Visualizing GitHub Social Networks: 6.S079 Final Project

Emma Tysinger
MIT
tysinger@mit.edu

Michael Baumgartner
ETH/MIT
michbaum@ethz.ch

Massachusetts Institute of Technology

Initial Project & Approach

- Analyze possible correlation between r/wallstreetbets postings and stock market developments
- Recent Reddit API changes make structured scraping of large data volumes impossible
- Lost a team member
- Pivoted to current project

GitHub Social Graphs: Original goal

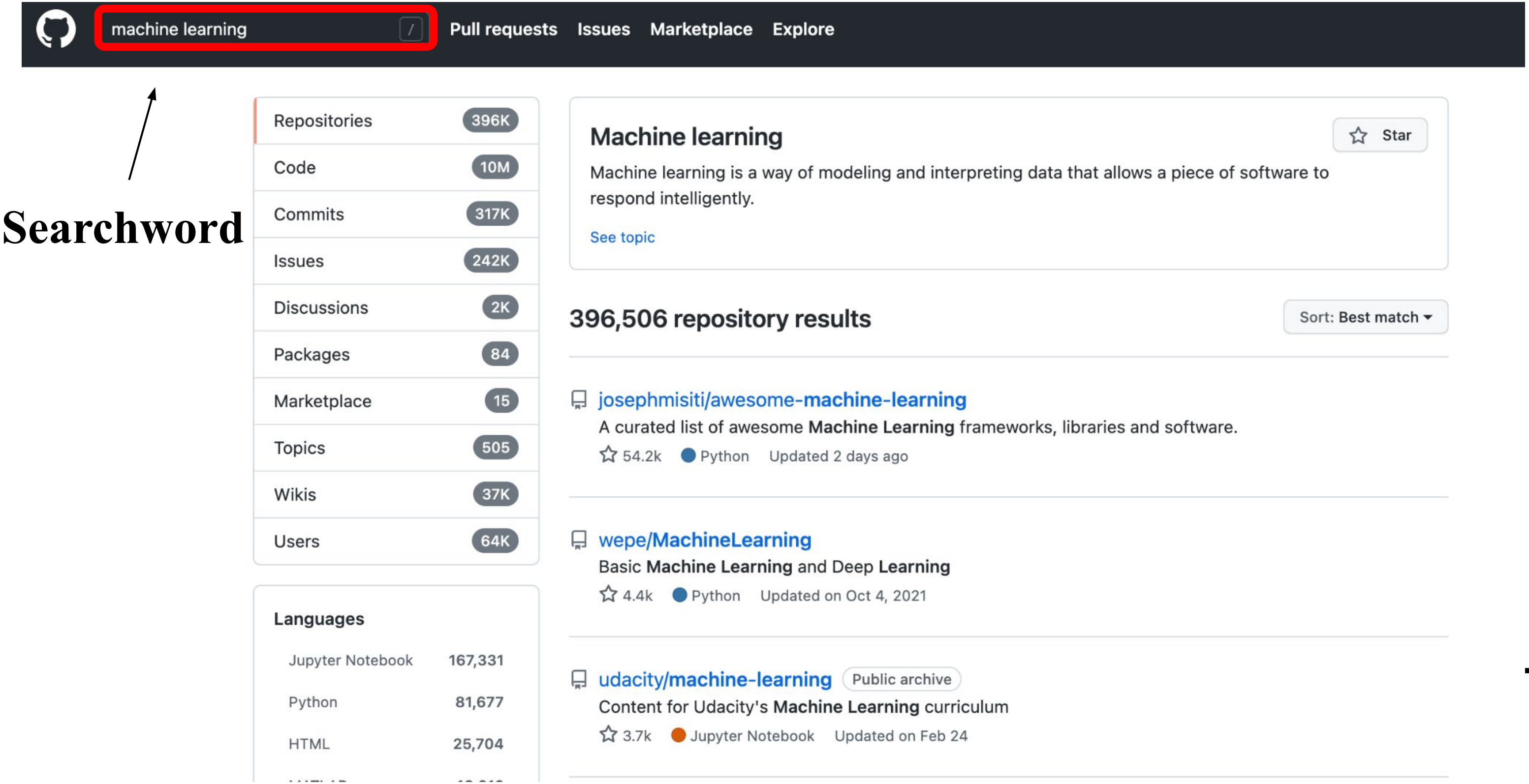
- Visualize & investigate the social connections in GitHub
- Can we discover underlying connections between different (successful) projects? Between different programming languages?
- Compare our visualization approach to common approaches seen in class

Data Scraping

- **38** unique search words
- Top **75** pages of results for each

GitHub

Rest API V3



machine learning

Repositories	396K
Code	10M
Commits	317K
Issues	242K
Discussions	2K
Packages	84
Marketplace	15
Topics	505
Wikis	37K
Users	64K

Languages

Jupyter Notebook	167,331
Python	81,677
HTML	25,704

Machine learning

Machine learning is a way of modeling and interpreting data that allows a piece of software to respond intelligently.

396,506 repository results

josephmisiti/awesome-machine-learning

A curated list of awesome **Machine Learning** frameworks, libraries and software.

☆ 54.2k ● Python Updated 2 days ago

wepe/MachineLearning

Basic **Machine Learning** and Deep **Learning**

☆ 4.4k ● Python Updated on Oct 4, 2021

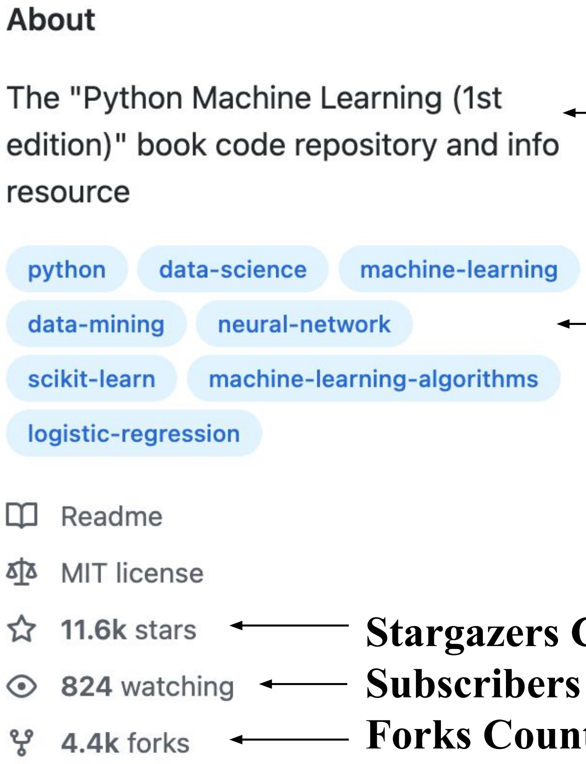
udacity/machine-learning

Content for Udacity's **Machine Learning** curriculum

☆ 3.7k ● Jupyter Notebook Updated on Feb 24

Searchword

Data Scraping : Repository Info



About

The "Python Machine Learning (1st edition)" book code repository and info resource

python data-science machine-learning data-mining neural-network scikit-learn machine-learning-algorithms logistic-regression

