EECS-317 Data Management and Information Processing

Lecture 13 – A Data Safari

Steve Tarzia

Spring 2019

Northwestern

Announcements

- Final project
  - Part 1 due May 22nd (next Wednesday)
  - Part 2 due June 12th (Wednesday of finals week)
- HW5 using MySQL due next Friday.
  - Please check that your login credentials work ASAP.

Last Lecture (1): Data Files

- Data is exchanged by data files (arrays of bits, zeros and ones).
- Several file formats are common:
  - CSV, XML, JSON, and less commonly SQL and proprietary formats.
- Many of these formats are text files with special syntax.
- Text files represent each character with a certain bit sequence.
  - ASCII uses 8 bits (one byte) for each character
  - UTF-8 uses 1-4 bytes for each character, is backward-compatible with ASCII
- CSV files store just one table & can be imported into SQL easily.
- JSON and XML files represent data with complex, nested relationships
  - However, no schema is defined ahead of time.
  - Data itself gives the structured (hence, we call it **semi-structured** data).
  - Python and R scripts can easily load these files.

Last Lecture (2): Data APIs

- Bulk access to data is simple, not always possible
  - Data may be too big, dynamic, or guarded by the owner
- Data is often exposed to users through **data APIs**, which allow users to request pieces of the data. In particular:
  - **REST APIs** use HTTP requests to get data from remote servers.
  - This involves web requests that return JSON data instead of HTML pages.

Chicago Park District – Event Permits

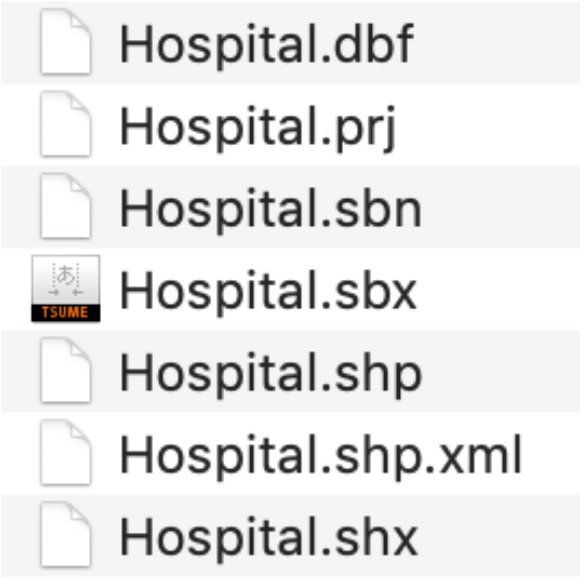
https://data.cityofchicago.org/Events/Chicago-Park-District-Event-Permits/pk66-w54g

- Data is just one big table of 78k rows.
- Can be downloaded as a CSV file.
- Also provides a "Data API," but it's really just a single url to fetch all the data in JSON format:
  - https://data.cityofchicago.org/resource/pk66-w54g.json

Cook County Hospitals

https://data.cityofchicago.org/Health-Human-Services/Cook-County-Hospitals/mkjv-t4kt

- Provides a zip file with several ArcGIS files:



- These provide geographic data (about the location and shape of hospital properties), and they require a GIS program to view.
- See also: Boundaries of Chicago Neighborhoods: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/9wp7-iasj

Chicago Crimes

- Dashboard: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g
- Data: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- Can be exported as one big CSV file with 6.8M rows.
- Or fetch with API:
  - All data:
    - https://data.cityofchicago.org/resource/crimes.json
  - Filtered data:
    - https://data.cityofchicago.org/resource/crimes.json?Primary+Type=THEFT

Stanford Dogs Dataset

- http://vision.stanford.edu/aditya86/ImageNetDogs/
- 20,850 images of 120 different dog breeds.
- Used for computer vision and machine learning research.
- Each image is accompanied by an XML file with annotations:

```
<annotation>
        <folder>02085620</folder>
        <filename>n02085620_7</filename>
        <source>
                <database>ImageNet database</datab
        </source>
        <size>
                <width>250</width>
                <height>188</height>
                <depth>3</depth>
        </size>
        <segment>0</segment>
        <object>
                <name>Chihuahua</name>
                <pose>Unspecified</pose>
                <truncated>0</truncated>
                <difficult>0</difficult>
                <bndbox>
                        <xmin>71</xmin>
                        <ymin>1</ymin>
                        <xmax>192</xmax>
                        <ymax>180</ymax>
                </bndbox>
        </object>
</annotation>
```

UW Madison Courses and grades

https://www.kaggle.com/Madgrades/uw-madison-courses

- 10 tables (10 CSV files)
- A sqlite file is provided

Lahman's Baseball Stats Database

- http://www.seanlahman.com/baseball-archive/statistics/
- Provides MS Access and MS SQL Server files
- Also provides 27 CSV files for other tools.

Google Books N-grams

- http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
- They scanned through lots of books and web pages to count the frequency with which words appeared in sequence.
- 2.2TB of tab-separated text data.

- You can play with the data here:
  - https://books.google.com/ngrams

Google Knowledge Graph API

- https://developers.google.com/knowledge-graph/
- Here's a sample request that asks for entities related to "taylor swift"
  - https://kgsearch.googleapis.com/v1/entities:search?query=taylor+swift&key=AIzaSyB9oolh0Sk_toyI6tVWzmlPKbEof1JwE8g&limit=10&indent=True
- Notice that spaces are converted to "+" characters in URLs
- Also, I had to register with Google and provide my secret key. So, I really should not be sharing this URL with you.

Google Geocoding API

- https://developers.google.com/maps/documentation/geocoding
- You provide an address in any format, like "2145 Sheridan Road, 60208" and it gives you the latitude and longitude coordinates:
  - https://maps.googleapis.com/maps/api/geocode/json?address=2145+Sheridan+Road,+60208&key=AIzaSyB9oolh0Sk_toyI6tVWzmlPKbEof1JwE8g
  - Again, I'm providing my secret API key in the URL above.
- This Data API is not just looking up an answer in a database.
- The request is processed in a complex way because it accepts addresses in many different formats, eg:
  - 2145 Sheridan Rd, Evanston
  - Mudd Library, Northwetsern University (Google tolerates the misspelling!)

Google Translate API

- https://cloud.google.com/translate/docs/reference/rest/v2/translate
- Provide text in one language and Google returns a translation:
  - "my name is Steve" translated to Spanish ("es" for español):
  - https://translation.googleapis.com/language/translate/v2?target=es&key=AIzaSyB9oolh0Sk_toyI6tVWzmlPKbEof1JwE8g&q=my+name+is+Steve.
  - Again, I'm providing my secret API key in the URL above.
  - This is an HTTP GET request.
- Version 3 of the API accepts requests in a JSON object using POST.
  - This allows longer texts and non-ASCII characters.

Data is provided in many different formats

- CSV files are common
- Geographic data uses special file formats ("shape files")
- A data set might include many files (eg., Stanford dogs)
- Multiple tables can be distributed as a single SQLite database file
- REST APIs allow fetching of data by providing query information in the URL (or in a POSTed JSON object).
  - Return value is usually a JSON object.
  - The data provider must provide a specification for the API, to tell users how to construct requests and how to interpret responses.