7) Text Analytics

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.
2. Create representation of documents by calculating Term Frequency and Inverse DocumentFrequency.

In [1]:

**import** nltk nltk**.**download('punkt') nltk**.**download('wordnet')

nltk**.**download('averaged\_perceptron\_tagger') nltk**.**download('stopwords')

**from** nltk **import** sent\_tokenize **from** nltk **import** word\_tokenize **from** nltk.corpus **import** stopwords

[nltk\_data] Downloading package punkt to

[nltk\_data] C:\Users\adesh\AppData\Roaming\nltk\_data... [nltk\_data] Unzipping tokenizers\punkt.zip.

[nltk\_data] Downloading package wordnet to

[nltk\_data] C:\Users\adesh\AppData\Roaming\nltk\_data... [nltk\_data] Downloading package averaged\_perceptron\_tagger to [nltk\_data] C:\Users\adesh\AppData\Roaming\nltk\_data... [nltk\_data] Unzipping taggers\averaged\_perceptron\_tagger.zip. [nltk\_data] Downloading package stopwords to

[nltk\_data] C:\Users\adesh\AppData\Roaming\nltk\_data... [nltk\_data] Unzipping corpora\stopwords.zip.

In [2]:

text**=**'Real madrid is set to win the UCL for the season . Benzema might win Balon dor . S

In [3]:

tokens\_sents **=** nltk**.**sent\_tokenize(text) print(tokens\_sents)

['Real madrid is set to win the UCL for the season .', 'Benzema might win Balon dor .', 'Salah might be the runner up']

In [4]:

tokens\_words **=** nltk**.**word\_tokenize(text) print(tokens\_words)

['Real', 'madrid', 'is', 'set', 'to', 'win', 'the', 'UCL', 'for', 'the', 'season', '.',

'Benzema', 'might', 'win', 'Balon', 'dor', '.', 'Salah', 'might', 'be', 'the', 'runner', 'up']

In [5]:

**from** nltk.stem **import** PorterStemmer

**from** nltk.stem.snowball **import** SnowballStemmer

**from** nltk.stem **import** LancasterStemmer

In [6]:

stem**=**[]

**for** i **in** tokens\_words: ps **=** PorterStemmer() stem\_word**=** ps**.**stem(i) stem**.**append(stem\_word)

print(stem)

['real', 'madrid', 'is', 'set', 'to', 'win', 'the', 'ucl', 'for', 'the', 'season', '.',

'benzema', 'might', 'win', 'balon', 'dor', '.', 'salah', 'might', 'be', 'the', 'runner', 'up']

In [7]:

*#Lemmatization*

**import** nltk

**from** nltk.stem **import** WordNetLemmatizer lemmatizer **=** WordNetLemmatizer()

In [8]:

lemmatized\_output **=** ' '**.**join([lemmatizer**.**lemmatize(w) **for** w **in** stem]) print(lemmatized\_output)

real madrid is set to win the ucl for the season . benzema might win balon dor . salah m ight be the runner up

In [9]:

leme**=**[]

**for** i **in** stem: lemetized\_word**=**lemmatizer**.**lemmatize(i) leme**.**append(lemetized\_word)

print(leme)

['real', 'madrid', 'is', 'set', 'to', 'win', 'the', 'ucl', 'for', 'the', 'season', '.',

'benzema', 'might', 'win', 'balon', 'dor', '.', 'salah', 'might', 'be', 'the', 'runner', 'up']

In [10]:

*#Part of Speech Tagging*

print("Parts of Speech: ",nltk**.**pos\_tag(leme))

Parts of Speech: [('real', 'JJ'), ('madrid', 'NN'), ('is', 'VBZ'), ('set', 'VBN'), ('t

o', 'TO'), ('win', 'VB'), ('the', 'DT'), ('ucl', 'NN'), ('for', 'IN'), ('the', 'DT'),

('season', 'NN'), ('.', '.'), ('benzema', 'NN'), ('might', 'MD'), ('win', 'VB'), ('balo

n', 'NN'), ('dor', 'NN'), ('.', '.'), ('salah', 'NN'), ('might', 'MD'), ('be', 'VB'), ('the', 'DT'), ('runner', 'NN'), ('up', 'RP')]

In [11]:

*#Stop Word*

sw\_nltk **=** stopwords**.**words('english') print(sw\_nltk)

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",

"you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himse

lf', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 't

hem', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that',

"that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'h

ave', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'bu

t', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'abo

ut', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'bel

ow', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'fu

rther', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'b

oth', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'onl

y', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don',

"don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'are

n', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "had

n't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",

'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'was

n', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

In [12]:

words **=** [word **for** word **in** text**.**split() **if** word**.**lower() **not in** sw\_nltk] new\_text **=** " "**.**join(words)

print(new\_text)

Real madrid set win UCL season . Benzema might win Balon dor . Salah might runner

In [ ]: