EXPERIMENT NO 3

PERFORMING PREPROCESSING TASK ON TWITTER DATA

1. DOWNLOADING THE TWITTER DATA

In [1]:
```python
# Importing Libraries

import tweepy as tw

import pandas as pd

import numpy as np
```

In [2]:
```python
# gathering the consumer key and secret

api_key = 'iR8RR78dKB5Cji5RikrPacFYC'
api_key_secret = '3VnG6gZRPzP4EF2bEZAeUXFblLVXZWZlJ4HYb4JuWAAK5LMNit'
access_token = '797899482566369280-EERRBmQbu2H8EqTAfUQoDJSweO8U3Kb'
access_token_secret = 'IxEx6OW5d7MtOaW8ZW57bkAlFf5t70IPoE01yldyzPwlB'
```

In [3]:
```python
# searching keyword name

search_words = "#Budget2022"

# searching number of tweets

number_posts = 1000
```

In [4]:
```python
# using the Oauth for accessing tweets

auth = tw.OAuthHandler(api_key, api_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

In [5]:
```python
# downloading the tweets

tweets = tw.Cursor(api.search_tweets, q=search_words,
                   lang='en').items(number_posts)

# list to store downloaded data

tweets_text = []

# downloading diffrent attributes

for tweet in tweets:
    tweets_text.append([
        tweet.user.screen_name, tweet.user.name, tweet.user.description,
        tweet.text, tweet.user.location, tweet.user.followers_count,
        tweet.source, tweet.created_at, tweet.user.friends_count
    ])
```

2.STORING THE DOWNLOADED DATA

In [6]:
```python
# storing tweet data

tweet_data = pd.DataFrame(tweets_text,
                          columns=[
                              "User_Name", "Name", "Description (Status)",
                              "Tweets Posted", "Location", "No of Followers",
                              "Device Used for Tweet",
                              "Date of Tweet Creation", "No of Friends"
                          ])

# starting few tweet data

tweet_data
```

Out[6]:

	User_Name	Name	Description (Status)	Tweets Posted	Location	No of Followers	Device Used for Tweet	Date of Tweet Creation	No of Friends
0	AdarshPallav	Adarsh Pallav		RT @satishacharya: Blueprint for 25 years! #Bu...		7	Twitter for Android	2022-02-07 07:11:43+00:00	38
1	bsindia	Business Standard	Latest news on the economy, companies, markets...	#BSMorningShow | Will #RBI go for a hike in re...	India	2178279	TweetDeck	2022-02-07 07:11:00+00:00	430
2	bsindia	Business Standard	Latest news on the economy, companies, markets...	#BSMornigShow | Will #RBI go for a hike in rev...	India	2178279	Twitter Web App	2022-02-07 07:10:43+00:00	430
3	bsindia	Business Standard	Latest news on the economy, companies, markets...	#BSMorningShow | Business Standard's @aruproyt...	India	2178279	TweetDeck	2022-02-07 07:10:00+00:00	430
4	orfonline	ORF	Non–partisan, independent analyses on security...	Tax concessions for co-operative societies is ...	India	106444	TweetDeck	2022-02-07 07:10:00+00:00	136
...	...	...	...	...	...	...	...	...	...
995	VikramB12Singh	विक्रम बहादुर सिंह	सहेयक आचार्य (Assistant Professor)	RT @IncomeTaxIndia: Budget Highlights:\nKey Di...	जांजगीरचांपा छत्तीसगढ, भरतखण्ड	467	Twitter for Android	2022-02-06 16:54:20+00:00	2205
996	TheRationalDesi	Rational One		#Budget2022 #Congress #BJPHataoDeshBachao #UPE...		0	Twitter Web App	2022-02-06 16:54:16+00:00	58
997	VikramB12Singh	विक्रम बहादुर सिंह	सहेयक आचार्य (Assistant Professor)	RT @cbic_india: Key highlights of Union Budget...	जांजगीरचांपा छत्तीसगढ, भरतखण्ड	467	Twitter for Android	2022-02-06 16:51:46+00:00	2205
998	Vishalk23312052	विशाल कुमार शाही	भारतीय जनता युवा मोर्चा (प्रदेश विशेष आमंत्रित...	RT @blsanthosh: 400 new generation Vande Bhara...	भिलाई छ.ग	78	Twitter for Android	2022-02-06 16:51:21+00:00	6
999	Vibhans28836256	Vibhanshu Mishra	Being sarcastic and truthful are deadly and cr...	RT @thebubblebuste1: My take on Financial Budg...	Seoni, India	79	Twitter for Android	2022-02-06 16:51:16+00:00	260

1000 rows × 9 columns

In [7]:
```python
# collected data information

