WebScraping

Mots clés

- HTTP
- HTML
- BeautifoulSoup

HTTP

- Déjà vu avec les API

Texte écrit par l'utilisateur

```
$ curl -v http://perdu.com
*   Trying 208.97.177.124...
* TCP_NODELAY set
* Connected to perdu.com (208.97.177.124) port 80 (#0)
> GET / HTTP/1.1
> Host: perdu.com
> User-Agent: curl/7.64.1
> Accept: */*
>
< HTTP/1.1 200 OK
< Date: Fri, 10 Jan 2020 13:23:47 GMT
< Server: Apache
< Upgrade: h2
< Connection: Upgrade
< Last-Modified: Thu, 02 Jun 2016 06:01:08 GMT
< ETag: "cc-5344555136fe9"
< Accept-Ranges: bytes
< Content-Length: 204
< Vary: Accept-Encoding
< Content-Type: text/html
<
<html><head><title>Vous Etes Perdu ?</title></head><body><h1>Perdu sur l'Internet ?</h1><h2>Pas de panique, on va vous aider</h2><strong><pre>    * <----- vous &ecirc;tes ici</pre></strong></body></html>
* Connection #0 to host perdu.com left intact
* Closing connection 0
```

Requête

Header de la requête

Status code de la réponse

Header de la réponse

Body de la réponse

HTML

- HyperText Markup Language
- Langage de balisage
  - <p>, , , <form>, …
- Décrit la structure d’une page
  - Pas son design
- Aujourd’hui : HTML5

HTML

- Interpréteurs HTML : navigateurs web (Firefox, Safari, Edge, Chrome, …)
- Possibilité de lire le code source de la page
- HTML = contenu non protégé

BeautifulSoup

- Analyse syntaxique de documents HTML
- $ conda install beautifulsoup4
- https://www.crummy.com/software/BeautifulSoup/

```
import requests
from bs4 import BeautifulSoup

html_doc = requests.get('https://perdu.com')
soup = BeautifulSoup(html_doc, ‘html.parser’)

title = soup.head.title
paragraphs = soup.find_all('p')

for paragraph in paragraphs:
    print(paragraph.text)
```

La Loi dans tout ça ?

- Droit d’auteur
- Propriété Intellectuelle
- Licenses (Creative Commons, etc…)
- Par défaut : interdiction de copier / ré-utiliser. Même en citant.