Assignment #1 -- Data Preprocessing and Visualization

Deadline: Oct 6 23:59 Grade: 10%

Data preprocessing and visualization are important skills of managing Internet services such as search engines and online social networks. In this assignment, you can touch real data and learn some practical techniques to analyze the data, and you will have an unique opportunity to access **2 weeks of search logs** from a large search engine. Using the data, we give you the assignment to help learn both basic and advanced techniques of data analysis. The things you can learn include:

- **Basic statistics:** coding to count the data in different ways, e.g., how many queries are served per minute. You can also use Power BI do it.
- **Visualization:** plot figures to show the data (e.g., line chart, histogram, and CDF). You can also use visualization tools (such as Power BI) to do it.

To finish the assignment, you will also learn to use several visualization tools and statistical methods. For example: matplotlib; exploratory visualization tool Power BI.

Background and Data

Once you submit a query to a search engine, the search engine will log some related attributes regarding this query, such as when the query is submitted (e.g., timestamp) and the search response time (SRT).

In this assignment, we use 2 weeks of search logs from a global top search engine. Each day of search logs are written into one log file, so we have 14 (7 * 2 weeks) log files in total. Note that, because there are more than one billion queries submitted every day, we do not log them all but only a small random part of them. The log files are of the CSV (Comma Separated Values) format, and each column represents one attribute. The first line in the file contains the names of each attribute, and the following lines are the specific values for each query. A sample of the raw log file is as follows:

```
Timestamp,#Images,UA,Ad,ISP,Province,PageType,Tnet,Tserver,Tbrowser,Tother,S
RT
1411315200,0,Chrome,noAD,CRTC,Heilongjiang,sync,1495.0,443.14,73.0,0.0,2011.
14
1411315200,13,Chrome,noAD,CHINANET,Guangdong,async,200.0,80.0,47.0,359.0,686
.0
1411315200,24,Chrome,noAD,UNICOM,Hunan,async,120.0,450.0,26.0,191.0,787.0
1411315200,4,Safari,noAD,OTHER,Guangdong,async,272.0,86.0,234.0,219.0,811.0
1411315200,25,Safari,AD,UNICOM,Jiangsu,async,582.0,449.0,297.0,875.0,2203.0
```

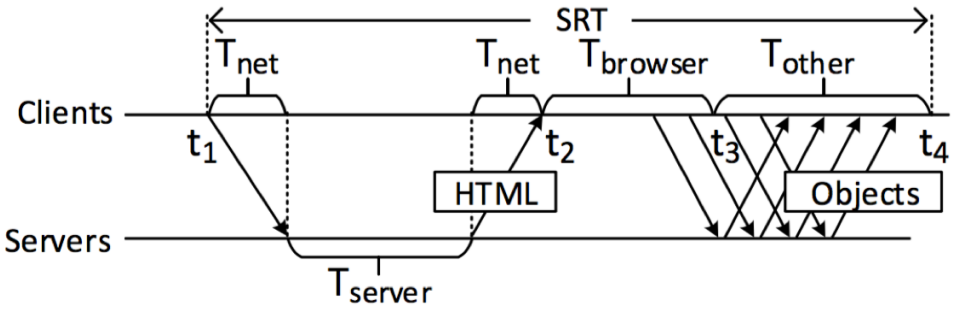
To be more clear, we show the above data in a table below (you can get the table view of the data by simply importing the CSV file into MS Excel, etc.).

Timestamp	#Images	UA	Ad	ISP	Province	PageType	Tnet	Tserver	Tbrowser	Tother	SRT
1411315200	0	Chrome	noAD	CRTC	Heilongjiang	sync	1495.0	443.14	73.0	0.0	2011.14
1411315200	13	Chrome	noAD	CHINANET	Guangdong	async	200.0	80.0	47.0	359.0	686.0
1411315200	24	Chrome	noAD	UNICOM	Hunan	async	120.0	450.0	26.0	191.0	787.0
1411315200	4	Safari	noAD	OTHER	Guangdong	async	272.0	86.0	234.0	219.0	811.0

The description of those attributes in the above table is as follows.

- **Timestamp:** the unix timestamp when the query is submitted. For example, "1411315200" represents "2014/9/22 0:0:0".
- **#Images:** the number of images embedded in the result page.
- **UA:** user agent, the type of the user's browser where the query submitted from.
- **Ad:** whether the result page contains ads or not, "AD" for yes and "noAD" for not.
- **ISP:** the ISP (Internet Service Provider) that the user or the query comes from.
- **Province:** the location of the user (32 provinces for this attribute since the logs only contain queries from China mainland).
- **PageType:** whether the page is loaded synchronously or asynchronously.
- **Tnet:** the 1st component of SRT (ms), the page transmission time over the network (see the figure below).
- **Tserver:** the 2nd component of SRT (ms), the server-side processing time of the query.
- **Tbrowser:** the 3rd component of SRT (ms), the DOM parsing time of the browser.
- **Tother:** the last component of SRT (ms), the remaining time for acquire other embedded elements in the page, such as images.
- **SRT:** search response time (ms), which is the sum of the above four SRT components.

The figure below shows a simplified timeline of a search, where you can find the details of the four SRT components and their relationship with SRT.



A very basic skill when dealing with the data is to count the numbers for different purposes, and then visualize them. In this assignment, you need to do the following jobs using the data:

- Calculate the average SRT of every 10 minutes, and plot the SRT with a **line chart** (x axis for date time and y axis for the average SRT).
- Calculate the average of each SRT component of every 10 minute, and plot the four SRT components together with a **stacked area chart** (x axis for date time and y axis for time) and also a **100% stacked area chart** (y axis for the percentage).
- Plot the **CDF (Cumulative distribution function) chart** of SRT.
- Plot the **CDF chart** of #Images.
- Count the number of queries (also called page views or PVs) of each minute, and plot the

minute-level PVs with a **line chart** (x axis for date time and y axis for the PVs).
- Count the PVs of each province, and plot it with a **histogram chart** (x axis for province and y axis for PVs).
- Count the PVs of each UA, and plot it with a **pie chart** (show the percentages in the chart).
- What are the differences among those charts (How to decide which one to use)
- Describe your experience or findings in doing those jobs. For example, experience of processing the data, observations from the charts, characteristics of the data, potential explanations, and any interesting things you would like to mention.

data

https://www.dropbox.com/s/akef557hnla0h9v/ANM-data.zip?dl=0

https://cloud.tsinghua.edu.cn/f/c8806b4c81ee45afa03c/?dl=1

submission

Please submit a report on web-learning. The result should contain the results of all above jobs.

Reference

[1] FOCUS: Shedding Light on the High Search Response Time in the Wild. INFOCOM 2016.