Contents

This repository contains the data, the jupyter notebooks and code to run this project.

The structure of the repository is as follows:

- `data`: contains the data for Reddit modelling
- `file-conversion`: contains the code which converts the dynamically generated data from Stack Overflow and Reddit to usable format in Python
- `modelling`: contains all the feature engineering and modelling codes
- `paper`: contains the LaTeX code
- `Reddit`: contains the Python code required to download data from Reddit
- `SESE`: contains the sql queries which were used to generate the data for Stack Overflow analysis

Requirements

A Python installation is required with the following packages and their dependencies:

- gensim
- pandas
- praw

This can be installed via `pip install`