Readme

MIT license

11.6k stars

824 watching

4.4k forks

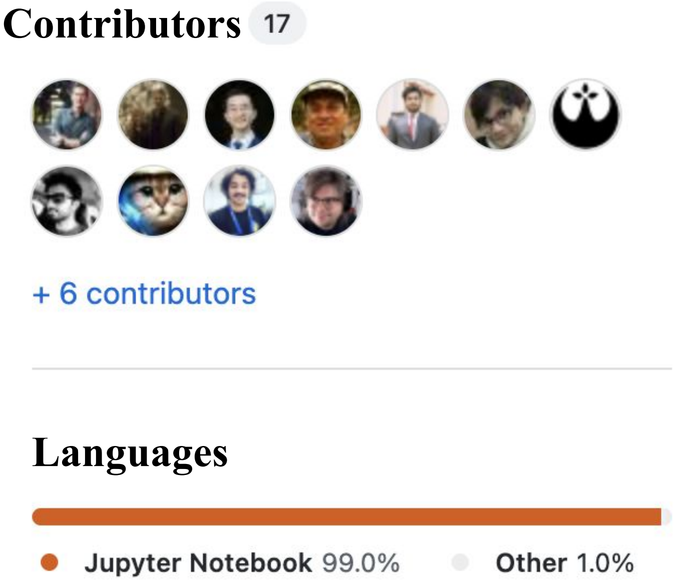
Bio

Topics

Stargazers Count

Subscribers Count

Forks Count



Contributors 17

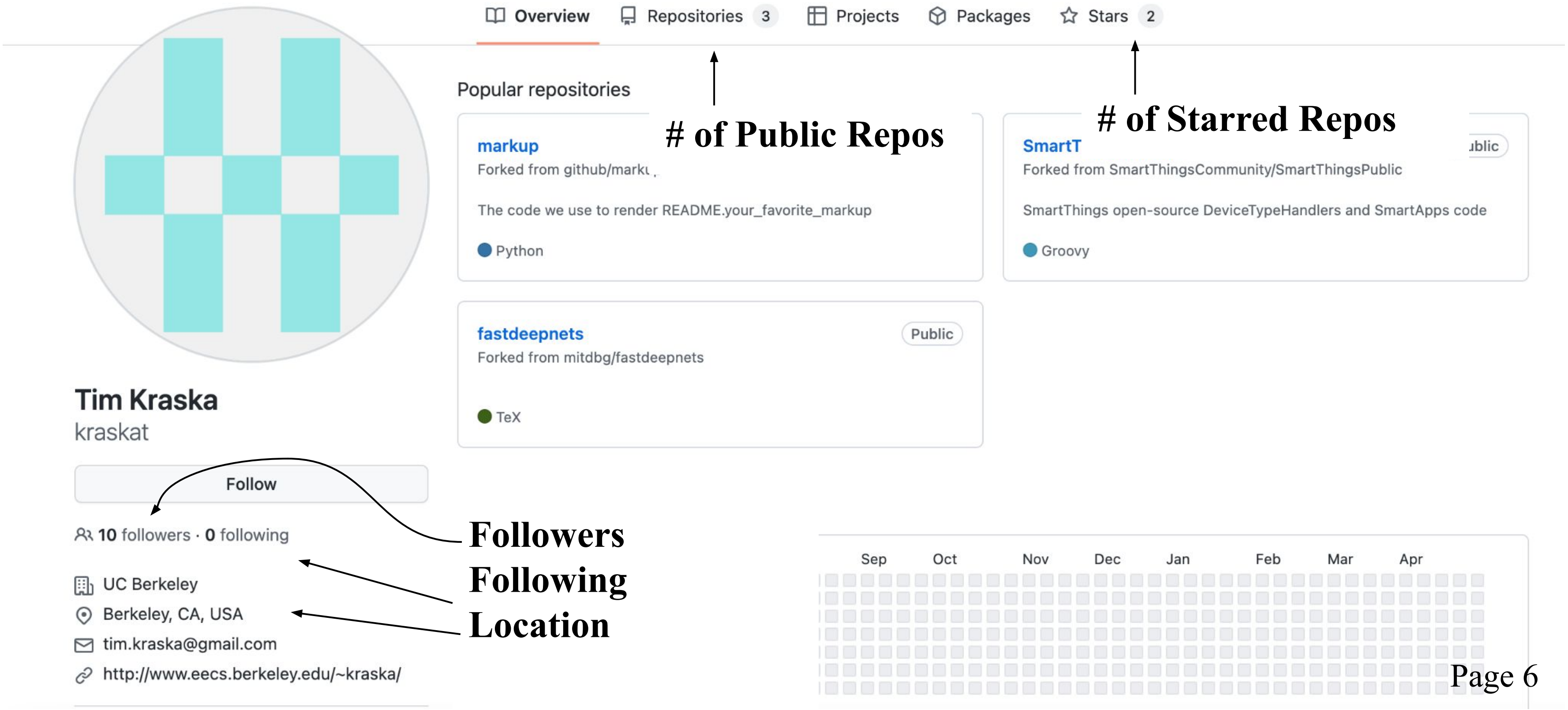
+ 6 contributors

Languages

● Jupyter Notebook 99.0% ● Other 1.0%

Data Scraping : Contributor Info

Top **10** contributors for each repository



Overview Repositories 3 Projects Packages Stars 2

Popular repositories

markup
Forked from github/markup

The code we use to render README.your_favorite_markup

● Python

SmartT
Public
Forked from SmartThingsCommunity/SmartThingsPublic

SmartThings open-source DeviceTypeHandlers and SmartApps code

● Groovy

fastdeepnets
Public
Forked from mitdbg/fastdeepnets

● TeX

Tim Kraska
kraskat

Follow

10 followers · 0 following

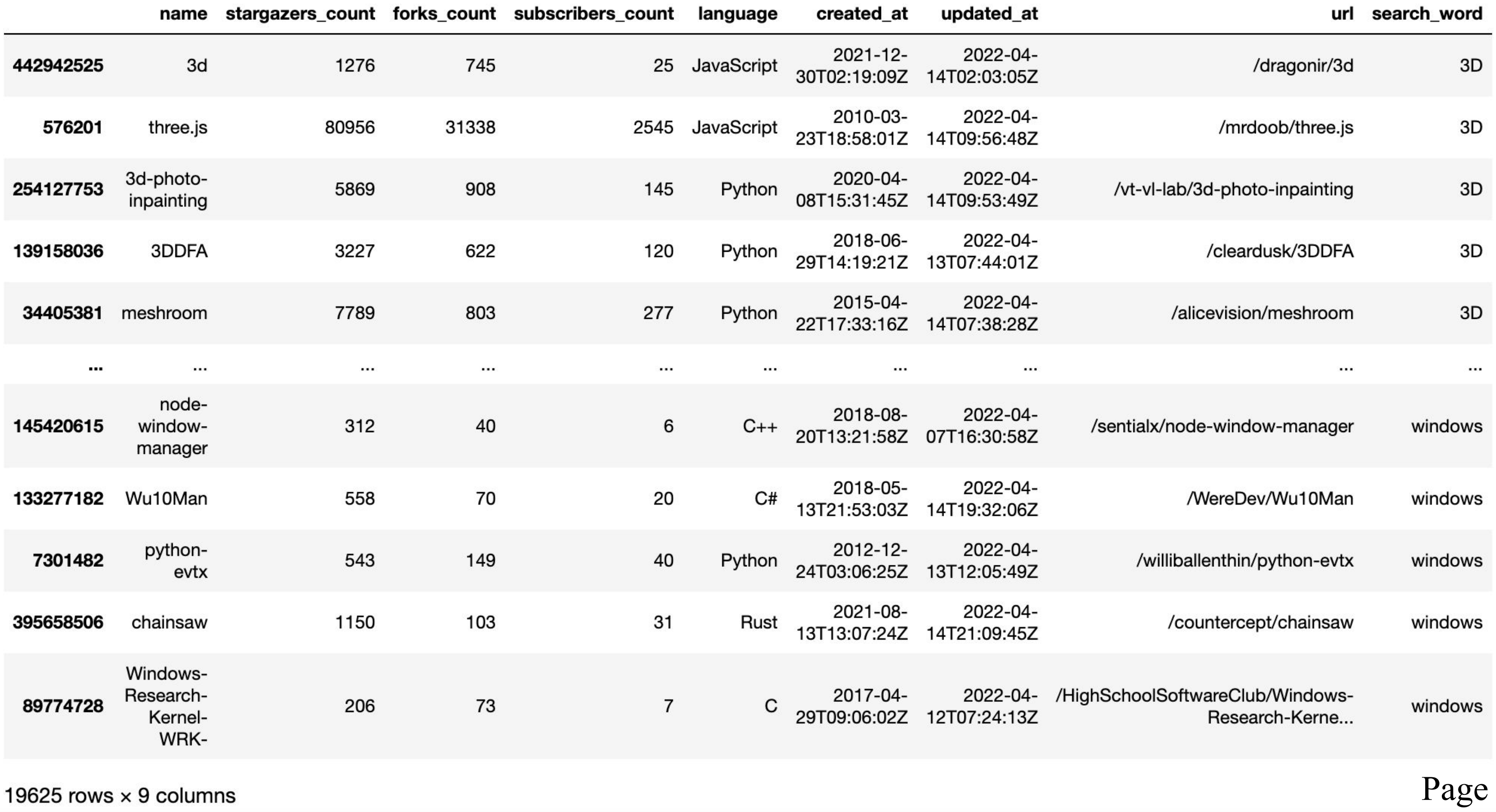
UC Berkeley
Berkeley, CA, USA
tim.kraska@gmail.com
http://www.eecs.berkeley.edu/~kraska/

# of Public Repos

# of Starred Repos

Followers
Following
Location

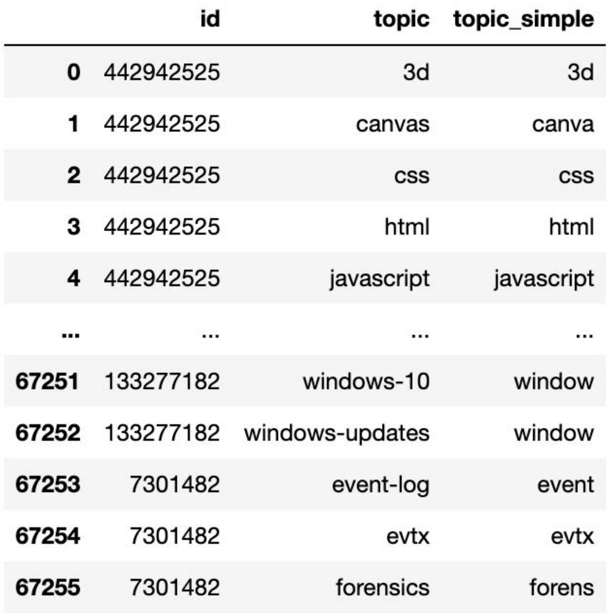
Datasets : 19625 Repositories



	name	stargazers_count	forks_count	subscribers_count	language	created_at	updated_at	url	search_word
