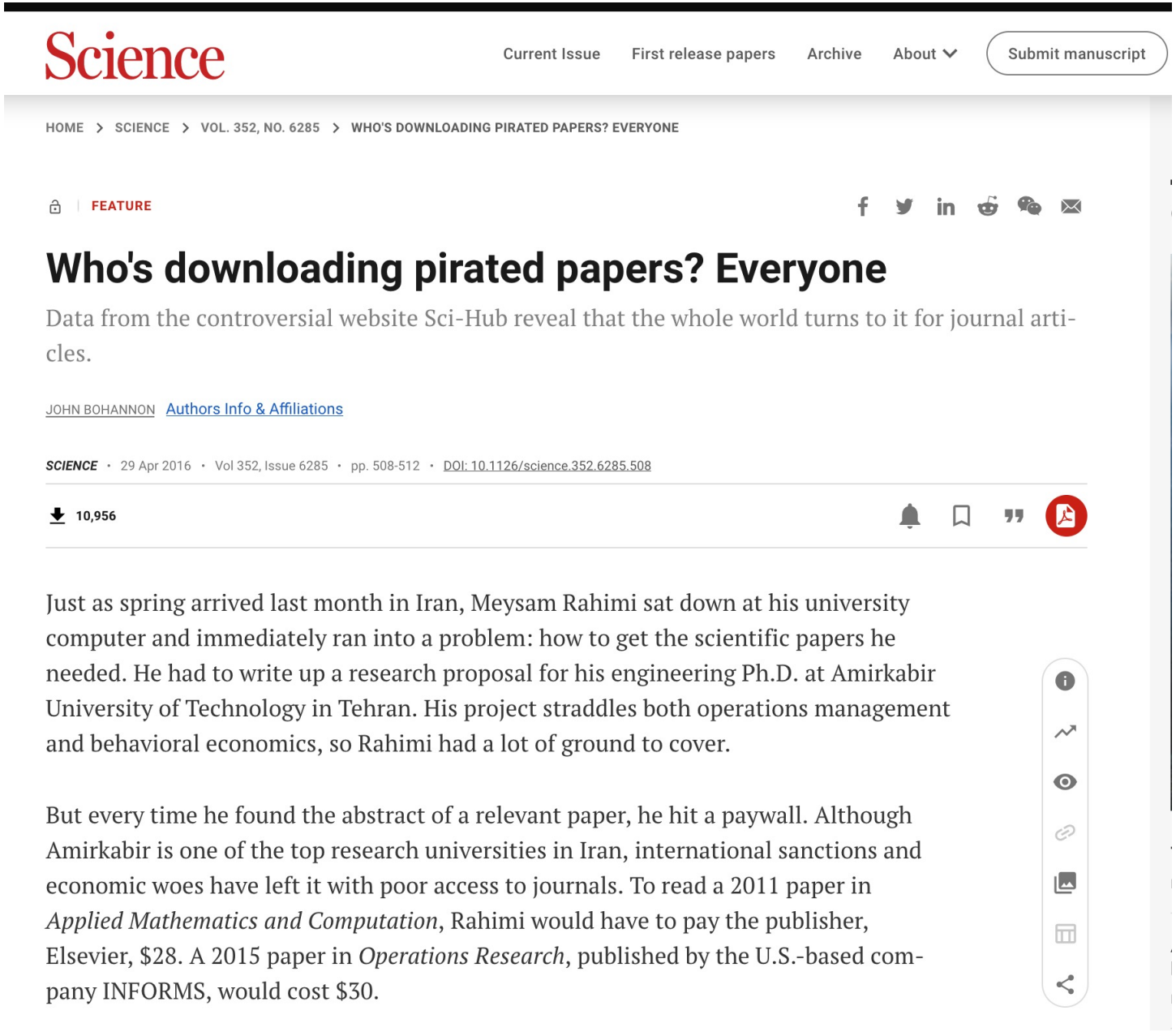
RDM in Jupyter

The Importance of Keeping your Data Reproducible

My Origin Story…

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



Science

Current Issue First release papers Archive About Submit manuscript

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

