



# Scraping

Julien Romero

# Introduction

# Qu'est-ce que le scrapping ?

- Le (web) scraping est un ensemble de techniques permettant d'**extraire des données** de sites webs
- Souvent associé au web crawling qui parcourt les pages du web de manière automatique

# Pourquoi faire du scraping ?

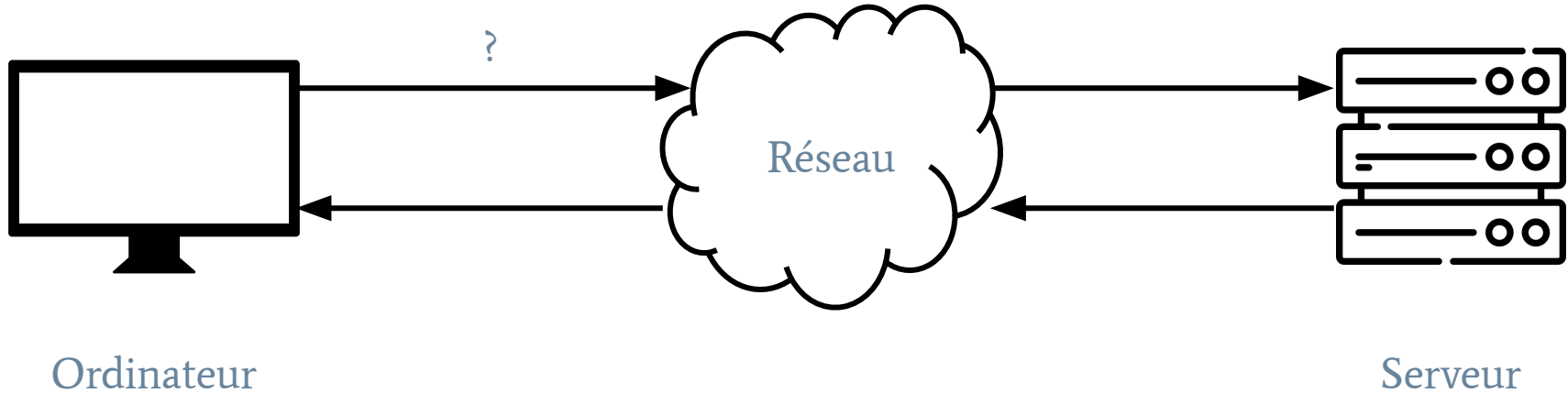
- Référencement : construction d'un moteur de recherche
- Extraction d'informations spécifiques
  - Comparateur de prix (billets d'avions, objets)
  - Agrégation d'évaluations de produits
  - Détection de changement dans un site web
  - Évaluation des tendances
  - Suivi de la présence en ligne
  - Détection de menaces
- Construction ou augmentation de jeu de données
- Test automatique de site web

# Comment faire du scrapping ?

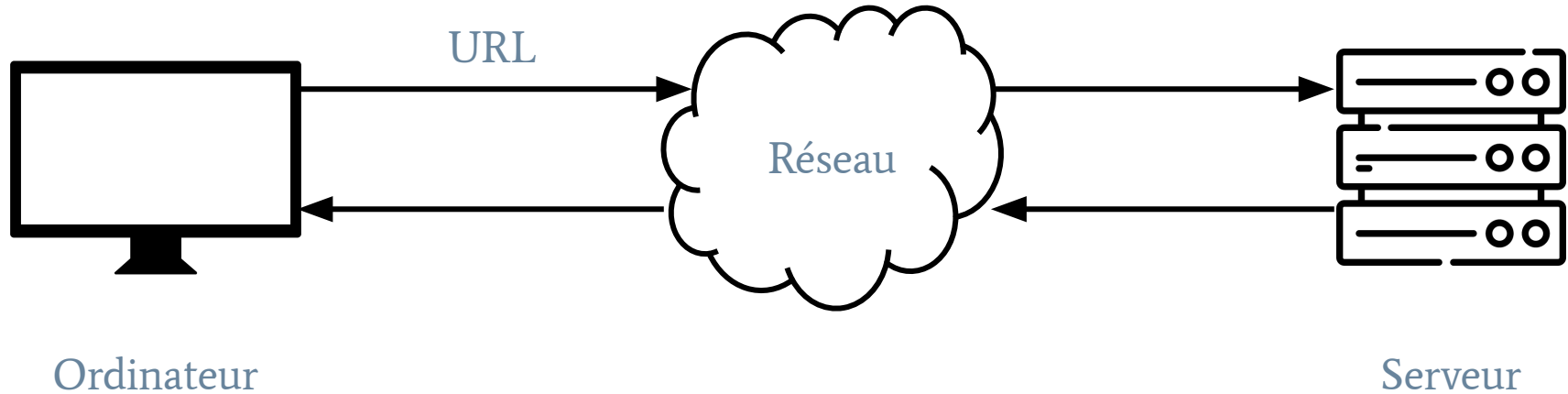
- Il est important de comprendre le fonctionnement du web et des différents composants
- Il faut “reverse-engineer” le site scrappé pour comprendre comment extraire au mieux l’information

# HTTP

# Comment obtenir une page web ?



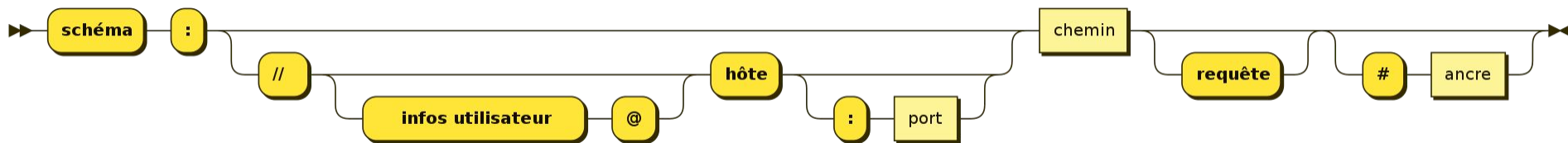
# Comment obtenir une page web ?





# URL

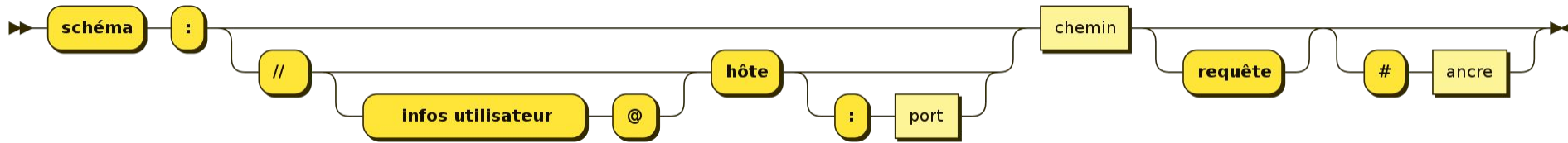
- URL = Uniform **Resource** Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



- Schéma = protocole, en général *http* ou *https*, mais aussi *ftp*, *mailto*, *irc*, ...

# URL

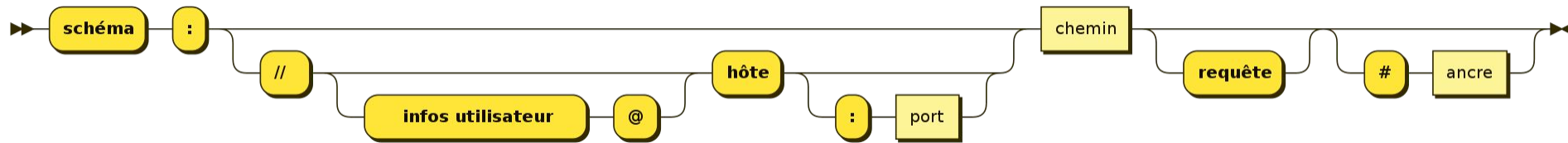
- URL = Uniform **R**esource Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



- Infos utilisateur : possibilité de mettre un nom d'utilisateur et un mot de passe

# URL

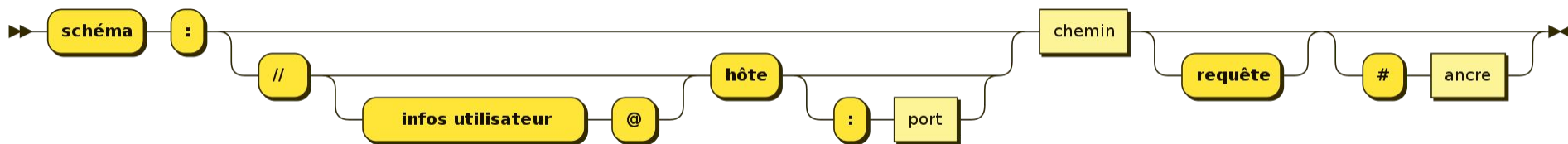
- URL = Uniform **Resource** Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



- Hôte : nom de domaine ou adresse IP permettant d'identifier la destination
  - Les noms de domaines sont transformés en IP par le DNS
  - <https://www-inf.telecom-sudparis.eu/COURS/CSC4538/Supports/>

# URL

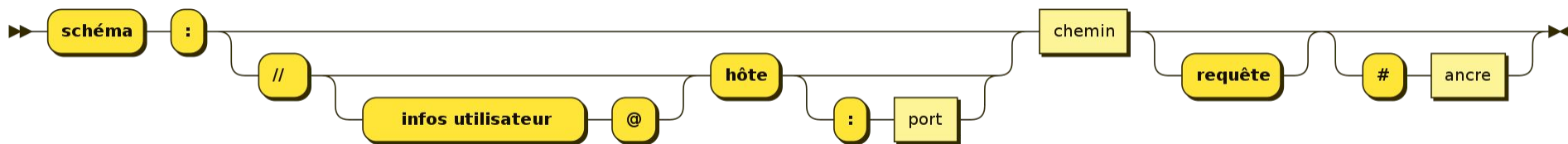
- URL = Uniform **R**esource Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



- Port : on peut indiquer le port de connexion
  - Par défaut, 80 pour HTTP, 443 pour HTTPS

# URL

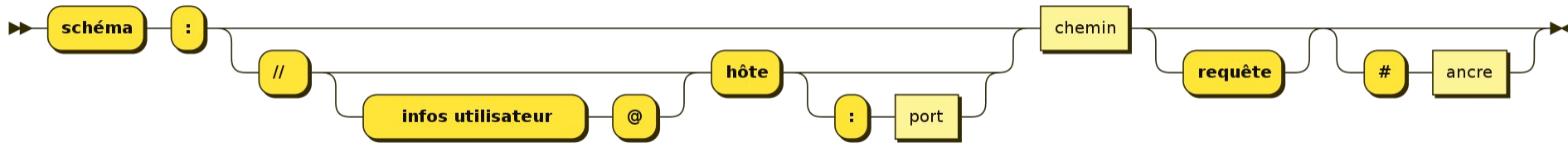
- URL = Uniform **R**esource Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



- Chemin : chemin absolu vers la ressource sur le serveur
  - Comme dans UNIX
  - <https://www-inf.telecom-sudparis.eu/COURS/CSC4538/Supports/>

# URL

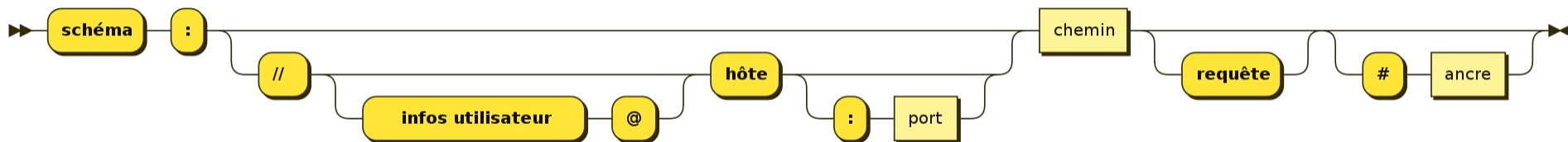
- URL = Uniform **Resource** Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



- Requête : Permet de passer des arguments ou données supplémentaires
  - Arguments souvent séparés par un délimiteur (&)
  - <https://www-inf.telecom-sudparis.eu/COURS/CSC4538/Supports/?page=exercices/python&wrap=true&soluce=true>

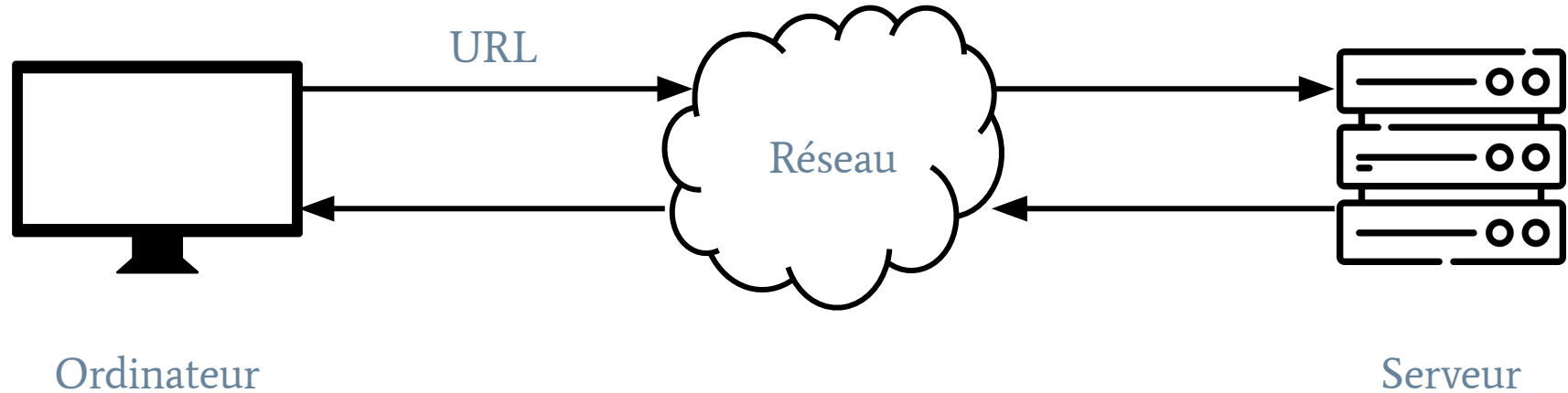
# URL

- URL = Uniform **R**esource Locator
  - Sert à localiser une ressource dans un réseau
- Cas spécifique d'une URI = Uniform Resource Identifier



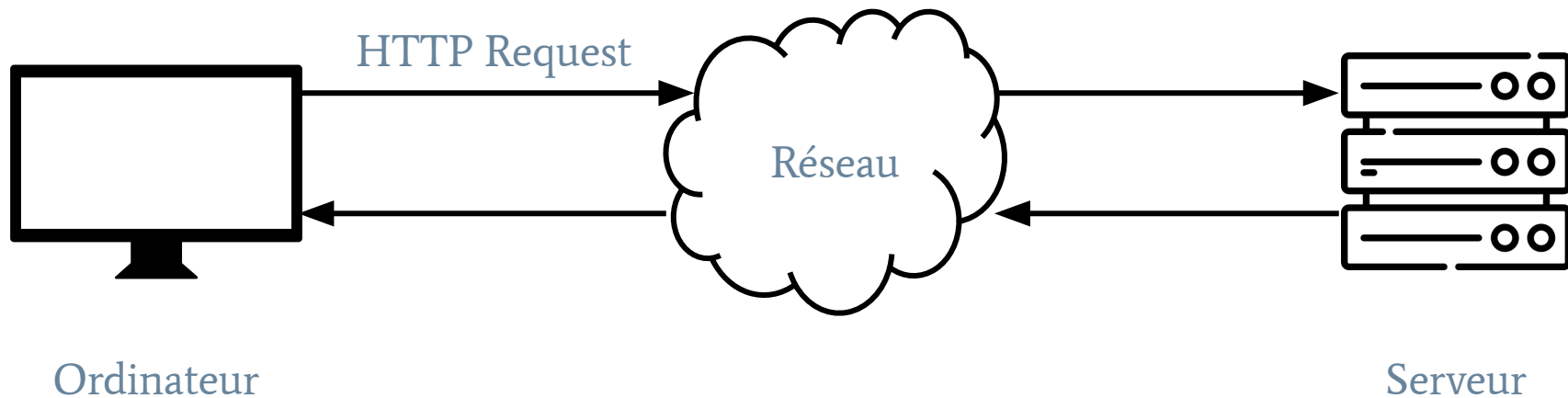
- Ancre : donnée supplémentaire utilisée une fois la réponse obtenue
  - Souvent un nom de section à laquelle se déplacer

# Comment obtenir une page web ?





# Comment obtenir une page web ?



# HTTP, le protocole du World Wide Web

- HTTP = Hypertext Transfer Protocol
- Protocole de communication client-serveur en mode *request-response* et *sans état*
- Pourquoi a-t-on besoin d'un protocole ?
  - Optimisation des requêtes (cache)
  - Ajout de fonctionnalités (authentification, session)
  - Donner des tailles sur la réponse attendue
  - Distribution du trafic

# Messages HTTP

- Trois composantes
  - La requête / le statut de la réponse
  - Un header (optionnel)
  - Un body (optionnel)

# La requête HTTP

- Une ligne indiquant ce que l'on demande de la forme

Méthode URL Protocole

- Le protocole est HTTP avec un numéro de version (HTTP/1.1 par exemple)
- La méthode indique l'action de l'on veut exécuter sur la ressource.
- Principales méthodes :
  - GET : On veut obtenir la ressource
  - POST : On envoie de l'information au serveur (message forum, formulaire)
  - PUT, DELETE, ...