tweet_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   User_Name               1000 non-null   object
 1   Name                    1000 non-null   object
 2   Description (Status)    1000 non-null   object
 3   Tweets Posted           1000 non-null   object
 4   Location                1000 non-null   object
 5   No of Followers         1000 non-null   int64
 6   Device Used for Tweet   1000 non-null   object
 7   Date of Tweet Creation  1000 non-null   datetime64[ns, UTC]
 8   No of Friends           1000 non-null   int64
dtypes: datetime64[ns, UTC](1), int64(2), object(6)
memory usage: 70.4+ KB
```

In [8]:
```python
# saving the data as csv file

tweet_data.to_csv(
    '/DOCUMENTS/COLLEGE/CLASSES/EXPERIMENT_NO_2/tweet_data_preprocess.csv')
```

3.PERFORMING PREPROCESSING TASK

In [9]:
```python
# importing the regex library

import re
```

In [10]:
```python
# function for removing the urls & special charcters

import re


def special_remove(string):
    temp = ''
    for character in string:
        if (character.isalnum() or character == ' '):
            temp += character
    return temp
```

In [11]:
```python
# performing the removal of url's & special charcaters using the url_remove & special_remove function

for i in range(len(tweet_data.columns)):
    if (tweet_data[tweet_data.columns[i]].dtype == np.object_):
        tweet_data[tweet_data.columns[i]].apply(lambda x: re.sub(
            r'^https?:\/\/.*[\r\n]*', '', x, flags=re.MULTILINE))
        tweet_data[tweet_data.columns[i]] = tweet_data[
            tweet_data.columns[i]].apply(special_remove)

# after the removal of url's & special charcaters starting few data

tweet_data.head()
```

Out[11]:

	User_Name	Name	Description (Status)	Tweets Posted	Location	No of Followers	Device Used for Tweet	Date of Tweet Creation	No of Friends
0	AdarshPallav	Adarsh Pallav		RT satishacharya Blueprint for 25 years Budget...		7	Twitter for Android	2022-02-07 07:11:43+00:00	38
1	bsindia	Business Standard	Latest news on the economy companies markets p...	BSMorningShow Will RBI go for a hike in rever...	India	2178279	TweetDeck	2022-02-07 07:11:00+00:00	430
2	bsindia	Business Standard	Latest news on the economy companies markets p...	BSMornigShow Will RBI go for a hike in revers...	India	2178279	Twitter Web App	2022-02-07 07:10:43+00:00	430
3	bsindia	Business Standard	Latest news on the economy companies markets p...	BSMorningShow Business Standards aruproytweet...	India	2178279	TweetDeck	2022-02-07 07:10:00+00:00	430
4	orfonline	ORF	Nonpartisan independent analyses on security s...	Tax concessions for cooperative societies is a...	India	106444	TweetDeck	2022-02-07 07:10:00+00:00	136

STEP 1: TOKENIZATION

In [12]:
```python
# importing the regex library

import re
```

In [38]:
```python
# installing the libraries

!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\hp\anaconda3\lib\site-packages (3.6.5)
Requirement already satisfied: click in c:\users\hp\anaconda3\lib\site-packages (from nltk) (8.0.3)
Requirement already satisfied: joblib in c:\users\hp\anaconda3\lib\site-packages (from nltk) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\hp\anaconda3\lib\site-packages (from nltk) (2021.8.
3)
Requirement already satisfied: tqdm in c:\users\hp\anaconda3\lib\site-packages (from nltk) (4.62.3)
Requirement already satisfied: colorama in c:\users\hp\anaconda3\lib\site-packages (from click->nltk) (0.4.4)
```

In [14]:
```python
# function for removing the urls & special charcters

