SCRAPY
XPATH
SELENIUM TESTS

LP Web & Mobile 2020-2021

IUT Orléans
Gérard Rozsavolgyi
roza@univ-orleans.fr

Framework crawling

- Tâches répétitives de crawling
- BeautifulSoup + Requests peut convenir dans des cas simples
- Automatisation : Framework plus avancé : scrapy

Requêtes CSS ou XPATH

- En HTML, on peut cibler n'importe quel élément d'une page en CSS ou en XPATH
- Requête CSS exemple :
  div > ul > li > input[name="email"]
- XPATH étend cette notion et permet de cibler au format XML n'importe quel contenu XML ou HTML

Exemple

```xml
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
<book category="COOKING">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>

<book category="CHILDREN">
  <title lang="en">Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>

<book category="WEB">
  <title lang="en">XQuery Kick Start</title>
  <author>James McGovern</author>
  <author>Per Bothner</author>
  <author>Kurt Cagle</author>
  <author>James Linn</author>
  <author>Vaidyanathan Nagarajan</author>
  <year>2003</year>
  <price>49.99</price>
</book>

<book category="WEB">
  <title lang="en">Learning XML</title>
  <author>Erik T. Ray</author>
  <year>2003</year>
  <price>39.95</price>
</book>

</bookstore>
```

XPATH

- Tous les titres: /bookstore/book/title
- Titre du premier livre: /bookstore/book[1]/title
- Tous les prix: /bookstore/book/price[text()]
- Tous les prix <30: /bookstore/book[price<30]/price

Scrapy

- scrapy startproject essai

- essai
  scrapy.cfg # fichier configuration
  essai/ # project's Python module
  __init__.py
  items.py # definition des items du projet
  pipelines.py # pipelines du projet
  settings.py # project settings
  spiders/ # araignées

Lancement et récupération données

- scrapy crawl essai-spider
- scrapy crawl essai-spider -o data.json

Exemples

- http://quotes.toscrape.com/
- Code : https://github.com/scrapy/quotesbot
- Crawlings de winestores