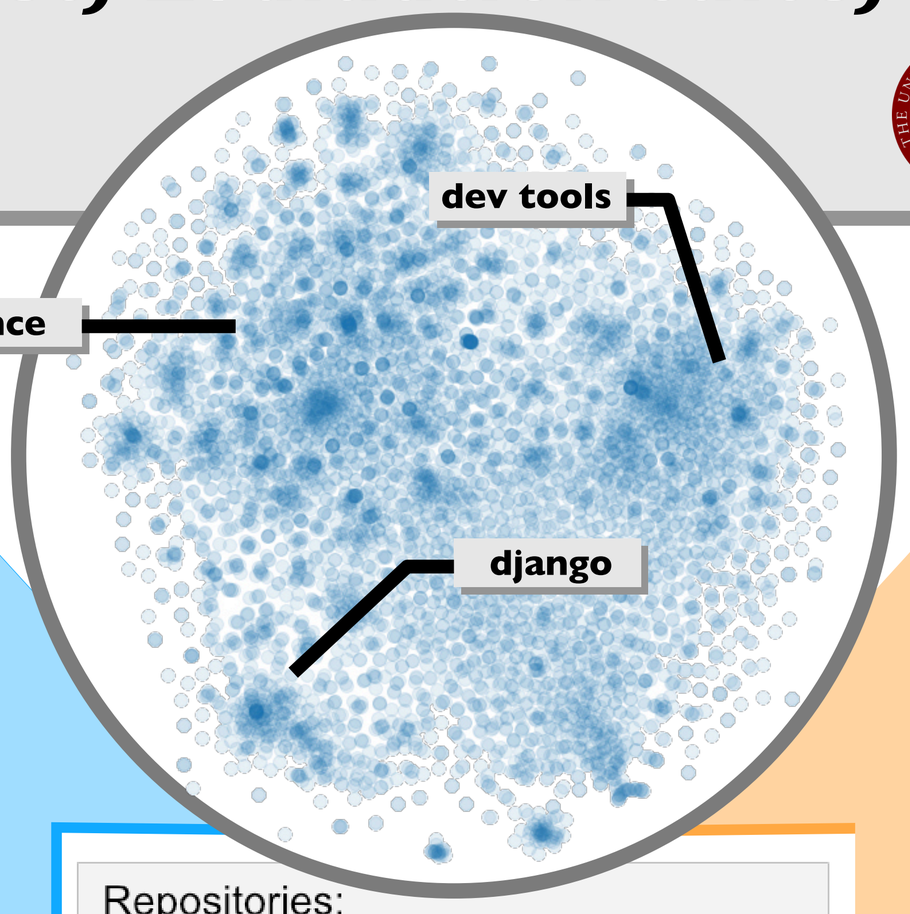
Representing Repositories in the Open-source Ecosystem: Dataset, Evaluation suite, and Baselines

J. Milbauer, Y. Chen,
D. Mukherjee, J. Evans

THE UNIVERSITY OF CHICAGO KNOWLEDGE LAB



Data

READMEs

To build representations based on project descriptions, we train the Doc2Vec model on text drawn from each repository's README file. Embeddings for repositories in the validation set are inferred from the pretrained model.

Library Imports

To represent repositories according to their technical implementation, we train a similar Doc2Vec-style model using 50.8M million pairs of imported libraries drawn from the codebase of each repository.

Social Network

We use the Node2Vec model on a social graph of collaborators. Repository nodes are joined by edges, each weighted by the number of shared contributors between the two repositories.

Repositories:

www.github.com/username/reponame

Fork Origin

Stars | # Forks | # Watches

tag_1 tag_2 • • • tag_n

Readme

The contents of the readme, after a set of basic cleaning and preprocessing steps have been performed.

$File_1$	$File_2$	• • •	$File_n$
$Library_1$	$Library_1$		$Library_1$
$Library_2$	$Library_2$		$Library_2$
• • •	• • •		• • •

$contributor_1$ • • • $contributor_2$

Tasks

Topic Tags

Users provide topic tags for some repositories on Github. We select for the top 500 tags, and manually sort them to a 20-cateogry label space. Embeddings are evaluated on this multi-label task using a nearest neighbors classifier.

