P5 p6 p7

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import csv

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

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

fields = ["review", "sentiment"]

df = pd.read\_csv('IMDB.csv', header = None, encoding='utf-8',names = fields)

df = df[:500]

cv = CountVectorizer()

x = cv.fit\_transform(df["transformed\_text"]).toarray()

y = df["sentiment"].values

#training

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

from sklearn.metrics import accuracy\_score, precision\_score

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size = 0.2,random\_state = 2)

from sklearn.naive\_bayes import GaussianNB, MultinomialNB, BernoulliNB

from sklearn.linear\_model import LogisticRegression, SGDClassifier

from sklearn.svm import SVC

encoder = LabelEncoder()

df['sentiment'] = encoder.fit\_transform(df.loc[:1500,'sentiment'])

df.head()

df["transformed\_text"] = df["review"].apply(transform\_text)

mnb = MultinomialNB()

bnb = BernoulliNB()

log\_r = LogisticRegression()

sgd = SGDClassifier()

sv = SVC()