GET example.com HTTP/1.1

# Le statut de la réponse HTTP

- De la forme :

Protocole CodeStatut Message

- Le code du statut est dans cinq catégories :
  - 1XX : informatif
  - 2XX : succès
  - 3XX : redirection
  - 4XX : erreur client
  - 5XX : erreur serveur
- En général, 200 et tout va bien

# Le header

- Ensemble de clefs-valeurs de la forme

Clef: valeur

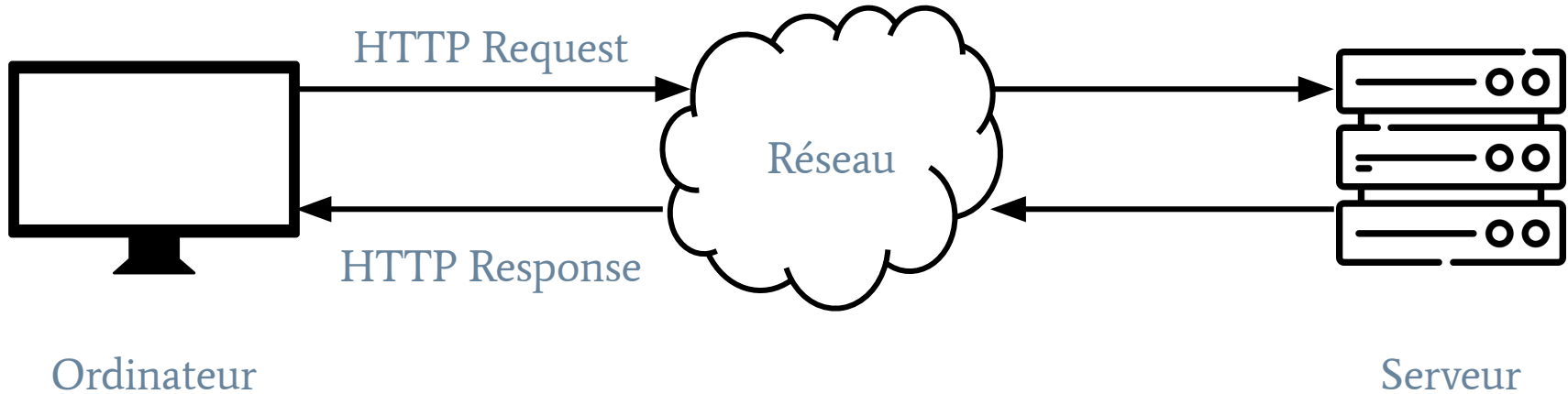
- Clefs populaires :

- Accept : Le type de ressource attendue (Accept: text/html)
  - Content-Type dans la réponse
- Accept-Encoding: L'encodage de la réponse (Accept-Encoding: gzip, deflate)
  - Content-Encoding dans la réponse
- Accept-Language: La langue de la réponse (Accept-Language: en-US)
- Authorization: Authentication (Authorization: Basic QWxhZGRpbjpvGVuIHNlc2FtZQ==)
- Cookie: Un cookie (Cookie: \$Version=1; Skin=new;)
- Host: Le nom de domaine (Host: en.wikipedia.org)
- User-Agent: Qui demande la ressource (User-Agent: Mozilla/5.0 (X11; Linux x86\_64; rv:12.0) Gecko/20100101 Firefox/12.0)

# Le body

- En général, le HTML ou le format demandé
  - Peut être compressé
- Les informations du POST

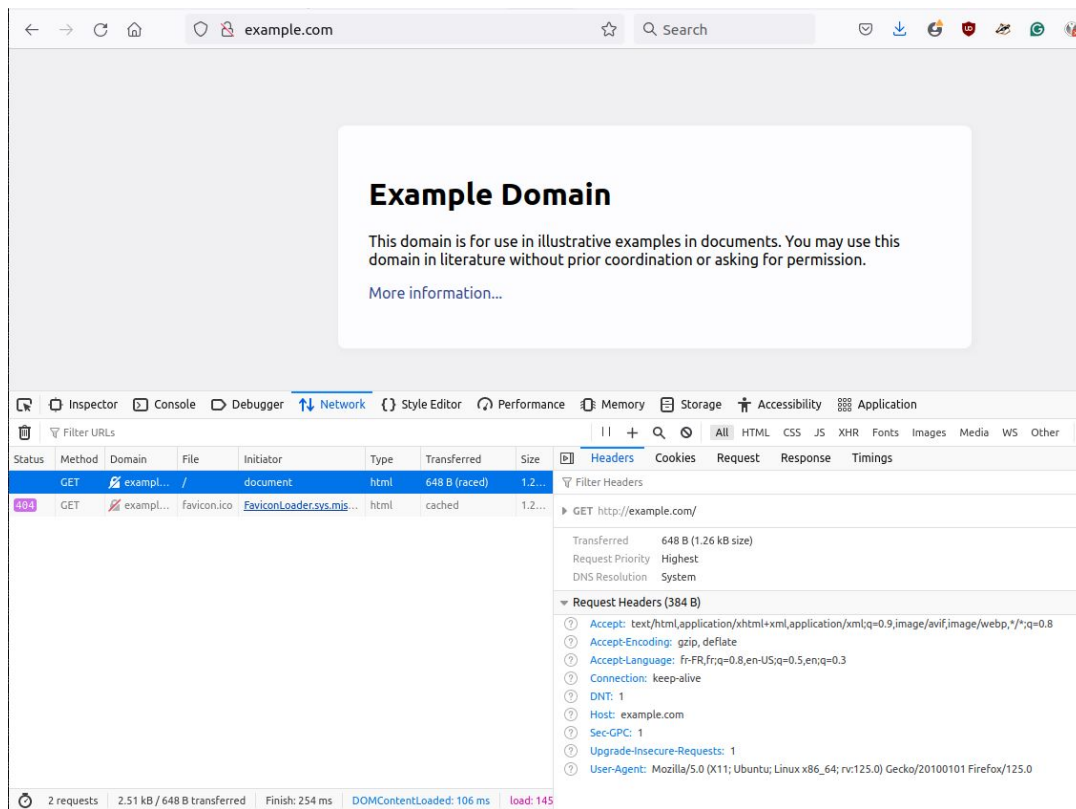
# Comment obtenir une page web ?





# Visualiser les requêtes

Dans la plupart des navigateurs, on peut accéder au trafic réseau et visualiser les requêtes



The screenshot shows a web browser window with the URL 'example.com'. The page content displays 'Example Domain' and a message: 'This domain is for use in illustrative examples in documents. You may use this domain in literature without prior coordination or asking for permission. More information...'. The developer tools network tab is open, showing a list of requests. The first request is highlighted:

Status	Method	Domain	File	Initiator	Type	Transferred	Size
	GET	exampl...	/	document	html	648 B (raced)	1.2...
304	GET	exampl...	favicon.ico	FaviconLoader.sys.mjs...	html	cached	1.2...

