

RDM in Jupyter

The Importance of Keeping your Data Reproducible

My Origin Story...

- Was lookup up Sci-Hub shenanigans
- Came across 'legitimate' research on it

Science

Current IssueFirst release papersArchiveAbout

Submit manuscript

HOME > SCIENCE > VOL. 352, NO. 6285 > WHO'S DOWNLOADING PIRATED PAPERS? EVERYONE

FEATURE

f t in r s e

Who's downloading pirated papers? Everyone

Data from the controversial website Sci-Hub reveal that the whole world turns to it for journal articles.

JOHN BOHANNON [Authors Info & Affiliations](#)

SCIENCE • 29 Apr 2016 • Vol 352, Issue 6285 • pp. 508-512 • DOI: 10.1126/science.352.6285.508

10,956

🔔 📖 🔒 📄

Just as spring arrived last month in Iran, Meysam Rahimi sat down at his university computer and immediately ran into a problem: how to get the scientific papers he needed. He had to write up a research proposal for his engineering Ph.D. at Amirkabir University of Technology in Tehran. His project straddles both operations management and behavioral economics, so Rahimi had a lot of ground to cover.

But every time he found the abstract of a relevant paper, he hit a paywall. Although Amirkabir is one of the top research universities in Iran, international sanctions and economic woes have left it with poor access to journals. To read a 2011 paper in *Applied Mathematics and Computation*, Rahimi would have to pay the publisher, Elsevier, \$28. A 2015 paper in *Operations Research*, published by the U.S.-based company INFORMS, would cost \$30.

i ↗ 👁 🔗 🖼 📅 ➦

My Origin Story...

- Found a link to a data record
- Some files!
- I could re-use those!

[Explore data](#) | [About](#) | [Help](#) | [Login](#)

Data from: Who's downloading pirated papers? Everyone

Elbakyan, Alexandra

Bohannon, John

john@johnbohannon.org

Publication date: August 16, 2021

Publisher: Dryad

<https://doi.org/10.5061/dryad.q447c>

Citation

Elbakyan, Alexandra; Bohannon, John (2021), Data from: Who's downloading pirated papers? Everyone, Dryad, Dataset, <https://doi.org/10.5061/dryad.q447c>

Abstract

In increasing numbers, researchers around the world are turning to Sci-Hub, the controversial website that hosts 50 million pirated papers and counting. Now, with server log data from Alexandra Elbakyan, the neuroscientist who created Sci-Hub in 2011 as a 22-year-old graduate student in Kazakhstan, Science addresses some basic questions: Who are Sci-Hub's users, where are they, and what are they reading? The Sci-Hub data provide the first detailed view of what is becoming the world's de facto open-access research library. Among the revelations that may surprise both fans and foes alike: Sci-Hub users are not limited to the developing world. Some critics of Sci-Hub have complained that many users can access the same papers through their libraries but turn to Sci-Hub instead—for convenience rather than necessity. The data provide some support for that claim. Over the 6 months leading up to March, Sci-Hub served up 28 million documents, with Iran, China, India, Russia, and the United States the leading requestors.

Usage notes

Sci-Hub download data

These data include 28 million download request events from the server logs of Sci-Hub from 1 September 2015 through 29 February 2016. The uncompressed 2.7 gigabytes of data are separated into 6 data files, one for each month, in tab-delimited text format.
scihub_data.zip

IPython Notebook for Sci-Hub raw data

IPython Notebook used to process the raw server log data (processing the GIS files into CSV, scraping DOI metadata, etc.).
Sci-Hub.html
Sci-Hub.ipynb

Sci-Hub publisher DOI prefixes

Data scraped from the CrossRef website which can be used to replicate the analysis of downloads by publisher.
publisher_DOI_prefixes.csv

Data files

 Download dataset


> April 22, 2017


Related Works

Article

<https://doi.org/10.11...science.352.6285.508>

Metrics

 51878 views

 8831 downloads

 5 citations

Keywords

[open access](#)

[scientific communication](#)

License

This work is licensed under a [CC0 1.0 Universal \(CC0 1.0\) Public Domain Dedication](#) license.



Jupyter!

