Instructions:

This project contains the implementation of ANN, SVM and Naive Bayes for OCR. Each algorithm implementation has two parts: classifier builder and classifier tester. Each part can be run directly.

1. Data Preparation:

To train and test the algorithm, the dataset is needed.
There is a file in the dataset folder called 'origin', which is the original dataset downloaded from Standford AI Lab's website. There is also a script called 'generate_dataset'. To generate the dataset for this project, run this script and pass the absolute path of 'origin' as the first parameter.

For example:

```
BinBos-MacBook-Pro:~ bigbug$ cd /Users/bigbug/ORC/OCR/dataset
BinBos-MacBook-Pro:dataset bigbug$ ./generate_dataset /Users/bigbug/ORC/OCR/data
set/origin
```

After that, some files should have been generated, and the dataset folder should like like this:

- dataset.py
- generate_dataset
- origin
- test_extra.arff
- test_extra.csv
- test.arff
- test.csv
- tiny_dataset_for_ANN
- training_extra.arff
- training_extra.csv
- training.arff
- training.csv

There are two different kinds of dataset, one with extra as the postfix and the other not. extra means this dataset (both for training and test) has got more features. ARFF files are used by SVM and Naïve Bayes while the CSV files are used by ANN only.

Note:
ANN is special because training a neural network takes huge amount of time on the local computer. To make life easy, the folder **tiny_dataset_for_ANN** is also uploaded with a tiny dataset that can be trained in less than an hour.

2. Path configuration

The **config.txt** file under the root directory is there to setup the paths used in this project. The following table describes the parameters used by this file.

Parameter	Description
USE_MORE_FEATURES	Set to true to use 152 features for training and testing; otherwise use 128 features dataset for training and testing.
TRAINING_DATA_PATH	Path for 128 features training data for SVM and Naïve Bayes.
TEST_DATA_PATH	Path for 128 features test data for SVM and Naïve Bayes.
TRAINING_DATA_MORE_FEATURES_PATH	Path for 152 features training data for SVM and Naïve Bayes.
TEST_DATA_MORE_FEATURES_PATH	Path for 152 features test data for SVM and Naïve Bayes.
ANN_TRAINING_DATA_PATH	Path for 128 features training data for ANN.
ANN_TEST_DATA_PATH	Path for 128 features test data for ANN.
ANN_TRAINING_DATA_MORE_FEATURES_PATH	Path for 152 features training data for ANN.
ANN_TEST_DATA_MORE_FEATURES_PATH	Path for 152 features test data for ANN.
ANN_TRAINED_MODEL_PATH	Path for saving and loading the trained model of ANN
NB_TRAINED_MODEL_PATH	Path for saving and loading the trained model of Naïve Bayes
SVM_TRAINED_MODEL_PATH	Path for saving and loading the trained model of SVM

Suppose you want to use dataset with 128 features to run the algorithm, what to do is:

1. Change USE_MORE_FEATURES from true to false
2. Set TRAINING_DATA_PATH to the path of training data with only 128 features (without the postfix of '_extra')
3. Set TEST_DATA_PATH to the path of test data with only 128 features
4. If you want to run ANN, then change ANN_TRAINING_DATA_PATH and ANN_TEST_DATA_PATH to the path of training data and test data with only 128 features (the same as the step 2 and 3)
5. Make sure the XX_TRAINED_MODEL_PATH is set for the algorithm you are running
6. Run the algorithm to train and test

Note:
ANN can only use CSV files for training and testing.

3. Run the algorithm

To run the algorithm, import the OCR as a Java project to Eclipse. As an example, open SVMClassifierBuilder.java in Eclipse and run it. After a few minutes, the trained model from SVM will appear in the **model** directory. Then run the SVMClassifierTester.java to test the model. The paths for the training data, test data and model should be pre-defined in **config.txt**.
For training ANN, the training won't stop until the error is below 0.1. It may take a great amount of time to training a ANN even if the dataset is relatively small.