import re


def url_remove(lst):
    for i in range(len(lst)):
        lst[i] = re.sub(r'^https?:\/\/.*[\r\n]*',
                        '',
                        lst[i],
                        flags=re.MULTILINE)
    return lst


def special_remove(lst):
    for i in range(len(lst)):
        temp = ''
        for character in lst[i]:
            if (character.isalnum()):
                temp += character
        lst[i] = temp
    return lst
```

In [15]:
```python
# importing the libraries

import nltk
```

In [16]:
```python
# we will perform toeknization for every attribute

tokenized_column_name = [
    'User_Name[Tokenized]', 'Name[Tokenized]',
    'Description (Status)[Tokenized]', 'Tweets Posted[Tokenized]',
    'Location[Tokenized]', 'No of Followers',
    'Device Used for Tweet[Tokenized]', 'Date of Tweet Creation[Tokenized]',
    'No of Friends'
]

# creating a datfame to store tokenized data of each attribute

tokenized_df = pd.DataFrame()
```

In [17]:
```python
# performing the tokenization using the nltk.word_tokenize function

for i in range(len(tokenized_column_name)):
    if (tweet_data[tweet_data.columns[i]].dtype == np.object_):
        tokenized_df[tokenized_column_name[i]] = tweet_data.apply(
            lambda row: nltk.word_tokenize(row[tweet_data.columns[i]]), axis=1)

    else:
        tokenized_df[tokenized_column_name[i]] = tweet_data[
            tweet_data.columns[i]]

# after performing tokenization starting few tokenized data

tokenized_df.head()
```

Out[17]:

	User_Name[Tokenized]	Name[Tokenized]	Description (Status) [Tokenized]	Tweets Posted[Tokenized]	Location[Tokenized]	No of Followers	Device Used for Tweet[Tokenized]	Date of Tw Creation[Tokeni
0	[AdarshPallav]	[Adarsh, Pallav]	[]	[RT, satishacharya, Blueprint, for, 25, years,...	[]	7	[Twitter, for, Android]	2022-02 07:11:43+0
1	[bsindia]	[Business, Standard]	[Latest, news, on, the, economy, companies, ma...	[BSMorningShow, Will, RBI, go, for, a, hike, i...	[India]	2178279	[TweetDeck]	2022-02 07:11:00+0
2	[bsindia]	[Business, Standard]	[Latest, news, on, the, economy, companies, ma...	[BSMornigShow, Will, RBI, go, for, a, hike, in...	[India]	2178279	[Twitter, Web, App]	2022-02 07:10:43+0
3	[bsindia]	[Business, Standard]	[Latest, news, on, the, economy, companies, ma...	[BSMorningShow, Business, Standards, aruproytw...	[India]	2178279	[TweetDeck]	2022-02 07:10:00+0
4	[orfonline]	[ORF]	[Nonpartisan, independent, analyses, on, secur...	[Tax, concessions, for, cooperative, societies...	[India]	106444	[TweetDeck]	2022-02 07:10:00+0

STEP 2: STOP WORD REMOVAL

In [18]:
```python
# we will perform stop word removal for every attribute

stop_word_removed_column_name = [
    'User_Name[Stop Word Removed]', 'Name[Stop Word Removed]',
    'Description (Status)[Stop Word Removed]',
    'Tweets Posted[Stop Word Removed]', 'Location[Stop Word Removed]',
    'No of Followers', 'Device Used for Tweet[Stop Word Removed]',
    'Date of Tweet Creation[Stop Word Removed]', 'No of Friends'
]

# creating a datfame to store stop word removed data of each attribute

stop_word_removed_df = pd.DataFrame(columns=stop_word_removed_column_name)
```

In [19]:
```python
# collecting all the 'english' language stop word

