Manual

This tool classifies bug reports using machine learning models and TF-IDF text vectorization. It compares Naive Bayes, Decision Tree, and Random Forest classifiers with two TF-IDF setups. The initial baseline was built upon from the lab1 solution.

Prerequisites and setting up

1. Install all dependencies from requirements.pdf'
2. Place your dataset (e.g. pytorch.csv) inside the datasets/ folder.
3. Make sure it has columns: Title, Body, class, and Number.
4. Set the dataset name in the script: (e.g. project = 'pytorch')
5. Open terminal and run (python br_classification.py)

This will run the program and output 9 results in the command line, saving the results into csv files, each combination is run e.g. original TF-IDF and Naive Bayes for the baseline, enchanced TF-IDF and Naive Bayes, etc.

Results

Raw data can be found in the results folder

The results folder creates a raw data folder for each run; and a mean folder that calculated the mean of all those runs

Result files are saved inside the results folder as "../pytorch_RF_ImprovedTFIDF.csv" - they follow the naming convention of dataset_classifier_TFIDF.csv.

Each file contains:

- Accuracy
- Precision
- Recall
- F1 Score
- AUC
- Average execution time
- AUC values from repeated runs