- You'll quickly learn that I am huge fan of Jupyter in all of its different flavours
- Works really good as both a teaching and research environment

Jupyter! – Use Cases

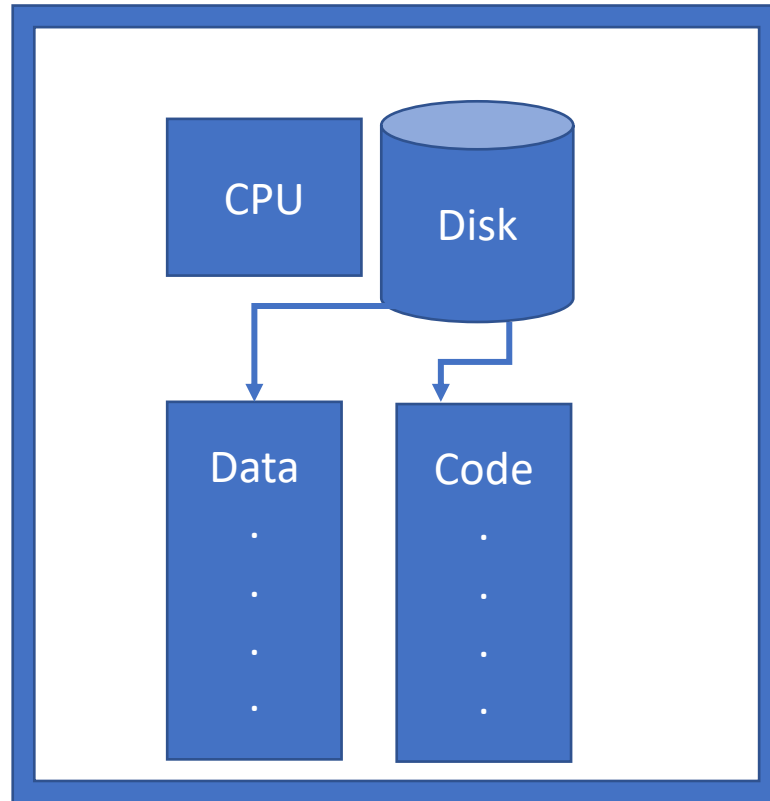
Teaching

- Provides a structured code environment that you can share with learners
- For example, can provide half completed code and ask them to fill in the details. (*ahem*)

Research

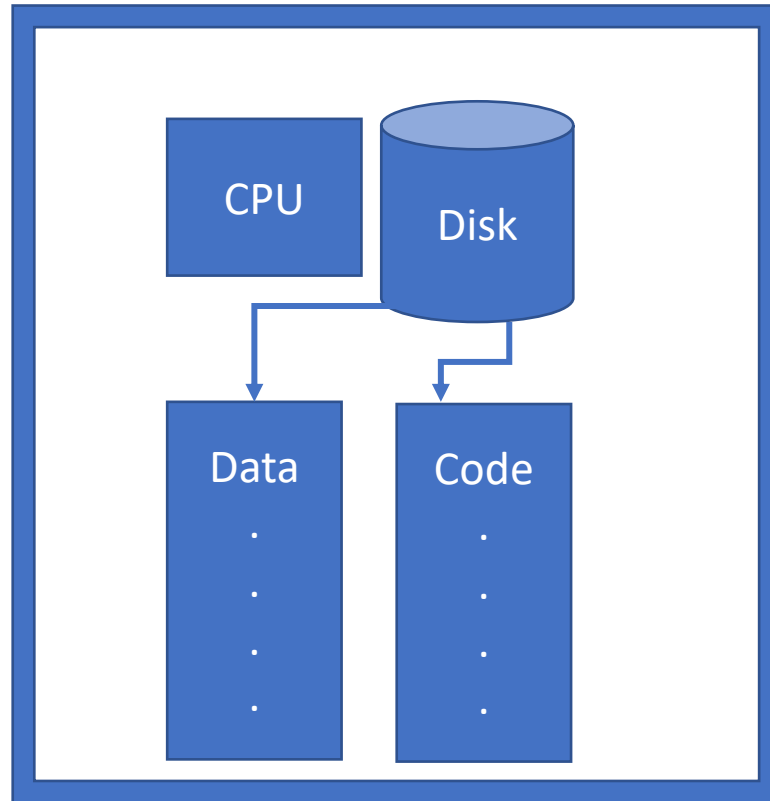
- You can do your analysis in a self-documenting process
- Mix of Markdown and Code creates a very readable end product.