The selected request details are shown on the right:

- Transferred: 648 B (1.26 kB size)
- Request Priority: Highest
- DNS Resolution: System
- Request Headers (384 B):
  - Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,\*/\*;q=0.8
  - Accept-Encoding: gzip, deflate
  - Accept-Language: fr-FR,fr;q=0.8,en-US;q=0.5,en;q=0.3
  - Connection: keep-alive
  - DNT: 1
  - Host: example.com
  - Sec-GPC: 1
  - Upgrade-Insecure-Requests: 1
  - User-Agent: Mozilla/5.0 (X11; Ubuntu; Linux x86\_64; rv:125.0) Gecko/20100101 Firefox/125.0

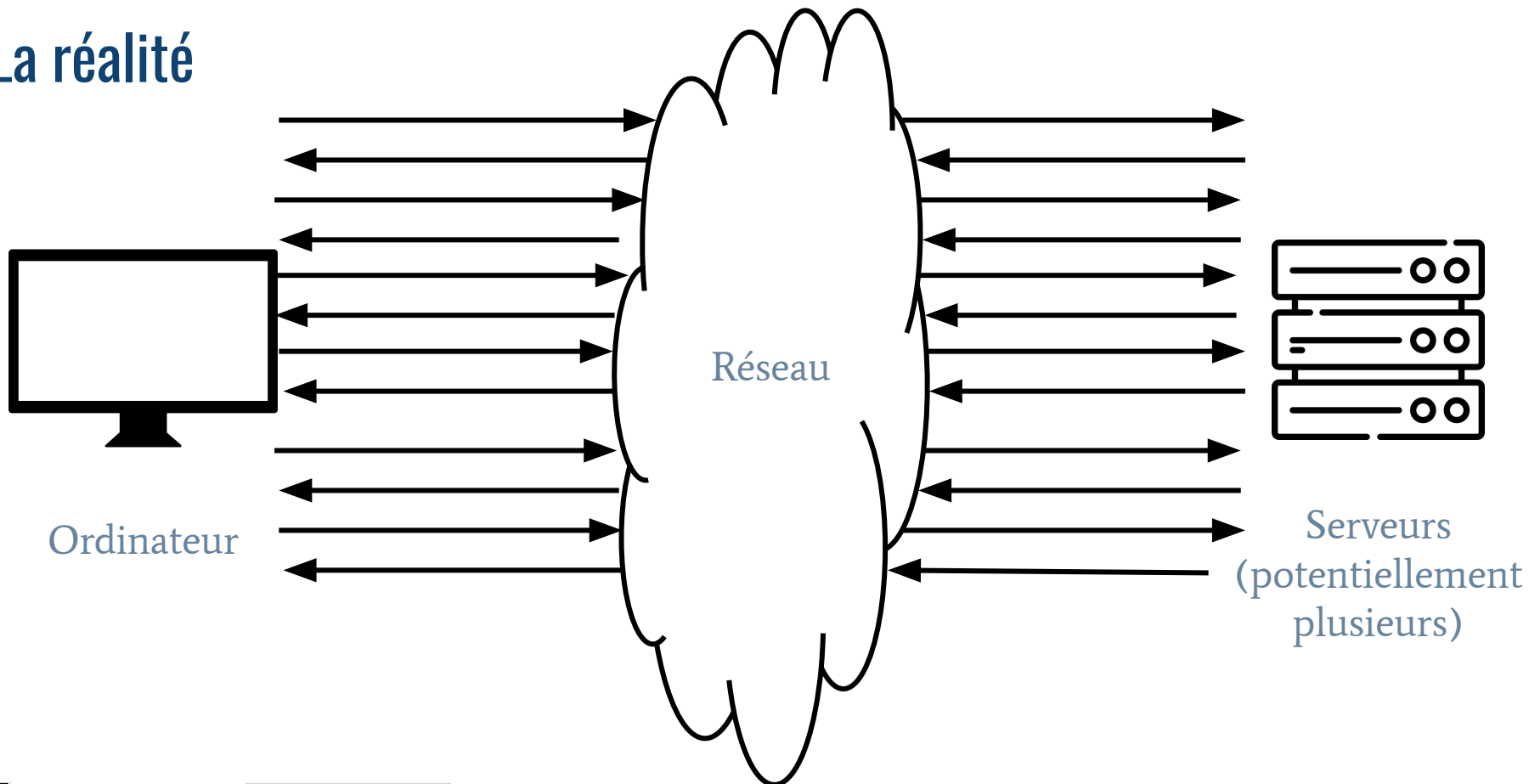
Summary at the bottom: 2 requests | 2.51 kB / 648 B transferred | Finish: 254 ms | DOMContentLoaded: 106 ms | load: 145

# Fonctionnement des sites webs en pratique

# Une page web a besoin de nombreuses ressources

- En général, une page web a besoin de charger de nombreuses ressources
  - Du HTML
  - Des images
  - Des scripts en Javascript
  - Des vidéos
  - Des données
  - Des polices d'écriture
  - Du tracking
- Exemple : Ouvrez un site comme twitter.com et regarder la quantité de ressources nécessaires à travers la console de réseau

# La réalité



# HTML

- HTML est un langage de balisage
  - On va annoter le contenu pour leur donner des propriétés
  - On a une structure d'arbre = des balises peuvent contenir d'autre balises
- Syntaxe d'une balise

```
<tag attribut1="valeur1" attribut2="valeur2">contenu</tag>
```

- Le tag définit le type de balise
- Les attributs donnent des propriétés à la zone balisée

# HTML - Tags classiques

- `<p>` : Paragraphe
- `<div>` : Division/section de code. Utilisé pour organiser et isoler des parties.
- `<head>` : Définit une zone de métadonnées
- `<a>` : Lien hypertexte
- `<input>` : Champ de texte
- `<h1>`, `<h2>`, `<h3>`, ... : Titre de taille 1, 2, 3, ...
- `<img>` : Image
- `<ul>`, `<ol>`, `<li>` : Listes
- `<table>`, `<th>`, `<td>`, `<tr>`, `<thead>`, `<tbody>` : tableaux

# HTML - Attributs classiques

- `href` : lien dans une balise `<a>`
- `src` : lien d'une image dans `<img>`
- `class` : utilisé pour identifier l'application d'un élément de style ou d'un script Javascript
- `id` : pareil que `class`, mais identifiant unique
- `label` : titre pour l'identification par les humains

Les attributs et leur utilisation peuvent grandement varier d'un site à l'autre.

# HTML en pratique pour le scrapping

- Il n'est pas nécessaire de connaître tous les tags et attributs possibles
  - Il faut être conscient de la structure imbriquée d'arbre
  - Il faut pouvoir lire et comprendre du HTML
- Nous utiliserons les tags et les attributs pour trouver de manière systématique des éléments dans la page
- Il est possible d'accéder au HTML d'une page grâce au navigateur
  - Soit en affichant le code source directement
  - Soit en utilisant l'outil inspecter qui permet un mode interactif pour l'exploration



# Exemple

h1.mb-0@xs.mb-s@sm.font-oswald.heading-m@xs.heading-xxl@md | 1152 x 115.2

## PSG – Toulouse : Une célébration fastueuse mais un réel malaise sur le cas Mbappé

MI FIGUE MI RAISIN · La fête pour le 12e titre de champion de France de l'histoire du club a été belle, mais la soirée a tout de même été rendue étrange par l'absence d'hommage pour le dernier match de Mbappé devant son public

[Les + lus](#) Les + lus Football

```
transform 0.4s;"></div> Flex
<div id="dialog-menu" class="c-modal js-modal c-menu" aria-hidden="true" aria-modal="true" tabindex="-1" role="dialog"></div> event
<div id="page-wrap" class="o-page">
  <div id="dfp_ban_atf" class="c-ad-placeholder c-ad-placeholder--ban c-ad-placeholder--ban--atf" data-dfp-name="ban_atf" data-dfp-path="/49926454/20minutes_desktop/article" data-delayed="false" data-distance-visible="0"></div>
  <article id="page-content" class="pb-xxl@xs.bg-grey-50 o-page-wrap u-arch-shadow" data-content-id="4090671" data-content-uri-social="/sport/football/ligue-1/4090671-20240513-psg-toulouse-celebration-fastueuse-reel-malaise-cas-mbappe"> overflow
    <div class="mt-m@xs.mt-xl@md.mb-l@xs.ml-m@xs.ml-0@md.c-slider js-navigation-container c-breadcrumb" data-slider-variant="default" data-slider-auto="false" data-slider-delay="3000"></div>
    <header class="mb-l@xs.flex@xs.items-start@xs.justify-between@xs"> Flex
      <div class="">
        <h1 class="mb-0@xs.mb-s@sm.font-oswald.heading-m@xs.heading-xxl@md"> MI FIGUE MI RAISIN</h1>
        <span class="color-theme-sport font-weight-bold@xs text-transform-uppercase text-m@xs.c-label.c-label--no-background">MI FIGUE MI RAISIN</span>
        <span class="mx-xxs@xs.color-grey-400 text-s@xs"></span>
        <span class="font-source-serif-pro text-xxl@xs"></span>
      </div>
    <div class="mb-l@xs.flex@xs.gap-xl@xs.o-lt-show"> Flex
      <div id="dfp_ban_mtf" class="mb-xl@xs.c-ad-placeholder c-ad-placeholder--ban" data-dfp-name="ban_mtf" data-dfp-path="/49926454/20minutes_desktop/article" data-delayed="true" data-distance-visible="700"></div> event
      <div class="mb-xxl-2@xs.flex@xs.gap-xl@xs.o-lt-show"> Flex
      <div class="mb-xxl-2@xs">
      <div class="mb-xxl-2@xs"></div>
    </div>
  </article>
</div>
```

```
Filter Styles show cls +
element {} inl
  articlePage-desktop-critical-1.20.1.css
@media screen and (min-width: 767px) {
  .heading-xxl@md {} {
    font-size: var(--typography-heading-xxl-font-size);
  }
}
.heading- articlePage-desktop-critical-1.20.1.css
m\@xs {} {
  font-size: var(--typography-heading-m-font-size);
}
.font-oswald {}
  articlePage-desktop-critical-1.20.1.css
{
  font-family: var(--font-family-oswald);
}
articlePage-desktop-critical-1.20.1.css
@media screen and (min-width: 479px) {
  .mb-s@sm {} {
```

