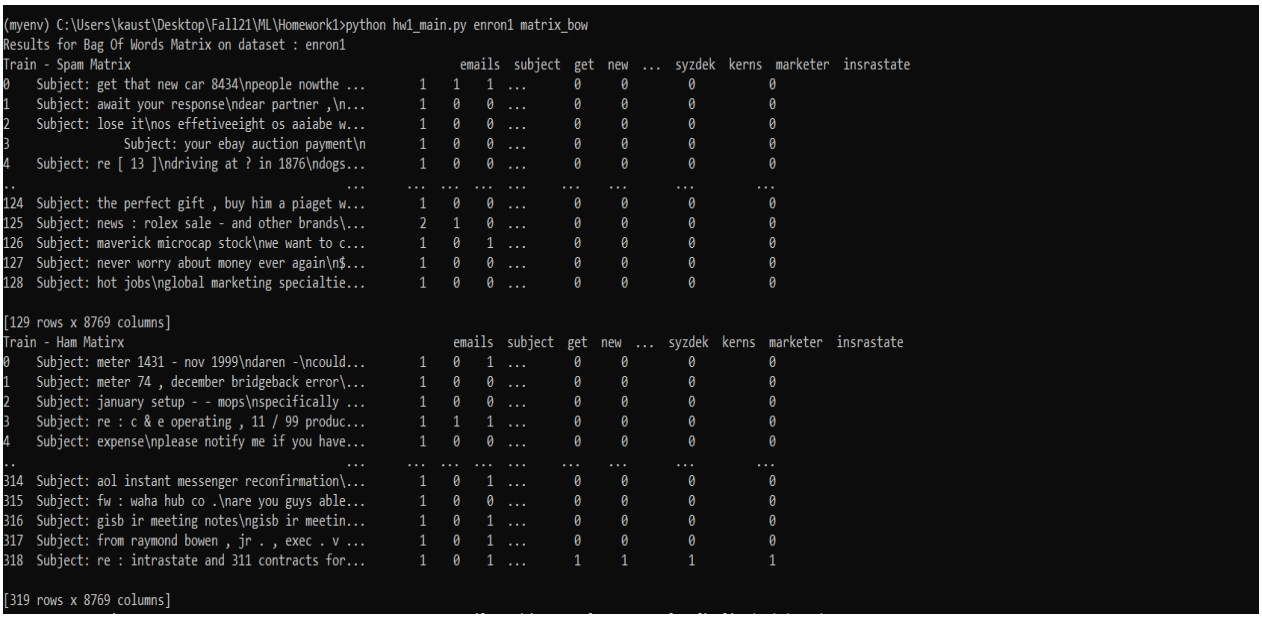
HOMEWORK 1 REPORT

Name: Kaustubh Rajendra Kulkarni

Net ID: kxk200003

In this assignment, I have implemented 4 algorithms for Spam/Ham Text Classification, namely Multinomial Naïve Bayes, Discrete Naïve Bayes, Logistic Regression and Stochastic Gradient Descent. All 4 algorithms were run on 3 different datasets. All datasets were converted to 2 models, namely bag of words model and Bernoulli model. Detailed results obtained like accuracy, precision, recall and F1 score have been mentioned in the table given below.

- Converting the text data into matrix of features X examples.
  Below snippet shows the result for bag of words representation matrix. Some files were not utf-8 encoded. I have ignored those files.



```
(myenv) C:\Users\kaust\Desktop\Fall21\ML\Homework1>python hw1_main.py enron1 matrix_bow
Results for Bag Of Words Matrix on dataset : enron1
Train - Spam Matrix                                          emails  subject  get  new  ...  syzdek  kerns  marketer  insrastate
0    Subject: get that new car 8434\npeople nowthe ...        1    1    1  ...       0      0         0          0
1    Subject: await your response\ndear partner ,\n...        1    0    0  ...       0      0         0          0
2    Subject: lose it\nos effetiveeight os aaiabe w...        1    0    0  ...       0      0         0          0
3               Subject: your ebay auction payment\n          1    0    0  ...       0      0         0          0
4    Subject: re [ 13 ]\ndriving at ? in 1876\ndogs...        1    0    0  ...       0      0         0          0
..                                                  ...      ...  ...  ...  ...     ...    ...       ...        ...
124  Subject: the perfect gift , buy him a piaget w...        1    0    0  ...       0      0         0          0
125  Subject: news : rolex sale - and other brands\...        2    1    0  ...       0      0         0          0
126  Subject: maverick microcap stock\nwe want to c...        1    0    1  ...       0      0         0          0
127  Subject: never worry about money ever again\n$...        1    0    0  ...       0      0         0          0
128  Subject: hot jobs\nglobal marketing specialtie...        1    0    0  ...       0      0         0          0

[129 rows x 8769 columns]
Train - Ham Matirx                                           emails  subject  get  new  ...  syzdek  kerns  marketer  insrastate
0    Subject: meter 1431 - nov 1999\ndaren -\ncould...        1    0    1  ...       0      0         0          0
1    Subject: meter 74 , december bridgeback error\...        1    0    0  ...       0      0         0          0
2    Subject: january setup - - mops\nspecifically ...        1    0    0  ...       0      0         0          0
3    Subject: re : c & e operating , 11 / 99 produc...        1    1    1  ...       0      0         0          0
4    Subject: expense\nplease notify me if you have...        1    0    0  ...       0      0         0          0
..                                                  ...      ...  ...  ...  ...     ...    ...       ...        ...
314  Subject: aol instant messenger reconfirmation\...        1    0    1  ...       0      0         0          0
315  Subject: fw : waha hub co .\nare you guys able...        1    0    0  ...       0      0         0          0
316  Subject: gisb ir meeting notes\ngisb ir meetin...        1    0    1  ...       0      0         0          0
317  Subject: from raymond bowen , jr . , exec . v ...        1    0    1  ...       0      0         0          0
318  Subject: re : intrastate and 311 contracts for...        1    0    1  ...       1      1         1          1

[319 rows x 8769 columns]
```

- Results for MNB on Bag of words model

Multinomial Naive Bayes	Accuracy	Precision	Recall	F1
hw1	94.33	98.11	80.62	88.51
enron1	91.85	97.41	76.87	85.93
enron4	94.21	92.53	100	96.12

- Results for DNB on Bernoulli model

Discrete Naive Bayes	Accuracy	Precision	Recall	F1
hw1	97.48	96.06	94.57	95.31
enron1	93.39	89.79	89.79	89.79
enron4	90.85	91.56	96.09	93.77

- Results for Logistic Regression on Bag of words model

Logistic Regression	Accuracy	Precision	Recall	F1
hw1	90.14	96.59	65.89	78.34
enron1	87.0	90.0	67.3	77.04
enron4	95.70	94.56	99.73	97.08

- Results for Logistic Regression on Bernoulli model

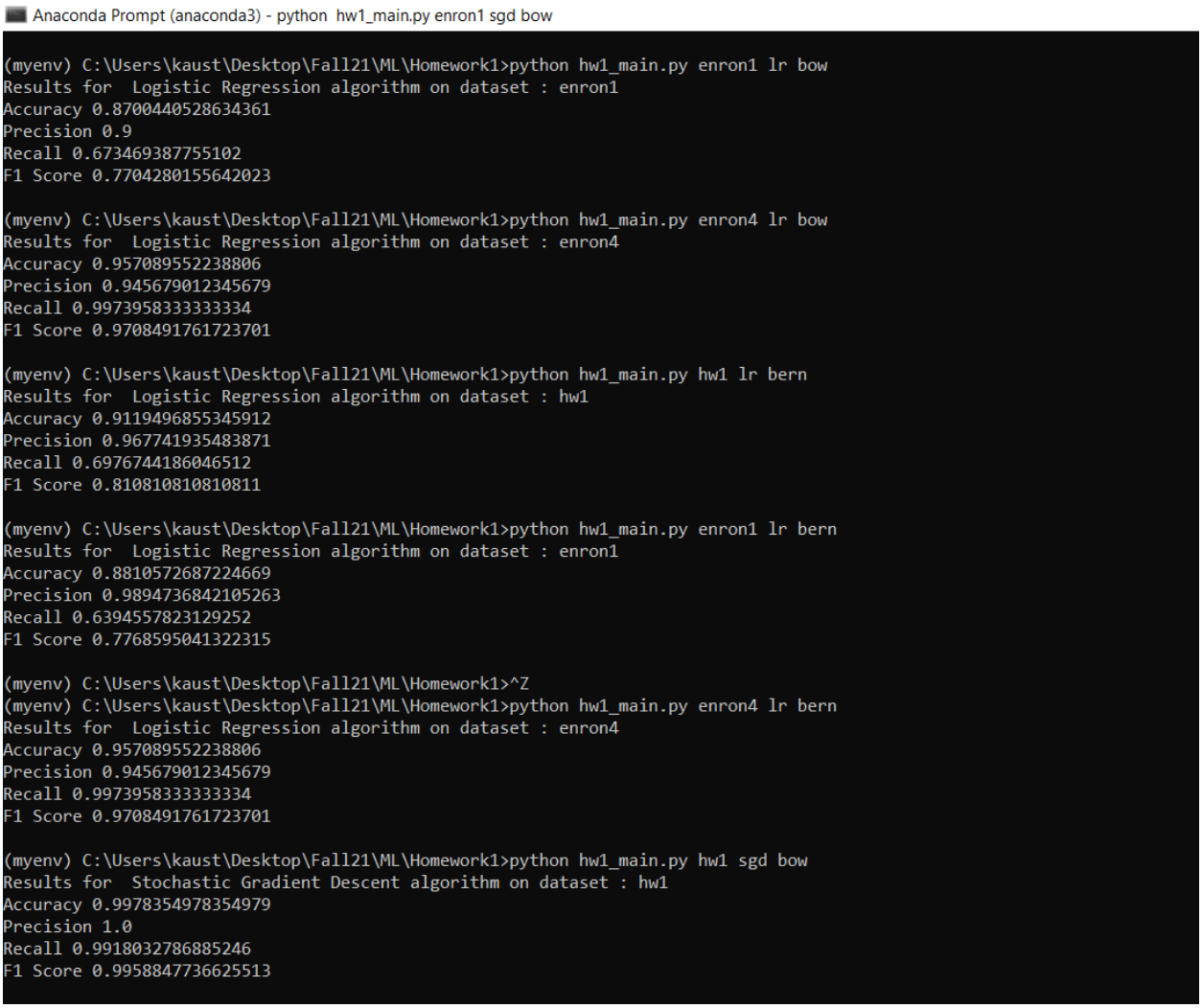
Logistic Regression	Accuracy	Precision	Recall	F1
hw1	91.19	96.77	69.76	81.08
enron1	88.10	98.94	63.94	77.65
enron4	95.70	94.56	99.73	97.08

- Results for SGD on Bag of words model

SGD Classifier	Accuracy	Precision	Recall	F1
hw1	99.7	1	99.18	99.5
enron1	1	1	1	1
enron4	99.81	99.74	1	99.87

- Results for SGD on Bernoulli

SGD Classifier	Accuracy	Precision	Recall	F1
hw1	1	1	1	1
enron1	1	1	1	1
enron4	1	1	1	1



Parameter Tuning for Logistic Regression:

For finding the best value of lambda, I performed the following steps.

- First, divide the training data into train data and validation data.
- Set learning rate eta to 0.01
- Set initial lambda value to 2.
- For lambda values ranging 1 to 10 with an increment of 2, calculate the model weights using the learning rate and weights of each feature in the validation data. Train the model for 25 iterations.
    - For each document in validation data, calculate the conditional log likelihood using the above model weights.
    - If the conditional log likelihood is maximized, set the corresponding lambda value as the final lambda parameter.

Parameter Tuning for SGD Classifier:

For parameter tuning for SGD Classifier, I performed the following steps.

- Alpha value set to 0.01 or 0.05
- Max_iter ranges from 500 to 3000
- Learning rate is optimal, invscaling and adaptive, which calculates the value of eta0
- Eta0, the initial learning rate is 0.03 or 0.07
- The stopping criteria tol is set to 0.001 or 0.005
- All these hyper parameters are fed to GridSearchCV, and the classifier is fitted on the validation data. The classifier is then used to train the model on the training data, and the predictions is done on test data.

Answer the following questions:

1. Which data representation and algorithm combination yields the best performance (measured in terms of the accuracy, precision, recall and F1 score) and why?

As observed from the above results, clearly Stochastic Gradient Descent Algorithm on Bernoulli Model and the Bag of Words model performed the best and gave the best accuracy. SGD performs the best since hyper parameter optimization is used, and it works better than logistic regression on large data. Also, one major advantage of Stochastic Gradient Descent is that it performs calculations faster that Gradient Descent/Ascent used in Logistic Regression.

2. Does Multinomial Naive Bayes perform better (again performance is measured in terms of the accuracy, precision, recall and F1 score) than LR and SGDClassifer on the Bag of words representation? Explain your yes/no answer.

Multinomial Naive Bayes performed better than LR for smaller datasets (hw1 and enron1). For large dataset enron4, LR performed better than MNB. SGD outperformed both LR and MNB for bag of words.

3. Does Discrete Naive Bayes perform better (again performance is measured in terms of the accuracy, precision, recall and F1 score) than LR and SGDClassifer on the Bernoulli representation? Explain your yes/no answer.

Discrete Naive Bayes performed better than LR on Bernoulli model. Naive Bayes assumes that the features are conditionally independent. So, it performs better on the dataset. LR predicts probability using a function form while Naïve Bayes figures out how data was generated given the results, which works out better on the Bernoulli model.

4. Does your LR implementation outperform the SGDClassifer (again performance is measured in terms of the accuracy, precision, recall and F1 score) or is the difference in

performance minor? Explain your yes/no answer.

No, SGD outperforms LR in terms of accuracy, precision, recall and F1 score. By default, SGD does not perform well, but after hyper parameter tuning and setting appropriate learning rates and number of iterations, SGD gave the best results. The difference is performance is large, and as the size of the dataset increases, SGD will continue to outperform LR, as observed in the enron4 dataset.