SoMEF: A Framework for Capturing Software Metadata from its Documentation

Allen Mao, Daniel Garijo, Shobeir Fakhraei

University of Southern California, Information Sciences Institute

2019 IEEE BigData REU Symposium

Contact: allenmao AT berkeley DOT edu

Computational Sciences have increasingly become a fundamental scientific approach

- But the continuous development of new software makes it hard to keep track of or evaluate different software (or even versions)
- As a result, scientists spend much time poring through software documentation and code to understand how to use it or cite it.
- This process is time consuming and unappealing to scientists.

Easier Understanding, Reuse, and Attribution of Scientific Software

- We present SoMEF, a *Software Metadata Extraction Framework* that automatically extracts relevant software metadata from its documentation.

$$\text{README} \xrightarrow{SoMEF} \{\text{description, installation, invocation, citation}\}$$

- Examples[1]:
  - **Description**: A Python package for pore pressure prediction...
  - **Installation**: `pip install pygeopressure`
  - **Invocation**: `import pygeopressure as ppp`
  - **Citation**: Yu, (2018). PyGeoPressure: Geopressure Prediction in Python. Journal of Open Source Software, 3(30), 992, https://doi.org/10.21105/joss.00992

[1] `https://github.com/whimian/pyGeoPressure`

Approach

- Corpus consisted of plain text manual annotations on READMEs from 74 GitHub repositories.
- Scientific software dominates and *Awesome* curations of provided links to curations of scientific projects from different fields[2].

Table: Byte fractions of languages in collected repositories

Language	Percent
C++	32.66%
Python	31.16%
Jupyter Notebook	23.39%
JavaScript	4.70%
HTML	2.79%
Lasso	1.00%
Go	0.70%
Other	3.61%
Total	100%

- Set of four binary classifiers with one for each category.

[2] `https://awesome.re/`

Corpus Preparation

- **Corpus Composition**: For each binary classifier, corpus transformed where category to be predicted becomes **True** and all others become **False**.
- Random sentences from Treebank also serve as control sentences to ensure classifiers do not devolve into code vs text.
- **Balancing the Corpus**: Per corpus of classifier, all negative categories contribute equally to the 50% negative class.

Table: Description Corpus Breakdown

Truth Value	Category	Apprx. Ratio	Count
True	Description	0.5	275
False	Installation	0.125	68
	Invocation	0.125	68
	Citation	0.125	68
	Treebank	0.125	68
	Total	1.0	547

Data Preparation

- Tf-idf Vectorizer with unigram features.
- Since command line inputs and computer language lexicons are precise, no stemming or stop words used.
- TF-IDF matrix with 1509 features.
- Each feature is a "word", i.e. space-delimited string of characters.

Classifiers

- Two classifiers from the *Scikit-learn* package.
- Logistic Regression *liblinear* solver and balanced class weights because of small corpus size and imbalances that may arise from undersampling.
- Multinomial Naive Bayes (MNB) additive smoothing parameter of $\alpha = 1$ as default fail-safe probability.

Results

- Five-fold cross validation
- $\geq$ 0.90 ROC AUC and $\geq$ 0.92 Average Precision across all categories
- Logistic Regression and MNB performed similarly per same category
- On average, citation and installation classifiers performed best.

Extraction of Other Software Metadata

- Component that returns metadata from GitHub repositories with GitHub REST API as a JSON.

$$\text{GitHub Repository} \xrightarrow{SoMEF} \begin{bmatrix} \text{description} \\ \text{forks} \\ \text{license} \\ \text{programming languages} \\ \text{license} \\ \text{name} \\ \text{owner} \\ \text{README URL} \\ \text{releases} \\ \text{topics} \end{bmatrix}$$

Conclusion

Summary

- SoMEF is a novel approach that employs to extract software metadata.
- Promising initial steps: minimum average 0.92 precision and 0.90 ROC AUC.

Future Work

- Corpus Expansion
- Text Separation
- Other linguistic features
- Knowledge graphs of scientific software

Acknowledgements

A big thank you to the following:

- Dr. Daniel Garijo
- Dr. Shobeir Fakhraei
- Professor Yolanda Gil
- Professor Jelena Mirkovic
- The National Science Foundation
- The REU program

Questions?