Jupyter! – Home Machine



- Using something like Anaconda Navigator

Jupyter! – Hosted



- Using something like myBinder or Google Colab

Usability & Jupyter

- Our goal then is to save:
 - Code
 - Data
 - Anything else specific to our runtime
- So that data reusability is possible:
 - For verification
 - Further research
 - Long term preservation

Part 1 – Saving your project in GitHub

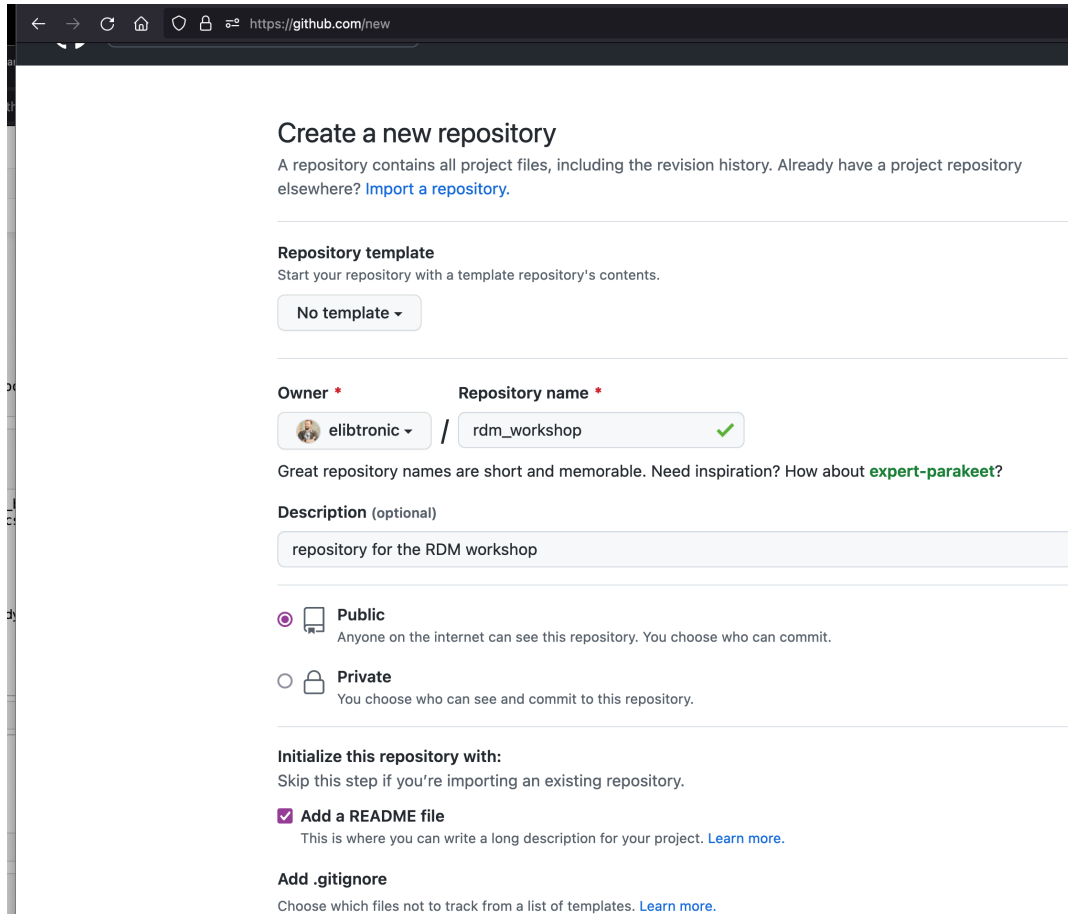
We'll start from scratch

The 'Research'

- We are interested in text sentiment (ie. providing numerical scores on blocks of text to measure how 'happy' or 'negative' they are)
- We are going to use a scoring systems called VADER
- Our 'corpus' will be a bunch of URLs that we identify

Over to Jupyter

GitHub - Creating a Repository




Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? [Import a repository](#).

Repository template
Start your repository with a template repository's contents.

