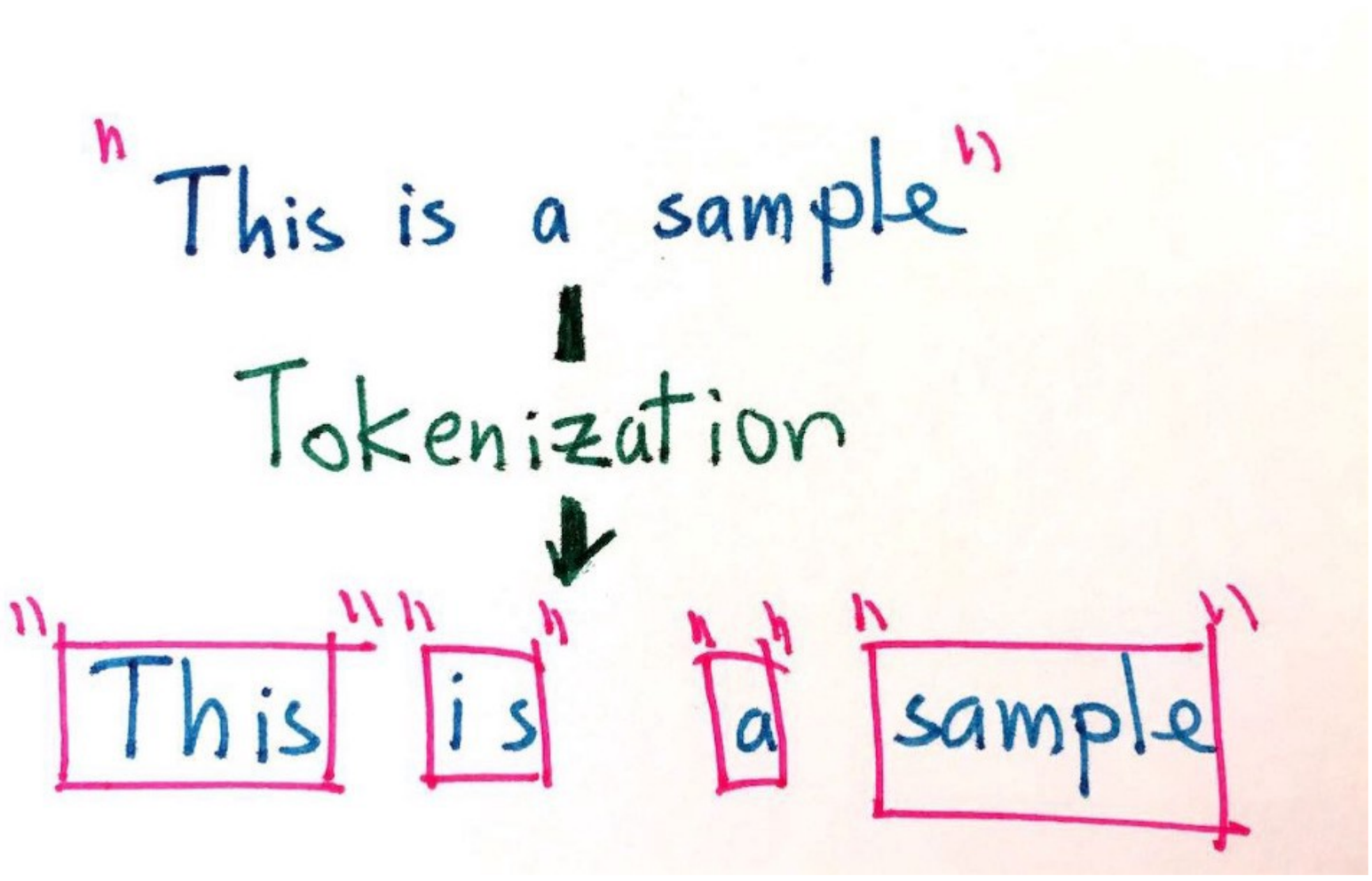
NLP

```
!pip install nltk
```

In [4]:
```
import nltk
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
```

Tokenization

Tokenization is the process of breaking down the given text in natural language processing into the smallest unit in a sentence called a token. Punctuation marks, words, and numbers can be considered tokens.




```
img source: https://medium.com/data-science-in-your-pocket/tokenization-algorithms-in-natural-language-processing-nlp-1fceab8454af
```

