Finding Data

Steve Pittard

February 2, 2016

Finding Some Interesting Data

This document attempts to identify sources of information that might be of interest for those needing data to analyze as part of a larger project. In general you can search for open datasets using Google with a search string like ("finding open data sets") so there really is no practical limit to what one can find. Note that some private data sets have crept out onto the Internet so don't assume that all data is open or free by default. This is especially true if you encounter data found on Bit Torrent sites. For example, The NetFlix Challenge involved the distribution of a large dataset to be used as part of a challenge. Unfortunately, NetFlix discovered that that data could be used to reverse engineer specific user ids of actual Netflix users so they withdrew it from distribtuion. But that didn't stop it from being copied a bunch of times prior to that. So, now it is not "legal" for one to use it but many still do and there are a variety of sites that still offer it for download. Proceed at your own risk.

Places like The Internet Movie Database use to have free data out there but since the rise of Big Data Analtics they have closed off access except for paying customers. They offer smaller subsets of data in unstructured format which makes it inconcvenient (though not impossible) to effectively mine that kind of data. Lot's of companies are moving into a direction wherein they protect their data so this isn't peculiar to IMDB.

Application Programming Interface ?

Some of these services offer Application Programming Interfaces (APIs) as well as built in visualization tools that can be helpful as good starting points for how one might present the data graphically. There is web site called The Programmable Web http://www.programmableweb.com/ that maintains a list of all data providers who offer an API. This is useful for identifying websites that allow you to programmatically access their data. They usually require some type of sign up or registration prior to use. Some of them might not be free but many are. Usually if they are free they will limit you to a specific number of access in a given time frame.

Formats ?

Lastly, keep in mind that not all of these sites offer so called "Big Datasets" but many do. Some of these require that you navigate to a specific section of the site to obtain raw or .CSV files. It might also be that the sites offer only a specific format such as XML or JSON in which case you will have to parse that information. We'll be looking at how to do that for some data formats later in the class. As an example of formats that are hard to parse the Million Song Database is provided in an HD5 format which can be parsed but not so easily. Amazon has offered space to host the entire 300GB dataset in a but it's too large for most people to download conveniently. Of course Amazon wants you to use their compute resources to analyze this data which is why they offer it for free. There is a subset of the MSD at http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset

Various Links

The list below is current to the best of my knowledge. That is all the links work as of this writing unless I've made a typo which is entirely possible. The list of interesting data grows all the time so just pick a domain of interest and Google it like "DNA Sequencing open data set" and I guarantee you that something

interesting will turn up. Also when perusing these sites if you find one of interest also search it to see if it offers an API which would simplify access. Sites are infamous for providing API access only to later revoke it usually because too many people use it. Or worse, they start charging lots of money for people to use it.

Description	Link
Medicare Data including Nursing Home information	http://data.medicare.gov
Amazon Public Data Sets	https://aws.amazon.com/public-data-sets/
Awesome Public Data Sets	https://github.com/caesar0301/awesome-public-datasets
Stanford Large Network Data Collection	http://snap.stanford.edu/data/
Machine Learning Data Repository	http://mldata.org/
UC Machine Learning Repository	http://archive.ics.uci.edu/ml/
City of Chicago	https://data.cityofchicago.org/
IMDB Movie Review Test Data	http://ai.stanford.edu/~amaas/data/sentiment/
City of San Francisco	https://data.sfgov.org/
Nee York City Data	https://nycplatform.socrata.com/
World Bank Data	http://econ.worldbank.org/datasets
US Bureau of Labor Stats	http://www.bls.gov/data/
US Census Data	http://www.census.gov/
US Energy Info Administration	http://www.eia.gov/opendata/
US EPA Data	http://developer.epa.gov/category/data/
USDA	http://www.ers.usda.gov/data-products/.aspx
US Open Data (Many types of Data)	http://www.data.gov/
UK Open Data	http://data.gov.uk/
European Open Data Portal	http://open-data.europa.eu/en/data/
US Health Data	http://www.healthdata.gov/
Canada Open Data	http://open.canada.ca/
France Open Data	http://www.data.gouv.fr
Gap Minder	http://www.gapminder.org/data/
Socrata Open Data	https://opendata.socrata.com/
World Health Organization	http://www.who.int/gho/en/
Enron Email Data	http://www.cs.cmu.edu/~enron/
CDC NHANES Data	http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm
Airline Challenge Data	http://stat-computing.org/dataexpo/2009/the-data.html
Kaggle	https://www.kaggle.com/
UC Berkeley Data Page	http://ucdata.berkeley.edu/
Links to Data on Reddit	https://www.reddit.com/r/datasets
Wiki Data Downloads	https://en.wikipedia.org/wiki/Wikipedia:Database_download
New York Times API	http://developer.nytimes.com/docs