CE 314/887 Assignment 2

Text classification

December 2021

Deadline: Jan 2rd 2022

Build a text classifier on the IMDB sentiment classification dataset, you can use any classification method, but you must training your model on the first 40000 instances and testing your model on the last 10000 instances. The IMDB dataset will be uploaded on the moodle page for you to download.

Your code should include:

1: Read the file, incorporate the instances into the training set and testing set.

2: Pre-processing the text, you can choose whether you need stemming, removing stop words, removing non-alphabetical words. (Not all classification models need this step, it is OK if you think your model can perform better without this step, and you can give some justification in the report.)

3: Analysing the feature of the training set, report the linguistic features of the training dataset.

4: Build a text classification model, train your model on the training set and test your model on the test set.

5: Summarize the performance of your model.

6: (Optional) You can speculate how you can improve your works based on your proposed model.

After you build such a model and test on the test set, you should write a report (no longer than two pages in A4, with Arial 11 fonts) to summarize your work.

(You can use the existing algorithms on github or kaggle, but you must not directly copy and paste their code!

However, you are not allowed to use the Naïve Bayes algorithm which practiced in Lab 4)

Suggestion: Have necessary comments on your code is a good habit!