In [63]:
```
paragraph = """ Narendra Damodardas Modi (Gujarati: [ˈnərendrə dɑmodərˈdɑs ˈmodiː] (listen); born 17 September 1950)[a] is an Indian politician serving as the 14th and current prime

Born and raised in Vadnagar, a small town in northeastern Gujarat, Modi completed his secondary education there. He was introduced to the RSS at age eight. He has drawn attention to
```

In [64]:
```
paragraph
```

Out[64]:
```
" Narendra Damodardas Modi (Gujarati: [ˈnərendrə dɑmodərˈdɑs ˈmodiː] (listen); born 17 September 1950)[a] is an Indian politician serving as the 14th and current prime minister of
India since 2014. Modi was the chief minister of Gujarat from 2001 to 2014 and is the Member of Parliament from Varanasi. He is a member of the Bharatiya Janata Party (BJP) and of
the Rashtriya Swayamsevak Sangh (RSS), a right-wing Hindu nationalist paramilitary volunteer organisation. He is the first prime minister to have been born after India's independen
ce in 1947 and the second prime minister not belonging to the Indian National Congress to have won two consecutive majorities in the Lok Sabha, or the lower house of India's parlia
ment. He is also the longest serving prime minister from a non-Congress party.\n\nBorn and raised in Vadnagar, a small town in northeastern Gujarat, Modi completed his secondary ed
ucation there. He was introduced to the RSS at age eight. He has drawn attention to having to work as a child in his father's tea stall on the Vadnagar railway station platform, a
description that has not been reliably corroborated. At age 18, Modi was married to Jashodaben Chimanlal Modi, whom he abandoned soon after. He left his parental home where she had
come to live. He first publicly acknowledged her as his wife more than four decades later when required to do so by Indian law, but has made no contact with her since. Modi has ass
erted he had travelled in northern India for two years after leaving his parental home, visiting a number of religious centres, but few details of his travels have emerged. Upon hi
s return to Gujarat in 1971, he became a full-time worker for the RSS. After the state of emergency was declared by prime minister Indira Gandhi in 1975, Modi went into hiding. The
RSS assigned him to the BJP in 1985 and he held several positions within the party hierarchy until 2001, rising to the rank of general secretary.[b]   "
```

In [54]:
```
!pip install textblob
```

```
Collecting textblob
  Using cached textblob-0.17.1-py2.py3-none-any.whl (636 kB)
Requirement already satisfied: nltk>=3.1 in c:\users\panka\anaconda3\envs\paper\lib\site-packages (from textblob) (3.7)
Requirement already satisfied: regex>=2021.8.3 in c:\users\panka\anaconda3\envs\paper\lib\site-packages (from nltk>=3.1->textblob) (2022.7.25)
Requirement already satisfied: click in c:\users\panka\anaconda3\envs\paper\lib\site-packages (from nltk>=3.1->textblob) (8.1.3)
Requirement already satisfied: tqdm in c:\users\panka\anaconda3\envs\paper\lib\site-packages (from nltk>=3.1->textblob) (4.64.0)
Requirement already satisfied: joblib in c:\users\panka\anaconda3\envs\paper\lib\site-packages (from nltk>=3.1->textblob) (1.1.0)
Requirement already satisfied: colorama in c:\users\panka\anaconda3\envs\paper\lib\site-packages (from click->nltk>=3.1->textblob) (0.4.5)
Installing collected packages: textblob
Successfully installed textblob-0.17.1
```

In [55]:
```
import textblob
from textblob import TextBlob
```

In [56]:
```
text = "Hello everyone!, I am studying Natural Language Processing."
```

In [57]:
```
TextBlob(text).words
```

Out[57]:
```
WordList(['Hello', 'everyone', 'I', 'am', 'studying', 'Natural', 'Language', 'Processing'])
```

In [58]:
```
import nltk
from nltk import sent_tokenize
from nltk import word_tokenize
```

In [60]:
```
tokens_sents = nltk.sent_tokenize(text)
print(tokens_sents)
```

```
['Hello everyone!, I am studying Natural Language Processing.']
```

In [61]:
```
tokens_words = nltk.word_tokenize(text)
print(tokens_words)
```

```
['Hello', 'everyone', '!', ',', 'I', 'am', 'studying', 'Natural', 'Language', 'Processing', '.']
```

In [7]:
```
## Tokenization -- convert paragraps-sentence-words
nltk.download('punkt')
sentences = nltk.sent_tokenize(paragraph)
```

