Overview

Data

This is a dataset for binary sentiment classification. We provide a set of 50,000 highly polar movie reviews for training and testing.

[Large Movie Review Dataset]

To do

Take the provided dataset and solve the binary classification task.

Target – sentiment pos/neg

Evaluation

Metric AUC-ROC with **visualisation**

Libraries

- scikit-learn
- NLTK

Criteria Scoring (15 max)

- Text preprocessing with explanations of all steps: 3
    - Cleaning - 1
    - Tokenization - 1
    - Normalisation (comparison of stemming and lemmatization) - 1
- Words importance: 2
- Hyperparameters tuning: 1
- Compare performance of models: 2
    - SGDClassifier
    - SVM
    - Naive Bayes
- Detailed conclusions: 2
- Quality of delivered work:
    - Analytical comments provided: 1
    - The experiment is structured (file is readable, pictures have titles): 1
    - Code is clear (reusable code in functions, comments, code is easy readable): 1
- Extra points for improvements not considered in the criteria: 2