**Exploring Data Sources (Project)**

List of Data Sources (Instructional Staff Curated)

This is a set of data sources curated by the instructional staff. Feel free to suggest new data sources in the forums. The initial list was provided by Kevyn Collins-Thomson from the University of Michigan School of Information.

Long general-purpose list of datasets:

* <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

The Academic Torrents site has a growing number of datasets, including a few text collections that might be of interest (Wikipedia, email, twitter, academic, etc.) for current or future projects.

* <http://academictorrents.com/browse.php?cat=6>

Google Books n-gram corpus

* external link: <http://books.google.com/ngrams>
* Dataset: external link: <http://aws.amazon.com/datasets/8172056142375670>

Common Crawl: • Currently 6 billion Web documents (81 Tb) • Amazon S3 Public Data Set

* <http://aws.amazon.com/datasets/41740>
* <https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set>
* Award project using Common Crawl: <http://norvigaward.github.io/entries.html>
* Python example: <http://www.freelancer.com/projects/Python-Data-Processing/Python-script-for-CommonCrawl.html>

Business/commercial data Yelp external link:

* <http://www.yelp.com/developers/documentation/v2/search_api>
* [Upcoming Deprecation of Yelp API v2](https://engineeringblog.yelp.com/2017/06/upcoming-deprecation-of-yelp-api-v2.html) on June 30, 2018 (Posted by Yelp Jun 28, 2017)

Internet Archive (huge, ever-growing archive of the Web going back to 1990s) external link:

* <http://archive.org/help/json.php>

WikiData:

* <https://www.wikidata.org/wiki/Wikidata:Main_Page>

World Food Facts

* <http://world.openfoodfacts.org/data>

Data USA - a variety of census data

* <https://datausa.io/>

Center for Disease Control - variety of data sets related to COVID

* <https://data.cdc.gov/browse>

U.S. Government open data - datasets from 75 agencies and subagencies

* <https://data.gov/>

NASA data portal - space and earth science

* <https://data.nas.nasa.gov/>
* <https://data.nasa.gov/>