github search limit results

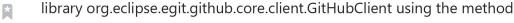Asked 5 years, 4 months ago   Active 3 years, 10 months ago   Viewed 11k times

▲

19

▼

🔖

6

↺

I need to do a very large search on Github for a statistic in my thesis.

For example, I need to explore a large number of Android projects on GitHub, but the site limits the search result to 1000 (ex. https://github.com/search?l=java&q=onCreate& ref=searchresults&type=Code&utf8=%E2%9C%93). Also using the Java GitHub API I tried the library org.eclipse.egit.github.core.client.GitHubClient using the method `GitHubClient.searchRepositories()` but even there the number of results is limited.

Does anyone know how to get all results?

github   github-api

Share Edit Follow

edited Jun 3 '16 at 2:22
bitoiu
5,938   5   35   51

asked Jun 2 '16 at 22:02
scott
191   1   1   3

---

2 Have you looked at the GitHub Archive? It could be a way to get your data without having to bother the live GitHub search API, which as you found out gives a limited number of results, and is also rate-limited. – Wander Nauta Jun 2 '16 at 22:12

Are you able to page through the results? You could get the first chunk of 1000, get the next chunk, and repeat until you have it all. – Kyle Falconer Jun 2 '16 at 22:19

This is not a Java question, or even a programming question. – shmosel Jun 2 '16 at 22:23

Correct, you're limited to 1000 results per search & 30 requests per minute: developer.github.com/v3 /search/#about-the-search-api – zapl Jun 3 '16 at 0:33 ✏️

Is your code publicly available? – Soubriquet Jan 13 '17 at 18:48

---

2 Answers

Active	Oldest	Votes

29

The Search API will return up to 1000 results per query (including pagination), as documented here:

https://developer.github.com/v3/search/#about-the-search-api

However, there's a neat trick you could use to fetch more than 1000 results when executing a repository search. You could split up your search into segments, by the date when the repositories were created. For example, you could first search for repositories that were created in the first week of October 2013, then second week, then September, and so on.

Because you would be restricting search to a narrow period, you will probably get less than 1000 results, and would therefore be able to get all of them. In case you notice that more than 1000 results are returned for a period, you would have to narrow the period even more, so that you can collect all results.

https://help.github.com/articles/searching-repositories/#search-based-on-when-a-repository-was-created-or-last-updated

You should be able to automate this via the API.

Share Edit Follow

answered Jun 5 '16 at 8:03

Ivan Zuzak
15.7k   2   61   54

Seems like you can't query the repository search api by date created. The following will search, but sort, order, and created are ignored: `curl -H 'Accept: application/vnd.github.v3.text-match+json'` `'https://api.github.com/search/repositories?q=language:Java&` `created>=2013-04-11T00:00:00Z&sort=created&order=asc' | grep created_at` – Soubriquet Jan 15 '17 at 23:09 ✏️

2 @Soubriquet You're not constructing that URL correctly. The "created" parameter should be a part of the query, not a parameter on its own. – Ivan Zuzak Jan 16 '17 at 10:06

1 Also, you can't sort by created -- the fields you can sort by are listed here: developer.github.com/v3 /search/#parameters – Ivan Zuzak Jan 16 '17 at 10:07

7 Awesome! Thanks! In case anyone else needs this: `https://api.github.com/search /repositories?q=language:Java+created:>=2013-04-11T00:00:00Z&order=asc` – Soubriquet Jan 16 '17 at 13:52

2 The `order=asc` applies on the `sort` field which can be stars, forks, updated or best_match(default). So `curl -G https://api.github.com/search/repositories  --data-urlencode "q=created:>2013-04-11" --data-urlencode "order=asc"` gets all repositories created after 2013-04-11 but not in the created order. We can fetch repositories within a range using `q=created:time1..time2`, but the results are not sorted by created time. – Alex Mar 20 '18 at 11:49

4

If you are searching for all files in Github with filename:your-file-name, you could also slice it with a query attribute : size.

For example, you are looking for all files named test.rb in Github, Github API may return more than 11M results, but you could only get 1000 of them because the GitHub Search API provides up to 1,000 results for each search. An url like : https://api.github.com/search /code?q=filename:test.rb+size:1000..1500 would be able to slice your search by changing size range.

Share Edit Follow   edited Dec 15 '17 at 15:01   answered Dec 15 '17 at 8:22

Frank Yang
73 5