Fachhochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Tooling for big data extraction

Robert Beilich
Bachelorarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 23.10.2020

Idea

Working with big datasets is a challenge, which reaches more and more aspects of our lives. Efficiently gathering information from unstructured data like web pages is an even more demanding process, as it also includes finding the relevant information instead of just utilising it.

As a practical example the CommonCrawl dataset is used to collect the JavaScript libraries that are or were used over time.

Based on this example, problems and solutions that arise, when working with big datasets of unstructured data, are presented.

Analysing the retrieved data could help to display, for example, the most used libraries of a given year or to make predictions about the success of new libraries, based on information about existing libraries. These insights could then enable developers to decide, if it is worth to invest time into a new library, or if it has no future.

Common Crawl

<script src=“jquery.js“></script>

Environmental setup

Starting with a small machine of 6GB of RAM for the database and laptops to run the parser and transitioning to a cluster of machines, provided by the Future SOC Lab, with 1TB of RAM and 40 cores each, presents different difficulties on each end of the spectrum. A common bottleneck of both approaches is the underlying network bandwidth. The processing power, provided by the Future SOC Lab, cannot get exhausted, because the bandwidth is the limiting factor.

Resulting dataset

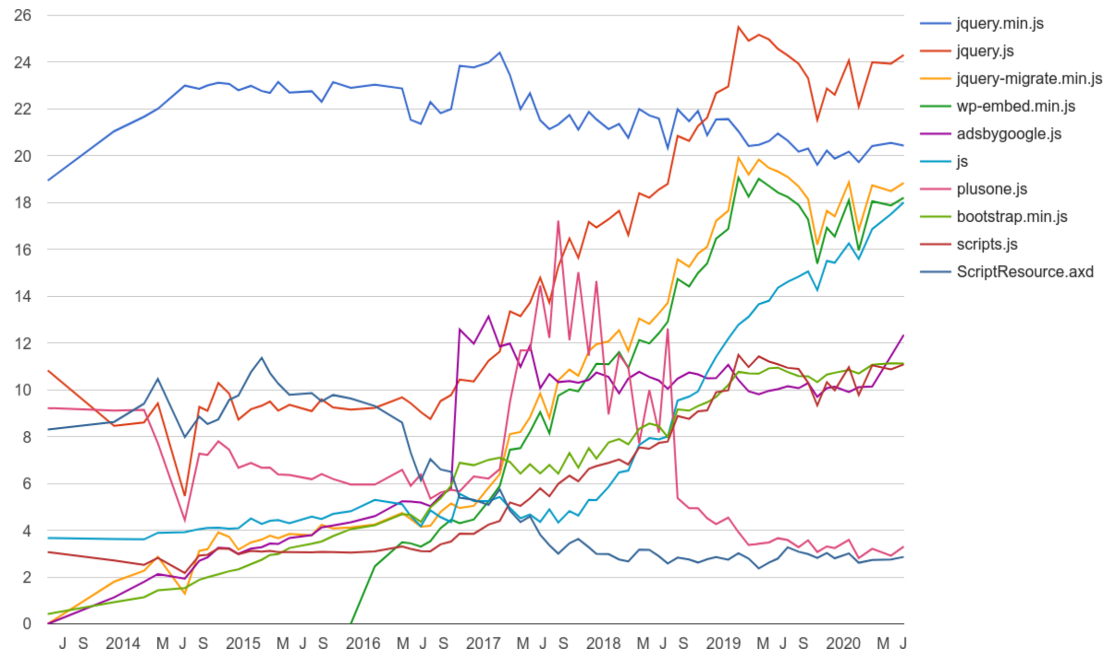
The parsed data is capped at the disk size of 1TB for the database, to get a exemplary portion of the complete dataset, which still just accounts to merely 0.0046% of it. This results in 7.7 billion unique sources from 829 million web pages.

Parser concept

Beautiful Soup may be the go to solution when it comes to parsing web pages in Python, but it also comes with some issues. Extrapolating from the time needed to extract the script tags from 500 web pages onto a complete crawl shows that it is too slow. The extraction needs 3.5 years for the data of one month, as shown in Table 1. Modifications on the extraction process are evaluated.

Method	time for 500 urls	time for 1 crawl	working
Beautiful Soup	28s	3.5y	yes
Beautiful Soup (filter)	23s	2.9y	yes
Regex	0.002s	9d	no
lxml	1s	46d	yes

Table 1: Comparision of source extraction methods



Analysis

Looking at the results of the parsing process shows problems in the parsing process, like misinterpreted names of libraries and some libraries representing the same library in different versions.

Conclusion

Analysing only this subset of the potential data, already reveals the fact, that the analysis is and will be a problem of big data extraction itself.

Going full circle, some of the measurements can get applied to this big data problem as well, with the additional benefit to include the process in the already existing one, connecting them, to simplify the overall process.

Sources

cog icon made by Becris from www.flaticon.com
CommonCrawl logo made by CommonCrawl Foundation

Betreuung Prof. Dr. Sven Buchholz • Technische Hochschule Brandenburg • Dipl. Informatik Ingo Boersch • Technische Hochschule Brandenburg