1

Geek Talents: Who are the Top Experts on GitHub and Stack Overflow?

Yijun Tian, Waii Ng, Jialiang Cao, Suzanne McIntosh
New York University
Courant Institute of Mathematical Sciences
yt1506, win205, jc8343, sm4971@nyu.edu

The 5th International Conference on Artificial Intelligence and Security (ICAIS 2019)
July 26-28, 2019

Overview

- Motivation
- Platform Display
- Datasets
- Application Design
- Experiments
- Conclusion

Motivation

Who are the users of this application?

- People who want to keep track of trending tech-topics and get in touch with talents related to the trending tech-topics.

Who will benefit from this application?

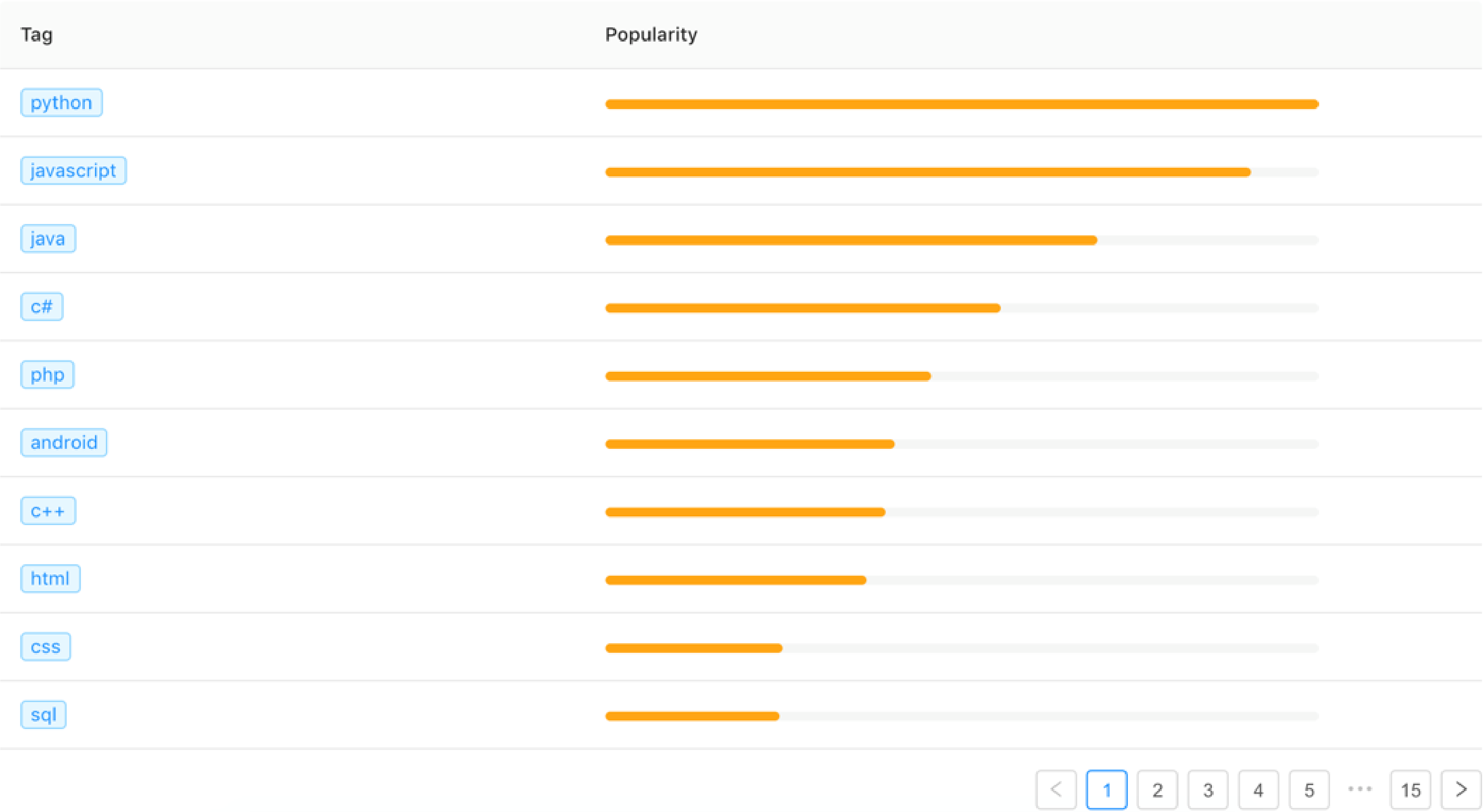
- Recruiters.
- Start-up companies.
- Everyone who wants to find and contact talents in a specific tech-domain.

