Homework 2 – Task 1: Dialogue Act Recognition (DAR)

Student UNI: em3907

Contents of the submission **folder** "task1":

1 Feature Extraction and analysis scripts

- Speech-based features:

  FeatureExtraction_speech_train.ipynb
  FeatureExtraction_speech_valid.ipynb
  FeatureExtraction_speech_test.ipynb

  - These three notebooks perform identical speech-based feature extraction processes using Parselmouth. Due to time and memory efficiency constraints, they are split into separate files for train/validation/test sets.

- Text-based features:

  FeatureExtraction_text.ipynb

  - This notebook extracts text-based features such as LIWC scores and utterance length.

- Feature Analysis:

  Feature_analysis.ipynb

  - This notebook conducts visual and statistical analysis of extracted speech and text features. It evaluates hypotheses about which features are informative for specific dialogue acts.

2 Classification script:

classification.ipynb

- This notebook performs preprocessing (e.g., normalization), model training, and evaluation using three feature setups:
  (1) Speech-based features only
  (2) Text-based features only
  (3) Speech + Text-based features
- Trained classifiers: Logistic Regression, Linear SVM, XGBoost, and MLPClassifier (Neural Network)
- The notebook includes the generation of classification reports, confusion matrices (raw, normalized by true labels, normalized by predicted labels), and model selection based on validation macro F1 score.

3 Extracted Feature Files
All CSVs follow the required format: “dialog_id”, “speaker”, “da_tag”, “start_time”, “end_time” followed by feature columns.

- Speech-based features:
  speech_features_train.csv, **speech_features_valid.csv**, **speech_features_test.csv**
- Text-based features:
  text_features_train.csv, **text_features_valid.csv**, **text_features_test.csv**

4 Final Prediction Outputs
Each file includes test set predictions using the respective model. Predictions are filled in the "da_tag" column, maintaining the original metadata fields.

- **test_em3907_speech.csv**, **test_em3907_text.csv**, **test_em3907_multi.csv**

5 Written Report and Readme
This document contains all answers and analysis for task 1, including hypotheses, feature choices, classification results, confusion matrix analysis, and improvement ideas.

- **em3907_task1_responses.pdf**, **readme.pdf**

Requirements

This project was implemented using Python in Jupyter Notebook. Run each notebook top-to-bottom in the order of your workflow.

Required Python packages (imported at the top of each notebook):

- numpy
- pandas
- matplotlib
- seaborn
- sklearn
- xgboost
- parselmouth (for speech feature extraction)

All packages can be installed via pip:

```
> pip install numpy pandas matplotlib seaborn scikit-learn xgboost praat-parselmouth nltk
```

Running the Code

1. Feature Extraction:

    - Run the four `**FeatureExtraction_*.ipynb**` notebooks to extract speech and text features.
    - Ensure the resulting CSVs are saved to the current working directory.

2. Feature Analysis:

    - Open and run `**Feature_analysis.ipynb**` to visualize feature patterns and validate hypotheses.

3. Model Training and Evaluation:

    - Open and run `**classification.ipynb**`.
    - This will load the features, train models, evaluate the validation set, output confusion matrices, and F1 scores, and save predictions for the test set.

Notes

- All notebook outputs (plots, metrics, tables) are retained and visible for evaluation.
- The test predictions are generated using the best-performing model (XGBoost, selected based on validation macro F1).
- Feature extraction and modeling strictly follow the task instructions.