442942525	3d	1276	745	25	JavaScript	2021-12-30T02:19:09Z	2022-04-14T02:03:05Z	/dragonir/3d	3D
576201	three.js	80956	31338	2545	JavaScript	2010-03-23T18:58:01Z	2022-04-14T09:56:48Z	/mrdoob/three.js	3D
254127753	3d-photo-inpainting	5869	908	145	Python	2020-04-08T15:31:45Z	2022-04-14T09:53:49Z	/vt-vl-lab/3d-photo-inpainting	3D
139158036	3DDFA	3227	622	120	Python	2018-06-29T14:19:21Z	2022-04-13T07:44:01Z	/cleardusk/3DDFA	3D
34405381	meshroom	7789	803	277	Python	2015-04-22T17:33:16Z	2022-04-14T07:38:28Z	/alicevision/meshroom	3D
...	...	...	...	...	...	...	...	...	...
145420615	node-window-manager	312	40	6	C++	2018-08-20T13:21:58Z	2022-04-07T16:30:58Z	/sentialx/node-window-manager	windows
133277182	Wu10Man	558	70	20	C#	2018-05-13T21:53:03Z	2022-04-14T19:32:06Z	/WereDev/Wu10Man	windows
7301482	python-evtx	543	149	40	Python	2012-12-24T03:06:25Z	2022-04-13T12:05:49Z	/williballenthin/python-evtx	windows
395658506	chainsaw	1150	103	31	Rust	2021-08-13T13:07:24Z	2022-04-14T21:09:45Z	/countercept/chainsaw	windows
89774728	Windows-Research-Kernel-WRK-	206	73	7	C	2017-04-29T09:06:02Z	2022-04-12T07:24:13Z	/HighSchoolSoftwareClub/Windows-Research-Kerne...	windows

