Internship's progress

First presentation

Johyn Papin

May 22, 2018

National Institute of Informatics

Table of contents

Introduction

The project

The project concerns a Linked Data database containing recipes and ingredients.

The project can be divided as follows:

- Extract prices and nutritional information from several sources
- Enrich the database with these data
- Use these data to calculate new information

Progress

What I've done

Here is what has been completed:

- Understand the linked data and its different formats
- Learn how to write SparQL queries
- Enrich the database with AGROVOC vocabulary
  - Read an n-triples file in GO
  - Represent an RDF graph in GO
  - Querying the AGROVOC API
  - Process n-triples as a stream
- Extract prices and nutritional information from monoprix
  - Bypass the security of "monoprix".
    - Use of TOR and request for a new circuit each time it is blocked
    - Use of a headless browser (firefox)
  - Scraping the menu, the categories and finally the products
  - Output the products as a stream
  - Scraping multiple pages at the same time (concurrently)

What I'm doing

Here is what I'm doing now:

- Dockerize the program
- Use an indexing system to search monoprix data
- Use sparql to process the triples of the database in a stream fashion
- Use sparql to add triples to the database

Problem encountered

- The security of monoprix. Solutions tested:
  - simple GET queries
  - as before but with cookies and random user-agents
  - as before but with TOR proxy
  - using a headless browser
  - as before but with TOR proxy
  - **as before but asking a new TOR circuit after each problem (and a new firefox profile)**
- Race conditions and other concurrent issues. Solutions:
  - Mutex (locks)
  - Work queue
  - State machine

Conclusion

Summary

Get the source of this presentation on

`github.com/johynpapin/yuuri`

Questions?