MPRI Web Data Management Project

Antoine Amarilli Liat Peterfreund

December 5, 2022

1 Overview of the project

The purpose of this project is to design and implement a novel application of Web data, that includes aspects such as data acquisition, extraction, cleaning, processing, integration, visualization, on one or several data sources from the Web (downloaded or used as services) such as:

- individual Web sites;
- existing crawls such as CommonCrawl or archives such as the Wayback Machine;
- social networking sites: Twitter, Facebook, Mastodon, etc.;
- knowledge bases such as Wikidata;
- Wikis such as Wikipedia, Web forums, QA Web sites such as StackExchange;
- scientific publication sources such as Crossref, BASE, DBLP, arXiv[1], SciHub, the S2ORC[2] citation dataset;
- geographical Web sources such as OpenStreetMaps, geolocation services;
- your own personal data: GPS traces, emails, calendar, contacts, bank statements, call logs, etc.;
- RSS feeds;
- open data sets: `data.gouv.fr`, `data.gov`, `data.gov.uk`, etc.
- legal data, e.g., the Légifrance data, or the DILA datasets[3]
- public transportation data, e.g., `data.ratp.fr`
- climate-related data sources and tools: `www.climatewatchdata.org`, the ADEME "Base carbone"[4], the CO2 Signal API at `docs.co2signal.com`, the IEA datasets at `www.iea.org/data-and-statistics` or the WRI datasets[5]

[1] For bulk download, see `https://archive.org/details/arxiv-bulk`

[2] `https://github.com/allenai/s2orc`

[3] `https://www.dila.premier-ministre.gouv.fr/repertoire-des-informations-publiques`

[4] `https://www.bilans-ges.ademe.fr/fr/accueil/contenu/index/page/decouverte/siGras/1`

[5] `https://www.wri.org/resources/data-platforms`

- semantic Web services and the linked open data cloud;
- microdata such as schema.org present on Web sites, e.g., via the Web Data Commons dumps at `webdatacommons.org`;
- online collaboration platforms (Etherpad, Moodle);
- linguistic resources (Wordnet, TLFi, Wiktionary);
- etc.

Other contributions to Web-related software such as Web browsers, servers, caches, etc., can also be considered.

2 Organization

Projects can be carried out by individual students, or (preferably) in groups of two. Projects chosen by different groups of students can integrate to each other (if relevant). The project chosen by each group of students will need to be submitted on the Moodle platform by **December 14** at the latest, for approval. Groups will defend their contributions, by giving an overall presentation and showing a demonstration of their system on February 27. The software needs to be developed on an open version control platform such as GitLab or GitHub and licensed as open source. A minimal documentation (README file) should be provided to explain the goal of the project, the structure of the code, the dependencies, and how to deploy and run the system.

3 Evaluation

The project is expected to be an implementation project; contributions that are more at the algorithmic level will also be accepted, but implementations of these algorithms are still required. The following elements will be particularly valued when evaluating a group's work:

- Depth of the contribution;
- Applicability value of the software and usability;
- Integration within existing software, services, platforms (in particular, contributions to existing open-source code bases are allowed and encouraged);
- Impact, wow effect of the demonstration;
- Initiative, creativity, originality;
- Good engineering practices, code quality;
- Quality of the presentation itself.

4 Questions

Questions about the project should be posted on the course forum on the Moodle platform.