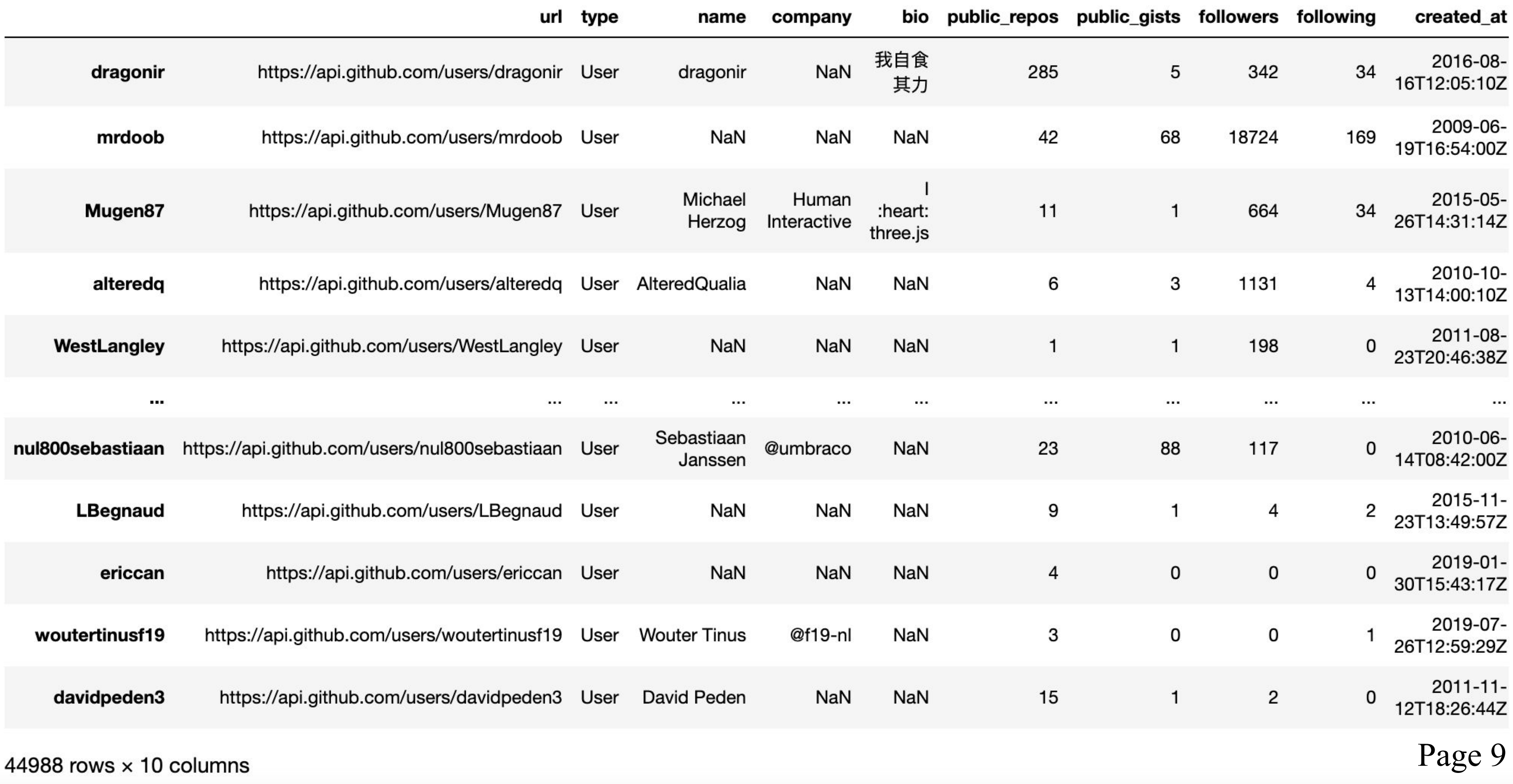
19625 rows × 9 columns

Datasets : Topic-Repo Relation Table



	id	topic	topic_simple
0	442942525	3d	3d
1	442942525	canvas	canva
2	442942525	css	css
3	442942525	html	html
4	442942525	javascript	javascript
...	...	...	...
67251	133277182	windows-10	window
67252	133277182	windows-updates	window
67253	7301482	event-log	event
67254	7301482	evtx	evtx
67255	7301482	forensics	forens

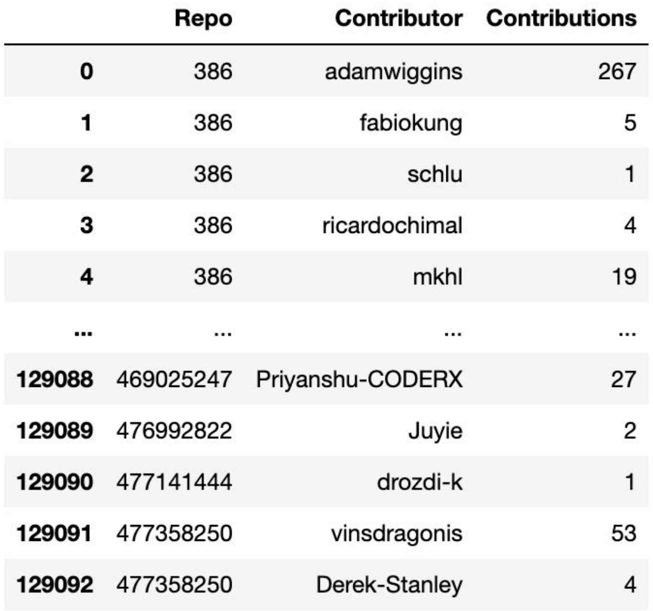
Datasets : 44988 Contributors



	url	type	name	company	bio	public_repos	public_gists	followers	following	created_at
dragonir	https://api.github.com/users/dragonir	User	dragonir	NaN	我自食其力	285	5	342	34	2016-08-16T12:05:10Z
mrdoob	https://api.github.com/users/mrdoob	User	NaN	NaN	NaN	42	68	18724	169	2009-06-19T16:54:00Z
Mugen87	https://api.github.com/users/Mugen87	User	Michael Herzog	Human Interactive	I :heart: three.js	11	1	664	34	2015-05-26T14:31:14Z
alteredq	https://api.github.com/users/alteredq	User	AlteredQualia	NaN	NaN	6	3	1131	4	2010-10-13T14:00:10Z
WestLangley	https://api.github.com/users/WestLangley	User	NaN	NaN	NaN	1	1	198	0	2011-08-23T20:46:38Z
...	...	...	...	...	...	...	...	...	...	...
nul800sebastiaan	https://api.github.com/users/nul800sebastiaan	User	Sebastiaan Janssen	@umbraco	NaN	23	88	117	0	2010-06-14T08:42:00Z
LBegnaud	https://api.github.com/users/LBegnaud	User	NaN	NaN	NaN	9	1	4	2	2015-11-23T13:49:57Z
ericcan	https://api.github.com/users/ericcan	User	NaN	NaN	NaN	4	0	0	0	2019-01-30T15:43:17Z
woutertinusf19	https://api.github.com/users/woutertinusf19	User	Wouter Tinus	@f19-nl	NaN	3	0	0	1	2019-07-26T12:59:29Z
davidpeden3	https://api.github.com/users/davidpeden3	User	David Peden	NaN	NaN	15	1	2	0	2011-11-12T18:26:44Z

44988 rows × 10 columns

Datasets : Repo-Contributor Rel Table

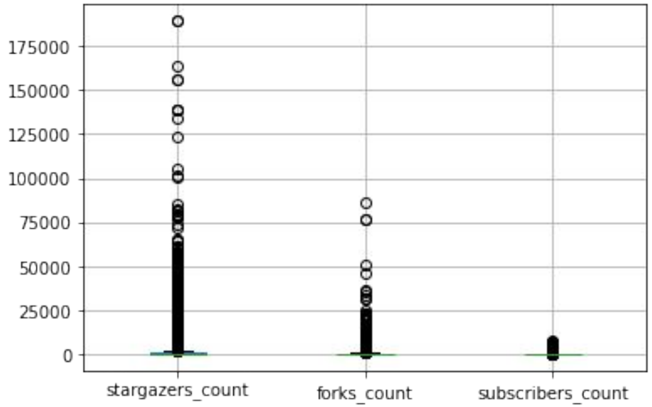


	Repo	Contributor	Contributions
0	386	adamwiggins	267
1	386	fabiokung	5
2	386	schlu	1
3	386	ricardochimal	4
4	386	mkhl	19
...	...	...	...
129088	469025247	Priyanshu-CODERX	27
129089	476992822	Juyie	2
129090	477141444	drozdi-k	1
129091	477358250	vinsdragonis	53
129092	477358250	Derek-Stanley	4