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

# stop word list

eng_stopwords
```

Out[19]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
 'its',
 'itself',
 'they',
 'them',
 'their',
 'theirs',
 'themselves',
 'what',
 'which',
 'who',
 'whom',
 'this',
 'that',
 "that'll",
 'these',
 'those',
 'am',
 'is',
 'are',
 'was',
 'were',
 'be',
 'been',
 'being',
 'have',
 'has',
 'had',
 'having',
 'do',
 'does',
 'did',
 'doing',
 'a',
 'an',
 'the',
 'and',
 'but',
 'if',
 'or',
 'because',
 'as',
 'until',
 'while',
 'of',
 'at',
 'by',
 'for',
 'with',
 'about',
 'against',
 'between',
 'into',
 'through',
 'during',
 'before',
 'after',
 'above',
 'below',
 'to',
 'from',
 'up',
 'down',
 'in',
 'out',
 'on',
 'off',
 'over',
 'under',
 'again',
 'further',
 'then',
 'once',
 'here',
 'there',
 'when',
 'where',
 'why',
 'how',
 'all',
 'any',
 'both',
 'each',
 'few',
 'more',
 'most',
 'other',
 'some',
 'such',
 'no',
 'nor',
 'not',
 'only',
 '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',
 'aren',
 "aren't",
 'couldn',
 "couldn't",
 'didn',
 "didn't",
 'doesn',
 "doesn't",
 'hadn',
 "hadn'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",
 'wasn',
 "wasn't",
 'weren',
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

In [20]:
```python
# function for removing stop word


def stop_remove(lst):
    for word in lst:
        if (word.lower() in eng_stopwords):
            lst.remove(word)
    return lst
```

In [21]:
```python
# performing the stop wprd removal using the stop_remove function

for i in range(len(stop_word_removed_column_name)):
    if (tokenized_df[tokenized_column_name[i]].dtype == np.object_):
        stop_word_removed_df[stop_word_removed_column_name[i]] = tokenized_df[
            tokenized_column_name[i]].copy(deep=True).apply(stop_remove)

    else:
        stop_word_removed_df[stop_word_removed_column_name[i]] = tokenized_df[
            tokenized_column_name[i]]

# after performing stop word removal starting few data

stop_word_removed_df.head()
```

Out[21]:

	User_Name[Stop Word Removed]	Name[Stop Word Removed]	Description (Status)[Stop Word Removed]	Tweets Posted[Stop Word Removed]	Location[Stop Word Removed]	No of Followers	Device Used for Tweet[Stop Word Removed]	Date of Tweet Creation[Stop Word Removed]	No of Friends
0	[AdarshPallav]	[Adarsh, Pallav]	[]	[RT, satishacharya, Blueprint, 25, years, Budg...	[]	7	[Twitter, Android]	2022-02-07 07:11:43+00:00	38
1	[bsindia]	[Business, Standard]	[Latest, news, the, economy, companies, market...	[BSMorningShow, RBI, go, a, hike, reverse, rep...	[India]	2178279	[TweetDeck]	2022-02-07 07:11:00+00:00	430
2	[bsindia]	[Business, Standard]	[Latest, news, the, economy, companies, market...	[BSMornigShow, RBI, go, a, hike, reverse, repo...	[India]	2178279	[Twitter, Web, App]	2022-02-07 07:10:43+00:00	430
3	[bsindia]	[Business, Standard]	[Latest, news, the, economy, companies, market...	[BSMorningShow, Business, Standards, aruproytw...	[India]	2178279	[TweetDeck]	2022-02-07 07:10:00+00:00	430
4	[orfonline]	[ORF]	[Nonpartisan, independent, analyses, security,...	[Tax, concessions, cooperative, societies, a, ...	[India]	106444	[TweetDeck]	2022-02-07 07:10:00+00:00	136

STEP3: PARTS OF SPEECH REMOVAL

Parts of Speech (POS) Tags Abbreviations Used

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP	Possessive wh-pronoun
WRB	Wh-adverb

In [22]:
```python
# list of nltk pos tags