No template ▾


Owner * **Repository name** *


 elibtronic ▾ / rdm_workshop ✓

Great repository names are short and memorable. Need inspiration? How about [expert-parakeet?](#)

Description (optional)

repository for the RDM workshop

☒  **Public**
Anyone on the internet can see this repository. You choose who can commit.

☐  **Private**
You choose who can see and commit to this repository.

Initialize this repository with:
Skip this step if you're importing an existing repository.

☒ **Add a README file**
This is where you can write a long description for your project. [Learn more](#).

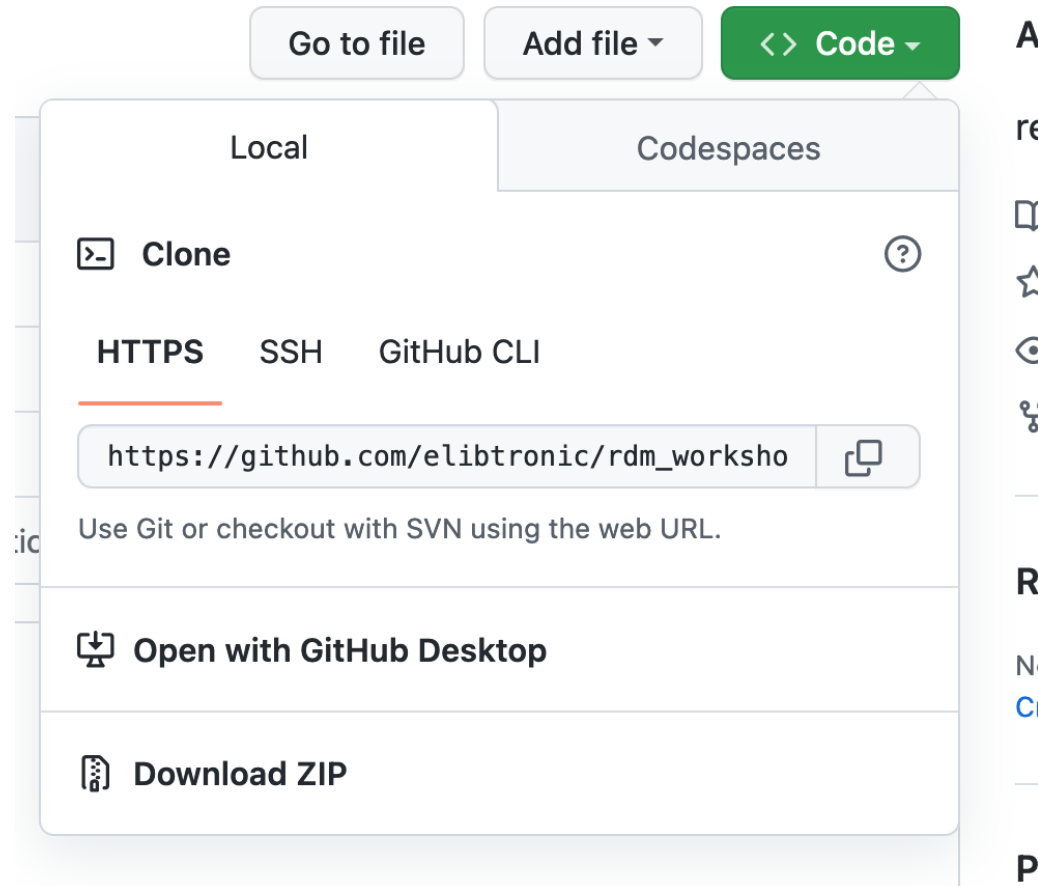
Add .gitignore
Choose which files not to track from a list of templates. [Learn more](#).

- New Repository

- Public

- Add a README file

GitHub – Creating a Repository



GitHub – Creating an access token

GitHub Apps

OAuth Apps

Personal access tokens

Fine-grained tokens Beta

Tokens (classic)

New fine-grained personal access token Beta

Create a fine-grained, repository-scoped token suitable for personal API use and for using Git over HTTPS.

Token name *

RDM_Workshop ✓

A unique name for this token. May be visible to resource owners.

Expiration *

30 days ▾


The token will expire on Thu, Mar 9 2023

Description

Just for Workshop on RDM and Jupyter

What is this token for?

