Final Project Data Ideas

Data Science 602: Introduction to Data Analysis and Machine Learning

Spring 2023

There are many resources available to obtain machine learning data. This document provides several resources across a number of domains to provide a starting point for finding real-world data. While many of these sources make available datasets for direct download, others require access to data through an Application Programming Interface (API).

This document is intended as a rough working list, and comments or ideas for other resources are welcomed.

1 General purpose datasets

Google Dataset Search (`https://datasetsearch.research.google.com/`) is a search engine for datasets that covers thousands of public repositories.

Wikidata (`https://www.wikidata.org/`) is an open knowledge graph of data in Wikimedia projects such as Wikipedia. Wikidata offers a powerful query service through which one can invoke queries against the knowledge graph.

2 Government data

Data.gov (U.S. government) (`https://data.gov/`) is a search engine for machine-readable datasets produced by the executive branch of the United States government. Other US Government datasets include:

- NASA (`https://data.nasa.gov/`)
- Smithsonian Institution (`https://www.si.edu/openaccess`) (API at `https://api.data.gov/docs/si/`)
- United States Congressional data API (`https://api.congress.gov/`)
- US Census Bureau (`https://data.census.gov/`)
- HealthData.gov (Public health datasets) (`https://healthdata.gov/`)

United Nations (`https://data.un.org/`) UNdata provides access to statistical resources compiled by the United Nations (UN) statistical system and other international agencies. Statistical themes include agriculture, crime, communication, development assistance, education, energy, environment, finance, gender, health, labour market, manufacturing, national accounts, population and migration, science and technology, tourism, transport and trade.

State and Local Government data Many state and local governments offer downloadable datasets. A small number of examples are included below, but information for other jurisdications are easily found. Examples include:

- New York City OpenData (`https://opendata.cityofnewyork.us/`)
- Maryland Open Data (`https://opendata.maryland.gov/`)
- Baltimore City (`https://data.baltimorecity.gov/`)
- Washington DC (`https://opendata.dc.gov/`)

3 Nonprofit Organizations

World Bank (https://data.worldbank.org/) The World Bank publishes data products that cover a wide range of development issues, much of which originates from the statistical systems of member countries.

Berkeley Library Listing of Non-Governmental Organization (NGO) data (https://guides.lib.berkeley.edu/c.php?g=496970&p=3401927) The Berkeley Library compiles a compendium of NGO sources spanning a wide variety of topics from NGOs.

ProPublica (https://www.propublica.org/datastore/) ProPublica publishes a wide variety of datasets covering topics of public interest, most freely downloadable.

Wikipedia REST API (https://en.wikipedia.org/api/rest_v1/) provides access to Wikipedia data.

4 Application Programming Interfaces and Industry Resources

Many e-commerce and tech companies provide application programming interfaces (APIs). API access is typically free, but typically require an API key and impose use limitations. Many APIs have Python wrappers available.

Yelp Open Dataset (https://www.yelp.com/dataset/) The Yelp dataset is a subset of businesses, reviews, and user data for use in personal, educational, and academic purposes. The dataset is designed for machine learning applications and contains nearly 7 million reviews for 150,000 businesses.

Twitter (https://developer.twitter.com/) The Twitter API enables programmatic access to Twitter and allows access to core elements of Twitter like: Tweets, Direct Messages, Spaces, Lists, and users.

Spotify (https://developer.spotify.com/) The Spotify Web API endpoints return metadata about music artists, albums, and tracks from the Spotify Data Catalogue.

Open datasets for healthcare (https://odsc.medium.com/15-open-datasets-for-healthcare-830b19980d9) This article in Open Data Science provides a number of commercial and public data repositories relating to healthcare