UTD ID - 2021459028
Net ID - SXA180157

Readme for Homework 1

CS 6375
October 3, 2019

Command to run the python code

```
python main.py [dataset_name] [algorithm_name] [type_of_model]
```

Please run the codes in python 2

For example to run the the discrete naive Bayes on enron1 dataset, you can run the following code on the command prompt.

```
python main.py enron1 -dnb
```

To run the the multi nomial naive Bayes on enron1 dataset, you can run the following code on the command prompt.

```
python main.py enron1 -mnnb
```

To run the the MCAP LR on enron1 dataset with bag of words, you can run the following code on the command prompt.

```
python main.py enron1 -mcap -bow
```

To run the the SGD classifier on enron1 dataset with Bernoulli model, you can run the following code on the command prompt. (Please install the sklearn library first)

```
python main.py enron1 -sgd -bm
```

Now let us elaborate on the command line parameters for the code:-

- **dataset_name**
  Provide the name of folder for the dataset (please keep the folder in the same directory as the code only)
  For example your folder in which the code is present should look like this :- The other

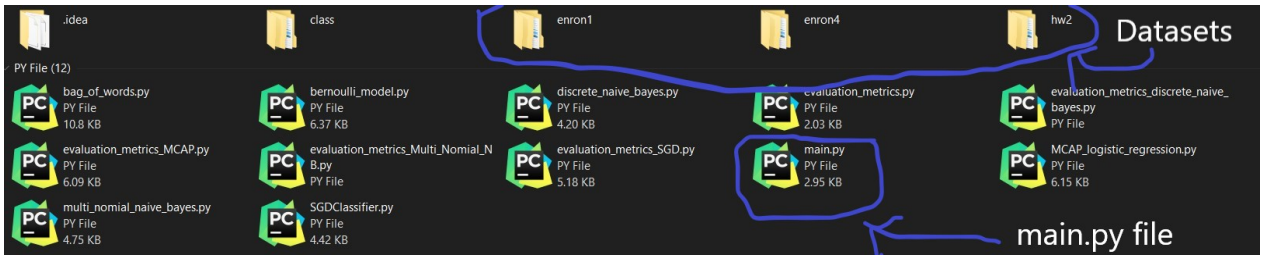

Figure 1: The directory for the code

folders like the test, train, ham and spam should have the same folder structure as mentioned in the hw2.pdf file provided on the course home page. For example for spam folder in train directory for enron1 datset should be like
directory_of_the_code/enron1/train/spam.

- **algorithm_name**
  -mnnb - for the multi-nomial naive Bayes (without 3^{rd} argument)
  -dnb - for the discrete naive Bayes (without 3^{rd} argument)
  -mcap - for the Logistic Regression (MCAP) (with 3^{rd} argument)
  -sgd - for the SGD classifier (with 3^{rd} argument)

- **type_of_model**
  This is used only for the -mcap and the -sgd algorithms(2^{nd} parameter)
  -bow - use this parameter for choosing the bag of words model
  -bm - use this for the Bernoulli model

If you are using the Naive Bayes model(both) then the code doesn't require you to give the model parameter, but it will print a message for the same.

These are the modules/imports required for the code:-

```
import sys
import copy
import glob
import os
import re
from collections import Counter
import random
import numpy as np
from decimal import Decimal
from math import log10 as log
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import GridSearchCV
import warnings
```

These are the modules that need to be installed on the devices on which the code is run. Please install all these before running the code.
If the code does not run from the main.py file you can run it from the following files:-

1. evaluation_metrics_discrete_naive_bayes.py, using

```
    evaluate_discrete_NB(dataset_name)

```

   as the last line of the code(I have commented the same)

2. evaluation_metrics_MCAP.py, using

```
    evaluate_MCAP_bag_of_words(dataset_name)
```

for bag of words and

```
    evaluate_MCAP_bernoulli_model(dataset_name)

```

for Bernoulli model as the last line for the code. (I have commented the same)

3. evaluation_metrics_Multi_Nomial_NB.py, using

```
    evaluate_multinomial_NB(dataset_name)

```

4. evaluation_metrics_SGD.py, using

```
    evaluate_SGD_bag_of_words(dataset_name)

```

for bag of words and

```
   evaluate_SGD_bernoulli_model(dataset_name)

```

for Bernoulli model. (I have commented the same in the code)