'ai
FW-QL: foreign word: qualifier
    minus
FW-RB: foreign word: adverb
    bas assai deja um wiederum cito velociter vielleicht simpliciter

non zu
    domi nuper sic forsan olim oui semper tout despues hors
FW-RB+CC: foreign word: adverb + conjunction, coordinating
    forisque
FW-TO+VB: foreign word: infinitival to + verb, infinitive
    d'entretenir
FW-UH: foreign word: interjection
    sayonara bien adieu arigato bonjour adios bueno tchalo ciao o
FW-VB: foreign word: verb, present tense, not 3rd person singular, imperative or infinitive
    nolo contendere vive fermate faciunt esse vade noli tangere dites duces
    meminisse iuvabit gosaimasu voulez habla ksu'u'peli'afo lacheln miuchi
    say allons strafe portant
FW-VBD: foreign word: verb, past tense
    stabat peccavi audivi
FW-VBG: foreign word: verb, present participle or gerund
    nolens volens appellant seq. obliterans servanda dicendi delenda
FW-VBN: foreign word: verb, past participle
    vue verstrichen rasa verboten engages
FW-VBZ: foreign word: verb, present tense, 3rd person singular
    gouverne sinkt sigue diapiace
FW-WDT: foreign word: WH-determiner
    quo qua quod que quok
FW-WPO: foreign word: WH-pronoun, accusative
    quibusdam
FW-WPS: foreign word: WH-pronoun, nominative
    qui
HV: verb 'to have', uninflected present tense, infinitive or imperative
    have hast
HV*: verb 'to have', uninflected present tense or imperative, negated
    haven't ain't
HV+TO: verb 'to have', uninflected present tense + infinitival to
    hafta
HVD: verb 'to have', past tense
    had
HVD*: verb 'to have', past tense, negated
    hadn't
HVG: verb 'to have', present participle or gerund
    having
HVN: verb 'to have', past participle
    had
HVZ: verb 'to have', present tense, 3rd person singular
    has hath
HVZ*: verb 'to have', present tense, 3rd person singular, negated
    hasn't ain't
IN: preposition
    of in for by considering to on among at through with under into

    regarding than since despite according per before toward against
as
    after during including between without except upon out over ...
IN+IN: preposition, hyphenated pair
    f'ovuh
IN+PPO: preposition + pronoun, personal, accusative
    t'hi-im
JJ: adjective
    ecent over-all possible hard-fought favorable hard meager fit such
    widespread outmoded inadequate ambiguous grand clerical effective
    orderly federal foster general proportionate ...
JJ$: adjective, genitive
    Great's
JJ+JJ: adjective, hyphenated pair
    big-large long-far
JJR: adjective, comparative
    greater older further earlier later freer franker wider better
deeper
    firmer tougher faster higher bigger worse younger lighter nicer
slower
    happier frothier Greater newer Elder ...
JJR+CS: adjective + conjunction, coordinating
    lighter'n
JJS: adjective, semantically superlative
    top chief principal northernmost master key head main tops utmost
    innermost foremost uppermost paramount topmost
JJT: adjective, superlative
    best largest coolest calmest latest greatest earliest simplest
    strongest newest fiercest unhappiest worst youngest worthiest
fastest
    hottest fittest lowest finest smallest staunchest ...
MD: modal auxillary
    should may might will would must can could shall ought need wilt
MD*: modal auxillary, negated
    cannot couldn't wouldn't can't won't shouldn't shan't mustn't
musn't
MD+HV: modal auxillary + verb 'to have', uninflected form
    shouldda musta coulda must've woulda could've
MD+PPSS: modal auxillary + pronoun, personal, nominative, not 3rd
person singular
    willya
MD+TO: modal auxillary + infinitival to
    oughta
NN: noun, singular, common
    failure burden court fire appointment awarding compensation Mayor
    interim committee fact effect airport management surveillance jail
    doctor intern extern night weekend duty legislation Tax Office ...
NN$: noun, singular, common, genitive
    season's world's player's night's chapter's golf's football's
    baseball's club's U.'s coach's bride's bridegroom's board's

