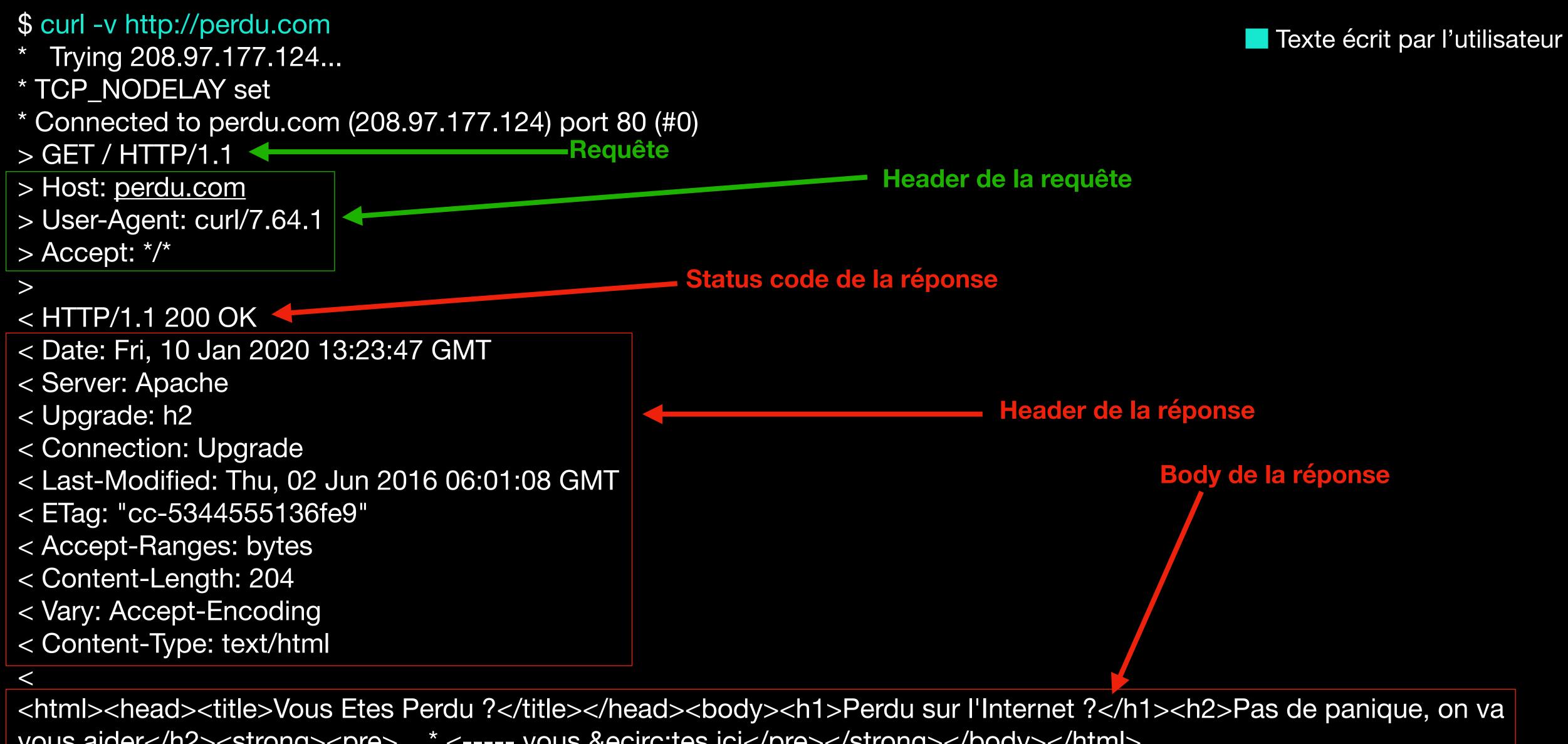
WebScraping

Mots clés

- HTTP
- HTML
- BeautifoulSoup

Déjà vu avec les API



^{*} Connection #0 to host perdu.com left intact

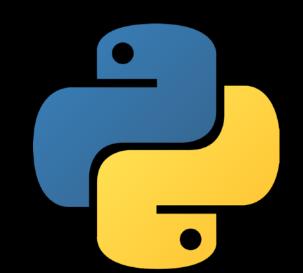
^{*} Closing connection 0

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

- 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.