# BeautifulSoup

```
from bs4 import BeautifulSoup
```

```
soup = BeautifulSoup(open("scraping_example.html"), 'html.parser')
```

```
# Un seul élément h1, facile
```

```
soup.find_all('h1')[0].getText()
```

```
# PSG - Toulouse : Une célébration fastueuse mais un réel malaise sur le cas Mbappé
```

# Exemple

🏠 Sport Football Ligue 1 Ligue des champions CAN 2024 Euro 2024 Equipe de France

## PSG – Toulouse : Une célébration fastueuse mais un réel malaise sur le cas Mbappé

MI FIGUE MI RAISIN · La fête pour le 12e titre de champion de France de l'histoire du club a été belle, mais la soirée a tout de même été rendue étrange par l'absence d'hommage pour le dernier match de Mbappé devant son public

Les + lus Les + lus Football

```
transform 0.4s;"/>
</div> [flex]
  <div id="dialog-menu" class="c-modal js-modal c-menu" aria-hidden="true" aria-modal="true" tabindex="-1" role="dialog">
</div> [event]
  <div id="page-wrap" class="o-page">
    <div id="dfp_ban_atf" class="c-ad-placeholder c-ad-placeholder--ban c-ad-placeholder--ban--atf" data-dfp-name="ban_atf" data-dfp-path="/49926454/20minutes_desktop/article" data-delayed="false" data-distance-visible="0"></div>
    <article id="page-content" class="pb-xxl&xs bg-grey-50 o-page-wrap u-arch-shadow" data-content-id="4090671" data-content-uri-social="/sport/football/ligue-1/4090671-20240513-psg-toulouse-celebration-fastueuse-reel-malaise-cas-mbappe">
      <div class="mt-mqxs mt-xl&md mb-l&xs ml-mqxs ml-0&md c-slider js-navigation-container c-breadcrumb" data-slider-variant="default" data-slider-auto="false" data-slider-delay="3000"></div>
      <header class="mb-l&xs flex&xs items-start&xs justify-between&xs">
        <div class="">
          <h1 class="mb-0&xs mb-s&sm font-oswald heading-mqxs heading-xxl&md">
            <span class="color-theme-sport font-weight-bold&xs text-transform-uppercase text-mqxs c-label c-label--no-background">Mi figue mi raisin</span>
            <span class="mx-xx&xs color-grey-400 text-s&xs"></span>
          </h1>
          <span class="font-source-serif-pro text-xxl&xs">
            La fête pour le 12e titre de champion de France de l'histoire du club a été belle, mais la soirée a tout de même été rendue étrange par l'absence d'hommage pour le dernier match de Mbappé devant son public
          </span>
        </div>
      </header>
    </div>
  <div class="mb-l&xs flex&xs gap-xl&xs o-lt-show">
    <div id="dfp_ban_mtf" class="mb-xl&xs c-ad-placeholder c-ad-placeholder--ban" data-dfp-name="ban_mtf" data-dfp-path="/49926454/20minutes_desktop/article" data-delayed="true" data-distance-visible="700"></div> [event]
```

```
Filter Styles show cls +
element {} {
  .text-xxl&xs {
    font-size: var(--typography-text-xxl-font-size);
  }
  .font-source-serif-pro {
    font-family: var(--font-family-source-serif-pro);
  }
  ++before, ++after {} {
    box-sizing: inherit;
  }
Inherited from body
body {} {
  --fbc-blue-60: #0060df;
```

# BeautifulSoup

Plus compliqué, de nombreux tags `<span>`

```
len(soup.find_all('span'))
```

```
# 217
```

Si on a de la chance, notre span est toujours au même endroit. Sinon, il faut l'identifier de manière unique, soit grâce à ces attributs, soit dans l'arborescence.

```
soup = BeautifulSoup(open("scraping_example.html"), 'html.parser')
```

```
len(soup.find_all('span', {"class": "font-source-serif-pro text-xxl@xs"}))
```

```
# 1
```

# BeautifulSoup

Plus compliqué, de nombreux tags `<span>`

```
len(soup.find_all('span'))
```

```
# 217
```

Si on a de la chance, notre span est toujours au même endroit. Sinon, il faut l'identifier de manière unique, soit grâce à ces attributs, soit dans l'arborescence.

```
for div in soup.find_all("div"):
    if div.find() is not None and div.find().name == "h1":
        print(div.find_all("span")[-1].getText())
```

```
# La fête pour le 12e titre de champion
```

# Les images, vidéo

- Chaque image a aussi une URL associée
- On peut constituer des jeux de données en utilisant l'image et les attributs pour une description (`alt`)
  - Parfois, le texte autour peut servir
- Souvent difficile ou cher d'exploiter l'information dans une image, encore pire pour une vidéo

# Javascript

- Javascript est un langage de script très utilisé sur le web
- Ce qui nous intéresse :
  - Javascript rend le pages dynamiques
  - Javascript peut télécharger des données et les insérer dans la page
  - Javascript peut effectuer des opérations à la volée
- Les scripts sont en clair sur notre machine et sont téléchargés comme toutes les ressources. Par contre, ils sont souvent “compressé”, voire obfusqués pour réduire la taille de scripts et cacher leur contenu.
  - Exploitation directe très compliquée

# Exemple de page dynamique - Twitter/X

Presque aucun texte dans le code source !

```
<!DOCTYPE html><html dir="ltr" lang="fr"><head>
[Contenu header métadonnées/chargement de scripts]
</head><body style="background-color: #FFFFFF;"><noscript><style>
  [Éléments de style]
</style>
  <div class="errorContainer">
    [message d'erreur si Javascript n'est pas disponible]
  </div></noscript>
[Plein de div vides pour plus tard]
[Une image SVG]
[Appels de scripts]
```



# Que faire dans ce cas ?

- Il va falloir comprendre les informations échangées en regardant les communications réseau

la chaîne **météo**

Rechercher une ville, une station, un pays, ...

Palaiseau 20 | Évry 21

Devenez VIP

Se connecter

ALERTE MÉTÉO **FRANCE** MONTAGNE MONDE VOYAGE ACTUALITÉS MARINE TV PLAGE +

Profitez d'une navigation sans pub à partir de 1,50€ par mois

Météo > France > Île-de-France > Essonne > Évry

MÉTÉO  
**ÉVRY** ☆  
91000 - Essonne - France

Lun 13	Mar 14	Mer 15	Jeu 16	Ven 17	Sam 18	Dim 19	Lun 20	Mar 21	Mer 22	Jeu 23	Ven 24	Sam 25	Dim 26	Lun 27	+
21° 18°	16° 14°	19° 11°	18° 11°	19° 11°	17° 12°	18° 11°	19° 12°	21° 11°	23° 13°	22° 14°	22° 16°	19° 14°	20° 13°	18° 13°	

# La chaine météo - HTML source

```
<span class="tt-day">lun</span>
```

```
<span class="tt-day-num">13</span>
```

```
<h2 class="tt-img">
```

```

```

```
</h2>
```

```
<span class="tt-tempe-max">-</span>
```

```
<span class="tt-tempe-min">-</span>
```

Pas de température !

