The assignment is an NLP problem. You are required to build a machine learning model that predicts the labels of a given paragraph of text. Please download data.zip file from google drive link below.

The zip file contains 4 files as described:

1) train\_data.csv : This training file contains two fields < *text*, *id*>.

The *text* is the input data and *id* is a number that maps the *text* to its corresponding labels in *train\_label.csv*

2) train\_label.csv : This label file contains two fields < *id*, *label*>.

The *id* should be used to map the *label* to corresponding input *text* in *train\_data.csv*

3) test\_data.csv : This test file contains two fields < *text*, *id*>.

The *text* here is the test data for which classification probabilities need to estimated.

4) sample\_submission.csv : This is file that needs to be submitted. Columns are the labels and rows are *id's* from test\_data. Please replace the zeros with probability outputs of the model.

Please share the python code and a detailed write up of the following steps:

- noise removal

- data visualization

- model fitting

- performance metric evaluation

- analyzing mistakes and improving the model