county's
    firm's company's superintendent's mob's Navy's ...
NN+BEZ: noun, singular, common + verb 'to be', present tense, 3rd person singular
    water's camera's sky's kid's Pa's heat's throat's father's money's
    undersecretary's granite's level's wife's fat's Knife's fire's name's
    hell's leg's sun's roulette's cane's guy's kind's baseball's ...
NN+HVD: noun, singular, common + verb 'to have', past tense
    Pa'd
NN+HVZ: noun, singular, common + verb 'to have', present tense, 3rd person singular
    guy's Knife's boat's summer's rain's company's
NN+IN: noun, singular, common + preposition
    buncha
NN+MD: noun, singular, common + modal auxillary
    cowhand'd sun'll
NN+NN: noun, singular, common, hyphenated pair
    stomach-belly
NNS: noun, plural, common
    irregularities presentments thanks reports voters laws legislators
    years areas adjustments chambers $100 bonds courts sales details raises
    sessions members congressmen votes polls calls ...
NNS$: noun, plural, common, genitive
    taxpayers' children's members' States' women's cutters' motorists'
    steelmakers' hours' Nations' lawyers' prisoners' architects' tourists'
    Employers' secretaries' Rogues' ...
NNS+MD: noun, plural, common + modal auxillary
    duds'd oystchers'll
NP: noun, singular, proper
    Fulton Atlanta September-October Durwood Pye Ivan Allen Jr. Jan.
    Alpharetta Grady William B. Hartsfield Pearl Williams Aug. Berry J. M.
    Cheshire Griffin Opelika Ala. E. Pelham Snodgrass ...
NP$: noun, singular, proper, genitive
    Green's Landis' Smith's Carreon's Allison's Boston's Spahn's Willie's
    Mickey's Milwaukee's Mays' Howsam's Mantle's Shaw's Wagner's Rickey's
    Shea's Palmer's Arnold's Broglio's ...
NP+BEZ: noun, singular, proper + verb 'to be', present tense, 3rd person singular
    W.'s Ike's Mack's Jack's Kate's Katharine's Black's Arthur's Seaton's
    Buckhorn's Breed's Penny's Rob's Kitty's Blackwell's Myra's Wally's
    Lucille's Springfield's Arlene's
NP+HVZ: noun, singular, proper + verb 'to have', present tense, 3rd

person singular
    Bill's Guardino's Celie's Skolman's Crosson's Tim's Wally's
NP+MD: noun, singular, proper + modal auxillary
    Gyp'll John'll
NPS: noun, plural, proper
    Chases Aderholds Chapelles Armisteads Lockies Carbones French
Marskmen
    Toppers Franciscans Romans Cadillacs Masons Blacks Catholics
British
    Dixiecrats Mississippians Congresses ...
NPS$: noun, plural, proper, genitive
    Republicans' Orioles' Birds' Yanks' Redbirds' Bucs' Yankees'
Stevenses'
    Geraghtys' Burkes' Wackers' Achaeans' Dresbachs' Russians'
Democrats'
    Gershwins' Adventists' Negroes' Catholics' ...
NR: noun, singular, adverbial
    Friday home Wednesday Tuesday Monday Sunday Thursday yesterday
tomorrow
    tonight West East Saturday west left east downtown north northeast
    southeast northwest North South right ...
NR$: noun, singular, adverbial, genitive
    Saturday's Monday's yesterday's tonight's tomorrow's Sunday's
    Wednesday's Friday's today's Tuesday's West's Today's South's
NR+MD: noun, singular, adverbial + modal auxillary
    today'll
NRS: noun, plural, adverbial
    Sundays Mondays Saturdays Wednesdays Souths Fridays
OD: numeral, ordinal
    first 13th third nineteenth 2d 61st second sixth eighth ninth
twenty-
    first eleventh 50th eighteenth- Thirty-ninth 72nd 1/20th twentieth
    mid-19th thousandth 350th sixteenth 701st ...
PN: pronoun, nominal
    none something everything one anyone nothing nobody everybody
everyone
    anybody anything someone no-one nothin
PN$: pronoun, nominal, genitive
    one's someone's anybody's nobody's everybody's anyone's everyone's
PN+BEZ: pronoun, nominal + verb 'to be', present tense, 3rd person
singular
    nothing's everything's somebody's nobody's someone's
PN+HVD: pronoun, nominal + verb 'to have', past tense
    nobody'd
PN+HVZ: pronoun, nominal + verb 'to have', present tense, 3rd person
singular
    nobody's somebody's one's
PN+MD: pronoun, nominal + modal auxillary
    someone'll somebody'll anybody'd
PP$: determiner, possessive