nltk.help.upenn_tagset()
```

```
$: dollar
    $ -$ --$ A$ C$ HK$ M$ NZ$ S$ U.S.$ US$
'': closing quotation mark
    ' ''
(: opening parenthesis
    ( [ {
): closing parenthesis
    ) ] }
,: comma
    ,
--: dash
    --
.: sentence terminator
    . ! ?
:: colon or ellipsis
    : ; ...
CC: conjunction, coordinating
    & 'n and both but either et for less minus neither nor or plus so
    therefore times v. versus vs. whether yet
CD: numeral, cardinal
    mid-1890 nine-thirty forty-two one-tenth ten million 0.5 one forty-
    seven 1987 twenty '79 zero two 78-degrees eighty-four IX '60s .025
    fifteen 271,124 dozen quintillion DM2,000 ...
DT: determiner
    all an another any both del each either every half la many much nary
    neither no some such that the them these this those
EX: existential there
    there
FW: foreign word
    gemeinschaft hund ich jeux habeas Haementeria Herr K'ang-si vous
    lutihaw alai je jour objets salutaris fille quibusdam pas trop Monte
    terram fiche oui corporis ...
IN: preposition or conjunction, subordinating
    astride among uppon whether out inside pro despite on by throughout
    below within for towards near behind atop around if like until below
    next into if beside ...
JJ: adjective or numeral, ordinal
    third ill-mannered pre-war regrettable oiled calamitous first separable
    ectoplasmic battery-powered participatory fourth still-to-be-named
    multilingual multi-disciplinary ...
JJR: adjective, comparative
    bleaker braver breezier briefer brighter brisker broader bumper busier
    calmer cheaper choosier cleaner clearer closer colder commoner costlier
    cozier creamier crunchier cuter ...
JJS: adjective, superlative
    calmest cheapest choicest classiest cleanest clearest closest commonest
    corniest costliest crassest creepiest crudest cutest darkest deadliest
    dearest deepest densest dinkiest ...
LS: list item marker
    A A. B B. C C. D E F First G H I J K One SP-44001 SP-44002 SP-44005
    SP-44007 Second Third Three Two * a b c d first five four one six three
    two
MD: modal auxiliary
    can cannot could couldn't dare may might must need ought shall should
    shouldn't will would
NN: noun, common, singular or mass
    common-carrier cabbage knuckle-duster Casino afghan shed thermostat
    investment slide humour falloff slick wind hyena override subhumanity
    machinist ...
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 ...
NNPS: noun, proper, plural
    Americans Americas Amharas Amityvilles Amusements Anarcho-Syndicalists
    Andalusians Andes Andruses Angels Animals Anthony Antilles Antiques
    Apache Apaches Apocrypha ...
NNS: noun, common, plural
    undergraduates scotches bric-a-brac products bodyguards facets coasts
    divestitures storehouses designs clubs fragrances averages
    subjectivists apprehensions muses factory-jobs ...
PDT: pre-determiner
    all both half many quite such sure this
POS: genitive marker
    ' 's
PRP: pronoun, personal
    hers herself him himself hisself it itself me myself one oneself ours
    ourselves ownself self she thee theirs them themselves they thou thy us
PRP$: pronoun, possessive
    her his mine my our ours their thy your
RB: adverb
    occasionally unabatingly maddeningly adventurously professedly
    stirringly prominently technologically magisterially predominately
    swiftly fiscally pitilessly ...
RBR: adverb, comparative
    further gloomier grander graver greater grimmer harder harsher
    healthier heavier higher however larger later leaner lengthier less-
    perfectly lesser lonelier longer louder lower more ...
RBS: adverb, superlative
    best biggest bluntest earliest farthest first furthest hardest
    heartiest highest largest least less most nearest second tightest worst
RP: particle
    aboard about across along apart around aside at away back before behind
    by crop down ever fast for forth from go high i.e. in into just later
    low more off on open out over per pie raising start teeth that through
    under unto up up-pp upon whole with you
SYM: symbol
    % & ' '' ''. ) ). * + ,. < = > @ A[fj] U.S U.S.S.R * ** ***
TO: "to" as preposition or infinitive marker
    to
UH: interjection
    Goodbye Goody Gosh Wow Jeepers Jee-sus Hubba Hey Kee-reist Oops amen
    huh howdy uh dammit whammo shucks heck anyways whodunnit honey golly
    man baby diddle hush sonuvabitch ...
VB: verb, base form
    ask assemble assess assign assume atone attention avoid bake balkanize
    bank begin behold believe bend benefit bevel beware bless boil bomb
    boost brace break bring broil brush build ...
VBD: verb, past tense
    dipped pleaded swiped regummed soaked tidied convened halted registered
    cushioned exacted snubbed strode aimed adopted belied figgered
    speculated wore appreciated contemplated ...
VBG: verb, present participle or gerund
    telegraphing stirring focusing angering judging stalling lactating
    hankerin' alleging veering capping approaching traveling besieging
    encrypting interrupting erasing wincing ...
VBN: verb, past participle
    multihulled dilapidated aerosolized chaired languished panelized used
    experimented flourished imitated reunifed factored condensed sheared
    unsettled primed dubbed desired ...
VBP: verb, present tense, not 3rd person singular
    predominate wrap resort sue twist spill cure lengthen brush terminate
    appear tend stray glisten obtain comprise detest tease attract
    emphasize mold postpone sever return wag ...
VBZ: verb, present tense, 3rd person singular
    bases reconstructs marks mixes displeases seals carps weaves snatches
    slumps stretches authorizes smolders pictures emerges stockpiles
    seduces fizzes uses bolsters slaps speaks pleads ...
WDT: WH-determiner
    that what whatever which whichever
WP: WH-pronoun
    that what whatever whatsoever which who whom whosoever
WP$: WH-pronoun, possessive
    whose
WRB: Wh-adverb
    how however whence whenever where whereby whereever wherein whereof why
``: opening quotation mark
    ` ``
```

In [23]:
```python
# we will perform parts of speech (POS) tagging for every attribute