# Communications

Il semble y avoir une requête par jour qui retourne un JSON



Status	Method	Domain	File	Initiator	Type	Transferred	Size
200	GET	www.lachainemeteo.com	101-121767d=2024-05-25	previsions-meteo-evry-aij...	json	665 B	742 B
200	GET	www.lachainemeteo.com	101-121767d=2024-05-27	previsions-meteo-evry-aij...	json	648 B	694 B
200	GET	www.lachainemeteo.com	101-12176	previsions-meteo-evry-aij...	json	514 B	224 B
Blocked	GET	tagger.opecloud.com	uid	oneplux.9954a157fbb4b...		Blocked By uBlock Ori...	
200	GET	www.lachainemeteo.com	101-121767d=2024-05-13	previsions-meteo-evry-aij...	json	651 B	700 B
200	GET	www.lachainemeteo.com	101-121767d=2024-05-14	previsions-meteo-evry-aij...	json	662 B	740 B
200	GET	www.lachainemeteo.com	101-121767d=2024-05-15	previsions-meteo-evry-aij...	json	645 B	688 B
200	GET	www.lachainemeteo.com	101-121767d=2024-05-16	previsions-meteo-evry-aij...	json	654 B	702 B
200	GET	www.lachainemeteo.com	101-121767d=2024-05-17	previsions-meteo-evry-aij...	json	653 B	712 B
200	GET	www.lachainemeteo.com	101-11782	previsions-meteo-evry-aij...	json	748 B	859 B
200	GET	www.lachainemeteo.com	101-12176	previsions-meteo-evry-aij...	json	754 B	859 B

Headers	Cookies	Request	Response	Timings	Stack Tr
Filter properties					
JSON					
picto_img: "https://static1.mclm.net/lcm2018/int/picto/four/c0070.png"					
picto_label: "Ciel se dégageant"					
tempe_min: 18					
tempe_max: 21					
wind_icon_svg: '<svg class="icon icon-DirectionArrow5" aria-hidden="true" ><us arrow-s" /></svg>'					
wind_class: "scaleWindOrange"					
wind_speed: 15					
icon_weather_svg: '<svg class="icon icon-WeatherCloudy" aria-hidden="true" ><cloudy" /></svg>'					
weather_label: "Ciel se dégageant"					

# Détails de la requête

Il semble n'y avoir qu'un seul champs dans la requête (la date), mais il faut l'identifiant de la ville (101-12176)

New Request		Search	Blocking	St...	M...	Domain	File	Initiator	Type	Tran...	Size
GET	https://www.lachainemeteo.com/ajax/forecast/day/101-12176?d=2024-05-13			200	GET	ww...	101-1217	previs...	json	665 B	742 B
URL Parameters				200	GET	ww...	101-1217	previs...	json	648 B	694 B
<input checked="" type="checkbox"/> d	2024-05-13			200	GET	ww...	101-1217	previs...	json	514 B	224 B
<input checked="" type="checkbox"/> name	value			🚫	GET	tag...	uid	onepl...		Bloc...	
Headers				200	GET	ww...	101-1217	previs...	json	651 B	700 B
<input checked="" type="checkbox"/> Host	www.lachainemeteo.com			200	GET	ww...	101-1217	previs...	json	662 B	740 B
<input checked="" type="checkbox"/> Accept-Encoding	gzip, deflate, br			200	GET	ww...	101-1217	previs...	json	645 B	688 B
<input checked="" type="checkbox"/> Referer	https://www.lachainemeteo.com/meteo-france/ville-12176/previsions-meteo-evry-aujourd'hui			200	GET	ww...	101-1217	previs...	json	654 B	702 B
<input checked="" type="checkbox"/> Connection	keep-alive			200	GET	ww...	101-1217	previs...	json	653 B	712 B
<input checked="" type="checkbox"/> Cookie	cmp_v2_uuid=cmp1715595094467.7354; lcm_firstparty_ppid="none"; savedProfile={"id":0,"date":"M...			200	GET	ww...	101-1178	previs...	json	748 B	859 B

# Comment obtenir l'identifiant de la ville

The screenshot shows a web browser displaying the weather page for Évry. The search bar contains 'Évry' and a dropdown menu lists nearby locations: Évry - 91000, Évreux - 27000, Évires - 74570, and Évian-les-Bains - 74500. Below the search bar, a weather forecast for Évry (91000 - Essonne - France) is shown, including a 7-day forecast with temperatures ranging from 11°C to 23°C. The browser's developer tools are open to the Network tab, showing a list of requests. The first request is a GET request to 'www.lachainemeteo.com' with a status of 200. The second request is a GET request to 'www.lachainemeteo.com' with a status of 200. The third request is a GET request to 'www.lachainemeteo.com' with a status of 200. The fourth request is a GET request to 'tagger.opecloud.com' with a status of 200. The fifth request is a GET request to 'www.lachainemeteo.com' with a status of 200. The sixth request is a GET request to 'www.lachainemeteo.com' with a status of 200. The seventh request is a GET request to 'lefigaro.papi-public.eu-central-1.tagger...' with a status of 200. The eighth request is a POST request to 'lefigaro.dcapitagger.opecloud.com' with a status of 200. The ninth request is a GET request to 'pdmp.papi-public.eu-central-1.tagger...' with a status of 200. The tenth request is a GET request to 'www.lachainemeteo.com' with a status of 200. The eleventh request is a GET request to 'www.lachainemeteo.com' with a status of 200. The twelfth request is a GET request to 'www.lachainemeteo.com' with a status of 200. The thirteenth request is a GET request to 'www.lachainemeteo.com' with a status of 200.

Status	Method	Domain	File	Initiator	Type	Size
200	GET	www.lachainemeteo.com	101-121767d-2024-05-16	previsions-meteo-evry-aujourd'hui:117 (fe...	json	654 B
200	GET	www.lachainemeteo.com	101-121767d-2024-05-17	previsions-meteo-evry-aujourd'hui:117 (fe...	json	653 B
200	GET	www.lachainemeteo.com	alert	previsions-meteo-evry-aujourd'hui:117 (fe...	json	415 B
200	GET	tagger.opecloud.com	uid	one-lux.9954a1577bb4b4a84e10.js:1 (xhr)	Blocked By uBlock Origin	
200	GET	www.lachainemeteo.com	101-11782	previsions-meteo-evry-aujourd'hui:117 (fe...	json	748 B
200	GET	www.lachainemeteo.com	101-12176	previsions-meteo-evry-aujourd'hui:117 (fe...	json	755 B
200	GET	lefigaro.papi-public.eu-central-1.tagger...	targeting?url=https://www.lachainemeteo.com	ope-lefigaro.js:1 (fetch)	Blocked By Privacy Badger	
200	POST	lefigaro.dcapitagger.opecloud.com	visit?fpid=94679133-421d-4cee-9177-0165a6e	ope-lefigaro.js:1 (fetch)	Blocked By Privacy Badger	
200	GET	pdmp.papi-public.eu-central-1.tagger...	targeting?url=https://www.lachainemeteo.com	ope-lefigaro.js:1 (fetch)	Blocked By Privacy Badger	
200	GET	www.lachainemeteo.com	search-entity-redesign?e=12176_101	previsions-meteo-evry-aujourd'hui:117 (fe...	json	630 B
200	GET	www.lachainemeteo.com	search-autocomplete-redesign?q=	previsions-meteo-evry-aujourd'hui:117 (fe...	json	NS_BINDING_ABORTED
200	GET	www.lachainemeteo.com	search-autocomplete-redesign?q=ev	previsions-meteo-evry-aujourd'hui:117 (fe...	json	767 B
200	GET	www.lachainemeteo.com	search-autocomplete-redesign?q=evry	previsions-meteo-evry-aujourd'hui:117 (fe...	json	805 B

# L'API de l'autocomplétion

Les appels à l'API de l'autocomplétion retournent un identifiant que l'on peut utiliser pour avoir accès à l'identifiant

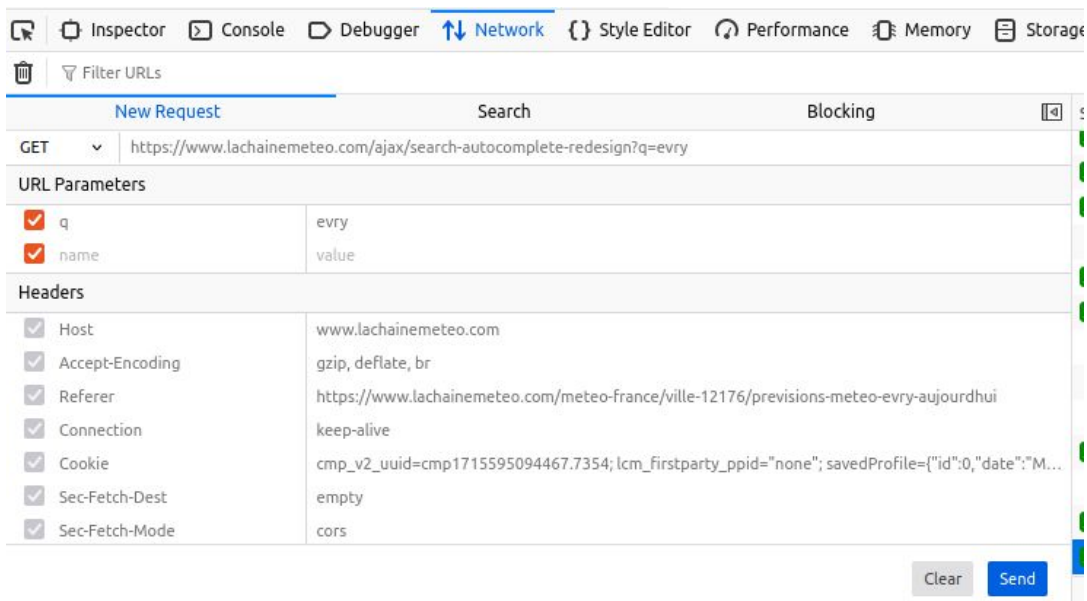
The screenshot displays the Chrome DevTools Network tab. The left pane shows a list of network requests. The right pane shows the response for the selected request, which is a JSON object containing an 'entity' field with a unique ID and a 'search' field with a search term.