Data Cleaning

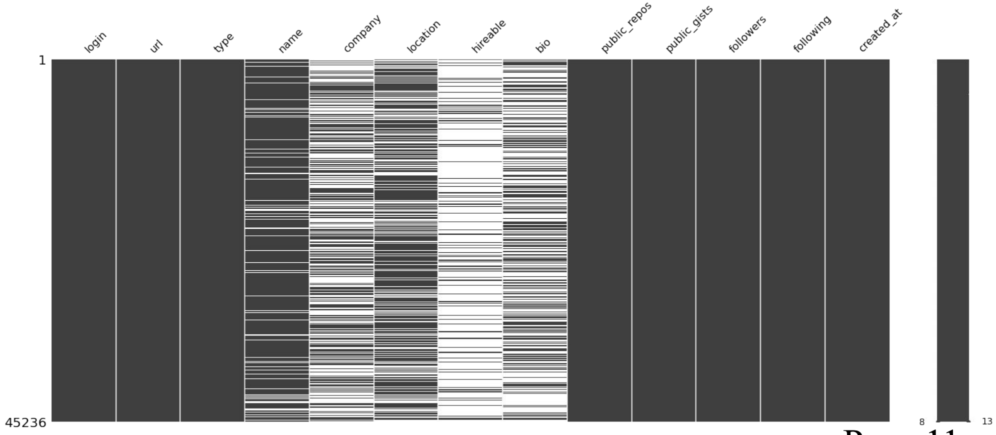
Repository Info

- Listwise deletion for Null Language values
- Verified outliers

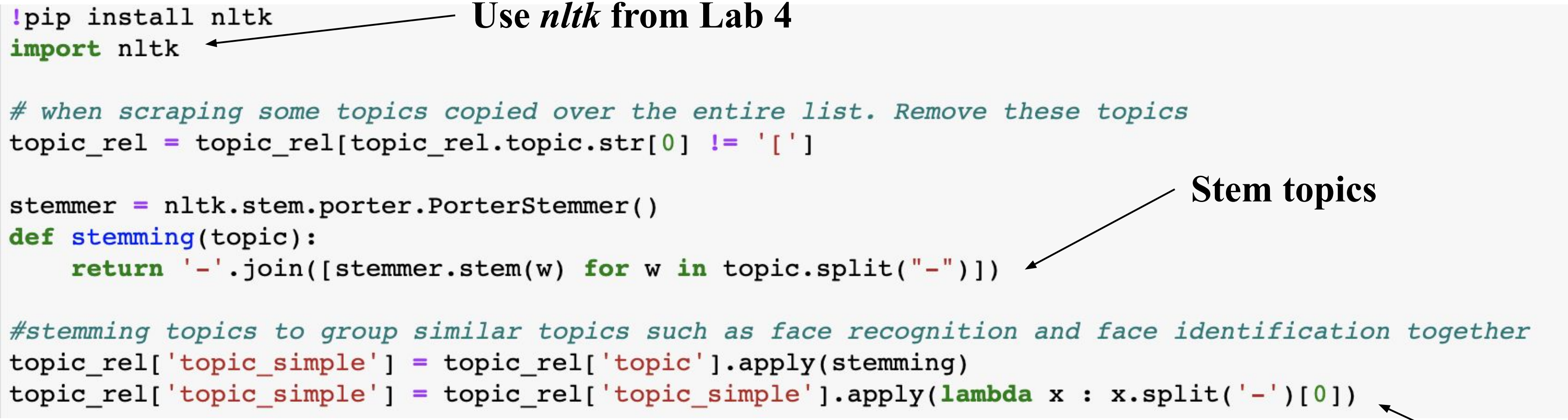


Contributor Info

- No correlation between missing values -> Pairwise deletion



Data Cleaning : Topic Merging



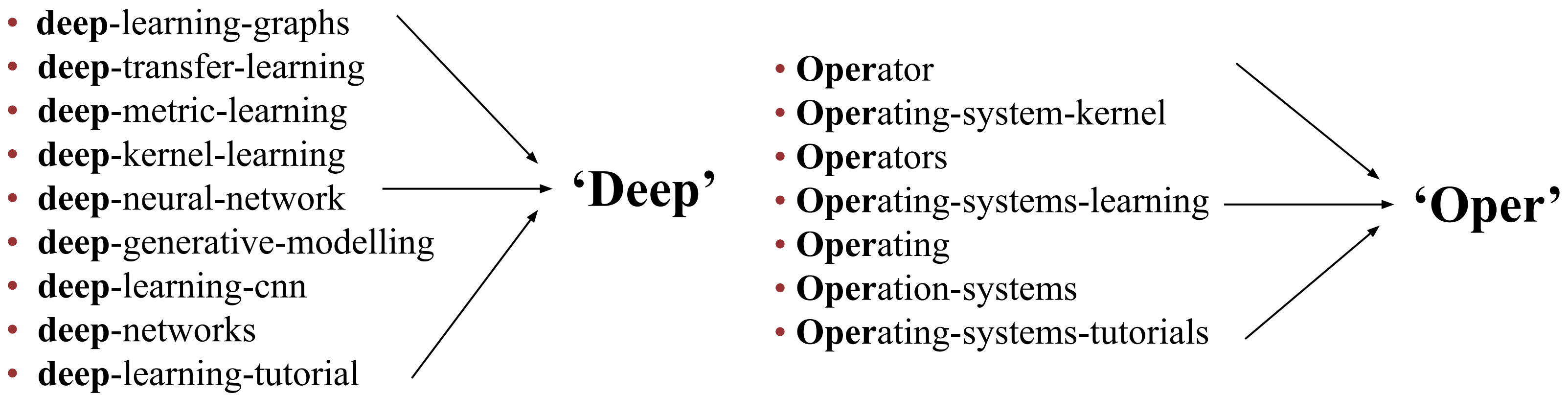
```
!pip install nltk
import nltk

# when scraping some topics copied over the entire list. Remove these topics
topic_rel = topic_rel[topic_rel.topic.str[0] != '[']

stemmer = nltk.stem.porter.PorterStemmer()
def stemming(topic):
    return '-'.join([stemmer.stem(w) for w in topic.split("-")])

#stemming topics to group similar topics such as face recognition and face identification together
topic_rel['topic_simple'] = topic_rel['topic'].apply(stemming)
topic_rel['topic_simple'] = topic_rel['topic_simple'].apply(lambda x : x.split('-')[0])
```

Use *nltk* from Lab 4

Stem topics

Use only first word of topic

20314 Topics → 10193 Topics



- **deep**-learning-graphs
- **deep**-transfer-learning
- **deep**-metric-learning
- **deep**-kernel-learning
- **deep**-neural-network
- **deep**-generative-modelling
- **deep**-learning-cnn
- **deep**-networks
- **deep**-learning-tutorial

‘Deep’

