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

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)

Future Work

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

Problem Statement

Given a README excerpt, e.g.

<table>
<thead>
<tr>
<th>pyGeoPressure</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="pyGeoPressure.png" alt="Image" /></td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Citation</td>
</tr>
<tr>
<td>Installation</td>
</tr>
<tr>
<td>Invocation</td>
</tr>
</tbody>
</table>

- 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

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