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

In [8]:
```
sentences
```

Out[8]:
```
[' Narendra Damodardas Modi (Gujarati: [ˈnərendrə dɑmodərˈdɑs ˈmodiː] (listen); born 17 September 1950)[a] is an Indian politician serving as the 14th and current prime minister of India since 2014.',
 'Modi was the chief minister of Gujarat from 2001 to 2014 and is the Member of Parliament from Varanasi.',
 'He is a member of the Bharatiya Janata Party (BJP) and of the Rashtriya Swayamsevak Sangh (RSS), a right-wing Hindu nationalist paramilitary volunteer organisation.',
 "He is the first prime minister to have been born after India's independence in 1947 and the second prime minister not belonging to the Indian National Congress to have won two consecutive majorities in the Lok Sabha, or the lower house of India's parliament.",
 'He is also the longest serving prime minister from a non-Congress party.',
 'Born and raised in Vadnagar, a small town in northeastern Gujarat, Modi completed his secondary education there.',
 'He was introduced to the RSS at age eight.',
 "He has drawn attention to having to work as a child in his father's tea stall on the Vadnagar railway station platform, a description that has not been reliably corroborated.",
 'At age 18, Modi was married to Jashodaben Chimanlal Modi, whom he abandoned soon after.',
 'He left his parental home where she had come to live.',
 'He first publicly acknowledged her as his wife more than four decades later when required to do so by Indian law, but has made no contact with her since.',
 'Modi has asserted he had travelled in northern India for two years after leaving his parental home, visiting a number of religious centres, but few details of his travels have emerged.',
 'Upon his return to Gujarat in 1971, he became a full-time worker for the RSS.',
 'After the state of emergency was declared by prime minister Indira Gandhi in 1975, Modi went into hiding.',
 'The RSS assigned him to the BJP in 1985 and he held several positions within the party hierarchy until 2001, rising to the rank of general secretary.',
 '[b]']
```

In [9]:
```
type(sentences)
```

Out[9]:
```
list
```

In [10]:
```
sentences[1]
```

Out[10]:
```
'Modi was the chief minister of Gujarat from 2001 to 2014 and is the Member of Parliament from Varanasi.'
```

In [11]:
```
sentences[5]
```

Out[11]:
```
'Born and raised in Vadnagar, a small town in northeastern Gujarat, Modi completed his secondary education there.'
```

In [13]:
```
stemmer = PorterStemmer()
```

In [14]:
```
stemmer.stem('going')
```

Out[14]:
```
'go'
```

In [15]:
```
stemmer.stem('thinking')
```

Out[15]:
```
'think'
```

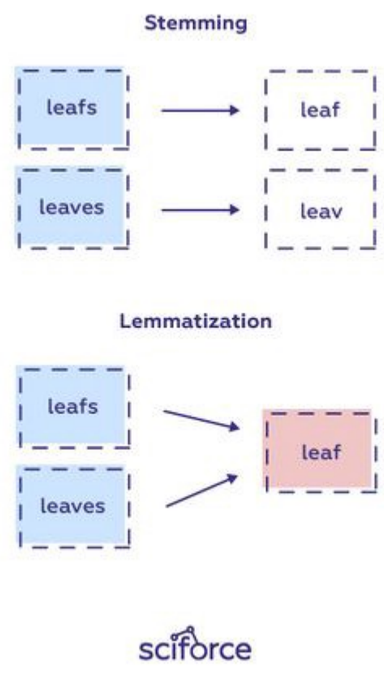
In [17]:
```
stemmer.stem('football')
```

Out[17]:
```
'footbal'
```

In [18]:
```
stemmer.stem('History')
```

Out[18]:
```
'histori'
```

Stemming and Lemmatization




```
img source :https://tr.pinterest.com/pin/706854104005417976/
```

Lemetization

Lemmatization is the process of finding the form of the related word in the dictionary. It is different from Stemming. It involves longer processes to calculate than Stemming.

In [25]:
```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
```

In [29]:
```
import nltk
nltk.download('wordnet')
nltk.download()
```

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

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

Out[29]:
```
True
```

In [30]:
```
lemmatizer.lemmatize('drinking')
```

Out[30]:
```
'drinking'
```

In [31]:
```
len(sentences)
```

Out[31]:
```
16
```

In [32]:
```
import re
```

In [36]:
```
corpus = []
for i in range(len(sentences)):
    review = re.sub('[^a-zA-Z]',' ',sentences[i])
    review = review.lower()
    corpus.append(review)
```

In [37]:
```
corpus
```

Out[37]:
```
[' narendra damodardas modi  gujarati    n  end   d mod   d s  modi    listen   born    september       a  is an indian politician serving as the   th and current prime minister of
india since      ',
 'modi was the chief minister of gujarat from      to      and is the member of parliament from varanasi ',
 'he is a member of the bharatiya janata party  bjp  and of the rashtriya swayamsevak sangh  rss   a right wing hindu nationalist paramilitary volunteer organisation ',
 'he is the first prime minister to have been born after india s independence in      and the second prime minister not belonging to the indian national congress to have won two co
nsecutive majorities in the lok sabha  or the lower house of india s parliament ',
 'he is also the longest serving prime minister from a non congress party ',
 'born and raised in vadnagar  a small town in northeastern gujarat  modi completed his secondary education there ',
 'he was introduced to the rss at age eight ',
 'he has drawn attention to having to work as a child in his father s tea stall on the vadnagar railway station platform  a description that has not been reliably corroborated ',
 'at age     modi was married to jashodaben chimanlal modi  whom he abandoned soon after ',
 'he left his parental home where she had come to live ',
 'he first publicly acknowledged her as his wife more than four decades later when required to do so by indian law  but has made no contact with her since ',
 'modi has asserted he had travelled in northern india for two years after leaving his parental home  visiting a number of religious centres  but few details of his travels have em
erged ',
 'upon his return to gujarat in       he became a full time worker for the rss ',
 'after the state of emergency was declared by prime minister indira gandhi in       modi went into hiding ',
 'the rss assigned him to the bjp in      and he held several positions within the party hierarchy until       rising to the rank of general secretary ',
 ' b ']
```

In [43]:
```
stopwords.words('english')
```

Out[43]:
```
 'ma',
 'mightn',
 "mightn't",
 'mustn',
 "mustn't",
 'needn',
 "needn't",
 'shan',
 "shan't",
 'shouldn',
 "shouldn't",
 'wasn',
 "wasn't",
 'weren',
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

Stemming

Stemming is the process of finding the root of words.

Overstemming: Overstemming occurs when words are over-truncated. In such cases, the meaning of the word may be distorted or have no meaning.

Understemming: Understemming occurs when two words are stemmed from the same root that is not of different stems.

In [41]:
```
for i in corpus:
    words = nltk.word_tokenize(i) #converts sentences into a words
    for word in words:
        if word not in set(stopwords.words('english')):
            print(stemmer.stem(word))
```

Out[41]:
```
rss
right
wing
hindu
nationalist
paramilitari
volunt
organis
first
prime
minist
born
india
independ
second
prime
minist
belong
indian
nation
```

Lemmatization

Lemmatization is the process of finding the form of the related word in the dictionary. It is different from Stemming. It involves longer processes to calculate than Stemming

In [42]:
```
for i in corpus:
    words = nltk.word_tokenize(i)
    for word in words:
        if word not in set(stopwords.words('english')):
            print(lemmatizer.lemmatize(word))
```

Out[42]:
```
narendra
damodardas
modi
gujarati
n
end
mod
modi
listen
born
september
indian
politician
serving
th
current
prime
minister
india
```

In [46]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()  # binary=False
```

In [47]:
```
X=cv.fit_transform(corpus)
```

In [48]:
```
cv.vocabulary_
```

```
 'eight': 39,
 'has': 55,
 'drawn': 37,
 'attention': 11,
 'having': 57,
 'work': 173,
 'child': 22,
 'father': 43,
 'tea': 145,
 'stall': 141,
 'on': 105,
 'railway': 117,
 'station': 143,
 'platform': 112,
 'description': 34,
 'that': 148,
 'reliably': 121,
 'corroborated': 29,
 'married': 89,
 'jashodaben': 77,
```

In [49]:
```
corpus[0]
```

Out[49]:
```
' narendra damodardas modi  gujarati    n  end   d mod   d s  modi    listen   born    september      a  is an indian politician serving as the   th and current prime minister of india since      '
```

In [50]:
```
X[0].toarray()
```

Out[50]:
```
array([[0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
        0, 0, 0, 1, 1, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
        0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]],
      dtype=int64)
```

In [51]:
```
22*8
```

Out[51]:
```
176
```

In [ ]:
```

```