Machine Learning
Report of Assignment - 4
George Koshy - gxk140830

Question:

You will use the 20 Newsgroups dataset. It contains newsgroup documents relating to 20 different topics. It can be downloaded from: http://qwone.com/~jason/20Newsgroups/ There are 3 different versions available for download. It is recommended that you use the "bydate" version. It has documents split up into training and testing sets. You can limit yourself to any 5 topics that are most interesting to you. The topics can be inferred from the names of the directories.

Language used:
Python

Libraries Used:
Latest dev version of sklearn (Latest dev version 0.18DEV)

IDE used:
Jupyter

The support is the number of occurrences of each class in y_true

Comparison table of the classifiers:

```
The accuracy after running Stochastic Gradient Descent Algorithm is 0.908413

The table metrics for Stochastic Gradient Descent Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.92      0.77      0.84       319
         comp.graphics       0.85      0.98      0.91       389
               sci.med       0.95      0.87      0.91       396
soc.religion.christian       0.88      0.95      0.91       398
 talk.politics.mideast       0.96      0.95      0.95       376

           avg / total       0.91      0.91      0.91      1878

************************************************************
The accuracy after running Support Vector Classifier Algorithm is 0.885517

The table metrics for Support Vector Classifier Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.83      0.77      0.80       319
         comp.graphics       0.88      0.96      0.92       389
               sci.med       0.90      0.89      0.90       396
soc.religion.christian       0.85      0.93      0.89       398
 talk.politics.mideast       0.98      0.85      0.91       376

           avg / total       0.89      0.89      0.88      1878

************************************************************
```

```
****************************************************************
The accuracy after running Multinomial Naive Bayes Algorithm is 0.849308

The table metrics for Multinomial Naive Bayes Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.97      0.59      0.73       319
         comp.graphics       0.97      0.89      0.93       389
               sci.med       0.97      0.82      0.89       396
soc.religion.christian       0.63      0.99      0.77       398
 talk.politics.mideast       0.95      0.92      0.93       376

           avg / total       0.89      0.85      0.85      1878

****************************************************************
The accuracy after running Random Forest Algorithm is 0.823216

The table metrics for Random Forest Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.88      0.66      0.76       319
         comp.graphics       0.70      0.97      0.81       389
               sci.med       0.86      0.69      0.77       396
soc.religion.christian       0.82      0.93      0.87       398
 talk.politics.mideast       0.96      0.83      0.89       376

           avg / total       0.84      0.82      0.82      1878

****************************************************************
The accuracy after running Passive Aggressive Classifier Algorithm is 0.923855

The table metrics for Passive Aggressive Classifier Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.90      0.84      0.87       319
         comp.graphics       0.92      0.97      0.94       389
               sci.med       0.95      0.91      0.93       396
soc.religion.christian       0.89      0.97      0.92       398
 talk.politics.mideast       0.97      0.91      0.94       376

           avg / total       0.93      0.92      0.92      1878

****************************************************************
The accuracy after running Decision Tree Algorithm is 0.685304

The table metrics for Decision Tree Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.59      0.67      0.63       319
         comp.graphics       0.71      0.78      0.74       389
               sci.med       0.59      0.54      0.56       396
soc.religion.christian       0.74      0.77      0.75       398
 talk.politics.mideast       0.83      0.66      0.73       376

           avg / total       0.69      0.69      0.69      1878

****************************************************************
```

```
************************************************************
The accuracy after running Nearest Centroid Algorithm is 0.722577

The table metrics for Nearest Centroid Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.87      0.56      0.68       319
         comp.graphics       0.56      0.93      0.70       389
               sci.med       0.81      0.58      0.68       396
soc.religion.christian       0.68      0.77      0.72       398
 talk.politics.mideast       0.94      0.74      0.83       376

           avg / total       0.77      0.72      0.72      1878

************************************************************
The accuracy after running K nearest neighbor Algorithm is 0.748669

The table metrics for K nearest neighbor Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.56      0.83      0.67       319
         comp.graphics       0.93      0.77      0.84       389
               sci.med       0.89      0.51      0.65       396
soc.religion.christian       0.74      0.83      0.78       398
 talk.politics.mideast       0.76      0.82      0.79       376

           avg / total       0.78      0.75      0.75      1878

************************************************************
The accuracy after running Perceptron Algorithm is 0.897764

The table metrics for Perceptron Algorithm is
                        precision    recall  f1-score   support

           alt.atheism       0.86      0.81      0.84       319
         comp.graphics       0.93      0.94      0.93       389
               sci.med       0.94      0.89      0.91       396
soc.religion.christian       0.85      0.96      0.90       398
 talk.politics.mideast       0.92      0.86      0.89       376

           avg / total       0.90      0.90      0.90      1878

************************************************************
```

Observations:

Please look at the **avg/total** row in the tables for the appropriate metrics as precision, recall and f1- score respectively.

The **support** is the number of occurrences of each class in y_true.

We observe that Stochastic Gradient Descent, SVM, Multinomial Naive Bayes, Random forest and Perceptron give good results for text document classification.

github link of the assignment 4:

https://github.com/gkoshyk/MachineLearning/tree/master/assignment4