quoras: A Python API for Quora Data Collection to Increase Multi-Language Social Science Research

Dipto Das
ddas05@syr.edu
Syracuse University
Syracuse, New York

Bryan Semaan
bsemaan@syr.edu
Syracuse University
Syracuse, New York

ABSTRACT

Quora[1] is a fast growing crowdsourced Q/A site that also creates online social networks and community practices among the users. Operating in several regional languages, it catalyzes more contextual discussions on local incidents and issues. To understand how language-specific social communities conduct Q/A-based discussions on online forums, we need to study Quora platform. As the first step to that, we need a data collection API. We introduce **quoras**, a Python API for collecting data from Quora. The API relies on Selenium [3], which is an open-source cross platform web automation framework. The API operates by creating custom HTTPS requests to Quora and parsing responses from it. It has the ability to perform many types of advanced searches that are otherwise only available on the Quora website, and not through any other existing APIs. The **quoras** API is released under an open-source MIT license and available along with the full API reference on GitHub[2]. The latest stable release is also available on Python Package Index (PyPI).

[1]https://www.quora.com/

[2]https://github.com/DiptoDas8/quoras

CSCW '20 Companion, October 17–21, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8059-1/20/10.
https://doi.org/10.1145/3406865.3418333

INTRODUCTION

With the rise of scholarship focusing on better understanding social media interactions (e.g. on Twitter), scholars have increasingly developed toolkits for collecting and analyzing this data. Given the broad dissemination and adoption of these platforms globally, people from across the world are beginning to use social media to interact for myriad reasons. Research on social media has found that people use these platforms for information seeking, social interaction, and expression of opinions [21, 29]. While social media has become a space for people to engage with one another and discuss potentially sensitive topics and issues [14, 18, 29], social media platforms with known identity, e.g. Facebook, do not often support these practices because of the presence of family and friends [6, 16]. This can motivate people to adopt and/or use discussion forums or question-and-answer (Q/A) sites like Reddit, Stack Exchange, and more.

Though increased attention has been paid to moderated online community spaces like Reddit [11, 12], these platforms are typically designed around the English language, and do not always support people who wish to engage in discourse using their native language. According to prior work, forums that better support users' native languages encourage more engaging discussions [1, 13]. In this paper, we focus on a moderated online discourse community that supports a range of languages– Quora. Quora is a Q/A website with an embedded social network structure and practice of following users. The community structure of Quora is organized around questions and emergent discussions centered around those questions. Though the platform welcomes Q/A threads on various topics, it encourages discussion-based questions that promotes exchange of opinions. Users of this platform can participate in these discussions–ask, answer, and edit questions. However, any internet user, irrespective of having an account on Quora, can view the questions and answers from web links, subject to a daily limit on each browser session. Currently, Quora has forums in 17 languages [8]. These forums are available under the Quora domain, using corresponding two-character language codes as the subdomain ids[3].

Though Quora as a Q/A forum has been studied to some extent [17, 28], the research community are yet to explore its language-dedicated forums. In addition to how platforms, like Quora, often change data collection policies, most APIs and machine learning tools are not designed for non-English languages. We believe this has created a dearth of scholarship exploring the growing uses of online community spaces amongst non-native English speaking populations. Tools that can help researchers collect data in languages other than English would allow researchers access to data that would better inform the design of sociotechnical systems that better accomodate different cultural and linguistic communities and contexts. Quora does not provide an official API for collecting data from its platform. However, their terms of services allow collecting data from their platform with proper credits to the users and the Quora platform [9]. On the other hand, the Python scripting language has demonstrated

[3]For example, Bengali (language code: 'bn') Quora can be accessed at: http://bn.quora.com/.

IMPLEMENTATION

There are three main classes in **quoras** that have various roles: (a) Quoras, (b) Browser, and (c) Scraper. The API user can only interact with the library through the functions available under the Quoras class. The API user has to call the constructor method of this class with their account credentials and optional language code of the forum which is set to 'en' by default. Initialization of an instance of this class creates an instance of the Browser class. Then, it calls the login() method of the browser instance with the provided credentials. This creates a web browser session that initiates at Quora's base URL and logs into Quora. Then, it sets the domain URL based on the value of language code parameter. For example, the browser goes to the Bengali Quora forum (https://bn.quora.com/), if the language code in the Quoras constructor method was set to 'bn'. This instance of Browser class models the coordination, sequencing, transactions of data from the Quora platform. The methods offered by our API falls into two broader categories: phrase-based searching and URL-based searching. In case of phrase-based searching, the browser instance makes GET requests with custom queries based on the search phrases and the types. The Selenium web driver (e.g. ChromeDriver) executes scripts on the loaded page to scroll it a certain times, specified by the user. For the URL-based searches, the browser makes GET requests with the given URLs. The final loaded page in cases of both phrase-based query and URL search, are sent to the Scraper class which parses the page and then sends the data objects back to Quoras class and the data collection methods return the data values to the user.

usefulness in unifying different data sources and analysis tools [5, 23, 25]. Among these tools, there are some independently developed Python Quora APIs available [4, 15, 19, 24]. However, most of them are out of date [15, 19, 24] and none of these available APIs supports access to non-English language-based forums. Moreover, these available APIs cannot collect full answers from the JavaScript-enabled web pages of Quora. We developed a selenium-based Python API for collecting full Q/A threads' data from Quora along with support for other languages besides English. In this poster, we describe the development of the **quoras** API, and explain how scholars can use it.

RELATED WORKS

Social networking sites and Q/A forums have become increasingly important in advancing social science research. Whereas a lot of work exists exploring the design of data collection tools for more popular platforms like Twitter and Stack Exchange, there is a dearth of literature exploring Quora or in developing an API to help with social science research. Prior work on Quora has explored its community practices and network structures via manually collected datasets and web crawlers [17, 28]. In other cases, authors have not described their data collection approach [2, 22]. Given that scholars often have different research goals, the data collection methods have been constrained by the nature of any given Quora related project. Whereas Quora users generate several topics for discussion, prior approaches have relied on research teams developing a predefined list topics to inform the search queries that would generate the data corpus [20], which does not allow for exhaustive data exploration as users may utilize a range of terms that may not be identified by the research team to describe a topic or phenomenon. Similarly, work by Wang [17, 28] started with a randomly selected set of questions and used a breadth first search (BFS) through the related questions links to collect their datasets, which is biased by Quora's recommendation algorithm [30]. Q/A threads often cover several topics and include multiple user-assigned tags. Moreover, previous toolkits have used Ruby on Rails [17, 28] which limits their utility amongst social science scholars who may not have programming expertise [10, 27], or may be more familiar with Python [7, 26]. This suggests that prior data collection methods are limiting the kinds of data that can subsequently be used for social science inquiry, and our work seeks to address this gap. We argue that our API will encourage computational social scientists to study Quora, especially its diverse regional discussions by providing multilingual support and making data collection easier. We explain the implementation of **quoras** in the column on the left.

USAGE

The **quoras** API is available on PyPI and can be installed by running the following command: `pip install -U quoras`. The API user needs to have a Quora account. To collect data from the platforms in other languages, the users need to have those languages added to their profiles. We have used Chrome

[4]Downloadable from: https://sites.google.com/a/chromium.org/chromedriver/home

Table 1: quoras Data collection functions.