Why is this application important?

- It is a novel platform that automatically identifies geek talents in specific field from GitHub, Stack Overflow and across two platforms.
- It has a new method to deal with user extraction problem, containing SO-based approach, GH-based approach as well as the approach to join them with particular weighing factor.
- It is a complete system with a carefully designed User Interface that visualize the result, which makes the exploration of large, complex user dataset easier.

4

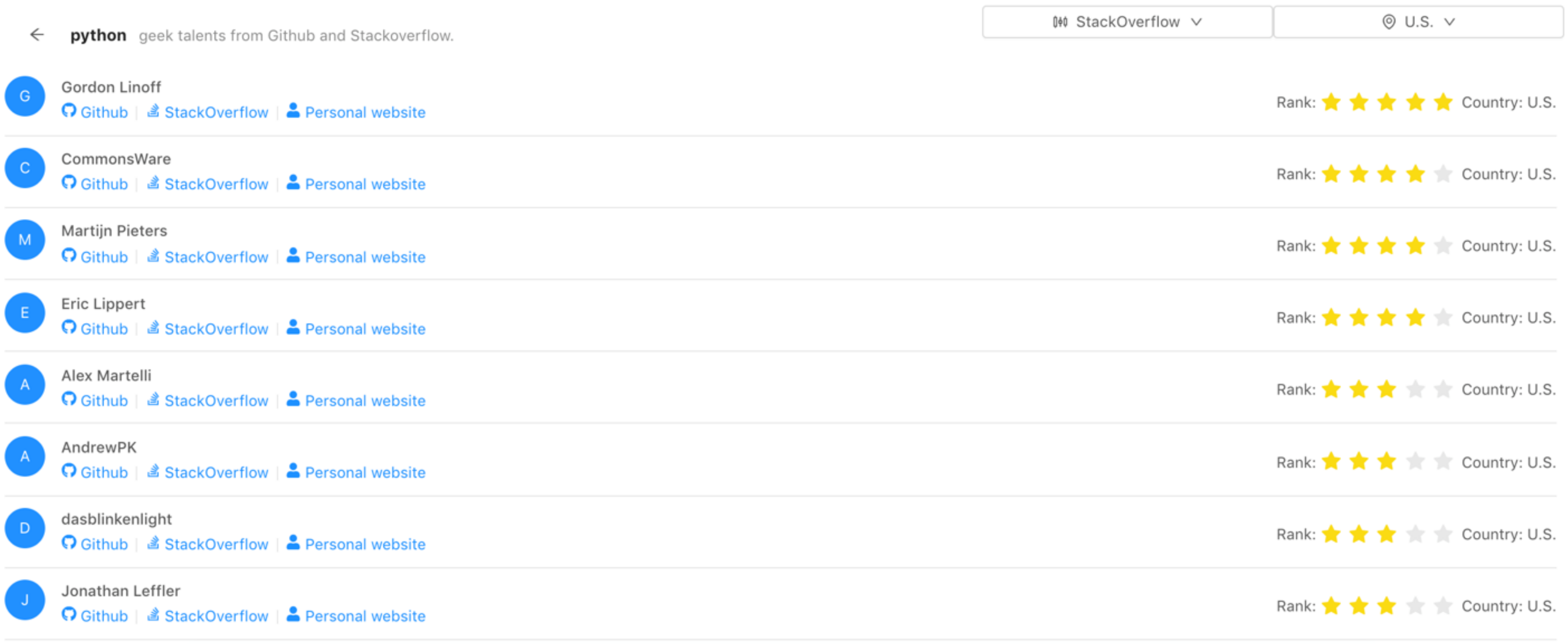
Platform Display



Tag	Popularity
python	
javascript	
java	
c#	
php	
android	
c++	
html	
css	
sql	

5

Platform Display



← **python** geek talents from Github and Stackoverflow.

