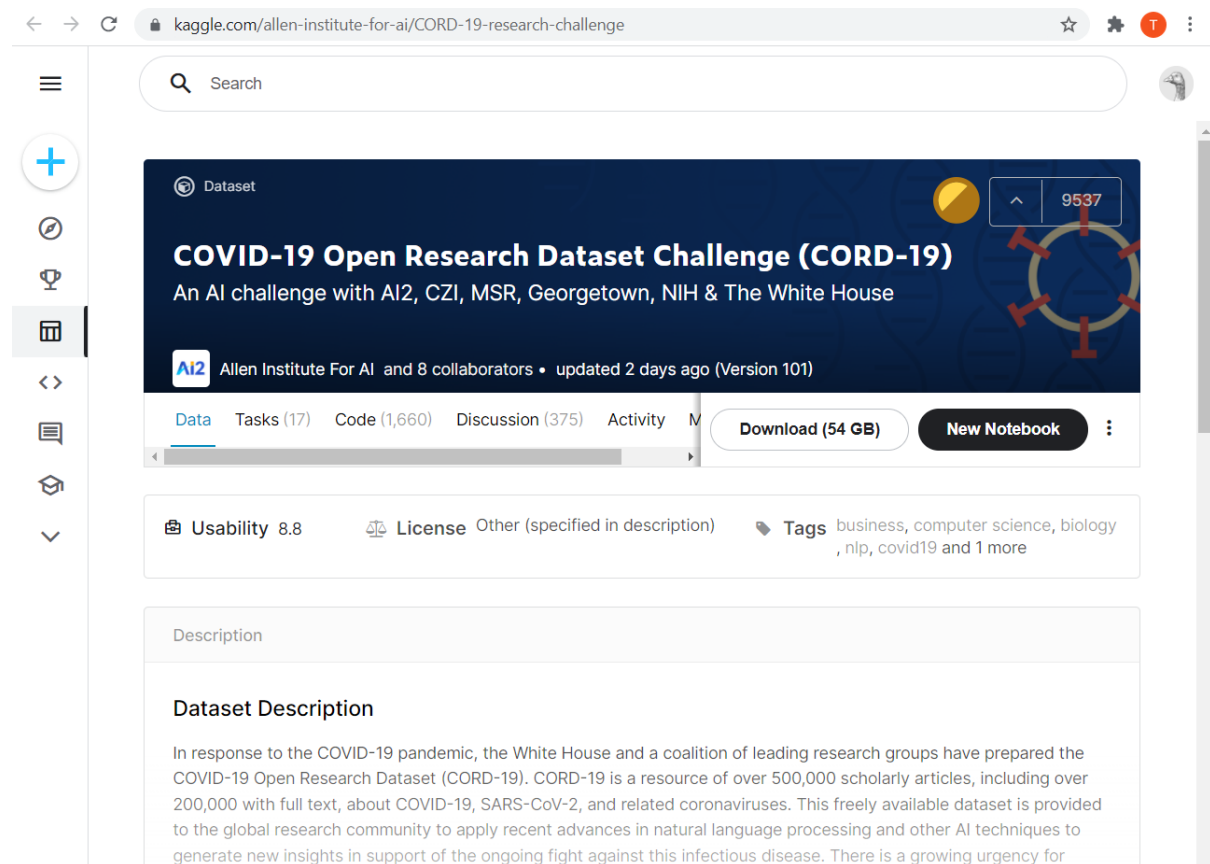


The instructions below show how to download the CORD-19 dataset

Step 1: Visit <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>



The screenshot shows the Kaggle dataset page for the COVID-19 Open Research Dataset Challenge (CORD-19). The page features a dark blue header with the dataset title and a gold medal icon. Below the header, there are tabs for Data, Tasks (17), Code (1,660), Discussion (375), and Activity. A 'Download (54 GB)' button and a 'New Notebook' button are visible. The 'Usability' is 8.8, and the 'License' is 'Other (specified in description)'. The 'Tags' include 'business, computer science, biology, nlp, covid19 and 1 more'. The 'Description' section states that the dataset is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses.

**COVID-19 Open Research Dataset Challenge (CORD-19)**  
An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

AI2 Allen Institute For AI and 8 collaborators • updated 2 days ago (Version 101)

Data Tasks (17) Code (1,660) Discussion (375) Activity M

Download (54 GB) New Notebook

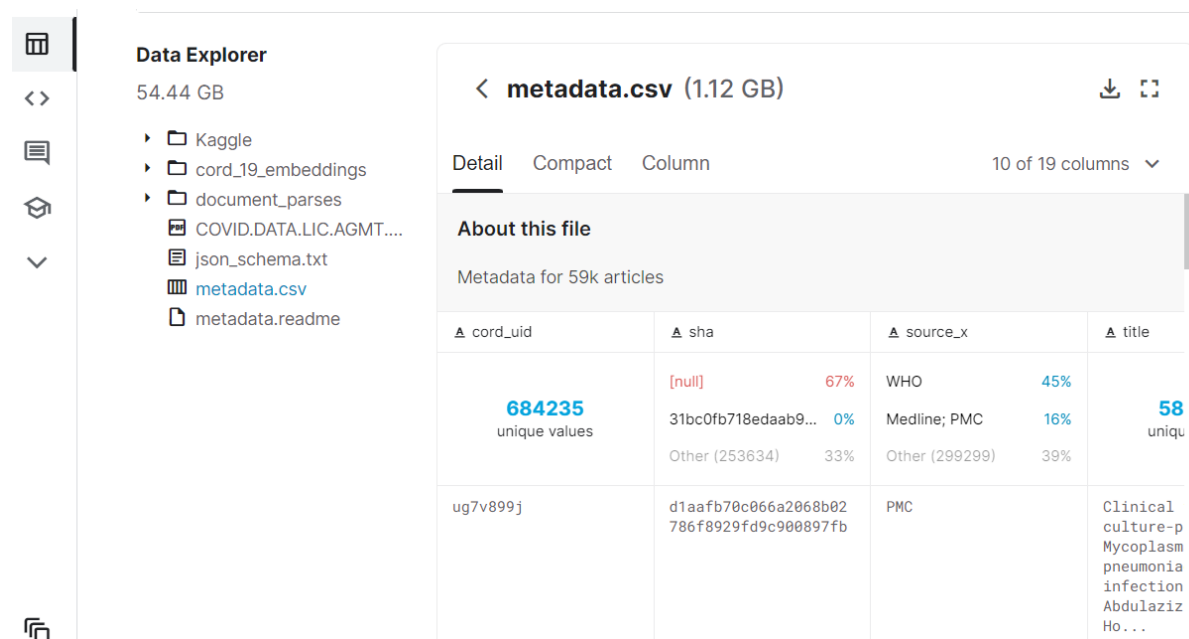
Usability 8.8 License Other (specified in description) Tags business, computer science, biology, nlp, covid19 and 1 more

**Description**

**Dataset Description**

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for

Step 2: This is a large dataset (about 54 GB). You can have an overview of the dataset by click on metadata files:



The screenshot shows the Kaggle Data Explorer interface for the 'metadata.csv' file (1.12 GB). The left sidebar shows the file structure, including 'Kaggle', 'cord\_19\_embeddings', 'document\_parses', 'COVID.DATA.LIC.AGMT...', 'json\_schema.txt', 'metadata.csv', and 'metadata.readme'. The main area displays the 'About this file' section, which states 'Metadata for 59k articles'. Below this, a table shows the distribution of values for 'cord\_uid', 'sha', 'source\_x', and 'title'.

**Data Explorer**  
54.44 GB

- Kaggle
- cord\_19\_embeddings
- document\_parses
- COVID.DATA.LIC.AGMT...
- json\_schema.txt
- metadata.csv
- metadata.readme

**metadata.csv (1.12 GB)**

Detail Compact Column 10 of 19 columns

**About this file**  
Metadata for 59k articles

cord_uid	sha	source_x	title
684235 unique values	[null] 67% 31bc0fb718edaab9... 0% Other (253634) 33%	WHO 45% Medline; PMC 16% Other (299299) 39%	58 unique
ug7v899j	d1aafb70c066a2068b02 786f8929fd9c900897fb	PMC	Clinical culture-p Mycoplasm pneumonia infection Abdulaziz Ho...


Step 3: Click the Download button. You need to Sign In before being able to download. If you do not have an account, you can Register and then Sign In:





## Download Datasets


Welcome to Kaggle! Join our community of over 6 million data scientists. Find datasets and code as well as free access to compute on our platform.

**Sign In**   Register

 Sign in with Google

 Sign in with your email

 Sign in with Facebook

 Sign in with Yahoo

No Account? [Create one.](#)