pos_tag_column_name = [
    'User_Name[POS Tagged]', 'Name[POS Tagged]',
    'Description (Status)[POS Tagged]', 'Tweets Posted[POS Tagged]',
    'Location[POS Tagged]', 'No of Followers',
    'Device Used for Tweet[POS Tagged]', 'Date of Tweet Creation[POS Tagged]',
    'No of Friends'
]

# creating a datfame to store parts of speech (POS) tagged data of each attribute

pos_tag_df = pd.DataFrame()
```

In [24]:
```python
# performing the parts of speech (POS) tagging using the nltk.pos_tag function

for i in range(len(pos_tag_column_name)):
    if (stop_word_removed_df[stop_word_removed_column_name[i]].dtype ==
            np.object_):
        pos_tag_df[pos_tag_column_name[i]] = stop_word_removed_df.apply(
            lambda row: nltk.pos_tag(row[stop_word_removed_column_name[i]]),
            axis=1)

    else:
        pos_tag_df[pos_tag_column_name[i]] = stop_word_removed_df[
            stop_word_removed_column_name[i]]

# after performing tokenization starting few tokenized data

pos_tag_df.head()
```

Out[24]:

	User_Name[POS Tagged]	Name[POS Tagged]	Description (Status)[POS Tagged]	Tweets Posted[POS Tagged]	Location[POS Tagged]	No of Followers	Device Used for Tweet[POS Tagged]	Date of Tweet Creation[POS Tagged]	No of Friends
0	[(AdarshPallav, NN)]	[(Adarsh, NNP), (Pallav, NNP)]	[]	[(RT, NNP), (satishacharya, VBD), (Blueprint, ...	[]	7	[(Twitter, NNP), (Android, NNP)]	2022-02-07 07:11:43+00:00	38
1	[(bsindia, NN)]	[(Business, NNP), (Standard, NNP)]	[(Latest, NNP), (news, NN), (the, DT), (econom...	[(BSMorningShow, NNP), (RBI, NNP), (go, VBP), ...	[(India, NNP)]	2178279	[(TweetDeck, NN)]	2022-02-07 07:11:00+00:00	430
2	[(bsindia, NN)]	[(Business, NNP), (Standard, NNP)]	[(Latest, NNP), (news, NN), (the, DT), (econom...	[(BSMornigShow, NNP), (RBI, NNP), (go, VBP), (...	[(India, NNP)]	2178279	[(Twitter, NNP), (Web, NNP), (App, NNP)]	2022-02-07 07:10:43+00:00	430
3	[(bsindia, NN)]	[(Business, NNP), (Standard, NNP)]	[(Latest, NNP), (news, NN), (the, DT), (econom...	[(BSMorningShow, NNP), (Business, NNP), (Stand...	[(India, NNP)]	2178279	[(TweetDeck, NN)]	2022-02-07 07:10:00+00:00	430
4	[(orfonline, NN)]	[(ORF, NN)]	[(Nonpartisan, NNP), (independent, JJ), (analy...	[(Tax, NNP), (concessions, NNS), (cooperative,...	[(India, NNP)]	106444	[(TweetDeck, NN)]	2022-02-07 07:10:00+00:00	136

STEP 4: STEMMING & LEMMATIZATION

In [27]:
```python
# we will perform stemming & lemmatization for every attribute