Function	Description
`search (phrase, type='post', scroll_count=1)`	Searches for posts, topic RSS feeds or users containing specified phrase and returns a list of URLs. The valid values for the parameter `type` are: 'post', 'topic', and 'user'. For these values, this function searches Quora for keyword matching.
`search_topic (topic, scroll_count=1)`	Searches for Q/A threads with exact user-assigned topic tags.
`search_url (url)`	Searches details about a Q/A thread or a user. If a URL to a Q/A thread is provided, it returns a dictionary containing question, user-assigned topics, participating users' IDs, answers' URLs, and Quora-recommended related questions' URLs. If a URL to a user profile is provided, a dictionary containing the user's statistics (follower count, following count, numbers of posts, questions, answers, and shares) and the URLs to the user's top posts is returned.
`get_full_answer (url)`	Searches for an answer with a URL, retrieves the full text of the answer, if available, and returns it.

```
from quoras import Quoras
eq = Quoras('user-email-address','password','en')
res0 = eq.search('history', 'post', scroll_count=3)
res1 = eq.search_topic('finance', scroll_count=1)
bq = Quoras('user-email-address','password','bn')
res2 = bq.search_url('https://bn.quora.com/চীন-কিভাবে-তিব্বত-দখল')
res3 = bq.get_full_answer('https://bn.quora.com/\
                          বিজ্ঞানীদের-মধ্যেও-কি/answers/150612153')
```

Figure 1: Example quoras workflow in an iPython notebook.

web driver for our implementation and expect the users to have Google Chrome installed on their systems. We recommend that the API user's working directory has a subfolder called "chrome_path" containing the web driver executable[4].

To collect large volume of data avoiding Quora's per-session limit on anonymous browsing, **quoras** requires the users to login by calling the constructor method with their account credentials as: `quora = Quora (email, password, language='en')`. The currently supported languages by the API are: English ('en'), Bengali ('bn'), French ('fr'), Hindi ('hi'), Japanese ('jp'), and Spanish ('es'). We can call different data collection functions to make queries. The **quoras** API can query the Quora platform for posts, i.e., Q/A threads, topic RSS feeds, and users using search phrases. It can also retrieve details about any Q/A thread from its URL, including question, full answers, participating users, and related questions. Additionally, **quoras** can collect users' top posts and statistics (e.g., followers-following count) from their user IDs. Table 1 shows the data collection functions available in **quoras**. An example code illustrating the uses of different data collection functions from English and Bengali forums of Quora is shown in Figure 1 and is available at the project's GitHub repository.

Among the other available packages, [15, 19, 24] do not work at present. The only other working package as off now is [4] which only allows users to search on Quora with keywords. However, it cannot retrieve the full text of the discussions in the threads, besides lacking support for several other common social media data collection use cases. Our proposed **quoras** library solves that issue and can fetch the full available data in a thread. Additional to that, our package can collect the related Q/A threads suggested by Quora's recommendation algorithm [30]; identify the participating users and get their basic statistics which are some commonly used functionalities of any social media API. Moreover, **quoras** has multilingual support, which none of the existing Quora data collection packages offers [4, 15, 19, 24].

CONCLUSION AND FUTURE WORK

Over time, the API functionalities can be extended to collect multimodal data, reply thread data, and perform adaptive switching among forum languages. The full code is available on GitHub, and we encourage forks and pull requests for extending data collection features and support for the Quora forums in other languages. We believe that a tool like **quoras** will enable more broadscale social science research in understanding other cultures and cultural contexts, given that many of Quora's users are using their native languages. Our platform was built to encourage more work that better understands the use of social media in non-English contexts, and given the breadth of languages supported by Quora, we believe this is fantastic space to explore how people from different cultures and contexts engage in discourses of importance, how different cultures view anonymity, and how cultural and language inform how people present and discuss topics, just to name a few.

REFERENCES

[1] Pushkal Agarwal, Kiran Garimella, Sagar Joglekar, Nishanth Sastry, and Gareth Tyson. 2020. Characterising User Content on a Multi-lingual Social Network. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 2–11.

[2] Ahmad Aghaebrahimian. 2017. Quora question answer dataset. In *International Conference on Text, Speech, and Dialogue*. Springer, 66–73.

[3] Satya Avasarala. 2014. *Selenium WebDriver practical guide*. Packt Publishing Ltd.

[4] Zacchaeus Bolaji. 2019. quorapy. https://github.com/djunehor/quorapy. PyPI: https://pypi.org/project/quorapy/, last accessed: June 18, 2020.

[5] Marco Bonzanini. 2016. *Mastering social media mining with Python*. Packt Publishing Ltd.

