GitHub 爬虫工作进度介绍

陆绿

日期：2020 年 8 月 6 日

1 Github API 使用介绍

Github API可以帮助用户对自己 GitHub 账户、repository 进行管理，也可以使用它对 GitHub 内进行搜索，本次我们筛选出 GitHub 中的 jupyter notebook 需要中到其中的Github Search API，根据文档，这个 API 基本的使用方式为：

```
1.构建header(包含一些必要认证信息和查询参数)
2.向GitHub发送get请求
3.如果请求成功，网页返回为json格式，里面包括了需要的查询信息
4.将json加载为字典，通过'items'这个key找到一个list，这个list的每一个元素为一个
    repo的metadata
```

1.1 header 的构造

目前使用的一个头部如下，这个是按照文档示例构造的，必不可少的是 User-Agent 和 Authorization（和 Authorization 用来提高下载速度）。

```
    {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit
            /537.36 (KHTML, like Gecko) Chrome/64.0.3282.140 Safari/537.36 Edge
            /18.17763', #必须有
        'Authorization':'token '+token, #token字段需要使用自己的GitHub生成
        'Content-Type':'application/json',
        'method':'GET',
        'Accept':'application/json'
    }
```

1.2 发送请求

文档中说明了请求存在一定时间内次数的限制，使用 token 之后，这样的限制是每分钟最多 30 次 request。

所以访问得太频繁将爬不到数据，因此在代码中加了一个 time.sleep() 控制发送请求次数。另外我在尝试的过程中因为请求有点频繁被封了 ip

1.3 认证的获取

认证可以有几种方式，比如用户名 + 密码，或者使用自己 GitHub 的 token。token 生成和使用都比较简单，所以在代码里我使用了这种方式。token 的生成：GitHub -> Settings -> Developer settings -> Personal access tokens -> Generate new token. 生成之后，直接将 token 粘贴到代码中的 token='' 中就可以。

1.4 参数的构造

目前使用的请求语句是

```
https://api.github.com/search/repositories?q=+language:jupyter+notebook&sort=
    stars&order=desc

基本参数:
language指定了repo的编程语言
sort:排序方式
order:降序 or 升序
```

其他 parameter 如何使用可以在参数的文档中查看（下文提到的问题需要靠控制 parameter 中的时间参数来解决）

1.5 请求的处理

目前使用 request 库 +api 直接发送请求，得到网页后使用 loadjson 将返回内容解析为字典。如果返回正确，那么就可以将每个 repo 的元信息获取。repo 的元信息包含得比较多。目前我将 repo 元信息保存了以下几项

```
repoId,name,fullname,repoUrl,zipUrl,size,starCnt,forksCnt,watchCnt,isForked

repoId:GitHub中返回的ID号
name: repository的名字
fullname: 作者/name
repoUrl: 对应GitHub中这个repository的主页
zipUrl: 下载地址
size:repo的size, 单位为KB
starCnt: star 数目
forksCnt: 被fork多少次数
watchCnt: 被watch的次数
isForked: 如果是原创repo,这个值为0, 如果是从别人处fork的repository, 这个值为1
```

一个 repo 的元信息如下，我只保存了一些以后下载 repository 或者筛选 repository 需要用到的维度，像 repo 的作者信息就没管。文档最后我附上了一个 repo 元信息返回的示例。

2 下载 GitHub repository

下载 repository 的原理其实很简单，使用上文提到的 repoUrl 再加上一个后缀"/archive/master.zip"，zipUrl 正是这样生产的。zipUrl 也就是我们在 GitHub 某个 repo 的界面点击 download with zip 时会跳转的链接，其返回的是 repo 数据，将其保存即可。代码中的 download_github_repo 函数实现了下载的功能。

3 未解决的问题及可能的解决方案

3.1 返回结果数目限制

GitHub Search API 有返回结果数目的限制，目前查到的标记为 jupyter notebook 的数目为 82000+，但是 GitHub 只能返回 1000 个结果，这个问题比较棘手。限于时间，我还没解决，stack-overflow 上一个相同的问题下面给出了可能的解决方案。即通过时间参数分割原始查询，使得每一个查询的结果小于 1000 个，逐步将结果先存入 csv 文件中。

3.2 下载结果后进行筛选

在将 repo 信息记录入 csv 文件之后，可以根据一些规则筛选 repository。例如选取 fork=false 的 repo。在下载成功之后，可以按照 auto-suggest 里的方法筛选 ipynb 文件和数据集。这一部分尚未完成。