    our its his their my your her out thy mine thine
PP$$: pronoun, possessive
    ours mine his hers theirs yours
PPL: pronoun, singular, reflexive
    itself himself myself yourself herself oneself ownself
PPLS: pronoun, plural, reflexive
    themselves ourselves yourselves
PPO: pronoun, personal, accusative
    them it him me us you 'em her thee we'uns
PPS: pronoun, personal, nominative, 3rd person singular
    it he she thee
PPS+BEZ: pronoun, personal, nominative, 3rd person singular + verb 'to be', present tense, 3rd person singular
    it's he's she's
PPS+HVD: pronoun, personal, nominative, 3rd person singular + verb 'to have', past tense
    she'd he'd it'd
PPS+HVZ: pronoun, personal, nominative, 3rd person singular + verb 'to have', present tense, 3rd person singular
    it's he's she's
PPS+MD: pronoun, personal, nominative, 3rd person singular + modal auxillary
    he'll she'll it'll he'd it'd she'd
PPSS: pronoun, personal, nominative, not 3rd person singular
    they we I you ye thou you'uns
PPSS+BEM: pronoun, personal, nominative, not 3rd person singular + verb 'to be', present tense, 1st person singular
    I'm Ahm
PPSS+BER: pronoun, personal, nominative, not 3rd person singular + verb 'to be', present tense, 2nd person singular or all persons plural
    we're you're they're
PPSS+BEZ: pronoun, personal, nominative, not 3rd person singular + verb 'to be', present tense, 3rd person singular
    you's
PPSS+BEZ*: pronoun, personal, nominative, not 3rd person singular + verb 'to be', present tense, 3rd person singular, negated
    'tain't
PPSS+HV: pronoun, personal, nominative, not 3rd person singular + verb 'to have', uninflected present tense
    I've we've they've you've
PPSS+HVD: pronoun, personal, nominative, not 3rd person singular + verb 'to have', past tense
    I'd you'd we'd they'd
PPSS+MD: pronoun, personal, nominative, not 3rd person singular + modal auxillary
    you'll we'll I'll we'd I'd they'll they'd you'd
PPSS+VB: pronoun, personal, nominative, not 3rd person singular + verb 'to verb', uninflected present tense
    y'know
QL: qualifier, pre

    well less very most so real as highly fundamentally even how much
    remarkably somewhat more completely too thus ill deeply little
overly
    halfway almost impossibly far severly such ...
QLP: qualifier, post
    indeed enough still 'nuff
RB: adverb
    only often generally also nevertheless upon together back newly no
    likely meanwhile near then heavily there apparently yet outright
fully
    aside consistently specifically formally ever just ...
RB$: adverb, genitive
    else's
RB+BEZ: adverb + verb 'to be', present tense, 3rd person singular
    here's there's
RB+CS: adverb + conjunction, coordinating
    well's soon's
RBR: adverb, comparative
    further earlier better later higher tougher more harder longer
sooner
    less faster easier louder farther oftener nearer cheaper slower
tighter
    lower worse heavier quicker ...
RBR+CS: adverb, comparative + conjunction, coordinating
    more'n
RBT: adverb, superlative
    most best highest uppermost nearest brightest hardest fastest
deepest
    farthest loudest ...
RN: adverb, nominal
    here afar then
RP: adverb, particle
    up out off down over on in about through across after
RP+IN: adverb, particle + preposition
    out'n outta
TO: infinitival to
    to t'
TO+VB: infinitival to + verb, infinitive
    t'jawn t'lah
UH: interjection
    Hurrah bang whee hmpf ah goodbye oops oh-the-pain-of-it ha crunch
say
    oh why see well hello lo alas tarantara rum-tum-tum gosh hell
keerist
    Jesus Keeeerist boy c'mon 'mon goddamn bah hoo-pig damn ...
VB: verb, base: uninflected present, imperative or infinitive
    investigate find act follow inure achieve reduce take remedy re-
set
    distribute realize disable feel receive continue place protect
    eliminate elaborate work permit run enter force ...

VB+AT: verb, base: uninflected present or infinitive + article
    wanna
VB+IN: verb, base: uninflected present, imperative or infinitive + preposition
    lookit
VB+JJ: verb, base: uninflected present, imperative or infinitive + adjective
    die-dead
VB+PPO: verb, uninflected present tense + pronoun, personal, accusative
    let's lemme gimme
VB+RP: verb, imperative + adverbial particle
    g'ahn c'mon
VB+TO: verb, base: uninflected present, imperative or infinitive + infinitival to
    wanta wanna
VB+VB: verb, base: uninflected present, imperative or infinitive; hypenated pair
    say-speak
VBD: verb, past tense
    said produced took recommended commented urged found added praised
    charged listed became announced brought attended wanted voted defeated
    received got stood shot scheduled feared promised made ...
VBG: verb, present participle or gerund
    modernizing improving purchasing Purchasing lacking enabling pricing
    keeping getting picking entering voting warning making strengthening
    setting neighboring attending participating moving ...
VBG+TO: verb, present participle + infinitival to
    gonna
VBN: verb, past participle
    conducted charged won received studied revised operated accepted
    combined experienced recommended effected granted seen protected
    adopted retarded notarized selected composed gotten printed ...
VBN+TO: verb, past participle + infinitival to
    gotta
VBZ: verb, present tense, 3rd person singular
    deserves believes receives takes goes expires says opposes starts
    permits expects thinks faces votes teaches holds calls fears spends
    collects backs eliminates sets flies gives seeks reads ...
WDT: WH-determiner
    which what whatever whichever whichever-the-hell
WDT+BER: WH-determiner + verb 'to be', present tense, 2nd person singular or all persons plural
    what're
WDT+BER+PP: WH-determiner + verb 'to be', present, 2nd person singular or all persons plural + pronoun, personal, nominative, not 3rd person

