Lecture 1

Sparsh Agarwal

The lecture began with discussion about “Harvey’s energy toll” visualization diagram; the diagram can be found here : https://www.axios.com/harveys-energy-toll-1513305145-5c5826f5-60fa-4635-a617-19ce4a5135fd.html.

The main questions to be addressed were:

1. What does visualization show?
2. What methods it uses?
3. Its strengths/weakness.
4. The error in the visualization.

Suggested answers:

1. It shows the impact of hurricane Harvey on the operating conditions of the oil refineries situated along the Gulf Coast. The blue line in the figure describes the path of the hurricane—the bold line is the path already covered and the dashed line is the predicted path of the storm. The circles represents the oil refineries, with the capacity of the refinery increasing with the radius/size of the circle. The figure shows that as hurricane Harvey passes by an area, it results in large amount of rainfall in that particular region, resulting in shutdown of oil refineries in that area. Many oil refineries in Corpus-Christi, Houston can be seen to have got closed (indicated by the faded orange coloured circles) due to the extreme rain in those regions.

2. A. The varying amount of rainfall is shown by varying intensity of the blue colour (qualitative representation).
   B. The capacity of the refineries is positively correlated with size of the circle, and the circle also shows the location of the refinery.
   C. Shaded circles have been used to represent closed refineries, while the non-shaded ones are the ones that are still in operation.

3. A. It is difficult to distinguish between the closed and the opened refineries as the faded orange color is not that prominent, it sort of merges with the background blue color.
   B. The location of refinery is hard to predict just by looking at the circles, it seems ambiguous.
   C. The intersection of circles makes deciphering of the figure a bit difficult.

4. The capacity of the refineries in not proportional to the size of the circles. In other words, the rate of increase of the area of circles is not proportional to the square of the capacity of the refineries, it seems to be increasing at a larger rate that it is supposed to be.

Topics:

- Jupyter notebook: It was not ready so it will be discussed in the next class.

- Data formats:

  Background: Computers can only understand binary numbers, 1 and 0, so 1 and 0 are represented by the computers as on and off states respectively. The various combinations of 1 and 0 are used to make characters in a computer. A single binary digit is called a bit.

Lecture 1

Sparsh Agarwal

1. ASCII: It is an 8-bit character set and can represent 256 characters. 8 bits makes 1 byte.
2. CSV/TSV: CSV stands for comma separated values, and is used to represent tabular data in form of values separated by comma. Each value between 2 commas is part of a column, and every row represents a new row in the table. TSV is similar to CSV. The only difference is TSV has values separated by tabs instead of commas.
3. JSON: It is another data type which stores objects in itself, the objects being made up of human readable text. It stores the attribute-value pairs in form of objects. It is similar to XML.
4. HDF5: Hierarchical Data format. They allow different types of data stored in one file, and every group of data is self-describing. It was developed at UIUC.
5. PNG/JPEG,etc.: They are formats for compressed images.
6. Excel: Type of data stored in rows and columns when open with Microsoft Excel. This will open as a binary file if open without exporting it to a readable format.
7. Arrow: This is datatype related to pandas in python.
8. SQL: This is used in for making queries in database. It is a web-based format.
9. JSON/REST: REpresentational State Transfer - web architecture that keeps clients and servers independent and queries using flexible JSON formatting.

- Organization: Computer doesn't know what rows and columns are, thus, the way it stores tabular data is in linear fashion in the memory. In row-oriented storage, successive fields for a single record are adjacent. In column-oriented storage, successive records for a single field are adjacent. This helps in fast data processing.

CSVs: Can be separated by delimiters, like commas, spaces, etc. It is considered lowest common denominator format. Pretty much works for any type of data and for any type of software. Not every field need to have a value, it can be left blank.

Conversion of binary to decimal: Multiply every value in the array with 2 to the power index of the element/value. Sum up al of these values. Every number in computer is represented like this, be it integer, string, character, or anything, with help of ASCII characters or a string-float function.

JSON: Explained before. iPython files are also JSON formatted.

HDF5: Explained before. It is popular among data scientists. It allows you to jump to different field in data file without reading other fields. Thus, it is processed faster by computer. This format can have any type of data type in it, and everything is automatically converted to binary digits. Jumping from one part of file to other is also possible. Eg. In Mario image, it is possible to read all the white blocks and skip the black ones.

Filtering: It is extracting out information from data that is of interest.

Mutate: Converting from one unit to different unit, or like taking a log, etc. of the data.

Split: Turns the dataset into multiple datasets depending upon the characteristic we choose.

Nyquist limit: You should not sample your data less frequently than the highest frequency part of your data, otherwise, you will loose important details about the dataset. Eg. In the following graphs,

if the step size is increased, the curve is loosing its shape because of the loss in information about the dataset/curve.

Relationships:

- Equality operations can be used to compare whether two values belonging to same datatype.
- Greater than or less than signs can be used to compare quantitative measures or for order comparison.
- Set: Used to find the union or intersection between different sets of data.

The graphs have been discussed before. This was followed by a jupyter notebook section, link for which is this:

https://uiuc-ischool-dataviz.github.io/spring2019/nbv.html?notebook_name=%2Fspring2019%2Fweek02%2Fexamples_week02_pokemon.ipynb