StackOverflow ⌄ | U.S. ⌄

G	Gordon Linoff Github | StackOverflow | Personal website	Rank: ★★★★★	Country: U.S.
C	CommonsWare Github | StackOverflow | Personal website	Rank: ★★★★☆	Country: U.S.
M	Martijn Pieters Github | StackOverflow | Personal website	Rank: ★★★★☆	Country: U.S.
E	Eric Lippert Github | StackOverflow | Personal website	Rank: ★★★★☆	Country: U.S.
A	Alex Martelli Github | StackOverflow | Personal website	Rank: ★★★☆☆	Country: U.S.
A	AndrewPK Github | StackOverflow | Personal website	Rank: ★★★☆☆	Country: U.S.
D	dasblinkenlight Github | StackOverflow | Personal website	Rank: ★★★☆☆	Country: U.S.
J	Jonathan Leffler Github | StackOverflow | Personal website	Rank: ★★★☆☆	Country: U.S.

Datasets

- Stack Overflow Post data dump

  A post dataset containing questions and answers.

- Stack Overflow User data dump

  A user profile dataset containing user attributes.

- GitHub Search API

  Search for the specific projects we want to find under given tags.

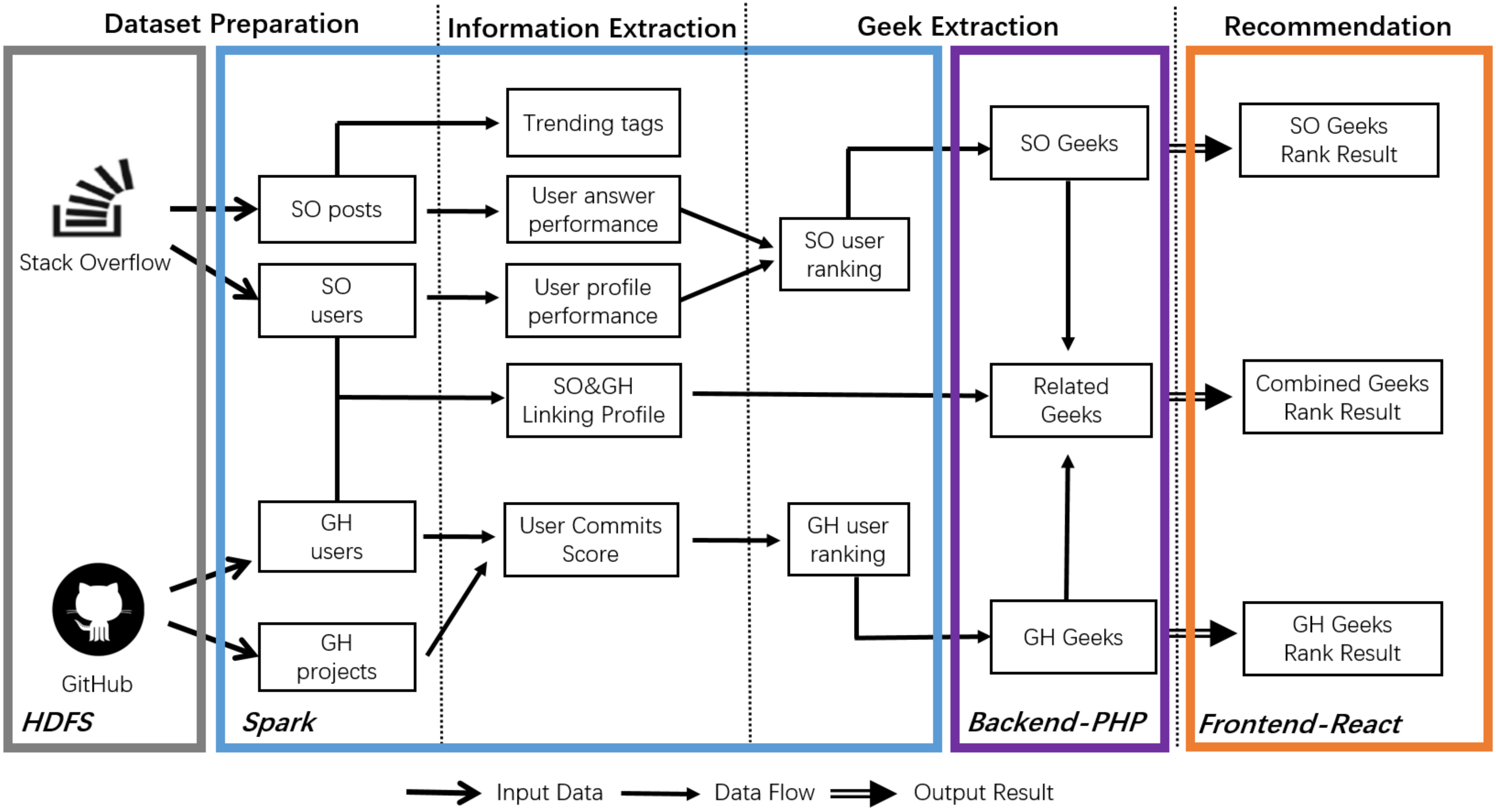
- GitHub Users API

  Get the information about currently authenticated user.