FEATURE

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

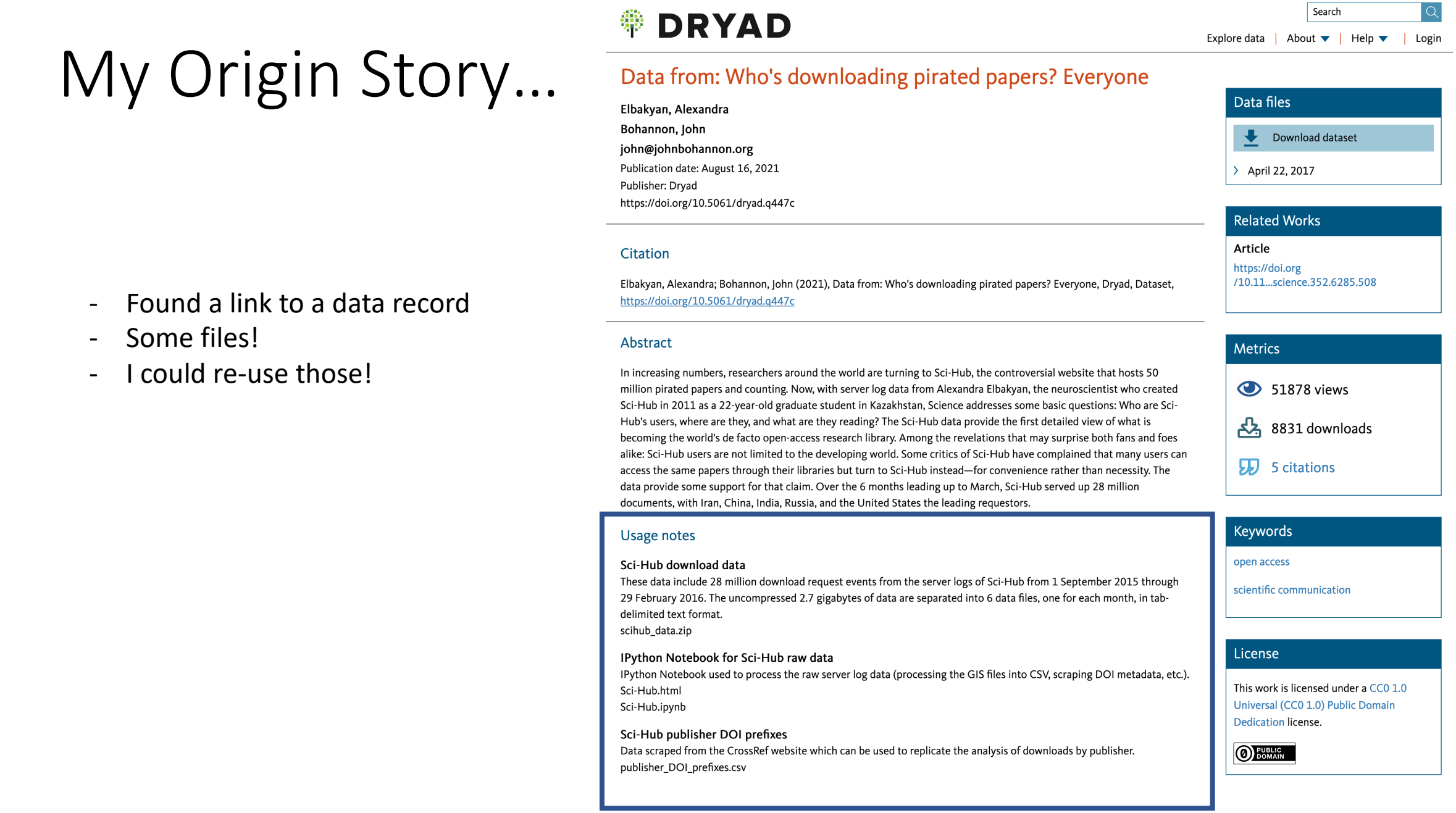
10,956

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.

My Origin Story…

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



DRYAD

Search

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.