Library Prediction

We evaluate whether embeddings can be used to predict imported libraries for repositories. We build a multi-label classification task for 7129 validation repositories, using the top 500 imported libraries on Github as labels. Embeddings are then evaluated with a nearest neighbors classifier.

Recommendation

We test the ability of embeddings to produce recommendations consistent with the co-collaboration signal on Github. For 1000 query repositories, we build gold rankings for 30 candidate repositories. Each embedding ranks repositories using the cosine similarity of each candidate to the query.

Collaborative activity in large communities is generally hard to follow; Github represents one of the few places where systemic collaboration occurs in a visible and recorded manner. Our dataset includes diverse features extracted from this ecosystem, spanning language, code, social networks, subject tags, and more.

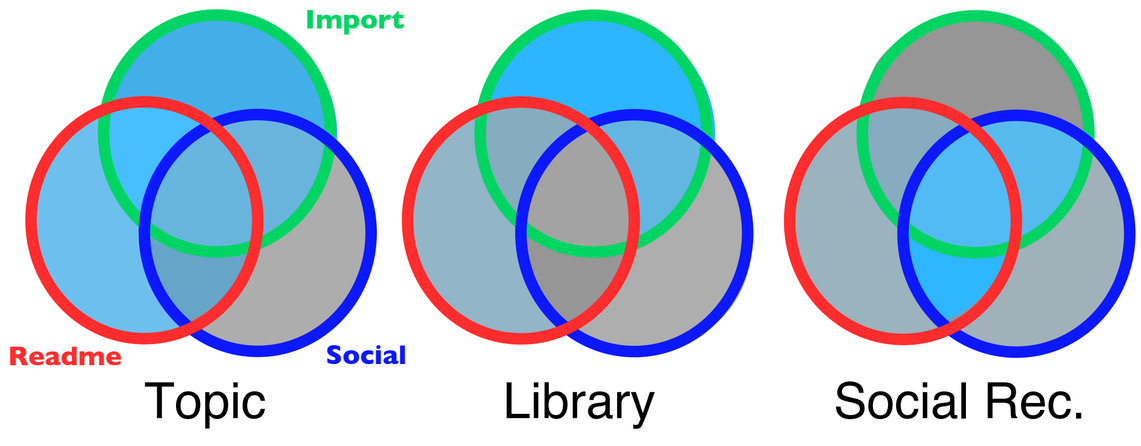
In addition to enabling the study of a dynamic and multimodal social community, this dataset empowers computational social scientists to:

Understand the individual developers, tools, and agendas at play in the open source ecosystem.

Examine the different **concepts** and **idioms** at play in different programming languages, and **track** the path of software developers' development to improve **pedagogical** practices.

Build systems capable of **recommending** repositories and code libraries to contributors, and **predict** the expansion (or collapse) of online resources.

Guide **policy** conversations on how to create **sustainable** platforms which support a productive and diverse open-source contributor base that is critical to software engineering everywhere.



Task →	Topic Classification			Library Prediction			Social Recommendation		
Model ↓ / Metric →	Precision	Recall	F1	Precision	Recall	F1	MRR	MAP	P@5
Readme	79.15	34.29	47.85	79.57	15.54	26.00	68.21	47.10	38.44
Import	75.21	35.20	47.96	92.85	**22.87**	**36.71**	60.24	42.53	34.14
Social	72.52	27.35	39.72	73.27	12.12	20.80	66.28	45.72	36.84
Readme+Import	77.23	**37.08**	**50.10**	**93.16**	15.02	25.87	69.30	48.36	39.36
Readme+Social	**80.16**	32.46	46.21	81.94	11.83	20.68	**79.89**	**61.97**	**54.16**
Import+Social	75.55	33.41	46.33	93.12	19.19	31.82	76.59	58.70	49.85
Readme+Import+Social	75.19	35.35	48.09	92.06	12.56	22.11	78.53	59.46	50.76

Relevant Citations:

Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. PMLR, 2014.

Theeten, Bart, Frederik Vandeputte, and Tom Van Cutsem. "Import2vec: Learning embeddings for software libraries." 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). IEEE, 2019.

We thank the Alfred P. Sloan Foundation for generously supporting this research to explore the relationship between programming, individual and collective problem-solving.