Finding and working with large datasets

Tom Theile, Max-Planck-Institute for demographic research, theile@demogr.mpg.de

2022

Finding useful datasets

Dataset search engines

https://datasetsearch.research.google.com/ - Google Dataset Search. Search by keywords. Good coverage, but bad sorting of the results.

https://archive.org/details/datasets - The internet Archive is a non-profit organization that tries to preserve the whole internet. They crawl the whole web, store the massive data and make it available to everyone for free. They also host some big datasets.

https://catalog.data.gov/dataset - The U.S. Government Catalog of Datasets. A lot of datasets that are not nice to work with…

https://www.kaggle.com/datasets - Kaggle is a free data science competition platform. Volunteers (mostly data scientists and machine learning engineers) upload datasets and "notebooks" that analyse those datasets. The focus is less on social siences. The advantage is that there will always be "notebooks" that work with exactly these datasets.

Noteworthy data dumps and data sources

Reddit comments and submissions

Reddit (almost) Every reddit comment from 2005 to June 2021: https://files.pushshift.io/reddit/comments/

(almost) Every reddit submission from 2005 to June 2021: https://files.pushshift.io/reddit/submissions/ (10 GB compressed json for one month)

(You can contact me if you want to work with this data or a subset of it. I have worked with it before)

Wikipedia text dumps

The Wikipedia project offers not only a dump of all wikipedia articles for every language, but also the complete edit-history of all articles. Another interesting resource are the wikipedia access-statistics (https://dumps.wikimedia.org/other/accessstats/) and the wikipedia page-views (https://dumps.wikimedia.org/other/pageviews/).

- overview: https://dumps.wikimedia.org/
- available:

  - whole text in many languages
  - edit history

- Wikipedia access statistics: https://dumps.wikimedia.org/other/accessstats/

Wikidata

Wikidata is a huge human-curated database of information about people, places, things and events - basically every "fact" of wikipedia (and much more) is stored in wikidata in a format that is easy to work with. The whole dump is around ~~40GB~~ 72GB of compressed json. There are more than 6 Million entries about persons, 2 Million entries about "administrative territorial entities" (e.g. countries, states, provinces, etc.), 22.5 Million entries about scholarly articles, 300 000 entries about movies, etc.

Examples:

- Barcelona - https://www.wikidata.org/wiki/Q1492
- Big City - https://www.wikidata.org/wiki/Q1549591

But it is not easy to handle the full dump in R. I was not able to decompress it partially, which is possible to do in Python. An easy start would be the library qwikidata.

But you can run SPARQL-queries online to generate a dataset. SPARQL is a query language and not super easy to master, but you can sift through the query-examples: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples and probably find something that you can adapt to your use case.

You can for example adapt the query to get all the movies of a person:

Exercise 2.1: Use https://query.wikidata.org/ to generate a dataset of all Hospitals in the world.

Stack Overflow

Stack Overflow is a question-answer site. Mostly for programmers and scientist. The whole dump can be downloaded and is about 40 GB compressed xml. It can also be accessed online here: https://data.stackexchange.com/stackoverflow/query/new

Project Gutenberg

A large collection of books that in the public domain. https://www.gutenberg.org/robot/harvest The books can also be downloaded independently and are a nice resource for e-readers.

OpenStreetMap

OpenStreetMap is the "Wikipedia of maps", a map of the whole world that is build by volunteers and available under the OpenStreetMap license (which is permissive).

You can download the whole dump or parts of it from https://archive.org/details/osmdata or https://wiki.openstreetmap.org/wiki/Planet.osm

you can also download it from:

Working with the data dumps of OpenStreetMap is not easy. Thankfully there is an API and a R-library to access it. You can read this article on how to use this to https://dominicroye.github.io/en/2018/accessing-openstreetmap-data-with-r/

Another possibility is to use the overpass-turbo web application to create a dataset from the OpenStreetMap data. The “wizard” makes it easy to create a query and download the data.

Please open http://overpass-turbo.eu/ and navigate in the map to a city of your liking.