Resource owner

 elibtronic ▾

GitHub – Creating an access token

Repository access

☐ **Public Repositories (read-only)**

☐ **All repositories**

This applies to all current *and* future repositories owned by the resource owner.
Also includes public repositories (read-only).

☒ **Only select repositories**

Select at least one repository. Max 50 repositories.
Also includes public repositories (read-only).

 **Select repositories ▾**

Selected 1 repository.

 elibtronic/rdm_workshop



GitHub – Creating an access token

Permissions

Read our [permissions documentation](#) for information about specific permissions.

Repository permissions 2 Selected

Repository permissions permit access to repositories and related resources.

Actions i

Workflows, workflow runs and artifacts.

Access: No access ▼

Administration i

Repository creation, deletion, settings, teams, and collaborators.

Access: Read and write ▼

Code scanning i

View and manage code scanning alerts.

Access: No access ▼

Codespaces i

Create, edit, delete and list Codespaces.

Access: No access ▼

GitHub – Creating an access token

[Settings](#) / [Developer settings](#) /
Personal access tokens

 GitHub Apps

 OAuth Apps

 **Personal access tokens** ^

Fine-grained tokens Beta

Tokens (classic)

Fine-grained personal access tokens Beta

Generate new token

These are fine-grained, repository-scoped tokens suitable for personal [API](#) use and for using Git over HTTPS.

Make sure to copy your personal access token now as you will not be able to see this again.

Never used

Delete

github_p[REDACTED]04df0



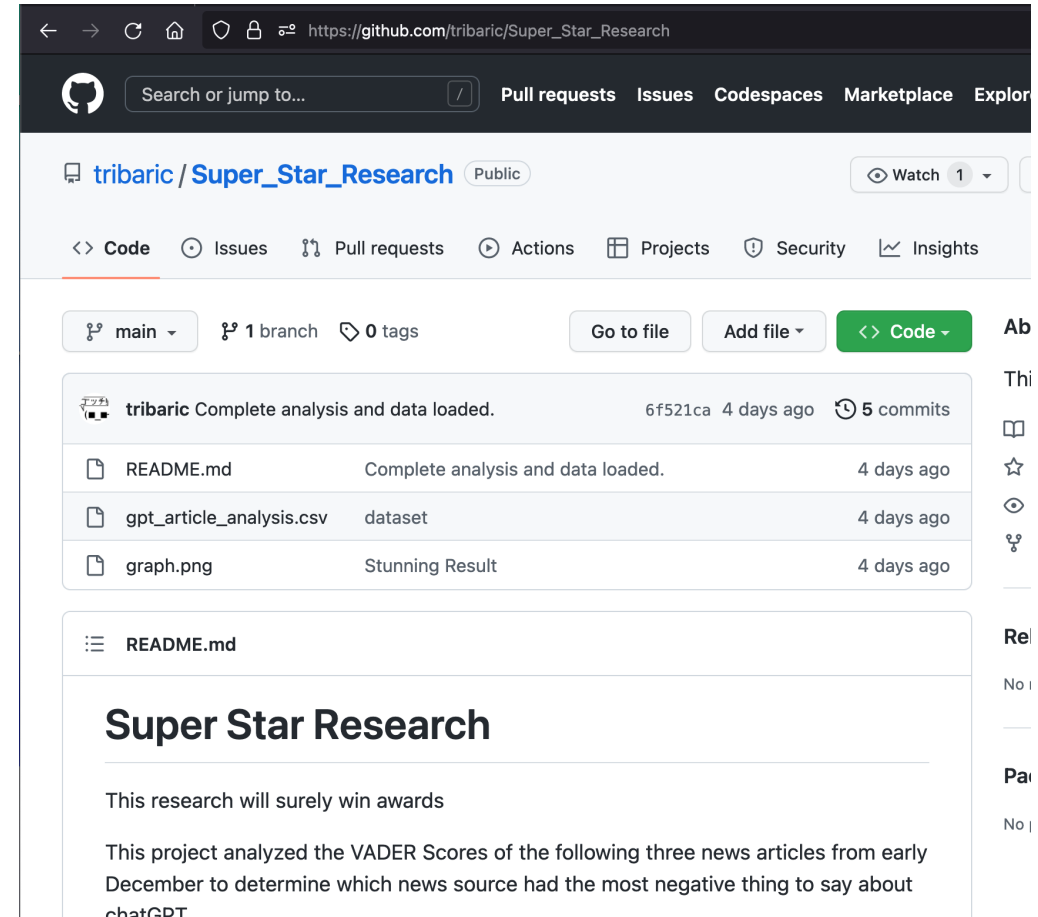
Expires *on Thu, Mar 9 2023.*

Part 2 — Reproducing a Project

And maybe submitting a pull request

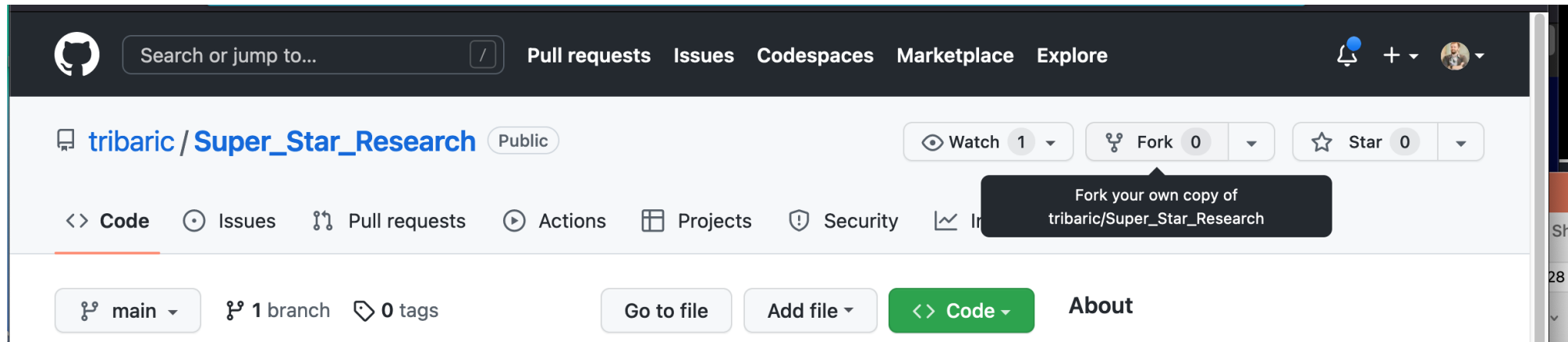
Someone's been bragging...

- You've heard from colleagues a rival research lab has found a result that looks suspicious
- Let's look at the repository and see what is going on



