```
# Importing necessary library
import pandas as pd
import numpy as np
import nltk
import os
import nltk.corpus
nltk.download('punkt')
```

we can see the text split into tokens. Words, comma, punctuations are called tokens.

```
# sample text for performing tokenization
text = "In Brazil they drive on the right-hand side of the road.
Brazil has a large coastline on the eastern side of South America"
# importing word_tokenize from nltk
from nltk.tokenize import word_tokenize
# Passing the string text into word tokenize for breaking the
sentences
token = word_tokenize(text)
print(token)
```

```
['In', 'Brazil', 'they', 'drive', 'on', 'the', 'right-hand', 'side',
'of', 'the', 'road', '.', 'Brazil', 'has', 'a', 'large', 'coastline',
'on', 'the', 'eastern', 'side', 'of', 'South', 'America']
```

Finding frequency distinct in the text

```
# Importing FreqDist library from nltk and passing token into FreqDist
from nltk.probability import FreqDist
fdist = FreqDist(token)
fdist
```

```
FreqDist({'.': 1,
          'America': 1,
          'Brazil': 2,
          'In': 1,
          'South': 1,
          'a': 1,
          'coastline': 1,
          'drive': 1,
          'eastern': 1,
          'has': 1,
          'large': 1,
          'of': 2,
          'on': 2,
          'right-hand': 1,
          'road': 1,
          'side': 2,
          'the': 3,
          'they': 1})
```

```
# To find the frequency of top 10 words
fdist1 = fdist.most_common(10)
fdist1
```

```
[('the', 3),
 ('Brazil', 2),
 ('on', 2),
 ('side', 2),
 ('of', 2),
 ('In', 1),
 ('they', 1),
 ('drive', 1),
 ('right-hand', 1),
 ('road', 1)]
```

```
#Find difference between LancasterStemmer and Porterstemmer

# Importing LancasterStemmer from nltk
from nltk.stem import LancasterStemmer
from nltk.stem import PorterStemmer
pst = PorterStemmer()
lst = LancasterStemmer()
stm = ["giving", "given", "given", "gave"]
print("LancasterStemmer")
for word in stm :
 print(word+ ":" +lst.stem(word))
print("\nPorterStemmer ")
for word in stm :
   print(word+ ":" +pst.stem(word))
```

```
LancasterStemmer
giving:giv
given:giv
given:giv
gave:gav

PorterStemmer
giving:give
given:given
given:given
gave:gave
```

“Stop words” are the most common words in a language like “the”, “a”, “at”, “for”, “above”, “on”, “is”, “all”. These words do not provide any meaning and are usually removed from texts. We can remove these stop words using nltk library

```
nltk.download('stopwords')
# importing stopwors from nltk library
from nltk import word_tokenize
from nltk.corpus import stopwords
a = set(stopwords.words("english"))
```

```
text = "Cristiano Ronaldo was born on February 5, 1985, in Funchal,
Madeira, Portugal."
text1 = word_tokenize(text.lower())
print(text1)
stopwords = [x for x in text1 if x not in a]
print(stopwords)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
['cristiano', 'ronaldo', 'was', 'born', 'on', 'february', '5', ',',
'1985', ',', 'in', 'funchal', ',', 'madeira', ',', 'portugal', '.']
['cristiano', 'ronaldo', 'born', 'february', '5', ',', '1985', ',',
'funchal', ',', 'madeira', ',', 'portugal', '.']
```

Part of speech tagging (POS)

Used to assign parts of speech to each word of a given text (such as nouns, verbs, pronouns, adverbs, conjunction, adjectives, interjection) based on its definition and its context.

There are many tools available for POS taggers and some of the widely used taggers are NLTK, Spacy, TextBlob, Standford CoreNLP, etc.

```
#first need to download the averaged_perceptron_tagger resource
through the NLTK downloader.
nltk.download('averaged_perceptron_tagger')
from nltk.tag import pos_tag
text = "vote to choose a particular man or a group (party) to
represent them in parliament"
#Tokenize the text
tex = word_tokenize(text)
for token in tex:
  print(nltk.pos_tag([token]))

#https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_po
s.html---->Link to check list of all tags
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[('vote', 'NN')]
[('to', 'TO')]
[('choose', 'NN')]
[('a', 'DT')]
[('particular', 'JJ')]
[('man', 'NN')]
[('or', 'CC')]
[('a', 'DT')]
[('group', 'NN')]
[('(', '(')]
[('party', 'NN')]
```

```
[(')', ')')]
[('to', 'TO')]
[('represent', 'NN')]
[('them', 'PRP')]
[('in', 'IN')]
[('parliament', 'NN')]
```