Then click the “Wizard”-Button and enter “Supermarket”

The wizard will create a query for you, that may look like this:

```
/*
This has been generated by the overpass-turbo wizard.
The original search was:
"supermarkt"
*/
[out:json][timeout:25];
// gather results
(
  // query part for: "supermarkt"
  node["shop"="supermarket"]({{bbox}});
  way["shop"="supermarket"]({{bbox}});
  relation["shop"="supermarket"]({{bbox}});
);
// print results
out body;
>;
out skel qt;
```

and it will display the results in the map. Please check whether it really catched all items - the query or the data might not be perfect. If it worked, you can download the data as json, which can be easily imported into R.

```
{
  "version": 0.6,
  "generator": "Overpass API 0.7.57.1 74a55df1",
  "osm3s": {
    "timestamp_osm_base": "2022-02-21T16:54:24Z",
    "copyright": "The data included in this document is from www.openstreetmap.org. The data is made ava
  },
  "elements": [

{
  "type": "node",
  "id": 183262186,
  "lat": 54.1125692,
  "lon": 12.1383369,
  "tags": {
    "addr:city": "Rostock",
    "addr:housenumber": "32",
    "addr:postcode": "18147",
    "addr:street": "Pablo-Picasso-Straße",
    "brand": "Netto Marken-Discount",
    "brand:wikidata": "Q879858",
    "brand:wikipedia": "de:Netto Marken-Discount",
    "name": "Netto Marken-Discount",
    "opening_hours": "Mo-Sa 07:00-21:00",
    "shop": "supermarket",
    "wheelchair": "yes"
  }
},
{
  "type": "node",
  "id": 253073176,
  "lat": 54.1527184,
  "lon": 12.0648400,
  "tags": {
    "addr:city": "Rostock",
    "addr:country": "DE",
    "addr:housenumber": "6a",
    "addr:postcode": "18109",
    "addr:street": "Güstrower Straße",
    "addr:suburb": "Lichtenhagen",
    "brand": "Rewe",
    "brand:wikidata": "Q16968817",
    "brand:wikipedia": "en:REWE",
    "name": "Rewe",
    "old_name": "sky",
    "opening_hours": "Mo-Sa 07:00-22:00;PH closed",
    "operator": "Supermärkte Nord Vertriebs GmbH & Co. KG",
    "organic": "yes",
    "shop": "supermarket",
    "wheelchair": "yes"
  }
},
{
  "type": "node",
  "id": 254511644,
  "lat": 54.1072999,
  "lon": 12.0359235,
  "tags": {
    "addr:city": "Sievershagen",
    "addr:housenumber": "1A",
    "addr:postcode": "18069",
    "addr:street": "Ostsee-Park-Straße",
    "brand": "Aldi",
    "brand:wikidata": "Q125054"
   }
}

    ]
    }
```

Exercise 2.2: Download the location data of all schools in a city of your choice.

Twitter

There are a number of tweet-data dumps somewhere in the internet with vastly different content. At the MPIDR we have a "1% sample" which contains 1 % of all tweets from 2012 to April 2018 with 20 to 50 GB per month = 2.6TB in total. ~~But I was not able to find it online (if you want to work with~~

~~it, you could get access through the MPIDR).~~ You can download this sample from the Internet Archive: https://archive.org/details/twitterstream

We also have a dump of all geo-located tweets from Europe from 2016-2018 (-> 150 GB total) at the MPIDR.

In general: The Twitter API is very good, as long as you don't hit the Limits (~10 Million Tweets per Month for Academics after application with an institutional email adress and a description of your project, 500K tweets for everyone), you should work with the API. But: 10 Million tweets is only a tiny fraction of Twitter!

Nevertheless, pushshift.io has some interesting and big twitter data:

https://files.pushshift.io/twitter/

https://files.pushshift.io/twitter/verified_feed/ - "This directory contains daily dumps of all publicly available verified tweets and retweets. Present daily volume from all verified accounts (~335k verified accounts) is around 2.5 million tweets per day = ~600MB per day. This should be ~100% of all verified tweets made for the day mentioned in the filename." But only available for October and November 2019.

How to Work with so much data? Again: Get a big computer!

Then: Only open a fraction at a time. Extract, filter, save.

Be patient.

Government data sources and sources from supranational organisations

Countries and governments produce and store a lot of data. Some of that is easily available through government portals. Some of it can be requested.

https://dataportals.org/ is a list of portals that provide access to government data.