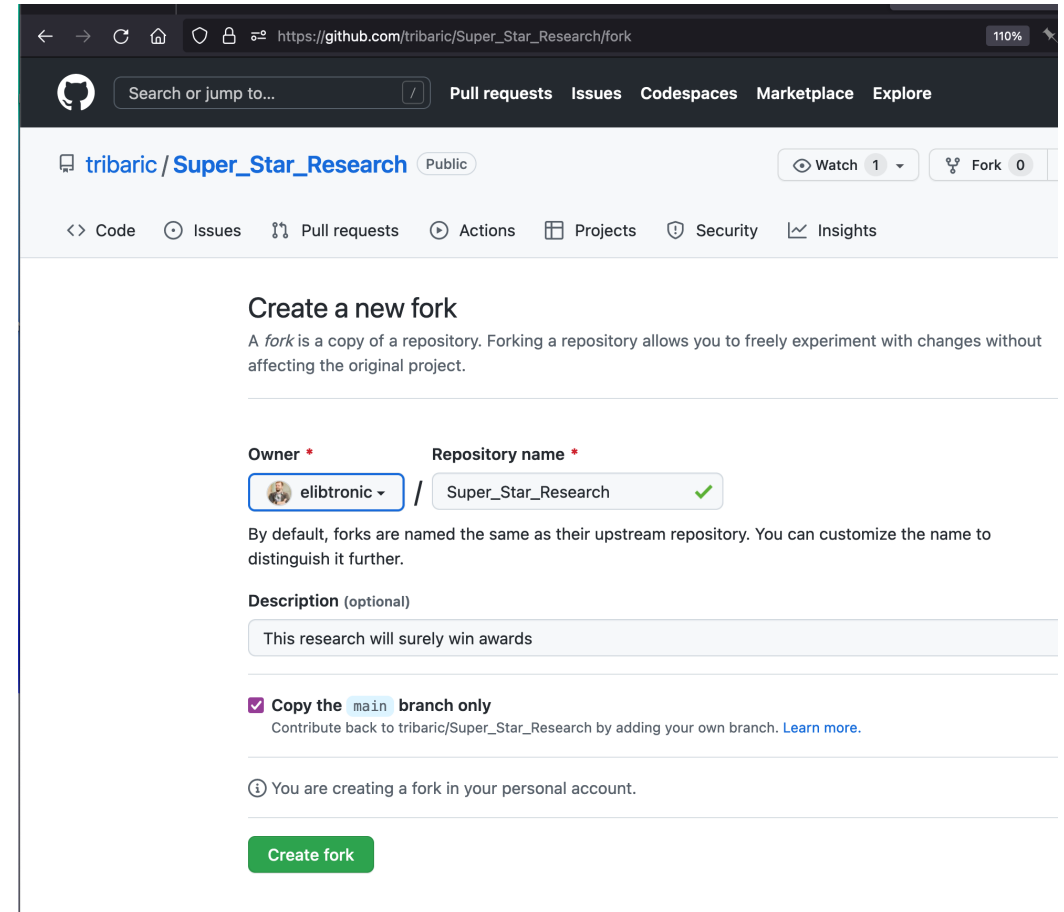
Let's fix the error

- We'll 'Fork' the Repository

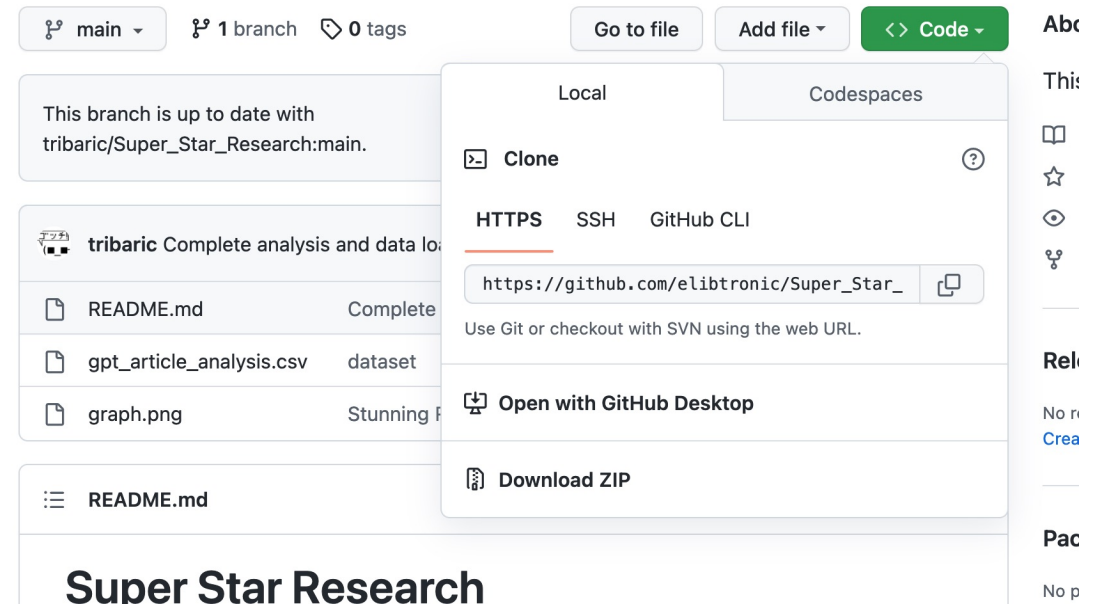
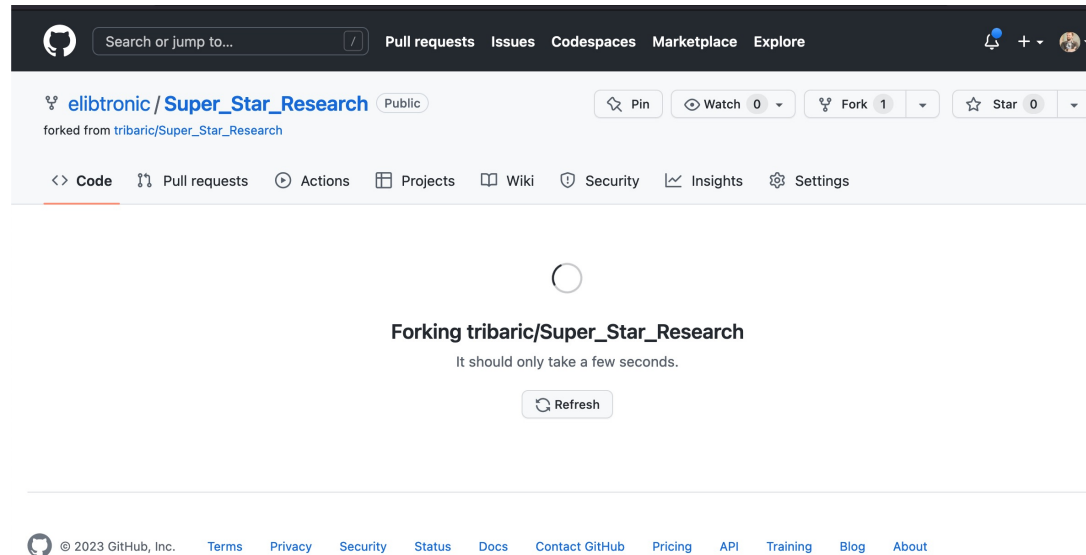


