Assignment No:4

Name: Ankitkumar Singh

Div: B

Seat no: B150694205

import re

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import CountVectorizer, TfidfTransformer

from sklearn.metrics import accuracy\_score

import warnings

warnings.filterwarnings("ignore")

def process\_tweet(tweet):

return " ".join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])", "",tweet.lower()).split())

def drop\_features(features,data):

data.drop(features,inplace=True,axis=1)

train\_data = pd.read\_csv("train\_tweets.csv")

train\_data['processed\_tweets'] = train\_data['tweet'].apply(process\_tweet)

drop\_features(['id','tweet'],train\_data)

x\_train, x\_test, y\_train, y\_test = train\_test\_split(train\_data["processed\_tweets"],train\_data["label"], test\_size = 0.2, random\_state = 42)

count\_vect = CountVectorizer(stop\_words='english')

transformer = TfidfTransformer(norm='l2',sublinear\_tf=True)

x\_train\_counts = count\_vect.fit\_transform(x\_train)

x\_train\_tfidf = transformer.fit\_transform(x\_train\_counts)

x\_test\_counts = count\_vect.transform(x\_test)

x\_test\_tfidf = transformer.transform(x\_test\_counts)

from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n\_estimators=200)

model.fit(x\_train\_tfidf,y\_train)

predictions = model.predict(x\_test\_tfidf)

print accuracy\_score(y\_test,predictions)

test\_data = pd.read\_csv('test\_tweets.csv')

test\_data['processed\_tweet'] = test\_data['tweet'].apply(process\_tweet)

drop\_features(['tweet'],test\_data)

train\_counts = count\_vect.fit\_transform(train\_data['processed\_tweets'])

test\_counts = count\_vect.transform(test\_data['processed\_tweet'])

train\_tfidf = transformer.fit\_transform(train\_counts)

test\_tfidf = transformer.transform(test\_counts)

model.fit(train\_tfidf,train\_data['label'])

predictions = model.predict(test\_tfidf)

final\_result = pd.DataFrame({'id':test\_data['id'],'label':predictions})

final\_result.to\_csv('output.csv',index=False)

/\*OUTPUT

C:\Users\HP\Desktop>python sentiment.py

0.9602690442671672 \*/