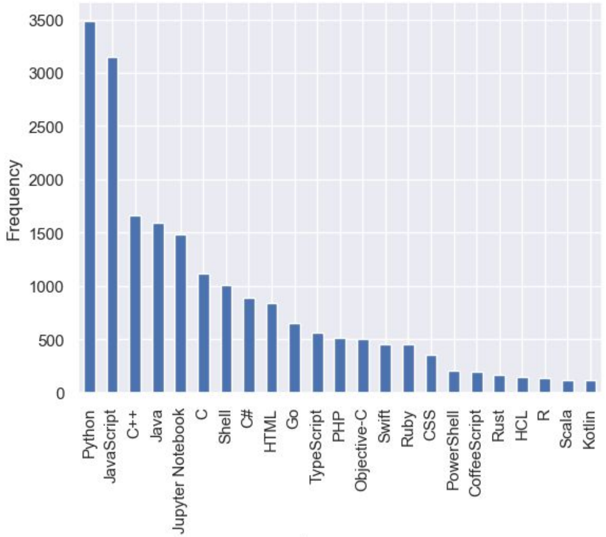
- **Oper**ator
- **Oper**ating-system-kernel
- **Oper**ators
- **Oper**ating-systems-learning
- **Oper**ating
- **Oper**ation-systems
- **Oper**ating-systems-tutorials

‘Oper’

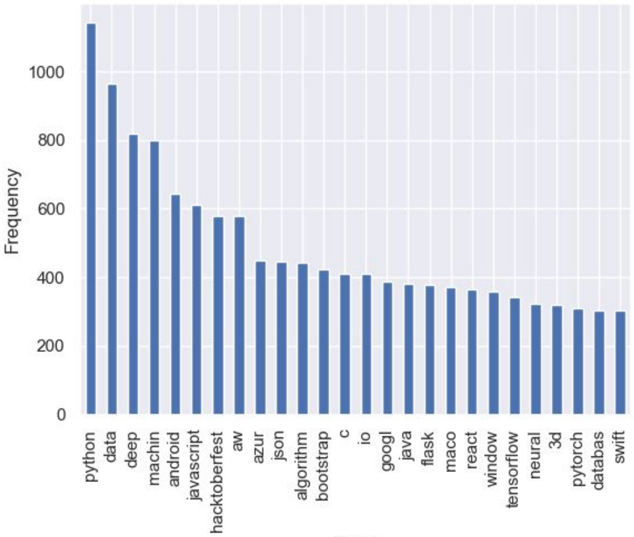
- Can’t determine conceptually topics (i.e machine-learning and ml)
- Use another similarity metric such as ML with Word2Vec embeddings

Visualizing Top Languages and Topics with Histograms

Top Languages

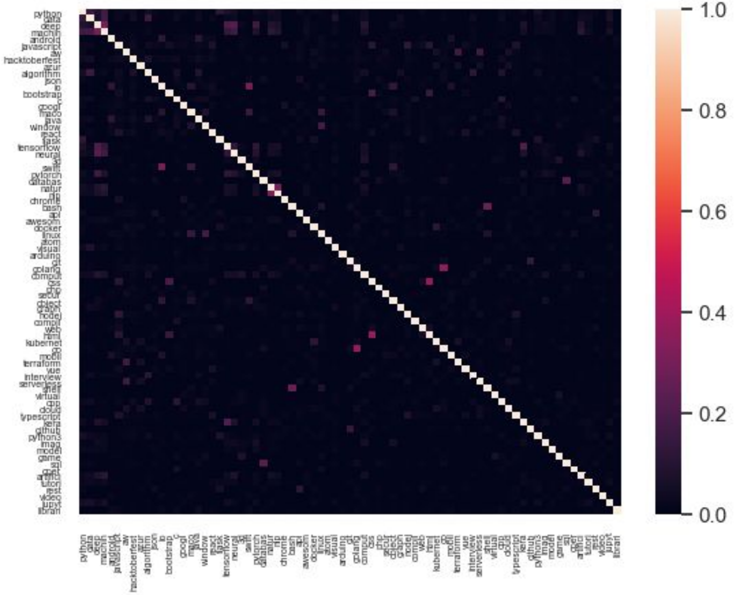


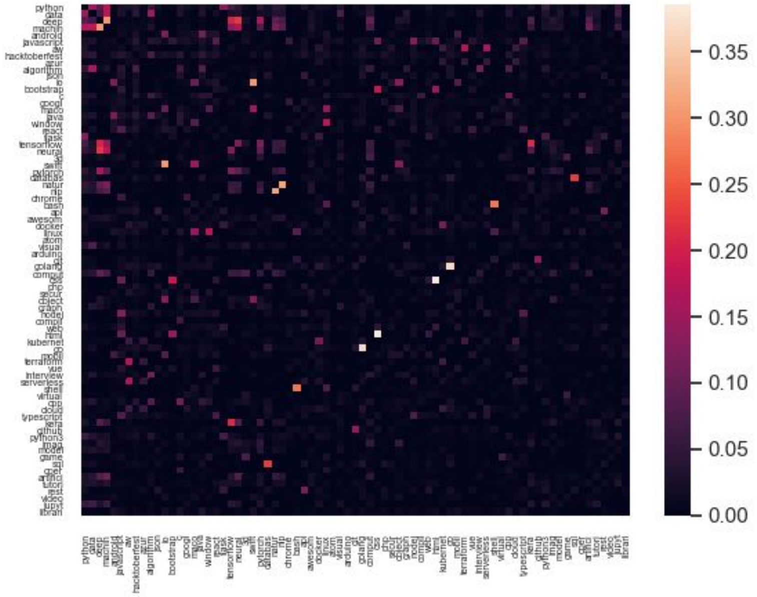
Top Topics



- Good to see general trends in GitHub popular language and topics
- Unable to see communities of languages/topics based on Repo→Contributor connections

Visualizing Jaccard Similarity between Topics based on Mutual Repos





- Remove Jaccard Similarity of 1 on diagonal to see if there are any underlying patterns
- Hard to spatially see communities of topics

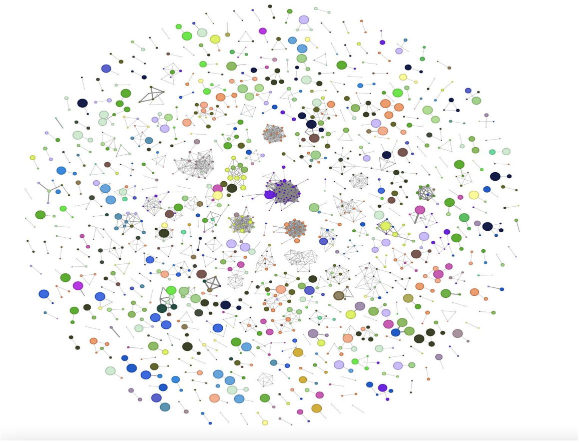
Network Mapping

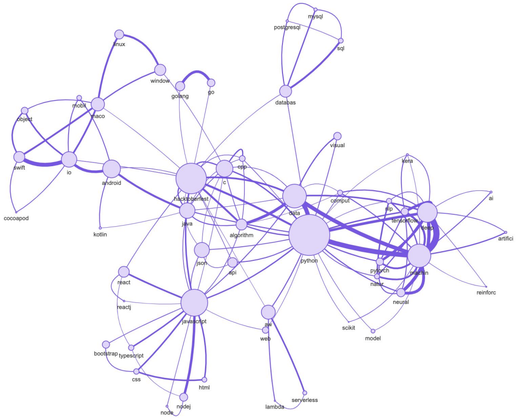
Nodes	Edges	Color
Simplified Topics	Mutual Repositories	N/A
Simplified Topics	Mutual Contributors	N/A
Repositories	Mutual Contributors	by Searchword
Repositories	Mutual Contributors	by Language