PUBLIC DOMAIN

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

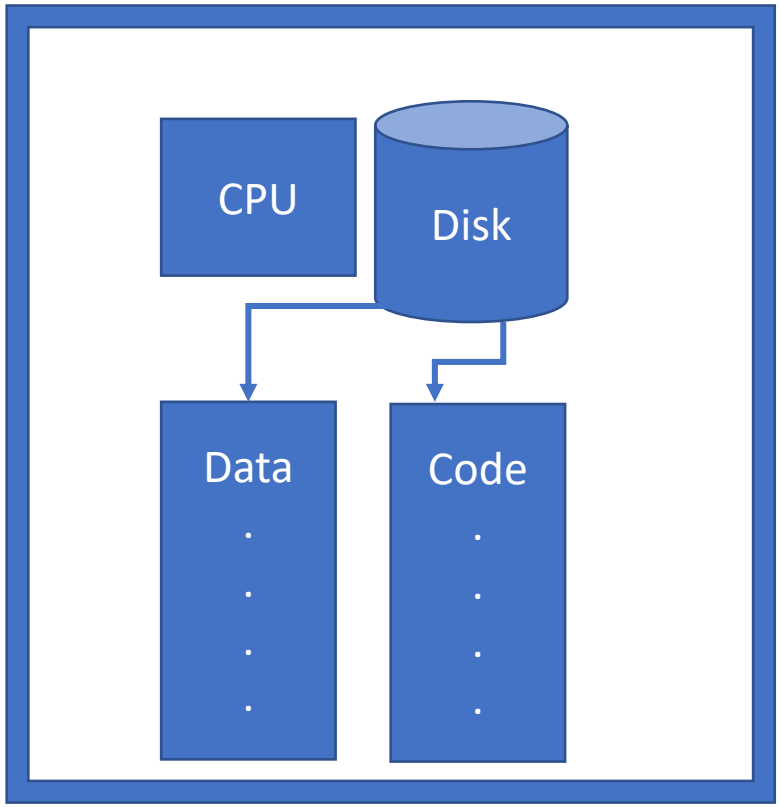
Teaching

- 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

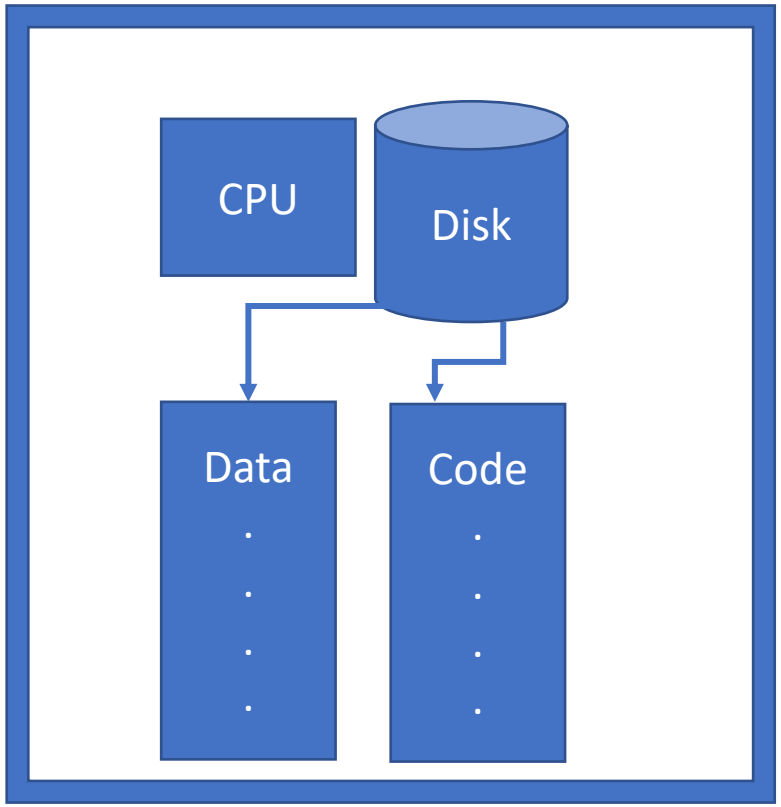
- You can do your analysis in a self-documenting process
- Mix of Markdown and Code creates a very readable end product.

