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

This website has dozens of public datasets - some fun, some a bit, well.. quirky. external link:

- https://rs.io/100-interesting-data-sets-for-statistics/

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: