Project: Centre Pompidou Library Data Analysis
February 2018

The Centre Pompidou Library in Paris http://www.bpi.fr/the-library is an immense public library open to the general public in the center of Paris. It receives over 120,000 visitors per month. There are 250 open access computers for the public to do searches, watch movies, write emails, etc.

We have received data that are logs of what URLs the users of these computers are looking at. The goal of the project is to analyse the data, and to create a visualization that expresses the data. The intent is to explore and see what each student will discover. It is an open-ended project but the goal can be to get a sense of who the public is, and possibly what their cultural interests and backgrounds may be.

I will be happy to work with any student who takes on this project as I can help translate and clarify any questions. The data is proprietary to the Pompidou library, so I will need to have an email agreement that the data will not be distributed publicly except as part of the project.

The DATA FILES
We have received 4 topic-based set of searches: News, searches, translation, and videos. The files sample 4 different dates to make comparisons between when schools are in session (March 2017), and summer (June, August 2017) as a way to see if student usage of the library may impact on the content in any way.

News viewing example: 10.6.7.21-PUB-2-INT-243;2017-03-13T14:37:29.000Z;nymag.com;http://nymag.com/daily/intelligencer/2017/03/breitbart-may-have-just-killed-trumpcare.html;Breitbart May Have Just Killed Trumpcare

logs-bpi-news-2017-03-13-2017-03-20.csv
logs-bpi-news-2017-03-20-2017-03-27.csv
logs-bpi-news-2017-06-15-2017-06-21.csv
logs-bpi-news-2017-08-14-2017-08-21.csv

A search example: 10.6.9.102-PUB-1-INT-125;2017-03-13T09:31:31.000Z;bing.com;http://www.bing.com/search?q=google&src=IE-SearchBox&FORM=IESR02;google

logs-bpi-search-2017-03-13-2017-03-20.csv
logs-bpi-search-2017-03-20-2017-03-27.csv
logs-bpi-search-2017-06-15-2017-06-21.csv
logs-bpi-search-2017-08-14-2017-08-21.csv

A translation example: 10.6.20.12-PUB-1-INT-101;2017-03-19T12:52:38.000Z;linguee.fr;http://www.linguee.fr/francais-anglais/search?qe=m&source=auto&cw=709&ch=822;francais;anglais;m

logs-bpi-traduction-2017-03-13-2017-03-20.csv
logs-bpi-traduction-2017-03-20-2017-03-27.csv
logs-bpi-traduction-2017-06-15-2017-06-21.csv
logs-bpi-traduction-2017-08-14-2017-08-21.csv

A video example: 10.6.20.44-PUB-2-PRS-209;2017-08-14T14:02:52.000Z;dailymotion.com;http://www.dailymotion.com/video/x2s61xj;Bombay Velvet Full Movie english subtitles by Atilla - Dailymotion

logs-bpi-video-2017-03-13-2017-03-20.csv
logs-bpi-video-2017-03-20-2017-03-27.csv
logs-bpi-video-2017-06-15-2017-06-21.csv
logs-bpi-video-2017-08-14-2017-08-21.csv

Data Processing: We have applied filters to these files (to delete pages from ads, pop up ...)
Each log has the following fields:

* @ version *: latest version of the log
* @ timestamp *: date of the request to the web page (day, month, year, hour)
* host *: name of the proxy through which the request is made
* ip *: virtual address of the post used
* req *: request type (GET or POST)
* url *: url of the web page consulted
* program *: connection protocol
* res *: return code of the request to the web page (200, 404 ...)
* catolfeo *: automatic categorization by the BPI proxy of the site visited by field of activity (see correspondence http://support.olfeo.com/kb/article/2717)
* domain *: domain name of the visited site
* domain2 *: aggregate domain name of the visited site (ie: video.lequipe.fr from come lequipe.fr)
* date *: simplified date of the request to the web page (year, month, day)
* post *: name of the internet post used
* geo1 *: floor in the BPI where the internet station is located (floor 1, 2 and 3)
* geo2 *: sector in the BPI where the internet station is located (eg: SCT for Space Science and Technology)
* geo3 *: number of the internet station used
* session_id *: unique identification of the session that visited the pageweb

---