stemmed_column_name = [
    'User_Name[Stemmed]', 'Name[Stemmed]', 'Description (Status)[Stemmed]',
    'Tweets Posted[Stemmed]', 'Location[Stemmed]', 'No of Followers',
    'Device Used for Tweet[Stemmed]', 'Date of Tweet Creation[Stemmed]',
    'No of Friends'
]

lemmatized_column_name = [
    'User_Name[Lemmatized]', 'Name[Lemmatized]',
    'Description (Status)[Lemmatized]', 'Tweets Posted[Lemmatized]',
    'Location[Lemmatized]', 'No of Followers',
    'Device Used for Tweet[Lemmatized]', 'Date of Tweet Creation[Lemmatized]',
    'No of Friends'
]

# creating a datfame to store stemmed & lemmatized data of each attribute

stem_lemma_df = pd.DataFrame()
```

In [28]:
```python
# function for stemming & lemmatizing the word


def stem_word(lst):
    for i in range(len(lst)):
        lst[i] = nltk.stem.PorterStemmer().stem(lst[i])
    return lst


def lemma_word(lst):
    for i in range(len(lst)):
        lst[i] = nltk.stem.WordNetLemmatizer().lemmatize(lst[i])
    return lst
```

In [29]:

```python
"""
performing the stemming & lemmatization using the nltk.stem.PorterStemmer().stem() &

nltk.stem.WordNetLemmatizer().lemmatize() function respectively.

"""
for i in range(len(stemmed_column_name)):
    if (stop_word_removed_df[stop_word_removed_column_name[i]].dtype ==
            np.object_):
        stem_lemma_df[stemmed_column_name[i]] = stop_word_removed_df[
            stop_word_removed_column_name[i]].copy(deep=True).apply(stem_word)

        stem_lemma_df[lemmatized_column_name[i]] = stop_word_removed_df[
            stop_word_removed_column_name[i]].copy(deep=True).apply(lemma_word)

    else:
        stem_lemma_df[stemmed_column_name[i]] = tokenized_df[
            tokenized_column_name[i]]
        stem_lemma_df[lemmatized_column_name[i]] = tokenized_df[
            tokenized_column_name[i]]

# after performing tokenization starting few tokenized data

stem_lemma_df.head()
```

Out[29]:

	User_Name[Stemmed]	User_Name[Lemmatized]	Name[Stemmed]	Name[Lemmatized]	Description (Status) [Stemmed]	Description (Status) [Lemmatized]	Tweets Posted[Stemmed]	Posted[
0	[adarshpallav]	[adarshpallav]	[adarsh, pallav]	[adarsh, pallav]	[]	[]	[rt, satishacharya, blueprint, 25, year, budge...	[rt, blue
1	[bsindia]	[bsindia]	[busi, standard]	[busi, standard]	[latest, news, the, economi, compani, market, ...	[latest, news, the, economi, compani, market, ...	[bsmorningshow, rbi, go, a, hike, revers, repo...	[bsmor go,
2	[bsindia]	[bsindia]	[busi, standard]	[busi, standard]	[latest, news, the, economi, compani, market, ...	[latest, news, the, economi, compani, market, ...	[bsmornigshow, rbi, go, a, hike, revers, repo,...	[bsmo go,
3	[bsindia]	[bsindia]	[busi, standard]	[busi, standard]	[latest, news, the, economi, compani, market, ...	[latest, news, the, economi, compani, market, ...	[bsmorningshow, busi, standard, aruproytweet, ...	[bs ar
4	[orfonlin]	[orfonlin]	[orf]	[orf]	[nonpartisan, independ, analys, secur, strateg...	[nonpartisan, independ, analys, secur, strateg...	[tax, concess, cooper, societi, a, step, the, ...	[tax, co societi,

----------X------------------------------X-----------------------------
X-----------------------------X------