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

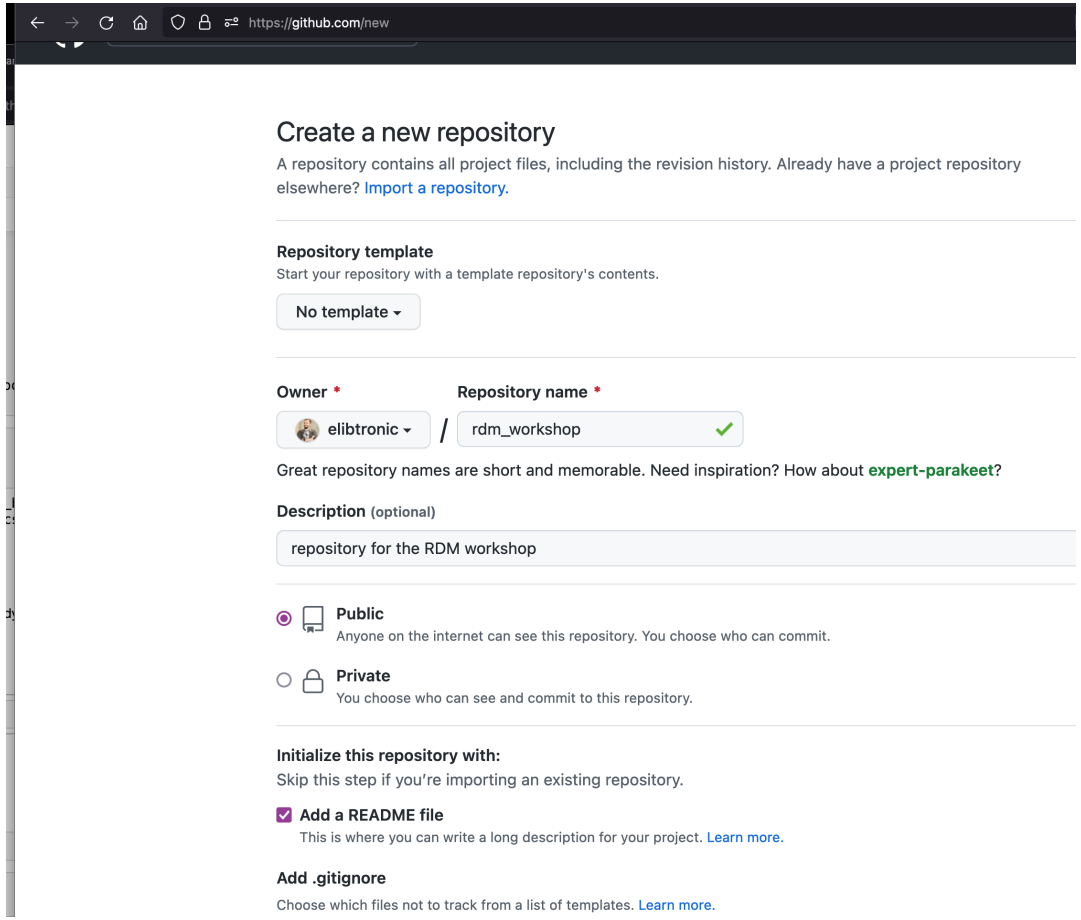
We’ll start from scratch

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

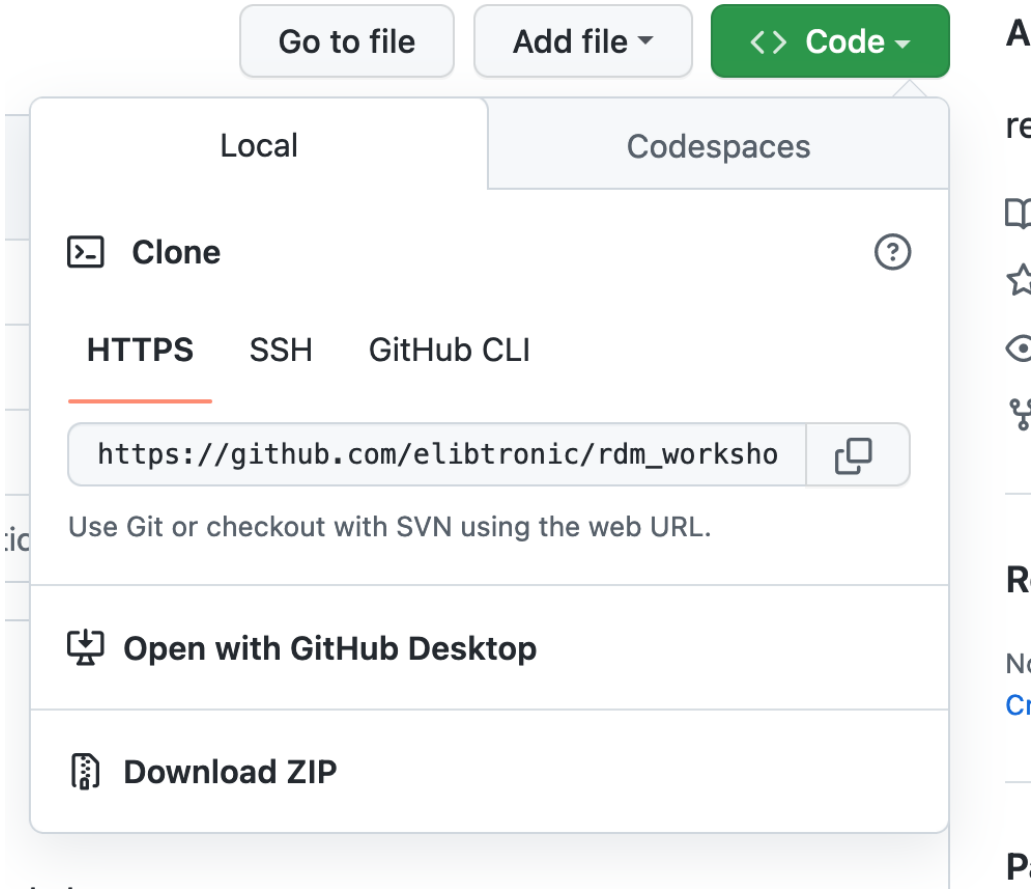
Over to Jupyter

GitHub - Creating a Repository

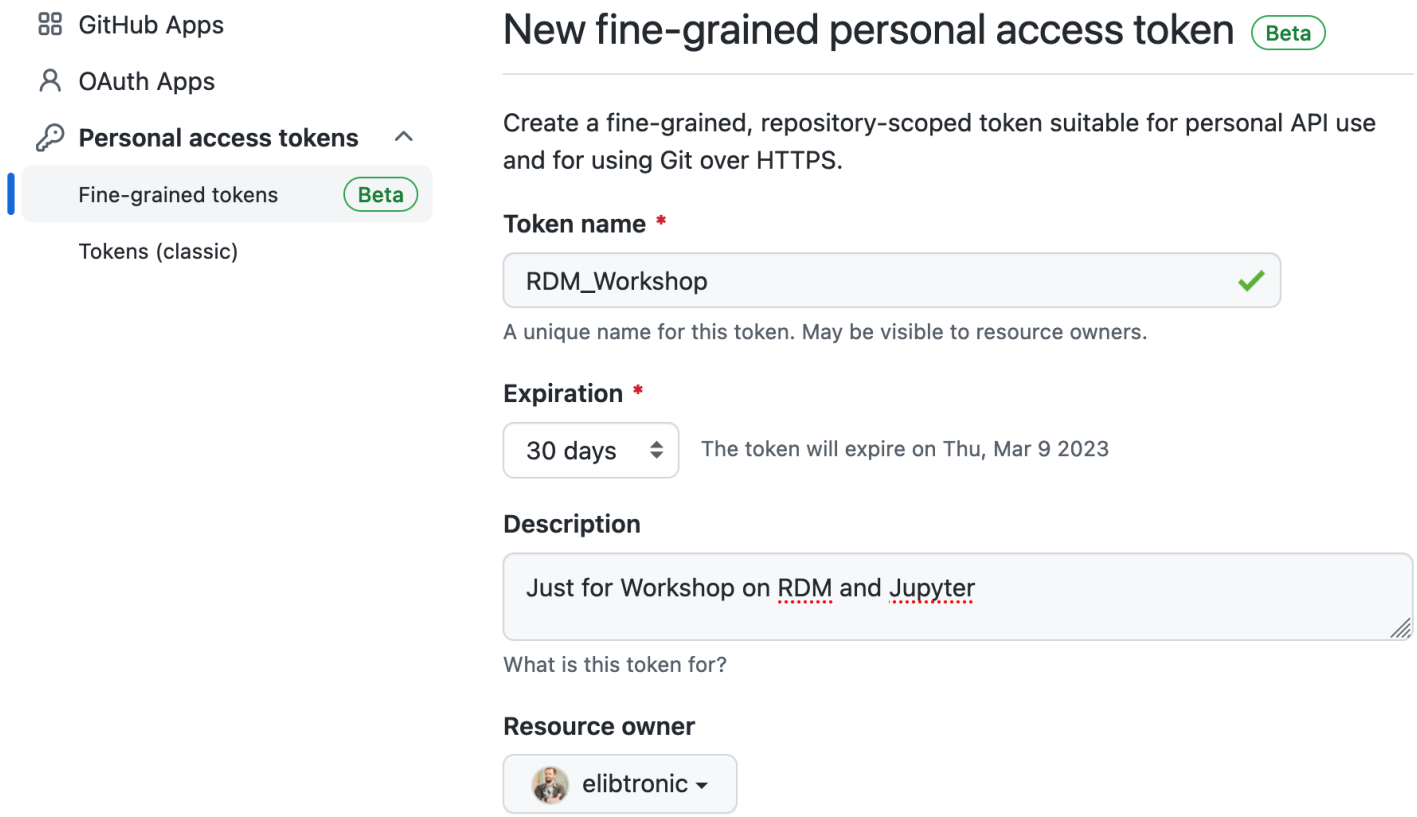


