# Importing Necessary Libraries

!pip install arff

import numpy as np

import pandas as pd

import arff # This is the ARFF library, which is used to create ARFF files.

import nltk

from collections import Counter

import re

!pip install liac-arff

# Reading the data

file\_path = 'english\_big.txt'

df = pd.read\_csv(file\_path, encoding='unicode\_escape',error\_bad\_lines=False, sep = None,header = None,names = ['first','second','third'])

print("Data Reading Successful")

df.head()

"""# Data Preprocessing"""

df['third'] = df['third'].replace(np.nan,' ')

df['third'].unique()

df['Text'] = df['first'] + df['second'] + df['third']

df = df[['Text']]

df.head()

df.shape

# Labelling the data based on spam or ham present in the entire text file

df['Label'] = df['Text'].str.contains('spam', case=False, regex=True)

df['Label'] = df['Label'].map({True: 'spam', False: 'ham'})

df['Text'] = df['Text'].str.replace('spam|ham', '.', case=False, regex=True)

df['Text'].iloc[0]

df.head(3)

# Creating the four different features as described in the project file

df['Character\_Count'] = df['Text'].apply(len)

currency\_symbols = re.compile(r'[$£€¥]')

df['Currency\_Symbol\_Count'] = df['Text'].apply(lambda x: len(re.findall(currency\_symbols, x)))

df['Numeric\_String\_Count'] = df['Text'].apply(lambda x: len(re.findall(r'\d+', x)))

nltk.download('punkt')

df['Most\_Popular\_Word\_Count'] = df['Text'].apply(lambda x: Counter(x.split()).most\_common(1)[0][1])

df.head(3)

df['Label'].value\_counts()

df.drop("Text",inplace = True,axis = 1)

df.columns

df

df['Label'].value\_counts()

df

df.columns

df.columns

df.to\_csv('sms\_features.CSV')

df

df.columns

df = df[['Character\_Count', 'Currency\_Symbol\_Count','Numeric\_String\_Count', 'Most\_Popular\_Word\_Count','Label']]

df.to\_csv('SMS\_Features.csv',index = False)

df.head(5)