4 附录：一个 repo 能返回的元信息示例

下面列出了一个 repo 能返回的元信息，加粗部分是代码中会提取的地方。{ **"id": 65388917,**
"node_id": "MDEwOlJlcG9zaXRvcnk2NTM4ODkxNw==",
"name": "PythonDataScienceHandbook",
"full_name": "jakevdp/PythonDataScienceHandbook",
"private": false,
"owner": {
"login": "jakevdp",
"id": 781659,
"node_id": "MDQ6VXNlcjc4MTY1OQ==",
"avatar_url": "https://avatars0.githubusercontent.com/u/781659?v=4",
"gravatar_id": "",
"url": "https://api.github.com/users/jakevdp",
"html_url": "https://github.com/jakevdp",
"followers_url": "https://api.github.com/users/jakevdp/followers",
"following_url": "https://api.github.com/users/jakevdp/following/other_user",
"gists_url": "https://api.github.com/users/jakevdp/gists/gist_id",

"starred_url": "https://api.github.com/users/jakevdp/starred/owner/repo",
"subscriptions_url": "https://api.github.com/users/jakevdp/subscriptions",
"organizations_url": "https://api.github.com/users/jakevdp/orgs",
"repos_url": "https://api.github.com/users/jakevdp/repos",
"events_url": "https://api.github.com/users/jakevdp/events/privacy",
"received_events_url": "https://api.github.com/users/jakevdp/received_events",
"type": "User",
"site_admin": false
},
"html_url": "https://github.com/jakevdp/PythonDataScienceHandbook", "description": "Python Data Science Handbook: full text in Jupyter Notebooks",
"fork": false,
"url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook",
"forks_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/forks",
"keys_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/keys/key_id",
"collaborators_url":
"https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/collaborators{/collaborator}",
"teams_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/teams",
"hooks_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/hooks",
"issue_events_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/issues/events{/number}",
"events_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/events",
"assignees_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/assignees/user",
"branches_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/branches/branch",
"tags_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/tags",
"blobs_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/git/blobs/sha",
"git_tags_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/git/tags/sha",
"git_refs_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/git/refs/sha",
"trees_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/git/trees/sha",
"statuses_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/statuses/sha",
"languages_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/languages",
"stargazers_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/stargazers",
"contributors_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/contributors",
"subscribers_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/subscribers",
"subscription_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/subscription",
"commits_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/commits/sha",
"git_commits_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/git/commits/sha",
"comments_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/comments/number",
"issue_comment_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/issues/comments/numb

"contents_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/contents/+path",
"compare_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/compare/base...head",
"merges_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/merges",
"archive_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/archive_format/ref",
"downloads_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/downloads",
"issues_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/issues/number",
"pulls_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/pulls/number",
"milestones_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/milestones/number",
"notifications_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/notifications?since,all,part
"labels_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/labels/name",
"releases_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/releases/id",
"deployments_url": "https://api.github.com/repos/jakevdp/PythonDataScienceHandbook/deployments",
"created_at": "2016-08-10T14:24:36Z",
"updated_at": "2020-08-06T11:30:32Z",
"pushed_at": "2020-07-15T23:02:21Z",
"git_url": "git://github.com/jakevdp/PythonDataScienceHandbook.git",
"ssh_url": "git@github.com:jakevdp/PythonDataScienceHandbook.git",
"clone_url": "https://github.com/jakevdp/PythonDataScienceHandbook.git",
"svn_url": "https://github.com/jakevdp/PythonDataScienceHandbook",
"homepage": "http://jakevdp.github.io/PythonDataScienceHandbook",
"size": 33922,
"stargazers_count": 24865,
"watchers_count": 24865,
"language": "Jupyter Notebook",
"has_issues": true,
"has_projects": true,
"has_downloads": true,
"has_wiki": true,
"has_pages": true,
"forks_count": 10886,
"mirror_url": null,
"archived": false,
"disabled": false,
"open_issues_count": 149,
"license": { "key": "other", "name": "Other", "spdx_id": "NOASSERTION", "url": null, "node_id":
"MDc6TGljZW5zZTA=" },
"forks": 10886,
"open_issues": 149,

"watchers": 24865,
"default_branch": "master",
"score": 1.0 }