Status	Method	Domain	File	Initiator	Type	Transferred	Size
200	GET	www.lachainemeteo.c...	101-121767d=2024-05-16	previsions-meteo-evry...	json	654 B	702 B
200	GET	www.lachainemeteo.c...	101-121767d=2024-05-17	previsions-meteo-evry...	json	653 B	712 B
200	GET	www.lachainemeteo.c...	alert	previsions-meteo-evry...	json	415 B	149 B
Blocked	GET	tagger.opecloud.com	uid	oneplux.9954a157fb4...	Blocked By uBlock Or...		
200	GET	www.lachainemeteo.c...	101-11782	previsions-meteo-evry...	json	748 B	859 B
200	GET	www.lachainemeteo.c...	101-12176	previsions-meteo-evry...	json	755 B	858 B
Blocked	GET	lefigaro.papi-public.e...	targeting?url=https://www.l...	ope-lefigaro.js:1 (fetch)	Blocked By Privacy B...		
Blocked	POST	lefigaro.dcap.i.tagger...	visit?fpid=94679133-421d-4...	ope-lefigaro.js:1 (fetch)	Blocked By Privacy B...		
Blocked	GET	www.lachainemeteo.c...	targeting?url=https://www.l...	ope-lefigaro.js:1 (fetch)	Blocked By Privacy B...		
200	GET	www.lachainemeteo.c...	search-entity-redesign?e=12	previsions-meteo-evry...	json	630 B	672 B
Blocked	GET	www.lachainemeteo.c...	search-autocomplete-redesi	previsions-meteo-evry...	json	NS_BINDING_ABORTED	0 B
200	GET	www.lachainemeteo.c...	search-autocomplete-redesi	previsions-meteo-evry...	json	767 B	2,67 kB
200	GET	www.lachainemeteo.c...	search-autocomplete-redesi	previsions-meteo-evry...	json	805 B	2,76 kB

```
12176_101: {"data-entity": "12176_101", "data-search": "12176_101_evry", "data-searchrs": "Évry - 91000", "data-url": "https://www.lachainemeteo.com/meteo-france/ville-12176/previsions-meteo-evry-aujourd'hui", "data-class": "text", "data-name": "Évry - 91000", "data-aria-hidden": "true", "data-xlinkhref": "/images/int/svg/icons-location-v1.1.0.svg#city"}
776835_101: {"data-entity": "776835_101", "data-search": "776835_101_evry", "data-searchrs": "Évry - 89140", "data-url": "https://www.lachainemeteo.com/meteo-france/ville-776835/previsions-meteo-evry-aujourd'hui", "data-class": "text", "data-name": "Évry - 89140", "data-aria-hidden": "true", "data-xlinkhref": "/images/int/svg/icons-location-v1.1.0.svg#city"}
402843_101: {"data-entity": "402843_101", "data-search": "402843_101_evry", "data-searchrs": "Évry-Grègy-sur-Yerre - 77166", "data-url": "https://www.lachainemeteo.com/meteo-france/ville-402843/previsions-meteo-evry-cregy-sur-yerre-aujourd'hui", "data-class": "text", "data-name": "Évry-Grègy-sur-Yerre - 77166", "data-aria-hidden": "true", "data-xlinkhref": "/images/int/svg/icons-location-v1.1.0.svg#city"}
795849_101: {"data-entity": "795849_101", "data-search": "795849_101_evry", "data-searchrs": "Évrypedo - Grèce (Macédoine - Thrace)", "data-url": "https://www.lachainemeteo.com/meteo-areco/ville-795849/previsions-meteo-evrypedo-aujourd'hui", "data-class": "text", "data-name": "Évrypedo - Grèce (Macédoine - Thrace)", "data-aria-hidden": "true", "data-xlinkhref": "/images/int/svg/icons-location-v1.1.0.svg#city"}
```

# L'API de l'autocomplétion

Un seul champ : la requête



The screenshot shows a browser's developer network tool with the 'Network' tab selected. A request is captured with the following details:

- Method:** GET
- URL:** `https://www.lachainemeteo.com/ajax/search-autocomplete-redesign?q=evry`
- URL Parameters:**

<input checked="" type="checkbox"/> q	evry
<input checked="" type="checkbox"/> name	value
- Headers:**

<input checked="" type="checkbox"/> Host	www.lachainemeteo.com
<input checked="" type="checkbox"/> Accept-Encoding	gzip, deflate, br
<input checked="" type="checkbox"/> Referer	https://www.lachainemeteo.com/meteo-france/ville-12176/previsions-meteo-evry-aujourd'hui
<input checked="" type="checkbox"/> Connection	keep-alive
<input checked="" type="checkbox"/> Cookie	cmp_v2_uuid=cmp1715595094467.7354; lcm_firstparty_ppid="none"; savedProfile={"id":0,"date":"M...
<input checked="" type="checkbox"/> Sec-Fetch-Dest	empty
<input checked="" type="checkbox"/> Sec-Fetch-Mode	cors

Buttons for 'Clear' and 'Send' are visible at the bottom right of the network tool interface.

# Plan d'action

Entrée : Le nom de la ville, une date

1. Utiliser l'autocomplétion pour obtenir l'identifiant de la ville
2. Obtenir la météo en utilisant la requête idoine



# En Python

```
import json
import urllib.request
from urllib.parse import quote, urlencode
```

```
URL_AUTOCOMPLETE = "https://www.lachainemeteo.com/ajax/search-autocomplete-redesign?q="
```

```
URL_METEO = "https://www.lachainemeteo.com/ajax/forecast/day/"
```

```
def get_id(city):
    url = URL_AUTOCOMPLETE + quote(city)
    with urllib.request.urlopen(url) as response:
        id_temp = list(json.loads(response.read().decode("utf-8")).keys())[0].split("_")
        return id_temp[1] + "-" + id_temp[0]
```

# En Python

```
def get_meteo(city, date):  
    id_city = get_id(city)  
    url = URL_METEO + id_city + "?" + urlencode({"d": date})  
    print(url)  
    with urllib.request.urlopen(url) as response:  
        data = json.loads(response.read().decode("utf-8"))  
        return data["picto_label"], data["wind_speed"], data["tempe_min"], data["tempe_max"]  
  
if __name__ == '__main__':  
    print(get_meteo("evry", "2024-05-17"))
```

# Mécanismes de protection contre le scrapping

# Le bon et le moins bon

Compromis entre :

- Expérience utilisateur et facilité d'accès pour les moteurs de recherche
- Prévention du scrapping non voulu

# Ralentir le scrapping

- Limiter le nombre d'accès par IP
  - Problème pour les réseaux publiques, les IPs non fixes, les VPNs
  - Parfois, beaucoup de requêtes même pour un utilisateur normal
- Captcha
  - Dur à mettre en place sur des appels à une API
- Détection d'activité inhabituelle
  - Difficile à mettre en place
  - Besoin de suivre plein de métriques et de les traiter (temps de remplissage formulaire, taille écran, polices installées, nombre de clicks par minute, ...)
- Forcer l'identification
  - Mauvais pour les moteurs de recherches et les utilisateurs anonymes
- Contrôler au maximum les informations accessibles
  - Pas trop d'informations d'un coup
- Donner l'information sous forme d'image
  - Pénible pour les vrais utilisateurs

# Ralentir le scrapping

- Contrôler le User-Agent
  - Facile à contourner
- Faire un mécanisme compliqué de contrôle d'accès
  - Génération de tokens d'authentification
  - Script Javascript difficile à lire
  - Une fois découvert et inversé, mécanisme inutile