UK data

https://ukdataservice.ac.uk/find-data/browse/ - "Funded by UK Research and Innovation (UKRI), the Service integrates and builds on investments the Economic and Social Research Council has made in UK research infrastructure for decades, including the UK Data Archive, Economic and Social Data Service, the Secure Data Service, Census Programme and Survey Question Bank. Now, on those foundations, we continue to house the largest collection of economic, social and population data in the UK."

https://www.data-archive.ac.uk/ - "UK Data Archive The UK's largest digital collection of social sciences and population research data"

https://data.gov.uk/ - "Since 2010 data.gov.uk has been helping people to find and use open government data, and supporting government publishers to maintain data. In March 2018, we re-designed the site and launched the Find open data service."

https://www.ons.gov.uk/ - Office for national statistics

Germany data

https://www.govdata.de/ - "Über GovData bieten öffentliche Stellen aus Bund, Ländern und Kommunen Daten der Verwaltung an. So soll insbesondere Verwaltungsmitarbeitern, Bürgern, Unternehmen und Wissenschaftlern die Möglichkeit gegeben werden, über einen zentralen Einstiegspunkt Daten und Informationen der öffentlichen Verwaltung in Deutschland ebenenübergreifend zugreifen zu können. Ziel ist es, dass diese „Datenschätze" aus der Verwaltung besser genutzt und weiterverwendet werden, so dass durch neue Ideen sowie Kombination und Analyse neue Erkenntnisse aus den vorhandenen Daten gewonnen und neue Anwendungsfelder erschlossen werden können."

https://www.destatis.de/DE/Service/Suche/suche_node.html - official website of the German Statistical Office

https://gdz.bkg.bund.de/index.php/default/open-data.html - mostly maps - “Kostenfreie Geodaten und Webdienste, INSPIRE-konforme Webdienste und Kartendownloads”

US data

https://www.data.gov/ - official site - “The home of the U.S. Government’s open data. Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.”

EU data

https://data.europa.eu/en - “The official portal for European data”

World Bank

https://datacatalog.worldbank.org/ search very good for world-wide economic indicators (e.g. GDP, unemployment, etc.) and for their world development indicators

The data is accessible through an online interface, but also downloadable as compressed csv-files. Another way to get the data is to use the packacke World Bank Development indicators for R, but I can’t recommend it.

Exercise 2.3: Download the GDP data of all countries in the world for all available years.

OECD

https://stats.oecd.org/ - official site - “OECD.Stat includes data and metadata for OECD countries and selected non-member economies.”

News datasets

We started to build our own news dataset. But as you can see, it will not be very large and we will not be able to look into the past. There are some other resources you can tap if you want to use news as a datasource:

https://components.one/datasets/all-the-news-2-news-articles-dataset/ - “2.7 million news articles and essays from 27 American publications. Includes date, title, publication, article text, publication name, year, month, and URL (for some). Articles mostly span from 2013 to early 2020.” 500MB compressed text.

https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research#News_articles - a small list of some small news datasets.

Filtering the internet archive The internet archive regularly scrapes the whole internet and stores all the data. They provide the scraped data for free. They also provide a subset of the data that covers only news-websites. This is still a huge dataset that you can’t process on your laptop. But there is a nice script that you can use to download, filter and process the data: https://github.com/fhamborg/news-please/blob/master/newsplease/examples/commoncrawl.py

Freedom of information request

Most countries have a right similar to the [freedom of information act](https://en.wikipedia.org/wiki/Freedom_of_Information_Act_(United_States) of the United States. It gives its citizens the right to access and use the data of their country. But it is bound to forms, fees and waiting times. The answers may sometimes be in a hard to use format (pdf-scans…). But if you know that your government does have the information you want, this might be a really go way to get you valuable data. But make sure that you searched for this data in open data portals before.

I have personally never accessed data this way, but here are some resources to help you:

https://en.wikipedia.org/wiki/Freedom_of_information_laws_by_country

US-specific: https://foia.wiki/wiki/Making_a_FOIA_Request

Working with datasets

We already worked through some examples. Let's do some more

Converting data formats

Not every data provider serves the data in the format you want. But with some knowledge or R you can always convert the data to the format you need.

Exercise 2.4: Download more Data just like you did in exercise 2.2 and convert it to the following format and save it as a csv-file (use the countrycode package to convert the country names to country codes, if you need to):

Country, ISO3_countrycode, year, GDP, population, GDP per capita, GDP per capita (PPP)

–> See the file exercise_2.4.R