Let's fix the error

- Default options should be fine



Let's fix the error








GH Token – again

- We will generated a new token for this part of the workshop as well.
- Navigate to your profile
- Click ‘Developer settings’

<> Developer settings

GH Token - again

[Settings](#) / Developer settings

-  **GitHub Apps**
-  OAuth Apps
-  Personal access tokens 
 - Fine-grained tokens 
 - Tokens (classic)

GitHu

Want to
develop
[docume](#)

GH Token - again

Fine-grained personal access tokens Beta

Generate new token

These are fine-grained, repository-scoped tokens suitable for personal [API](#) use and for using Git over HTTPS.



RDM_Workshop

Expires *on Thu, Mar 9 2023.*

Last used within the last week

Delete

GH Token - again

New fine-grained personal access token Beta

Create a fine-grained, repository-scoped token suitable for personal API use and for using Git over HTTPS.

Token name *

RDM_workshop_part_2



A unique name for this token. May be visible to resource owners.

Expiration *

30 days

The token will expire on Thu, Mar 16 2023

Description

For the RDM Jupyter Workshop, part 2

What is this token for?

Resource owner



elibtronic

GH – Token - again

Repository access

☐ **Public Repositories (read-only)**

☐ **All repositories**

This applies to all current *and* future repositories owned by the resource owner.
Also includes public repositories (read-only).

☒ **Only select repositories**

Select at least one repository. Max 50 repositories.
Also includes public repositories (read-only).

 **Select repositories** ▾

Selected 1 repository.

 elibtronic/Super_Star_Research

×

GH Token - again

Commit statuses ⓘ

Commit statuses.

Access: No access ▼

Contents ⓘ

Repository contents, commits, branches, downloads, releases, and merges.

Access: Read and write ▼

Dependabot alerts ⓘ

Retrieve Dependabot alerts.

Access: No access ▼

GH Token - again

Pages ⓘ

Retrieve Pages statuses, configuration, and builds, as well as create new builds.

Access: No access ▼

Pull requests ⓘ

Pull requests and related comments, assignees, labels, milestones, and merges.

Access: Read and write ▼

Repository announcement banners ⓘ

View and modify announcement banners for a repository.

Access: No access ▼

GH Token - again

U Account permissions

This token will expire **March 16, 2023**.

Generate token

[Cancel](#)

This token will be ready for use immediately.

GH Token - again

Fine-grained personal access tokens Beta

These are fine-grained, repository-scoped tokens suitable for personal [API](#) use and for using

Make sure to copy your personal access token now as you will not be able to see this again.

github_pat

[REDACTED]



Expires on *Thu, Mar 16 2023*.

Fire up the Notebook

part2.ipynb