In [1]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [2]:

```
#!pip install contractions
!pip install emoji
!pip install nltk
```

```
Collecting emoji
  Downloading emoji-2.8.0-py2.py3-none-any.whl (358 kB)
□[?25l     □[90m□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□[0m □[32m0.0/358.9 kB□[0m □[31m?□[0m eta □[36m-:--:--□[0m□[2K     □[91m□□□□□□□[0m□[91m□□[0m
□[?25hInstalling collected packages: emoji
Successfully installed emoji-2.8.0
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
```

In [ ]:

```
import pandas as pd
import numpy as np
from nltk.tokenize import word_tokenize
#import statistics
import nltk
```

In [ ]:

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('whitespace')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Error loading whitespace: Package 'whitespace' not found
[nltk_data]     in index
```

Out [4]:

```
False
```

In [ ]:

```
import json
import pandas as pd
import spacy
import nltk
from nltk import word_tokenize
from nltk.corpus import stopwords
import string
#import contractions
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
import re
#from nltk.tokenize import TweetTokenizer
import emoji
import regex
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
import numpy as np
```

In [ ]:

```

#dft=pd.read_csv("/content/drive/MyDrive/AI-ML lab /sample_text - Sheet1.csv")
dft=pd.read_csv("")
dft
```

Out [6]:

	tweet_id	text	task1
0	1123757263427186690	hate wen females hit ah nigga with tht bro 😂😂,...	HOF
1	1123733301397733380	RT @airjunebug: When you're from the Bay but y...	HOF
2	1123734094108659712	RT @DonaldJTrumpJr: Dear Democrats: The Americ...	NOT
3	1126951188170199049	RT @SheLoveTimothy: He ain't on drugs he just ...	HOF
4	1126863510447710208	RT @TavianJordan: Summer '19 I'm coming for yo...	NOT
...	...	...	...
995	1126798721025544193	RT @prodnose: Good morning, everyone.\nFollowi...	NOT
996	1126833089190219777	@cheezitking123 this what you get for tryna ge...	NOT
997	1130037092845670400	earphones ko 😭😭😭😭😭😭😭	NOT
998	1127028455651123201	RT @nj_linguist: @realgonegirl @elivalley I th...	NOT
999	1130285076858789889	i'm tired as fuck. and man, physically ain't S...	HOF

1000 rows × 3 columns

In [ ]:

```
dft['Text_Tokenized'] = dft['text'].str.lower().apply(word_tokenize)
dft
```

Out [7]:

	tweet_id	text	task1	Text_Tokenized
0	1123757263427186690	hate wen females hit ah nigga with tht bro 😂😂,...	HOF	[hate, wen, females, hit, ah, nigga, with, tht...
1	1123733301397733380	RT @airjunebug: When you're from the Bay but y...	HOF	[rt, @, airjunebug, :, when, you, 're, from, t...
2	1123734094108659712	RT @DonaldJTrumpJr: Dear Democrats: The Americ...	NOT	[rt, @, donaldjtrumpjr, :, dear, democrats, :,...
3	1126951188170199049	RT @SheLoveTimothy: He ain’t on drugs he just ...	HOF	[rt, @, shelovetimothy, :, he, ain, ’, t, on, ...
4	1126863510447710208	RT @TavianJordan: Summer ‘19 I’m coming for yo...	NOT	[rt, @, tavianjordan, :, summer, ‘, 19, i, ’, ...
...	...	...	...	...
995	1126798721025544193	RT @prodnose: Good morning, everyone.\nFollowi...	NOT	[rt, @, prodnose, :, good, morning, ,, everyon...
996	1126833089190219777	@cheezitking123 this what you get for tryna ge...	NOT	[@, cheezitking123, this, what, you, get, for,...
997	1130037092845670400	earphones ko 😭😭😭😭😭😭😭	NOT	[earphones, ko, 😭😭😭😭😭😭😭]
998	1127028455651123201	RT @nj_linguist: @realgonegirl @elivalley I th...	NOT	[rt, @, nj_linguist, :, @, realgonegirl, @, el...
999	1130285076858789889	i’m tired as fuck. and man, physically ain’t S...	HOF	[i, ’, m, tired, as, fuck, ., and, man, ,, phy...

1000 rows × 4 columns

In [ ]:

```
#Preprocessing
ps =PorterStemmer()
lemmatiser = WordNetLemmatizer()
english_stopwords = stopwords.words('english')
exclude = set(string.punctuation)
text=dft['text']

def preprocess(text):
  #text = contractions.fix(text.lower(), slang=True)
  text=text.lower()
  text= re.sub(r'\d+', '', text)
  text=re.sub(r'$', '', text)
  text=re.sub('<.*?>','',text)
  #text=re.sub(r'\W*\b\w{1,3}\b')
  text=re.sub(r'http\S+', '', text)
  text = text.encode("ascii", "ignore")
  text = text.decode()
  #text=re.sub(r"[\\p{Cf}]", "",text)
  text = ''.join(ch for ch in text if ch not in exclude)
  tokens = word_tokenize(text)
  #print("Tokens:", tokens)
  #tokens = [lemmatiser.lemmatize(t) for t in tokens]
  #tokens=[ps.stem(t) for t in tokens]
  tokens = [t for t in tokens if t not in english_stopwords]
  tokens = [t for t in tokens if len(t) > 2]
  text = " ".join(tokens)
  return text
```

In [ ]:

```
pre_data=dft['text'][:10].apply(lambda X: preprocess(X))
pre_data
```

Out [9]:

```
0    hate wen females hit nigga tht bro tryna make ...
1    airjunebug youre bay youre really nigga heart ...
2    donaldjtrumpjr dear democrats american people ...
3           shelovetimothy aint drugs bored shit bored
4    tavianjordan summer coming boring shit beach d...
5                                hermescxbin turn shit
6    spaceboykenny know fuck bout cel shading horny...
7    polo ones thats feeeling fly fly like bitch do...
8                                    fucking love life
9              nigbmt newspaper weak bro ending pissed
Name: text, dtype: object
```

In [ ]: