// Vaishakhi Kulkarni - vpk140230

Part A: **Perceptron DataSet1:**

Sr No	Iterations	Learning Rate	Accuracy without Stop Words	Accuracy with Stop Words
1	25	0.001	Ham: 76.14943 Spam: 88.46153	Ham: 76.43678 Spam:91.53846
2	25	0.09	Ham: 68.10345 Spam: 95.38461	Ham:79.02299 Spam: 90.76923
3	25	0.01	Ham: 77.29885 Spam: 96.15385	Ham:68.10345 Spam: 94.61538
4	25	0.1	Ham: 80.74713 Spam: 94.61538	Ham:58.908047 Spam: 96.92308
5	25	0.7	Ham: 87.93104 Spam: 92.30769	Ham: 88.21839 Spam:90.0
6	50	0.001	Ham: 76.72414 Spam: 86.15385	Ham: 77.58621 Spam: 88.46153
7	50	0.09	Ham: 94.25287 Spam: 87.69231	Ham: 93.965515 Spam: 86.15385
8	50	0.01	Ham: 93.10345 Spam: 88.46153	Ham: 95.97701 Spam: 85.38461
9	50	0.1	Ham: 94.25287 Spam: 88.46153	Ham: 94.54023 Spam: 88.46153
10	50	0.7	Ham: 93.965515 Spam: 90.76923	Ham: 96.264366 Spam: 85.38461
11	75	0.001	Ham: 78.44828 Spam: 84.61539	Ham: 79.5977 Spam: 87.69231
12	75	0.09	Ham: 93.965515 Spam: 84.61539	Ham:95.68965 Spam: 85.38461
13	75	0.01	Ham: 93.10345 Spam: 89.23077	Ham:94.54023 Spam: 85.38461
14	75	0.1	Ham: 93.10345 Spam:85.38461	Ham:95.97701 Spam: 84.61539
15	75	0.7	Ham: 93.39081 Spam: 86.15385	Ham: 95.4023 Spam: 91.53846
16	100	0.001	Ham: 88.505745 Spam: 89.23077	Ham:89.36782 Spam: 92.30769
17	100	0.09	Ham: 94.25287 Spam: 87.69231	Ham: 93.10345 Spam: 88.46153
18	100	0.01	Ham: 88.21839 Spam: 93.07692	Ham: 89.655174 Spam: 92.30769
19	100	0.1	Ham: 91.954025 Spam: 92.30769	Ham: 95.4023 Spam:86.92307
20	100	0.7	Ham: 93.39081 Spam: 90.76923	Ham: 95.11494 Spam: 86.92307

Perceptron Data Set2:Enron1

Sr No	Iterations	Learning Rate	Accuracy without Stop Words	Accuracy with Stop Words
1	25	0.001	Ham: 69.3811 Spam: 84.56376	Ham: 67.100975 Spam: 76.51006
2	25	0.09	Ham: 71.00977 Spam: 100.0	Ham:87.29642 Spam: 96.644295
3	25	0.01	Ham:81.43323 Spam:95.30202	Ham:79.1531 Spam:95.30202
4	25	0.1	Ham:75.89577 Spam:97.31544	Ham:84.69055 Spam:96.644295
5	25	0.7	Ham:80.78176 Spam:97.98658	Ham:85.01629 Spam:95.97315
6	50	0.001	Ham:77.52443 Spam:89.93289	Ham:74.91857 Spam:85.2349
7	50	0.09	Ham:90.55375 Spam:94.630875	Ham:91.856674 Spam:93.95973
8	50	0.01	Ham:88.92508 Spam:89.26175	Ham:89.576546 Spam:93.95973
9	50	0.1	Ham:91.20521 Spam:93.95973	Ham:91.530945 Spam:93.28859
10	50	0.7	Ham:91.856674 Spam:92.617455	Ham:91.530945 Spam:95.30202
11	75	0.001	Ham:78.175896 Spam:90.60403	Ham:78.50163 Spam:85.2349
12	75	0.09	Ham:89.90228 Spam:95.30202	Ham:90.87948 Spam:97.31544
13	75	0.01	Ham:89.576546 Spam:94.630875	Ham:90.22801 Spam:97.31544
14	75	0.1	Ham:90.22801 Spam:94.630875	Ham:90.55375 Spam:97.31544
15	75	0.7	Ham:90.55375 Spam:93.28859	Ham:91.530945 Spam:95.97315
16	100	0.001	Ham:84.69055 Spam:92.617455	Ham:86.31922 Spam:92.617455
17	100	0.09	Ham:88.92508 Spam:95.97315	Ham:92.50814 Spam:95.30202
18	100	0.01	Ham:88.27361 Spam:93.95973	Ham:90.87948 Spam:95.30202
19	100	0.1	Ham:90.22801 Spam:95.30202	Ham:91.856674 Spam:93.28859
20	100	0.7	Ham:89.25082 Spam:96.644295	Ham:92.18241 Spam:92.617455

Perceptron Data Set3:Enron4

Sr No	Iterations	Learning Rate	Accuracy without Stop Words	Accuracy with Stop Words
1	25	0.001	Ham: 65.789474 Spam: 87.723785	Ham: 57.236843 Spam: 84.14322
2	25	0.09	Ham: 92.10526 Spam: 93.60614	Ham:85.52631 Spam: 96.67519
3	25	0.01	Ham: 70.39474 Spam: 97.1867	Ham:60.526318 Spam: 98.46547
4	25	0.1	Ham: 91.44737 Spam: 93.09463	Ham: 87.5 Spam: 95.90793
5	25	0.7	Ham: 88.81579 Spam: 94.117645	Ham:86.84211 Spam: 95.396416
6	50	0.001	Ham: 63.815792 Spam: 95.652176	Ham: 56.578945 Spam: 96.16368
7	50	0.09	Ham: 91.44737 Spam: 93.09463	Ham: 87.5 Spam: 96.16368
8	50	0.01	Ham: 90.789474 Spam: 92.838875	Ham:86.18421 Spam: 97.69821
9	50	0.1	Ham: 90.789474 Spam: 94.117645	Ham:88.15789 Spam: 96.93095
10	50	0.7	Ham: 88.81579 Spam: 94.117645	Ham: 86.84211 Spam: 95.396416
11	75	0.001	Ham: 69.07895 Spam: 95.14066	Ham: 58.552628 Spam: 96.16368
12	75	0.09	Ham: 94.07895 Spam: 92.32737	Ham:86.18421 Spam: 96.93095
13	75	0.01	Ham: 86.18421 Spam: 95.14066	Ham:87.5 Spam: 97.953964
14	75	0.1	Ham: 92.76315 Spam: 94.117645	Ham: 88.15789 Spam: 96.41943
15	75	0.7	Ham: 92.76315 Spam: 91.81586	Ham:87.5 Spam: 94.62916
16	100	0.001	Ham: 74.34211 Spam: 96.16368	Ham: 63.815792 Spam: 97.1867
17	100	0.09	Ham: 90.131584 Spam: 93.35038	Ham:88.15789 Spam: 96.41943
18	100	0.01	Ham: 86.18421 Spam: 96.41943	Ham: 79.60526 Spam: 98.46547
19	100	0.1	Ham: 90.131584 Spam: 94.117645	Ham: 87.5 Spam: 97.44245
20	100	0.7	Ham: 91.44737 Spam: 92.838875	Ham:85.52631 Spam: 97.44245

Naive Bays :

Data Set 1:

Iterations: 100

Eta:0.01

Lambda:0.1

```
Accuracy of Ham without stopwords :96.55172413793103
Accuracy of Ham with stopwords :95.97701149425288

Accuracy of Spam without stopwords :97.6923076923077
Accuracy of Spam with stopwords :98.46153846153847
```

Dataset 2: Enron1

```
Iteration:100

Eta:0.01

Lambda:0.1

Accuracy of Ham without stopwords :96.74267100977198
Accuracy of Ham with stopwords :95.43973941368078

Accuracy of Spam without stopwords :93.95973154362416
Accuracy of Spam with stopwords :95.30201342281879
```

Data Set 3: Enron4

```
Iterations:100

Eta:0.01

Lambda:0.1

Accuracy of Ham without stopwords :90.13157894736842
Accuracy of Ham with stopwords :96.71052631578948

Accuracy of Spam without stopwords :86.95652173913044
Accuracy of Spam with stopwords :84.91048593350384
```

Logistic Regression:

Using Weka:

Data Set1 :

1) Learning rate: 0.01

Iterations :100

Eta:0.1

```
Correctly Classified Instances         436               91.2134 %
Incorrectly Classified Instances        42                8.7866 %
```

2) Learning rate: 0.1

Iteration :100

Eta:0.1

```
Correctly Classified Instances         442               92.4686 %
Incorrectly Classified Instances        36                7.5314 %
```

3) Learning rate: 0.1

Iteration :50

Eta:0.1

```
Correctly Classified Instances         454               94.9791 %
Incorrectly Classified Instances        24                5.0209 %
```

Data Set2 :Enron1

1) Learning rate: 0.01

Iteration :100

```
Correctly Classified Instances         426               93.4211 %
Incorrectly Classified Instances        30                6.5789 %
```

2) Learning rate: 0.1

Iteration :100

```
Correctly Classified Instances         430               94.2982 %

Incorrectly Classified Instances        26                5.7018 %
```

3) Learning rate: 0.1

Iteration :50

```
Correctly Classified Instances         431               94.5175 %

Incorrectly Classified Instances        25                5.4825 %
```

Data Set 3:Enron4

1) Learning rate: 0.01

Iterations :100

Eta:0.1

```
Correctly Classified Instances         525               96.6851 %

Incorrectly Classified Instances        18                3.3149 %
```

2) Learning rate: 0.1

Iteration :100

Eta:0.1

```
Correctly Classified Instances         526               96.8692 %

Incorrectly Classified Instances        17                3.1308 %
```

3) Learning rate: 0.1

Iteration :50

Eta:0.1

```
Correctly Classified Instances         526               96.8692 %
```

```
Incorrectly Classified Instances       17              3.1308 %
```

Perceptron Vs (Naive Bayes and Logistic Regression):

Accuracy by Perceptron Classifier is less when compared to accuracy by Naive Bayes and Logistic Regression classifier. In Perceptron we are using Perceptron training rule which will not converge if the data is not linearly separable, resulting in lesser accuracy. While in Naive Bayes and Logistic Regression , accuracy is calculated using the same data set values. All this analysis is done after comparing the reading which we are calculated. DataSet3 has highest accuracy after calculating using Logistic regression classifier as compared to other dataset. In Perceptron we can concluded that as the iterations are increased the accuracy has also increased in all DataSet. While as the learning rate value is very small we are getting High accuracy. To conclude, should have more iterations and small learning rate value then we will get highest accuracy.

Effects of StopWords:

Accuracy for DataSet 1 increases for some reading while decreases slightly for other reading. Accuracy for DataSet2 decreases for Spam while increases for Ham files. While in Dataset3 accuracy for Ham decreases and for Spam increases. At some places accuracy for Ham and Spam have increased slightly. This is because of stop words distribution on datasets. Ideally the accuracy should have increased but due to uneven distribution of stopwords on the given dataset we are getting varied results.

While in Naive Bayes due to stopwords filtering the accuracy for DataSet 1 and 2 is increasing for Spam and decreasing for Ham. On DataSet3 Ham is increasing and for Spam it is decreasing.

Part B:

Command to convert to ARFF format:

```
C:\>java -cp "C:\Program Files\Weka-3-6\weka.jar" weka.core.converters.TextDirec

toryLoader -dir C:\Users\Vaishakhi\Documents\Vibhav\Assignment3\enron1_train\enr

on1\train >C:\Users\Vaishakhi\Documents\Vibhav\Assignment3\enron1_train\enron1\t

rainn.arff

C:\>java -cp "C:\Program Files\Weka-3-6\weka.jar" weka.core.converters.TextDirec

toryLoader -dir C:\Users\Vaishakhi\Documents\Vibhav\Assignment3\enron1_train\enr

on1\test >C:\Users\Vaishakhi\Documents\Vibhav\Assignment3\enron1_train\enron1\te
```

```
st.arff

C:\>java -cp "C:\Program Files\Weka-3-6\weka.jar" weka.filters.unsupervised.attr
ibute.StringToWordVector -b -i C:\Users\Vaishakhi\Documents\Vibhav\Assignment3\e
nron1_train\enron1\trainn.arff -o C:\Users\Vaishakhi\Documents\Vibhav\Assignment
3\enron1_train\enron1\finaloutput.arff -r C:\Users\Vaishakhi\Documents\Vibhav\As
signment3\enron1_train\enron1\test.arff -s C:\Users\Vaishakhi\Documents\Vibhav\A
ssignment3\enron1_train\enron1\finaloutputtest.arff
```

<u>Neural Network :</u>

Data Set 1:

Change in Learning Rate

1.Hidden Layer:1

Momentum :0.1

Learning Rate:0.001

```
Correctly Classified Instances         448               93.7238 %
Incorrectly Classified Instances        30                6.2762 %
```

2.Hidden Layer:1

Momentum:0.1

Learning Rate:0.01

```
Correctly Classified Instances         411               85.9833 %
Incorrectly Classified Instances        67               14.0167 %
```

3.Hidden Layer:1

Momentum:0.1

Learning Rate:0.1

```
Correctly Classified Instances         437               91.4226 %
```

Incorrectly Classified Instances 41 8.5774 %

4.Hidden Layer:1

Momentum:0.1

Learning Rate:0.7

Correctly Classified Instances 349 73.0126 %

Incorrectly Classified Instances 129 26.9874 %

Change in Momentum

1.Hidden Layer:1

Momentum :0.004

Learning Rate:0.01

Correctly Classified Instances 444 92.887 %

Incorrectly Classified Instances 34 7.113 %

2.Hidden Layer:1

Momentum:0.01

Learning Rate:0.01

Correctly Classified Instances 444 92.887 %

Incorrectly Classified Instances 34 7.113 %

3.Hidden Layer:1

Momentum:0.9

Learning Rate:0.01

Correctly Classified Instances 454 94.9791 %

Incorrectly Classified Instances 24 5.0209 %

4.Hidden Layer:1

Momentum:1.0

Learning Rate:0.01

```
Correctly Classified Instances         348               72.8033 %
Incorrectly Classified Instances       130               27.1967 %
```

Change in Hidden Layer:

1.Hidden Layer:2,3

Momentum :0.1

Learning Rate:0.01

```
Correctly Classified Instances         446               93.3054 %
Incorrectly Classified Instances        32                6.6946 %
```

2.Hidden Layer:1,1,1

Momentum:0.01

Learning Rate:0.01

```
Correctly Classified Instances         348               72.8033 %
Incorrectly Classified Instances       130               27.1967 %
```

3.Hidden Layer:1,2,3,4

Momentum:0.09

Learning Rate:0.01

```
Correctly Classified Instances         348               72.8033 %
Incorrectly Classified Instances       130               27.1967 %
```

4.Hidden Layer:3,3

Momentum:0.1

Learning Rate:0.01

Correctly Classified Instances 447 93.5146 %

Incorrectly Classified Instances 31 6.4854 %

Data Set 2:

Change in Learning Rate

1.Hidden Layer:1

Momentum :0.1

Learning Rate:0.001

Correctly Classified Instances 430 94.2982 %

Incorrectly Classified Instances 26 5.7018 %

2.Hidden Layer:1

Momentum:0.1

Learning Rate:0.01

Correctly Classified Instances 430 94.2982 %

Incorrectly Classified Instances 26 5.7018 %

3.Hidden Layer:1

Momentum:0.1

Learning Rate:0.1

Correctly Classified Instances 433 94.9561 %

Incorrectly Classified Instances 23 5.0439 %

4.Hidden Layer:1

Momentum:0.1

Learning Rate:0.7

```
Correctly Classified Instances         307               67.3246 %
Incorrectly Classified Instances       149               32.6754 %
```

Change in Momentum:

1.Hidden Layer:1

Momentum :0.004

Learning Rate:0.01

```
Correctly Classified Instances         430               94.2982 %
Incorrectly Classified Instances        26                5.7018 %
```

2.Hidden Layer:1

Momentum:0.01

Learning Rate:0.01

```
Correctly Classified Instances         430               94.2982 %
Incorrectly Classified Instances        26                5.7018 %
```

3.Hidden Layer:1

Momentum:0.9

Learning Rate:0.01

```
Correctly Classified Instances         435               95.3947 %
Incorrectly Classified Instances        21                4.6053 %
```

4.Hidden Layer:1

Momentum:1.0

Learning Rate:0.01

```
Correctly Classified Instances         307               67.3246 %
Incorrectly Classified Instances       149               32.6754 %
```

Change in Hidden Layer

1.Hidden Layer:2,3

Momentum :0.1

Learning Rate:0.01

```
Correctly Classified Instances        428              93.8596 %
Incorrectly Classified Instances       28               6.1404 %
```

2.Hidden Layer:1,1,1

Momentum:0.01

Learning Rate:0.01

```
Correctly Classified Instances        307              67.3246 %
Incorrectly Classified Instances       149              32.6754 %
```

3.Hidden Layer:1,2,3,4

Momentum:0.09

Learning Rate:0.01

```
Correctly Classified Instances        307              67.3246 %
Incorrectly Classified Instances       149              32.6754 %
```

4.Hidden Layer:3,3

Momentum:0.1

Learning Rate:0.01

```
Correctly Classified Instances        430              94.2982 %
Incorrectly Classified Instances       26               5.7018 %
```

5. Hidden Layer:1

Momentum:0.1

Learning Rate:0.01

Nodes:5

```
Correctly Classified Instances        430              94.2982 %
Incorrectly Classified Instances       26               5.7018 %
```

Data Set 3:

Change in Learning Rate

1.Hidden Layer:1

Momentum :0.1

Learning Rate:0.001

```
Correctly Classified Instances        527              97.0534 %
Incorrectly Classified Instances       16               2.9466 %
```

2.Hidden Layer:1

Momentum:0.1

Learning Rate:0.01

```
Correctly Classified Instances        526              96.8692 %
Incorrectly Classified Instances       17               3.1308 %
```

3.Hidden Layer:1

Momentum:0.1

Learning Rate:0.1

```
Correctly Classified Instances        523              96.3168 %
Incorrectly Classified Instances       20               3.6832 %
```

4.Hidden Layer:1

Momentum:0.1

Learning Rate:0.7

Correctly Classified Instances 391 72.0074 %

Incorrectly Classified Instances 152 27.9926 %

Change in Momentum

1.Hidden Layer:1

Momentum :0.004

Learning Rate:0.01

Correctly Classified Instances 527 97.0534 %

Incorrectly Classified Instances 16 2.9466 %

2.Hidden Layer:1

Momentum:0.01

Learning Rate:0.01

Correctly Classified Instances 527 97.0534 %

Incorrectly Classified Instances 16 2.9466 %

3.Hidden Layer:1

Momentum:0.09

Learning Rate:0.01

Correctly Classified Instances 527 97.0534 %

Incorrectly Classified Instances 16 2.9466 %

4.Hidden Layer:1

Momentum:1.0

Learning Rate:0.01

Correctly Classified Instances 391 72.0074 %

Incorrectly Classified Instances 152 27.9926 %

Change in Hidden Layer

1.Hidden Layer:2,3

Momentum :0.1

Learning Rate:0.01

Correctly Classified Instances 528 97.2376 %

Incorrectly Classified Instances 15 2.7624 %

2.Hidden Layer:1,1,1

Momentum:0.01

Learning Rate:0.01

Correctly Classified Instances 391 72.0074 %

Incorrectly Classified Instances 152 27.9926 %

3.Hidden Layer:1,2,3,4

Momentum:0.09

Learning Rate:0.01

Correctly Classified Instances 391 72.0074 %

Incorrectly Classified Instances 152 27.9926 %

4.Hidden Layer:3,3

Momentum:0.09

Learning Rate:0.01

Correctly Classified Instances 528 97.2376 %

Incorrectly Classified Instances 15 2.7624 %

5. Hidden Layer:1

Momentum:0.1

Learning Rate:0.01

Nodes:5

```
Correctly Classified Instances        526              96.8692 %

Incorrectly Classified Instances       17               3.1308 %
```

Neural Network Analysis:

1.When learning rate is between 0.001 to 0.1 then the accuracy is maximum in all the dataset while accuracy drastically decreases when learning rate is more than 0.1 in all dataset.

2. When momentum is between 0.004 to 0.01 then the accuracy is maximum for all dataset while accuracy drastically decreases when momentum is more than 1.0 in all dataset.

3.When hidden layer is single layer the accuracy is high in all dataset. Also for 2 layer the accuracy is high. While for 3 layer the accuracy drastically decreases in all dataset.

In general for all dataset accuracy increases as number of epochs increases.

Part C:

Collaborative Filtering:

Dataset contains 1821 movies and 28978 users in all. The training set has 3.25 million ratings while test has 100000 records. We are using training set to predict the ratings provided in the test set. So it is taking more than 10hrs to run on whole data set. As we need to predict rating for every user from test set by learning from training set.

Mean Absolute Error formula:

Used to measure how close forecast or prediction are to eventual outcomes.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i| .$$

Where: f(i)=predicted weight; y(i)=actual or true weight

Root Mean Square Error formula:

Measure of the differences between value (Sample and population values) predicted by a model or an estimator and the values actually observed.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y)^2}{n}}.$$

Where: Y^(t)=predicted weight; Y= actual weight

K=normalizing factor used during the predication of weight is calculated by using following formula:

$$k = 1 / \sum_{u' \in U} |\,\mathrm{simil}(u, u')|$$

It's the reciprocal of summation of the summation of similarity/weight between two users.

Results on Netflix Dataset:

Mean Square Error	0.09885977203407334
Root Mean Absolute Error	0.11931616505703319

Reference:

http://en.wikipedia.org/wiki/Root-mean-square_deviation

http://en.wikipedia.org/wiki/Mean_absolute_error

http://en.wikipedia.org/wiki/Collaborative_filtering