- New Repository
- Public
- Add a README file

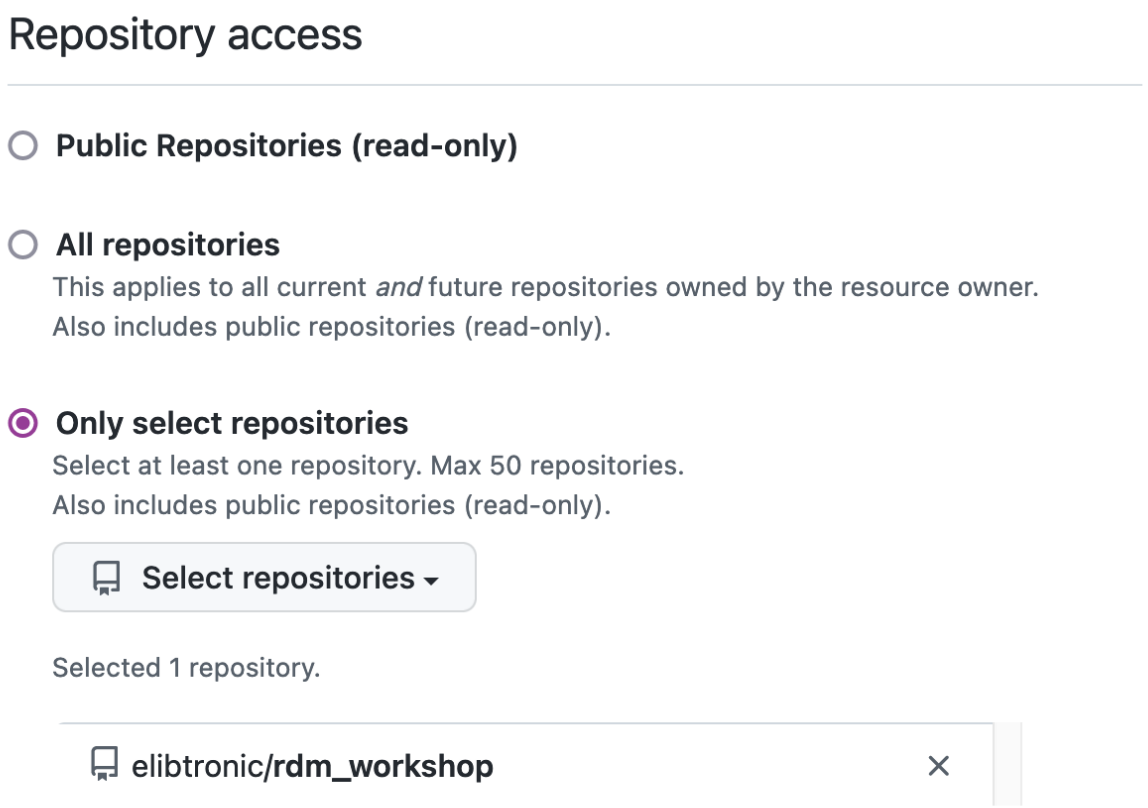
GitHub – Creating a Repository



