Integration of heterogeneous datasets

Arthur Imbert

September 2017

The internship took place in the Parietal team at INRIA (https://team.inria.fr/parietal/). My supervisors was Gaël Varoquaux.

1 Summary

2 Introduction

2.1 Heterogeneous data, producers and value

Nowadays, the production of data has considerably grown. Each organization exploits and issues data, using its own schema. Most of the time, these files are produced for specific purpose, taking into account some internal rules or constraints. That leads to a vast amount of data available online, often sharing the same file formats, but still deeply heterogeneous by their structure. Indeed, this heterogeneity complicates the integration of different files.

By the same time, the current enthusiasm around data science and its predictive models shows a high potential for crossing data.

2.2 A theoritical framework

Related problems

Our strategy

3 An application to Open Data

3.1 Open and heterogeneous data

In order to collect a vast amount of heterogeneous data we exploit the Open Data available online.

A French platform exists, gathering all the information needed to use the files issued by the French public organizations: data.gouv.fr. This website hosts and indexes multiple sources

Figure 1: Neighborhood examples

Figure 2: Overview of the process

(a) An example of file easy to clean

(b) An example of 'dirty data'

Figure 3: Cleaning data

3.2 How can we reuse the data

3.3 An academic angle

3.4 Internship's goal

4 From a bunch of heterogeneous data to a recommendation system

4.1 Build a database

The first step is to store a sufficient amount of dirty data in order to keep a statistically significant volume of data throughout all our process.

Collect the metadata As described above, the platform data.gouv.fr offers us the opportunity to easily collect a numerous files from different French public organizations. A RESTful API exists in the website. It allows us to get, from a HTTP request, several metadata about more than 25000 pages. These include producer's identity, a description of the page's content, some related tags, the different files belonging to the page and, the most important, an URL for each file[1].

Some URL lead us to another website which host files, some directly launch the download of file. We only focus on the latters. It already represents around 60000 URL.

Download the data We use an INRIA's server DRAGO and a parallelized process to efficiently download the maximum of data. We use DRAGOSTORAGE, another server, to store the results of our downloads. By the end it represents around 350 Gb of data for more than 55000 files.

4.2 Clean and filter the data

Among these files, there are several format: text, JSON, XML, spreadsheet, zip... Indeed, the second step consist in inferring the nature of file in order to filter it. The right process could then be applied in a attempt to clean the file and reshape its content in a tabular form.

Filter the formats According to the supposed format of the file, we apply a different strategy to clean it. That supposed to avoid the empty files and detect its MIME type. If we have a zip file, we repeat the procedure for every zipped elements. We mainly focus on three type of documents:

-

CSV files

Excel files

JSON and GEOJSON files

Errors and metadata

[1] A more detailed description of metadata is given in annexes

4.3 Build a file embedding

Text data preprocessing

Topic extraction

Metric learning

4.4 Find neighborhood

5 Results and validation

5.1 Metrics and experimental framework

5.2 Results

5.2.1 Presentation of the data

5.3 Wordclouds

5.3.1 Metric comparison

6 Discussion

6.1 Limitations

6.2 Problems

6.3 Future work

7 Conclusion

8 Annexes