Data Science Design Manual (Skiena) Exercises

Adam Catto

October 20, 2019

1 Chapter 1: What is Data Science?

1-1. [3] Identify where interesting data sets relevant to the following domains can be found on the web:

meta: good source in general is https://www.data.gov/

(a) Books:

- amazon.com
- https://www.goodreads.com/api/documentationcom – the goodreads API. Sign up for a developer key, then use HTTP GET requests to download XML files and use e.g. BeautifulSoup
- https://developer.nytimes.com/docs/books-product/1/overview – this is the NYT Books API. Can obtain list names, list data (e.g. "authors", "titles", "dates", etc.), and book reviews, serialized in JSON format. JSON files can be downloaded and processed using language of choice.

(b) Horse Racing:

- equibase.com
- http://www.horseracingdatasets.com/
- https://www.kaggle.com/noqcks/woodbine-races

(c) Stock Prices:

meta / pointer: https://www.investopedia.com/ask/answers/find-historical-stock-index-quotes/

- https://datahub.io/collections/stock-market-data
- https://iextrading.com/developer/
- https://www.marketwatch.com/tools/quotes/historical.asp

(d) Risks of Diseases:

- https://data.world/datasets/epidemiology
- https://catalog.data.gov/dataset?_organization_limit=0&organization=hhs-gov#topic=health_navigation
- http://www.spatialepidemiology.net/datasets/

(e) Colleges & Universities

- https://catalog.data.gov/dataset?tags=integrated-postsecondary-education-data-system&q=Integrated+Postsecondary+Education+Data+System#topic=education_navigation

- https://www.kaggle.com/theriley106/college-common-data-sets

- https://data-planet.libguides.com/c.php?g=454340&p=3103525

(f) Crime Rates

- https://data.cityofnewyork.us/Public-Safety/NYC-crime/qb7u-rbmr – NYC crime data. Can download in various formats, e.g. CSV, JSON, RDF, ...

- https://data.world/datasets/crime – example: US Mass Shootings datasets in xlsx format can be loaded in Microsoft Excel. (Or, converted in CSV format and processed as desired.)

- https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county – download CSV file.

(g) Bird Watching

- https://catalog.data.gov/dataset/idaho-birding-trail-routese136f – Idaho Birding Trail Routes data. Download geospatial data in CSV format or with ArcGIS.

- https://catalog.data.gov/dataset/wildlife-management-areasf8967 – Wildlife Management Areas. Arranged via township range section, geographic names, ITD roads, FS roads, hydrography, land ownership, county boundaries, etc.; can be downloaded in JSON format.

- https://catalog.data.gov/dataset/idaho-birding-trails-sites83614 – Idaho Birding Trail Sites data. Same as above.

1-2. [3] Propose relevant data sources for the following The Quant Shop prediction challenges. Distinguish between sources of data that you are sure somebody must have, and those where the data is clearly available to you.

1-3. [3] Visit http://data.gov, and identify ve data sets that sound interesting to you. For each write a brief description, and propose three interesting things you might do with them.

1-4. [3] For each of the following data sources, propose three interesting questions you can answer by analyzing them:

(a) Credit card billing data:

- What events occur at periods of time with the highest spending density? (e.g. holidays)?
- What does the probability distribution of prices look like?
- How much is spent on, say, goods/services corresponding to each level of Maslow's hierarchy? (follows an 80/20 rule, maybe?)

(b) Amazon Click Data:

- Given an ontology / data about object similarity, what is the average "distance" between consecutive objects clicked?
- Do people with ADHD have a significantly larger or significantly smaller average distance between objects / a significantly larger/smaller number of clicks per unit of time?
- Under what conditions do people tend to have the highest temporal click density?

(c) White Pages residential/commercial telephone directory:

- What area codes have the highest/lowest density of hospitals per capita?
- Has the intergenerational delta in name frequency changed significantly more/less in more progressive and/or afluent areas than in more conservative areas?
- Is there more homogeneity in the first/last names of people in regions that tend to vote Republican than there is in regions that tend to vote Democrat? (and, i suppose, how do we quantify homogeneity within a region, exactly?)