SM2KG: a Framework for Capturing Software Metadata from Documentation

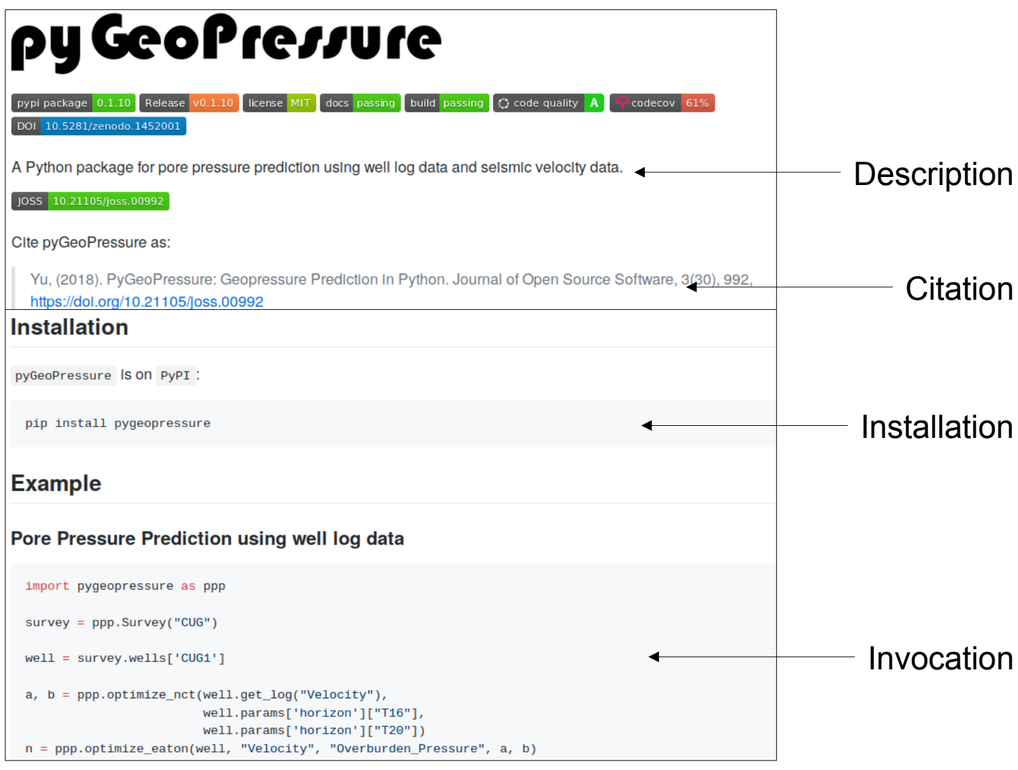
Allen Mao, Daniel Garijo, Shobeir Fakhraei
Please contact amao@isi.edu if interested

Motivation

- Reusable software:
  - ✓ reproduce computational methods
  - ✓ easy to integrate with other data and software
- Understanding software is time consuming
- → **Software metadata registries**
  - ✗ require manual curation
- SM2KG (Software Metadata to Knowledge Graphs)
  - ✓ extracts software metadata
  - ✓ organizes into knowledge graphs

Problem Statement

Given a README excerpt, e.g.



✔We aim to identify:

- ■ description (what does this software do?)
- ■ installation (how do I set it up?)
- ■ invocation (how do I invoke it?)
- ■ citation (who do I credit?)

➔Each class has its own linguistic characteristics

Approach

Corpus

- Default README of 74 Github repositories
- Plain text rendered Markdown
- Text split by newlines for convenience
- Each excerpt labeled by class
- 50% positive, 50% negative per classifier

Data Preparation

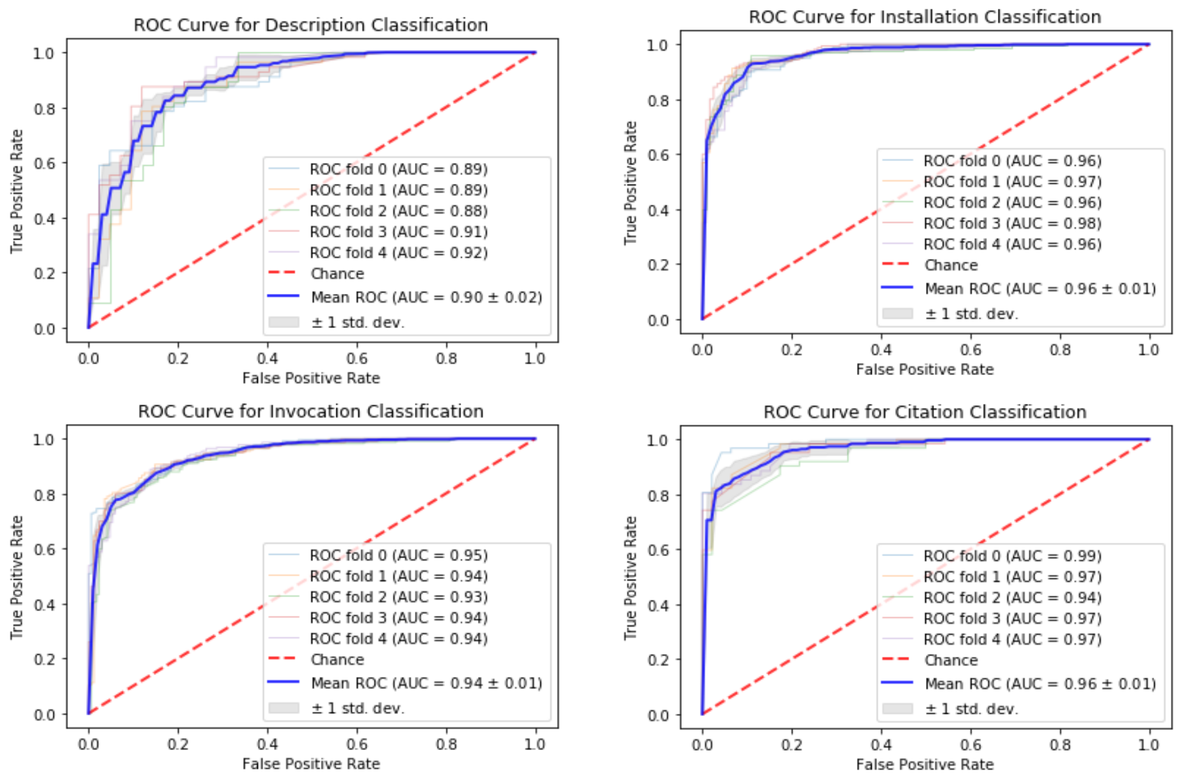
- Default scikit-learn tf-idf tokenizer without stemming

Classifiers

1) Logistic Regression, liblinear solver
2) Multinomial Naive Bayes Classifier

Evaluation

- Stratified 5-fold Cross Validation ROC
- Tf-idf + (Logistic Regression / Naive Bayes) results are promising (AUC > 0.89)



AUC curves for Logistic Regression

Future Work

- Expand corpus
- Use markdown metadata as a classification feature
- Test deep learning architectures

Project URL:
https://github.com/KnowledgeCaptureAndDiscovery/SM2KG
Work performed under REU Site program
supported by NSF grant #1659886

USC Viterbi
School of Engineering
Information Sciences Institute