Checkpoint 3 documentation

Environment Setup

Check for sudo access before you run the following commands. Also check if pip is configured on the system.

```
pip install nltk
pip install sklearn
pip install pandas
```

When you download the project zip file you'll find three files:
1) `Sentiment.py`
2) `dvc_test_data.txt`
3) `dvc_train_data.txt`

You would need to modify two lines in `sentiment.py` to update the location where `dvc_test_data.txt` and `dvc_train_data.txt` in your local system. You do this by updating the variables `test_data_file_name` and `train_data_file_name` with the corresponding location of the training and testing data in your system.

Running the program

To run the program, go to the directory where the files are stored and run:
`python sentiment.py`

We have used the following classifiers:
1) Logistic Regression
2) Multinomial Naive Bayes
3) Decision Tree Regressor
4) Gaussian Naive Bayes

For each of these classifiers we have evaluated the precision, recall, f1-score and the sample results of sentences using the classifier. We have also performed cross validation using four folds to check how the classifier performs. Finally at the end of the output you'll notice a list of classifiers with the corresponding accuracies. Logistic Regression performed the best with an accuracy of 99.13% and Gaussian Naive Bayes had a comparatively low performance with an accuracy of 94.26%.

Output

```
==========================================================================
                Cross Validation 4-folds
==========================================================================
Cross Validation with 4 Folds
4 Fold Scores     : [ 0.98966165  0.9943556   0.98964218  0.99152542]
('Scores after Mean: ', 0.99129621494688414)
()
==========================================================================
('\t\t\t', 'Logistic Regression')
==========================================================================
             precision    recall  f1-score   support

          0       0.99      0.99      0.99      1233
          1       0.99      1.00      0.99      1602

avg / total       0.99      0.99      0.99      2835

()
==========================================================================
                Sample Result
==========================================================================
('Sentiment', '\t\t', 'Text')
()
1     I LOVE YOU MIT HIEL MEINE HART(


1     And i've managed to cover all of my own food expenses as of thus far
by working as Harvard's random task bitch -- doubly freaking sweet(


0     MrsSatan Hillary Clinton sucked money out of Hollywood just like
Monica Lewinsky sucked...


0     harvard is dumb, i mean they really have to be stupid to have not
wanted her to be at their school.
```

```
0    The history here goes back to the summer of 1992, when the San
Francisco Giants were terrible.


()
===========================================================================
               Cross Validation 4-folds
===========================================================================
Cross Validation with 4 Folds
4 Fold Scores    : [ 0.94736842  0.93697084  0.9472693   0.93879473]
('Scores after Mean: ', 0.9426008221093789)
()
===========================================================================
('\t\t\t', 'Gaussian Naive Bayes')
===========================================================================
           precision    recall  f1-score    support

        0       0.95      0.91      0.93      1233
        1       0.93      0.97      0.95      1602

avg / total     0.94      0.94      0.94      2835

()
===========================================================================
               Sample Result
===========================================================================
('Sentiment', '\t\t', 'Text')
()
0    Boston can suck my fucking tits...


0    Everyone was saying Lakers sucks and Kobe sucks.....


1    yes i love the geico commercials i will admit it!


1    I like Honda...


0    stupid lakers!!!!!!!!!!!!
```

```
()
()
========================================================================
                Cross Validation 4-folds
========================================================================
Cross Validation with 4 Folds
4 Fold Scores    : [ 0.98590226  0.98588899  0.98022599  0.98022599]
('Scores after Mean: ', 0.98306080661377282)
()
========================================================================
('\t\t\t', 'Multinomial Naive Bayes')
========================================================================
            precision    recall  f1-score   support

        0        0.99      0.98      0.98      1233
        1        0.99      0.99      0.99      1602

avg / total      0.99      0.99      0.99      2835

()
========================================================================
                Sample Result
========================================================================
('Sentiment', '\t\t', 'Text')
()
0     A Honda Music Link is a piece of crap and a waste of money.


1     Love Story At Harvard [ awesome drama!


1     I love Tom Cruise!


1     hell ya i play on 2 basketball teams i love the lakers + pistons love
both!!!!..


0     As much as i love London, even London sucks during the winters.
```

```
()
==========================================================================
                    Cross Validation 4-folds
==========================================================================
Cross Validation with 4 Folds
4 Fold Scores    : [ 0.96568902  0.95811753  0.94257406  0.95375253]
('Scores after Mean: ', 0.95503328398993137)
()
==========================================================================
('\t\t\t', 'Decision Tree Regressor')
==========================================================================
              precision    recall  f1-score   support

          0       0.99      0.98      0.99      1233
          1       0.99      0.99      0.99      1602

avg / total       0.99      0.99      0.99      2835

()
==========================================================================
                    Sample Result
==========================================================================
('Sentiment', '\t\t', 'Text')
()
0.0   I hate Tom Cruise.


0.0   I hate the Lakers...


1.0   well i'm gonna go enjoy being in seattle.....


1.0   I also love the new rabbits, I still want an x-terra-but Luke's will
do, and I kinda like Honda Elements, but they got a bad safety rating: (..


1.0   But I bought my old car less because I liked it and more as a favor
because I've wanted a Toyota since high school ( I took drivers ed.
```

```
================================================================
            Classifiers Accuracy List
================================================================
('Logistic Regression::', 99.13, '%')
()
('Multinomial Naive Bayes::', 98.31, '%')
()
('Decision Tree Regressor::', 95.5, '%')
()
('Gaussian Naive Bayes::', 94.26, '%')
()
```