Task given at 12 March 2018

**Submission deadline: 18 March 2018.**

# The Task

* Building a model for movies reviews in English for classifying it into positive or negative.
* Test classifier on new reviews.

# Dataset

Movies reviews dataset can be found here (<http://www.cs.cornell.edu/people/pabo/movie-review-data>), consider using the file **polarity dataset v1.1**. This zipped file consists of two paths - tokens/pos/ and tokens/neg/, each contains 700 text documents with positive and negative reviews respectively.

# Take Detail

Use Naive Bayes classifier as mentioned in the table 6.2 in the text book. A positive and negative labeled movies reviews are already available. You should separate the dataset into two, so that most portion of them are used to train, and the rest are used to test your classifier.

*Case-1*: You should consider only first 100 review files from each category for training, and use next 10 review files for test and record the result.

*Case-2*: Update training/testing file counting for each category 200/10

*Case-3*: Update training/testing file counting for each category 300/10

**Prepare a small report** where you should discuss your experiment setup, methodology, and discuss your results that you received for three cases with below format.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Case # | Positive | Negative | TP | FP | TN | FN |
| 1 | 10 | 10 |  |  |  |  |
| 2 | 10 | 10 |  |  |  |  |
| 3 | 10 | 10 |  |  |  |  |

In the table, TP is True Positive: which indicates out of positive (in 2nd column), how may instances your classifier found true.

Similarly, FP indicates, how many instances your classifier found to be positive which are actually negative.

TN is true negative, and FN is false negative – meanings of which are self explanatory.

**Code submission:** Share your code by providing a Google drive link.