Linguistic models

This document explains the method and technicalities of the linguistic model construction.

The models are taken from a larger work, to appear, “The Corrective Commit Probability

Code Quality Metric“ (CCP).

For the models constructicon:

1. The model was constructed using domain knowledge, make it not sensitive to the train set. That enabled us to use data sets from
   1. “The Corrective Commit Probability Code Quality Metric“ (CCP).
   2. “Boosting Automatic Commit Classification Into Maintenance Activities By Utilizing Source Code Changes” provided by the courtesy of Stanislav Levin and Amiram Yehudai.
   3. Samples selected due to co-traning and active learning techniques, for performance improvement (to appear)
2. The test set samples are in “commit\_update5.csv”. There are 1,100 commits sampled uniformly from the active projects commit messages. These samples are marked as random.

The model language was generated using commit\_type\_model.py

The use of python enabled

1. Efficient and compact model representation (e.g., adding negative to all positive occurrences)
2. Ability to run the model locally on any text

In order to run the language model on the BigQuery Github repository, the models where transformed into regular expression.

The transformed model appear in the classify\_valid\_commit query

Labeling

1. Refactoring commits were labeled according to the protocol in Refactor labeling protocol
2. Corrective commits were labeled according to the protocol in Bug fix labeling protocol
3. Note the a commit might belong to few categories simultaneously (e.g., refactr and corrective)
4. We labeled a sample of the commits by some annotators to verify consistency.
5. Full details are part of a larger project and yet to be appear. Please reach out for more details in the meanwhile.