Read Me

The script program is created for the purpose of training a labeled data set for creating a machine

learning model that predicts the type of the document from a seen or unseen data point.

Setting up:

1.download the labeled data file, and the script into the destination directory
note: do not change the file name

Launch the program from command line:

1. set the current directory ‘cd: ~’
2. ‘python3 classifier.py’
"Enter labeled data file path: “ (eg. Enter labeled data file path: /Users/jbi/Downloads/)
“mode=” (eg. mode=predict)
“input=” (eg. input=3,1350,9450,Mrs. Jerri Larsen,turkey-kingstown-20190920-press, 1982-10-13T20:54:49.000Z)

Note: when entering unknown data point, please make sure that the features comma separated.

Output:
mode=train

outputs: list of validation accuracy scores on the train set and on the test set

mode=predict

output: the result of prediction eg. [‘other’]

Launch the program from an Editor:

Note: by launching the program from the editor, the attributes and the methods will be accessible, the classifier can thus be changed.

Classifier:

RandomForest is the chosen classifier. **Naive Bayes** with Gaussian distribution has been trained and compared with RandomForest. RF yields better results.

The classifier can be changed in the source code.
1. open script
2. go to ‘def main()’
3. instantiate Naive Bayes instead of RandomForest

Test Author:
Julian Seiderburg

Test Data source:

Mcarthy Finch

Testee:
Jianhui Bi

Script author:
Jianhui Bi