singular
    whaddya
WDT+BEZ: WH-determiner + verb 'to be', present tense, 3rd person singular
    what's
WDT+DO+PPS: WH-determiner + verb 'to do', uninflected present tense + pronoun, personal, nominative, not 3rd person singular
    whaddya
WDT+DOD: WH-determiner + verb 'to do', past tense
    what'd
WDT+HVZ: WH-determiner + verb 'to have', present tense, 3rd person singular
    what's
WP$: WH-pronoun, genitive
    whose whosever
WPO: WH-pronoun, accusative
    whom that who
WPS: WH-pronoun, nominative
    that who whoever whosoever what whatsoever
WPS+BEZ: WH-pronoun, nominative + verb 'to be', present, 3rd person singular
    that's who's
WPS+HVD: WH-pronoun, nominative + verb 'to have', past tense
    who'd
WPS+HVZ: WH-pronoun, nominative + verb 'to have', present tense, 3rd person singular
    who's that's
WPS+MD: WH-pronoun, nominative + modal auxillary
    who'll that'd who'd that'll
WQL: WH-qualifier
    however how
WRB: WH-adverb
    however when where why whereby wherever how whenever whereon wherein
    wherewith wheare wherefore whereof howsabout
WRB+BER: WH-adverb + verb 'to be', present, 2nd person singular or all persons plural
    where're
WRB+BEZ: WH-adverb + verb 'to be', present, 3rd person singular
    how's where's
WRB+DO: WH-adverb + verb 'to do', present, not 3rd person singular
    howda
WRB+DOD: WH-adverb + verb 'to do', past tense
    where'd how'd
WRB+DOD*: WH-adverb + verb 'to do', past tense, negated
    whyn't
WRB+DOZ: WH-adverb + verb 'to do', present tense, 3rd person singular
    how's
WRB+IN: WH-adverb + preposition
    why'n

```
WRB+MD: WH-adverb + modal auxillary
    where'd

[nltk_data] Downloading package tagsets to /home/dara/nltk_data...
[nltk_data]   Package tagsets is already up-to-date!

nltk.help.upenn_tagset('NNP')

NNP: noun, proper, singular
    Motown Venneboerger Czestochwa Ranzer Conchita Trumplane Christos
    Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin ODI Darryl
CTCA
    Shannon A.K.C. Meltex Liverpool ...
```

Stop words removal

For analyzing text and NLP, stopwords are removed from the text, as they do not add much value and meaning to the text. Stopwords, if added would bring in a lot of unnecessary noise and be of no use to the analytics process. Also, the removal of stopwords reduces the amount of data we have to process, thus reducing the number of tokens and makes everything faster.

Examples of Stopwords in English: ‘nor’, ‘me’, ‘were’, ‘her’, ‘more’, ‘himself’, ‘this’.

```
#Stop words are generally the most common words in a language.
#English stop words from nltk.

stopwords = nltk.corpus.stopwords.words('english')

words_new = []

#Now we need to remove the stop words from the words variable
#Appending to words_new all words that are in words but not in
stopwords

for word in words:
    if word not in stopwords:
        words_new.append(word)

len(words_new)

318305
```

Stemming and Lemmatization

Stemming just removes the last few characters of a word, often leading incorrect meanings and spelling.

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item.

Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

Lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

```
from nltk.stem import WordNetLemmatizer

wn = WordNetLemmatizer()

lem_words=[]

for word in words_new:
    word=wn.lemmatize(word)
    lem_words.append(word)

same=0
diff=0

for i in range(0,1832):
    if(lem_words[i]==words_new[i]):
        same=same+1
    elif(lem_words[i]!=words_new[i]):
        diff=diff+1

print('Number of words Lemmatized=', diff)
print('Number of words not Lemmatized=', same)
```

```
Number of words Lemmatized= 294
Number of words not Lemmatized= 1538
```

Now, with the Lemmatization done, we proceed to get the Frequency Distribution.