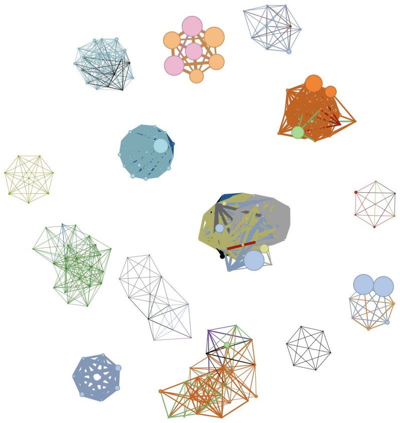
Network Visualization

NetworkX
Network Analysis in Python

PyVis : Interactive Network Visualization python package







Networks

Topic-Repo (k=2)
Topic-Repo (threshold=20)

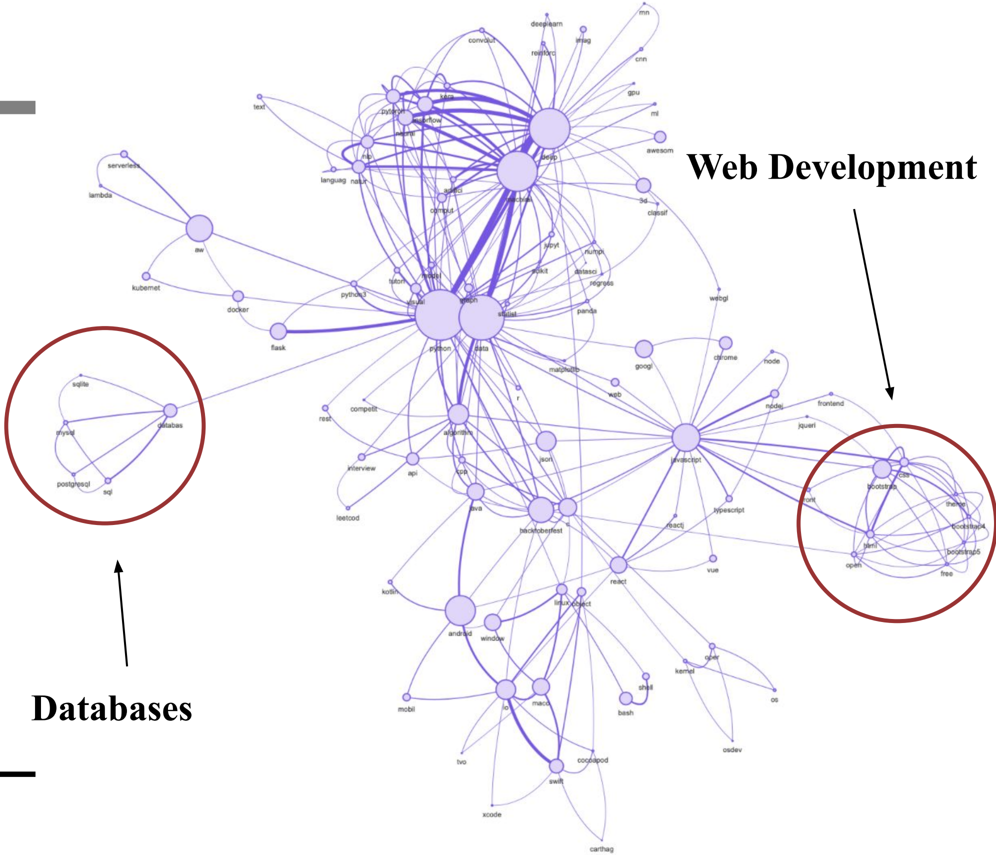
Topic-Contributor (threshold=200)
Topic-Contributor (k=4)

Repo-Contributor by Language (n=10)
Repo-Contributor by Language (n=5)

Repo-Contributor by Searchword (n=2)
Repo-Contributor by Searchword (n=10)

Patterns Found : Topic Repo

- There is one large connected component
- The top topics include python, data, machin, deep and javascript
- Tightly connected topics include python-data, machin-deep



Web Development

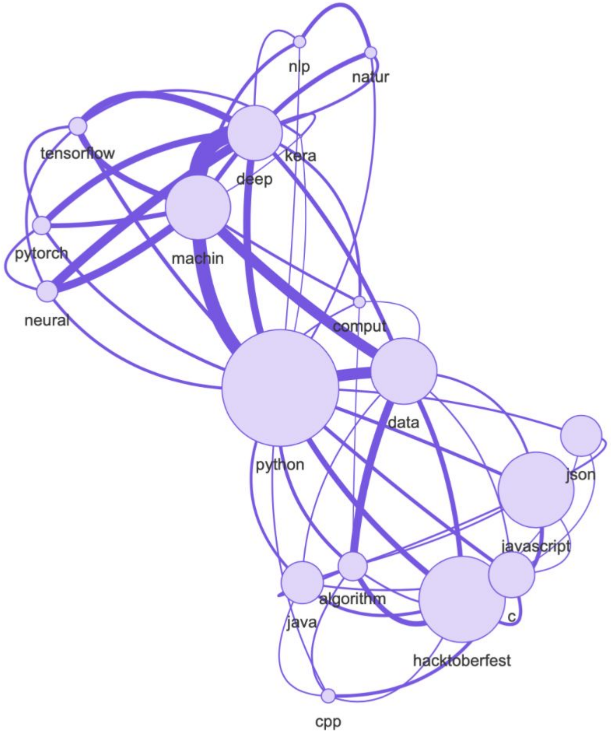
Databases

Massachusetts Institute of Technology

Patterns Found : Topic Contributor

- GitHub has tightly connected and popular community for machine learning/AI/Data science developers

K_core = 4

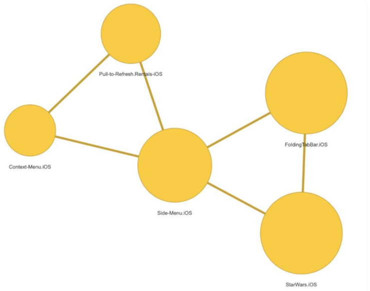


Massachusetts Institute of Technology

Patterns Found : Repo Contributor by Search words

Popular Repo and Unpopular Contributors

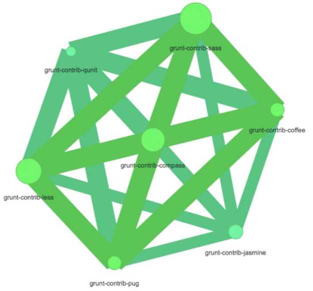
- Typically Repos part of an organization