- GitHub Repository API

  Access to repositories a user owns, repositories they contribute to, and repositories that they can access through an organization membership.

Application Design



Application Design

Stack Overflow Tag Selection

$$viewCount_{tag} > 1000, \forall tag \in Set_{Tag} \qquad (1)$$

$$viewCount_{tag} = \sum_{i=1}^{postNum_{tag}} viewCount_{post_i} \qquad (2)$$

$viewCount_{tag}$: the popularity of a tag
$viewCount_{post}$: the times a post been viewed
$postNum_{tag}$: the number of post under this tag

Application Design

Stack Overflow Expert Recommendation

$$S_{userSO} = 0.6 \times S_{UAP} + 0.4 \times S_{UPP} \qquad (3)$$

$$S_{UAP} = 0.5 \times S_{ans} + 0.5 \times S_{question} \qquad (4)$$

$$\begin{aligned} S_{question} &= 0.3 \times avgViewCount \\ &+ 0.3 \times avgFavoriteCount \\ &+ 0.3 \times avgAnsCount \end{aligned} \qquad (5)$$

$$S_{UPP} = 0.7 \times reputation + 0.3 \times viewCount_{user} \qquad (6)$$

S_{userSO}: score of user in Stack Overflow
S_{UAP}: score of user answer performance
S_{UPP}: score of user profile performance
S_{ans}: score of answer
$S_{question}$: score of question

Application Design

GitHub Expert Recommendation

$$S_{userGH} = \sum_{repo=0}^{repoNum} S_{repo,user} \tag{7}$$

$$S_{repo,user} = commits_{repo,user} \times S_{repo,PC} \tag{8}$$

$$S_{repo,PC} = Weight_{tag} \times \frac{watchers_{repo}}{commits_{repo}} \tag{9}$$

$$Weight_{tag} = \frac{BOC_{tag}}{BOC_{total}} \tag{10}$$

S_{userGH}: score of user in GitHub $S_{repo,user}$: the contributing score of user to repo
$S_{repo,PC}$: score of per commit of repo BOC: byte of code

Experiments

User Extraction

We extracted 1,295,622 users from SO users dataset and 7,953,512 users from GH users.

User Linking

We linked 332,362 users for our combined expert recommendation, including 309,735 using the hashed email address method and 28,294 using the GH username method.

Tag Selection

we extracted the latest month post tags from SO post dataset and found 12,370 different tags. Selecting those with posts been viewed by more than 1,000 times.

Experiment

User under Tag

2,548,505 pairs of (user, tag) tuples have been generated in GitHub under language tag, compared to 56,889 pairs of (user, tag) tuples under topic tag.

Regional Selection

There are 106,263 different cities in SO set, 31,205,585 different cities in GH set and 18,858 same cities between SO and GH. Therefore, we use ISO-3166-Countries-with-Regional-Codes table to map the relationship between country code and city names.

Analysis

We manually checked the profiles of the top hundred talents under the top ten trending topics. It turned out that:

(1) Their profiles are highly related to the given topic.

(2) The accuracy of cross platform profiles linking is high

Conclusion

- We addressed the problem of user recommendation in GitHub, Stack Overflow, and across both platforms.
- We proposed a novel methodology to deal with the user extraction problem, which make full use of different user attributes and related platform specific information.
- We build a complete system with a carefully designed User Interface that visualize the result, which makes the exploration of large, complex user dataset easier.

Q & A

THANKS FOR LISTENING

References

Balachandran, V. (2013): Reducing human effort and improving quality in peer code reviews using automatic static analysis and reviewer recommendation. *2013 35th International Conference on Software Engineering*, pp. 931-940.
Bosu, A.; Corley, C.S.; Heaton, D.; Chatterji, D.; Carver, J.C. et al. (2013): Building reputation in stackoverflow: an empirical investigation. *2013 10th Working Conference on Mining Software Repositories*, pp. 89–92.
Chen, C.; Xing, Z. (2016): Towards correlating search on google and asking on stack overflow. *2016 IEEE 40th Annual Computer Software and Applications Conference*, vol. 1, pp. 83–92.
Constantinou, E.; Kapitsaki, G.M. (2016): Identifying developers' expertise in social coding platforms. *2016 42th Euromicro Conference on Software Engineering and Advanced Applications*, pp. 63–67.
Hu, Y.; Wang, S.; Ren, Y.; Choo, K.K.R. (2018): User influence analysis for github developer social networks. *Expert Systems with Applications*, pp. 108-118.
Huang, W.; Mo, W.; Shen, B.; Yang, Y.K.; Li, N. (2016): Cpdscorer: Modeling and evaluating developer programming ability across software communities. *28th International Conference on Software Engineering and Knowledge Engineering*. pp. 87-92.
Kleinberg, J.M. (1999): Authoritative sources in a hyperlinked environment. *Journal of the ACM*, vol. 46, no. 5, pp. 604-632.
Liao, Z.; Jin, H.; Li, Y.; Zhao, B.; Wu, J. et al. (2017): Devrank: Mining influential developers in github. *2017 IEEE Global Communications Conference*. pp. 1-6.
Mao, K.; Yang, Y.; Wang, Q.; Jia, Y.; Harman, M. (2015): Developer recommendation for crowdsourced software development tasks. *2015 IEEE Symposium on Service-Oriented System Engineering*. pp. 347–356.
Montandon, J.E.; Silva, L.L.; Valente, M.T. (2019): Identifying experts in software libraries and frameworks among github users. *CoRR abs/1903.08113*, http://arxiv.org/abs/1903.08113.
Movshovitz-Attias, D.; Movshovitz-Attias, Y.; Steenkiste, P.; Faloutsos, C. (2013): Analysis of the reputation system and user contributions on a question answering website: stackoverflow. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. pp. 886–893.
Page, L.; Brin, S.; Motwani, R.; Winograd, T. (1998): The pagerank citation ranking: bringing order to the web. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

References (cont.)

Silvestri, G.; Yang, J.; Bozzon, A.; Tagarelli, A. (2015): Linking accounts across social networks: the case of stackoverflow, github and twitter. *1st International Workshop on Knowledge Discovery on the WEB*. pp. 41-52.

Sumanth, P.; K, R. (2018): Discovering top experts for trending domains on stack overflow. *8th International Conference on Advances in Computing Communications*. vol. 143, pp. 333-340.

Tsay, J.; Dabbish, L.; Herbsleb, J. (2014): Influence of social and technical factors for evaluating contribution in github. *Proceedings of the 36th International Conference on Software Engineering*. pp. 356–366.

Vasilescu, B.; Filkov, V.; Serebrenik, A. (2013): Stackoverflow and github: associations between software development and crowdsourced knowledge. *2013 International Conference on Social Computing*. pp. 188-195.

Venkataramani, R.; Gupta, A.; Asadullah, A.; Muddu, B.; Bhat, V. (2013): Discovery of technical expertise from open source code repositories. *Proceedings of the 22nd International Conference on World Wide Web*. pp. 97–98.

Wang, Z.; Sun, H.; Fu, Y.; Ye, L. (2017): Recommending crowdsourced software developers in consideration of skill improvement. *2017 32nd IEEE/ACM International Conference on Automated Software Engineering*. pp. 717–722.

Xuan, J.; Jiang, H.; Ren, Z.; Zou, W. (2012): Developer prioritization in bug repositories. *2012 34th International Conference on Software Engineering*. pp. 25-35.

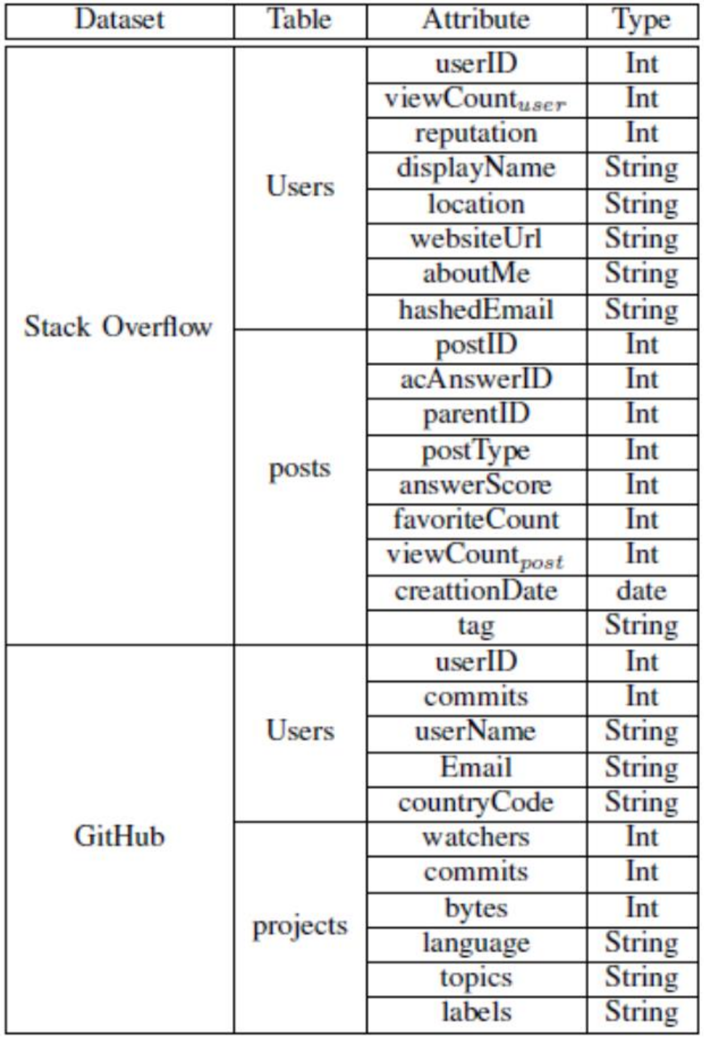
Yu, Y.; Wang, H.; Yin, G.; Ling, C.X. (2014): Reviewer recommender of pull-requests in github. *2014 IEEE International Conference on Software Maintenance and Evolution*. pp. 609-612.

Zhang, J.; Ackerman, M.S.; Adamic, L. (2007): Expertise networks in online communities: structure and algorithms. *Proceedings of the 16th International Conference on World Wide Web*. pp. 221–230.

Zhang, X.; Wang, T.; Yin, G.; Yang, C.; Yu, Y. et al. (2017): Devrec: a developer recommendation system for open source repositories. *Mastering Scale and Complexity in Software Reuse*, pp. 3–11.

Appendix

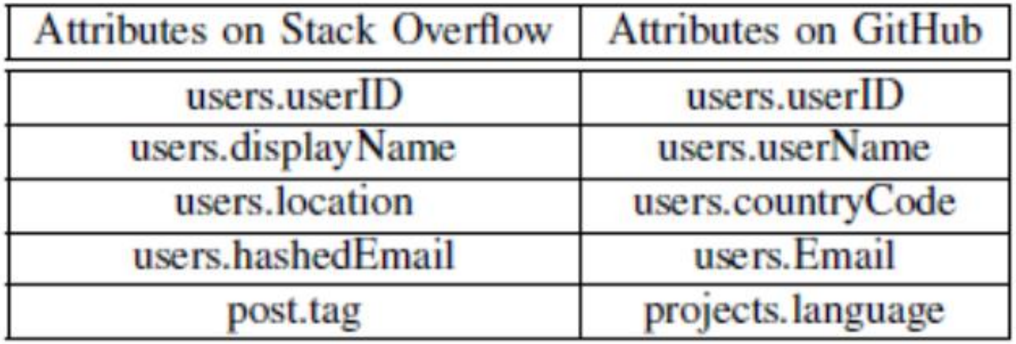
Table 1:selected attributes and content types for each dataset.



Dataset	Table	Attribute	Type
Stack Overflow	Users	userID	Int
		viewCount$_{user}$	Int
		reputation	Int
		displayName	String
		location	String
		websiteUrl	String
		aboutMe	String
		hashedEmail	String
	posts	postID	Int
		acAnswerID	Int
		parentID	Int
		postType	Int
		answerScore	Int
		favoriteCount	Int
		viewCount$_{post}$	Int
		creattionDate	date
		tag	String
GitHub	Users	userID	Int
		commits	Int
		userName	String
		Email	String
		countryCode	String
	projects	watchers	Int
		commits	Int
		bytes	Int
		language	String
		topics	String
		labels	String

Appendix

Table 2: Mapping between Stack Overflow and GitHub



Attributes on Stack Overflow	Attributes on GitHub
users.userID	users.userID
users.displayName	users.userName
users.location	users.countryCode
users.hashedEmail	users.Email
post.tag	projects.language