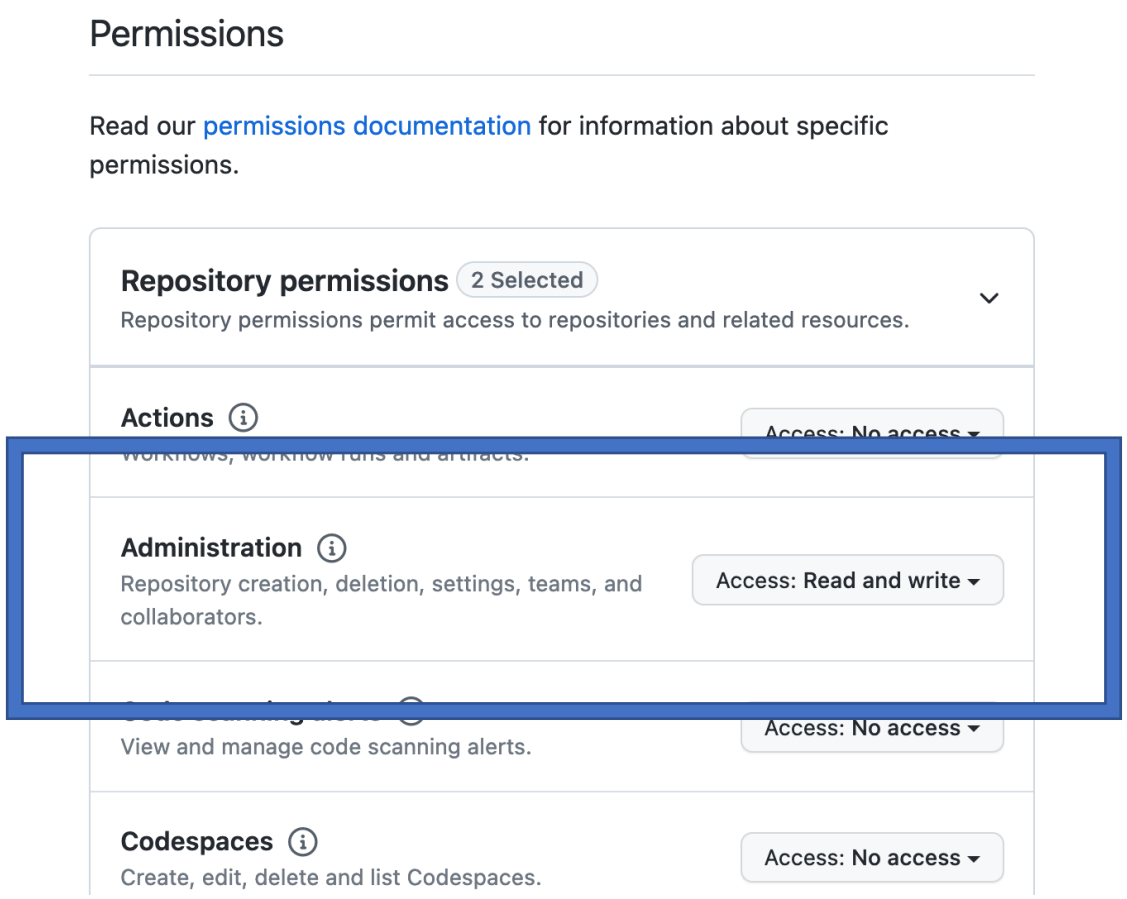
GitHub – Creating an access token



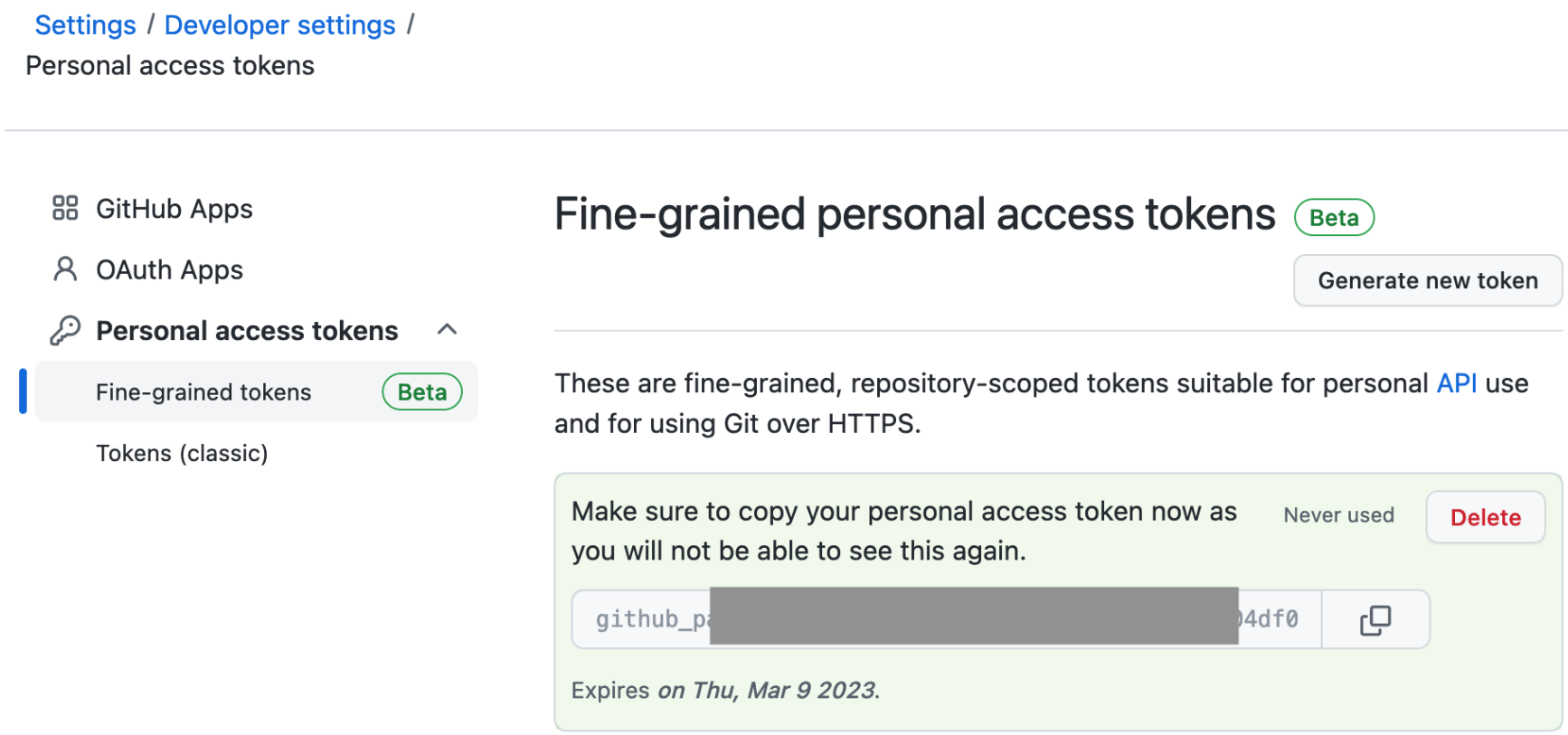
GitHub – Creating an access token



GitHub – Creating an access token



GitHub – Creating an access token



Part 2 – Reproducing a Project

And maybe submitting a pull request