[6] Petter Bae Brandtzæg, Marika Lüders, and Jan Håvard Skjetne. 2010. Too many Facebook "friends"? Content sharing and sociability versus the need for privacy in social network sites. *Intl. Journal of Human–Computer Interaction* 26, 11-12 (2010), 1006–1030.

[7] Pierre Carbonnelle. 2013. PYPL PopularitY of Programming Language Index. http://pypl.github.io/PYPL.html. Last accessed: June 29, 2020.

[8] Quora Help Center. 2019. What languages does Quora support? https://help.quora.com/hc/en-us/articles/360015662751-What-languages-does-Quora-support-. last accessed: April 23, 2020.

[9] Quora Help Center. December 20, 2019. Terms of Service. https://www.quora.com/about/tos. last accessed: June 26, 2020.

[10] Claudio Cioffi-Revilla. 2014. Introduction to computational social science. *London and Heidelberg: Springer* (2014).

[11] Bryan Dosono and Bryan Semaan. 2020. Decolonizing Tactics as Collective Resilience: Identity Work of AAPI Communities on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–20.

[12] Eric Gilbert. 2013. Widespread underprovision on Reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 803–808.

[13] Scott A Hale. 2016. User reviews and language: how language influences ratings. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1208–1214.

[14] Daniel Halpern and Jennifer Gibbs. 2013. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* 29, 3 (2013), 1159–1168.

[15] Hansika Hewamalage. 2016. scrape_quora. https://github.com/HansikaPH/pyquora. PyPI: https://pypi.org/project/scrape_quora/, last accessed: June 18, 2020.

[16] Maritza Johnson, Serge Egelman, and Steven M Bellovin. 2012. Facebook and privacy: it's complicated. In *Proceedings of the eighth symposium on usable privacy and security*. 1–15.

[17] Suman Kalyan Maity, Jot Sarup Singh Sahni, and Animesh Mukherjee. 2015. Analysis and Prediction of Question Topic Popularity in Community Q&A Sites: A Case Study of Quora.. In *ICWSM*. 238–247.

[18] Arjun Mukherjee and Bing Liu. 2012. Mining contentions from discussions and debates. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 841–849.

[19] Tapasweni Pathak. 2014. qtopic. https://github.com/tapaswenipathak/pyQTopic. PyPI: https://pypi.org/project/qtopic/, last accessed: June 18, 2020.

[20] Sharoda A Paul, Lichan Hong, and Ed H Chi. 2012. Who is authoritative? understanding reputation mechanisms in quora. *arXiv preprint arXiv:1204.3724* (2012).

[21] Andrew Perrin. 2015. Social media usage. *Pew research center* (2015), 52–68.

[22] Răzvan Rughiniş, Alina Petra Marinescu-Nenciu, Ştefania Matei, and Cosima Rughiş. 2014. Computer-supported collaborative questioning. Regimes of online sociality on Quora. In *2014 9th Iberian Conference on Information Systems and*

Technologies (CISTI). IEEE, 1–6.
[23] Matthew A Russell. 2013. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. O'Reilly Media, Inc.
[24] Christopher Su. 2016. pyquora. https://github.com/csu/pyquora. PyPI: https://pypi.org/project/quora/, last accessed: June 18, 2020.
[25] George Suciu, Corentin Boscher, Laura Prioux, Adrian Pasat, and Ciprian Dobre. 2017. Insights into Collaborative Platforms for Social Media Use Cases. *Studies in Informatics and Control* 26, 4 (2017), 435–440.
[26] TIOBE. 2003. TIOBE Programming Community Index. https://www.tiobe.com/tiobe-index/. Last accessed: June 29, 2020.
[27] Damian Trilling. 2018. Doing computational social science with python: An introduction. *Available at SSRN 2737682* (2018).
[28] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2013. Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*. 1341–1352.
[29] Anita Whiting and David Williams. 2013. Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal* (2013).
[30] Lei Yang and Xavier Amatriain. 2016. Recommending the World's Knowledge: Application of Recommender Systems at Quora. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 389–389.