Edit Quality Documentation

Release 0.5.0

Scoring Platform team

May 13, 2020

Contents

This package contains a set of utilities and assets for building and using "Edit Quality" prediction models. Here, Edit Quality represents aspects of edits in Wikipedia like "Did this edit cause damage to the article?" (damaging), "Does it like the editor meant well?" (goodfaith), and "Will this edit be reverted?" (reverted).

In this package, you'll find the `feature lists` used to train models for each the wikis supported. There's a set of `command-line utilities` that are used to perform data pipeline operations specific to training and testing edit quality models.

See Tutorials for more information about how to configure the data pipeline for training new models.

See the API Reference for low level details.

Project Info:

Changelog

All notable changes to this project will be documented in this file.

1.1 [0.5.0] - 2019-08-13

1.1.1 Added

- Added `needs_review` filter for `merge_labels`.
- RELEASE_CRITERIA.md
- CHANGELOG.md
- CODE OF CONDUCT.md
- Release automation using Travis
- Added modularized config structure
- New `join_observations` utility
- Ability to configure the number of extractors in Makefile
- Feature Lists for sqwiki, kowiki, fiwiki, rowiki, zhwiki and glwiki

1.1.2 Fixed

- `extractors` typo in generated parameters
- Flaky config test
- Notebook fixes for revscoring 2.x
- Silence mediawiki API errors in `autolabel`.
- `merge_labels` degenerate bug.

- Argument order for `merge_label` utility

- Minor issue with autolabel mw session args

1.1.3 Changed

- Bumped revscoring version to 2.5.1

- Refactored templates to be more modular

- Makefile now reads `trusted_edits` config.

- Use `min_samples_leaf` as a tuning param for GB.

- Interpret boolean string literals in label merge.

- Make option of auto-labeled dataset optional in `merge_labels`.

- Standardized `generate_make` utility and args.

- Use informals and badwords for wikidata comment parsing.

- Assume goodfaith in `fetch_labels`.

- Exclusions for huwiki badwords/informals

- Added `filter_rate_at_recall` metrics back to Makefile

- Migrate from nosetest to pytest

1.1.4 Removed

- XGBoost requirement

1.2 [v0.4.2] - 2017-01-07

1.2.1 Added

- Added `feature_lists` for svwiki and cswiki

- Added `--filter` param to `fetch_labels` utility

- Added "about.py" variables (`__version__`, `__author__`, etc.)

1.2.2 Fixed

- Data normalization issues in autolabel utility

1.2.3 Changed

- Utilities now use JSON-lines format

- Renamed `label_reverted` to `autolabel`

- Bumped revscoring requirement to 1.2.8 f9739b7

- `test_statistics` model param for damaging and goodfaith filtering

1.2.4 Removed

- XGBoost from models to tune

1.3 [v0.4.1] - 2016-04-27

1.3.1 Added

- huwiki feature lists for revert model
- ruwiki feature lists for damaging features

1.3.2 Changed

- Bumped revscoring requirement to >= 1.2.2
- Updated Revert detection demo notebook with more context.

1.4 [v0.4.0] - 2016-03-15

1.4.1 Added

- Added XGBoost to model tuning
- Added Revert detection demo notebook
- Added Wikibase feature lists
- Added separate Wikitext feature lists

1.4.2 Changed

- Update mediawiki features for comment new line regex and user registration.
- Update `wikidatawiki` to use new Wikibase features
- Update all damaging feature to use new Wikitext features
- Updated feature lists to handle linear models

1.5 [v0.3.0] - 2016-02-07

1.5.1 Added

- Added edit summary support for `label_reverted` util. New args: `exclude-reverted` and `excluded-reverting`

1.6 [v0.2.2] - 2016-01-21

1.6.1 Added

- dump reader

1.6.2 Changed

- Parallelized `label_reverted` util.
- Parallelized `prelabel` util.
- enwiki and wikidatawiki make commands use revscoring 1.0.1-rc1

1.6.3 Removed

- Remove old linear svc models from Makefile

1.7 [v0.2.1] - 2015-12-14

1.7.1 Added

- configs for models that exclude problematic user features

1.7.2 Removed

- informal words from itwiki's damaging feature

1.8 [v0.2.0] - 2015-12-14

1.8.1 Added

- Added `--include-revid` and cut -f2- to all feature extractions in Makefile.
- Configuration for model tuning
- Commands for tuning in Makefile

1.8.2 Fixed

- Import bug for idwiki features in Makefile

1.8.3 Changed

- Bumps revscoring requirement to 0.7.10
- Added back user models next to non-user models.

1.9 [v0.1.1] - 2015-11-21

1.9.1 Fixed

- Resolved `self_reverted` detection bug in `label_reverted`

1.9.2 Changed

- added handling for `reverted_to` condition in `label_reverted`

1.10 [v0.1.0] - 2015-11-12

1.10.1 Added

- New `rev_reverted` rule to makefile for all languages.
- Added revert labeling to all wikis

1.10.2 Changed

- Handles deleted users in `prelabel` utility
- Bumped revscoring dep to >= 0.7.2
- Updates makefile to include `rev_ids` in feature extraction

1.11 [v0.0.5] - 2015-10-17

1.11.1 Added

- New `fetch_labels` utility
- Added enwiki, fawiki and ptwiki feature lists

1.11.2 Fixed

- Model References in Makefile

1.12 [v0.0.4] - 2015-09-06

1.12.1 Fixed

- Minor cleanup in `prelabel` docstring

1.13 [v0.0.3] - 2015-09-06

1.13.1 Fixed

- Prelabel docstring fixes
- Authors typo

1.14 [v0.0.2] - 2015-09-06

1.14.1 Fixed

- Module error due to missing `__init__` file.

1.15 [v0.0.1] - 2015-09-06

1.15.1 Added

- First release containing working prelabel script

CHAPTER 2

Indices and tables

- genindex
- modindex
- search

e

Index

E