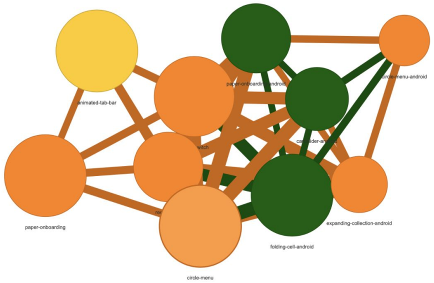
Yalantis

software engineering
and it consulting

Unpopular Repo and Popular Contributors

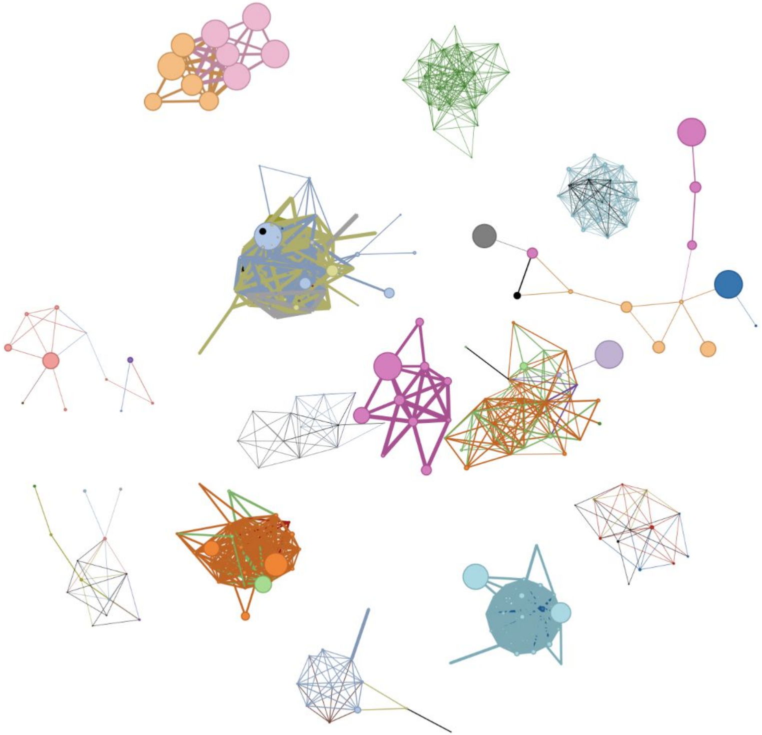


Popular Repo and Popular Contributors



Patterns Found : Repo Contributor by Languages

- Tightly connected languages: Java-Swift, C++-C, JavaScript-CoffeeScript
- Python not part of tightly connected repo communities



Future Possibilities

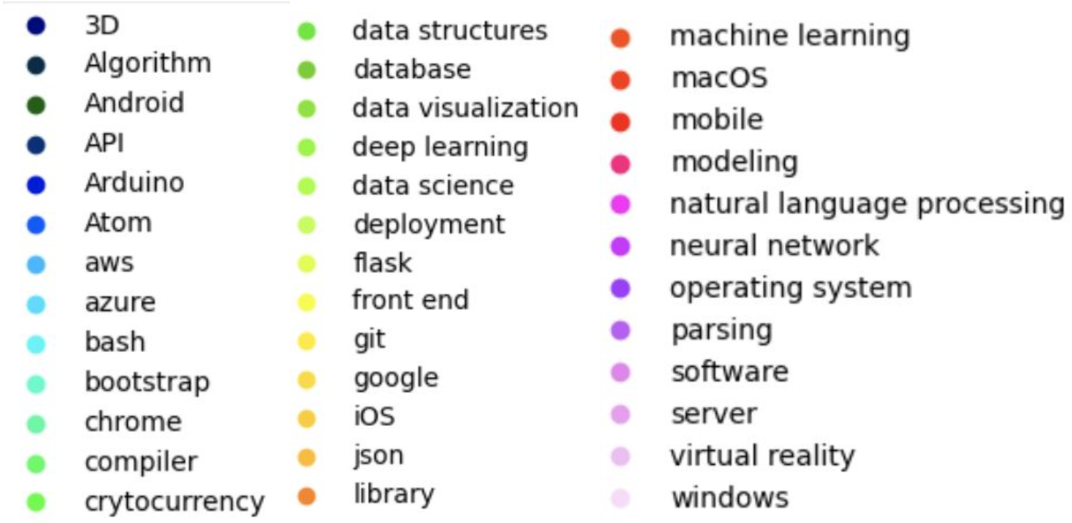
- **Optimize scraping efficiency** to get a larger sampling of the GitHub network
- **Perform network analysis algorithms** (i.e. centrality-degree and community-finding)
- **Explore other network modalities**(i.e. bipartite graphs between contributors → repositories)

Questions?

Appendix

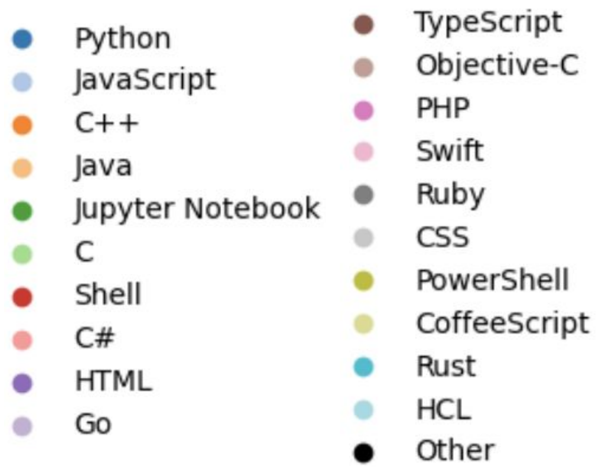
Repo-Contributor Networks Color Legends

Searchwords



- ● 3D
- ● Algorithm
- ● Android
- ● API
- ● Arduino
- ● Atom
- ● aws
- ● azure
- ● bash
- ● bootstrap
- ● chrome
- ● compiler
- ● crytocurrency
- ● data structures
- ● database
- ● data visualization
- ● deep learning
- ● data science
- ● deployment
- ● flask
- ● front end
- ● git
- ● google
- ● iOS
- ● json
- ● library
- ● machine learning
- ● macOS
- ● mobile
- ● modeling
- ● natural language processing
- ● neural network
- ● operating system
- ● parsing
- ● software
- ● server
- ● virtual reality
- ● windows

Languages



- ● Python
- ● JavaScript
- ● C++
- ● Java
- ● Jupyter Notebook
- ● C
- ● Shell
- ● C#
- ● HTML
- ● Go
- ● TypeScript
- ● Objective-C
- ● PHP
- ● Swift
- ● Ruby
- ● CSS
- ● PowerShell
- ● CoffeeScript
- ● Rust
- ● HCL
- ● Other

Topics

- 3D
- Algorithm
- Android
- API
- Arduino
- Atom
- Aws
- Azure
- Bash
- Bootstrap
- Chrome
- Compiler
- Cryptocurrency
- Data Structures
- Database
- Data Visualization
- Deep Learning
- Data Science
- Deployment
- Flask
- Front End
- Git
- Google
- iOS
- Json
- Library
- Machine Learning
- macOS
- Mobile
- Modeling
- NLP
- Neural Network
- Operating system
- Parsing
- Software
- Server
- Virtual Reality
- Windows