# Comment contourner sans effort ?

- Rendu de la page et exécution des scripts dans un navigateur contrôlé
  - En Python : Selenium
- Presque indétectable : on utilise le même outil qu'un humain
- Par contre, très lent !
- Possibilité d'interagir avec le site :
  - Click souris, clavier
- Souvent utilisé pour faire des tests automatiques de sites web

# Exemple

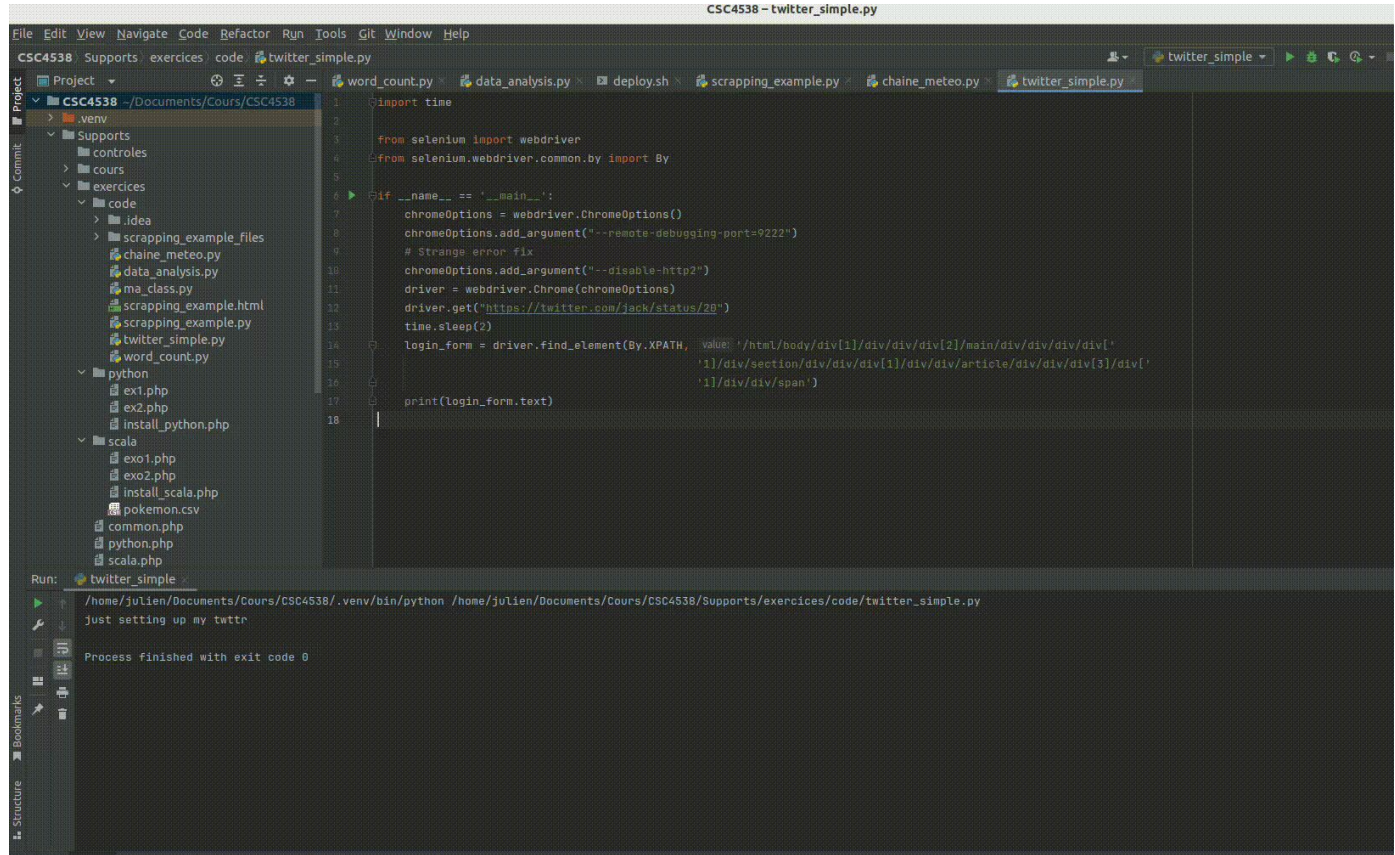
```
import time

from selenium import webdriver
from selenium.webdriver.common.by import By

if __name__ == '__main__':
    driver = webdriver.Chrome()
    driver.get("https://twitter.com/jack/status/20 ")
    time.sleep(2)
    login_form = driver.find_element(By.XPATH,
        '/html/body/div[1]/div/div/div[2]/main/div/div/div/div[
        1]/div/section/div/div/div[1]/div/div/article/div/div/div[3]/div[
        1]/div/div/span ')
    print(login_form.text)
```



# Exemple



```
CSC4538 - twitter_simple.py
File Edit View Navigate Code Refactor Run Tools Git Window Help
CSC4538 Supports exercices code twitter_simple.py
Project - word_count.py data_analysis.py deploy.sh scrapping_example.py chaine_meteo.py twitter_simple.py
CSC4538 ~/Documents/Cours/CSC4538
  .venv
  Supports
  controles
  cours
  exercices
  code
  .idea
  scrapping_example_files
  chaine_meteo.py
  data_analysis.py
  ma_class.py
  scrapping_example.html
  scrapping_example.py
  twitter_simple.py
  word_count.py
  python
  ex1.php
  ex2.php
  install_python.php
  scala
  exo1.php
  exo2.php
  install_scala.php
  pokemon.csv
  common.php
  python.php
  scala.php

1 import time
2
3 from selenium import webdriver
4 from selenium.webdriver.common.by import By
5
6 if __name__ == '__main__':
7     chromeOptions = webdriver.ChromeOptions()
8     chromeOptions.add_argument("--remote-debugging-port=9222")
9     # Strange error fix
10    chromeOptions.add_argument("--disable-http2")
11    driver = webdriver.Chrome(chromeOptions)
12    driver.get("https://twitter.com/jack/status/28")
13    time.sleep(2)
14    login_form = driver.find_element(By.XPATH, value='//html/body/div[1]/div/div/div[2]/main/div/div/div[1]'
15                                          '/div/section/div/div/div[1]/div/div/article/div/div/div[3]/div['
16                                          '1]/div/div/span')
17
18    print(login_form.text)
```

Run: twitter\_simple

```
/home/julien/Documents/Cours/CSC4538/.venv/bin/python /home/julien/Documents/Cours/CSC4538/Supports/exercices/code/twitter_simple.py
just setting up my tattr

Process finished with exit code 0
```

# Considérations utiles

# Comment obtenir toutes les pages webs d'un site ?

- Web crawling
  - On part d'une page racine et on suit tous les liens récursivement
  - Fait par les moteurs de recherche
  - Mais toutes les pages ne sont pas accessibles par un lien (exemple : champ de recherche)

# Robots.txt

- Fichier souvent placé à la racine d'un site web
  - Exemple : <https://twitter.com/robots.txt>
- Décrit les pages autorisées ou non pour les robots crawler de site web
- Donne parfois une idée de pages intéressantes à considérer (ex. Le sitemap)

# Le sitemap

- Souvent un ou plusieurs fichiers XML sitemap.xml à la racine d'un site
  - Exemple : <https://www.lachainemeteo.com/sitemaps/www-fr-fr/sitemap-index.xml>
  - Trouvé dans le robots.txt
  - Ou parfois avec une recherche Google : site:lachainemeteo.com filetype:xml
- Contient une liste de pages de site
  - Ou redirection vers d'autres sitemaps
  - Exemple : [https://www.lachainemeteo.com/sitemaps/www-fr-fr/sitemap-cities\\_001.xml](https://www.lachainemeteo.com/sitemaps/www-fr-fr/sitemap-cities_001.xml)
- À télécharger une seule fois

# Exemple

```
</url>
-<url>
  -<loc>
    https://www.lachainemeteo.com/meteo-france/ville-33/previsions-meteo-paris-aujourd'hui
  </loc>
  <lastmod>2023-12-20</lastmod>
  <changefreq>hourly</changefreq>
</url>
-<url>
  -<loc>
    https://www.lachainemeteo.com/meteo-france/ville-33/previsions-meteo-paris-heure-par-heure
  </loc>
  <lastmod>2023-12-20</lastmod>
  <changefreq>hourly</changefreq>
</url>
-<url>
```

identifiant

ville

On peut éviter un appel à l'autocomplétion !

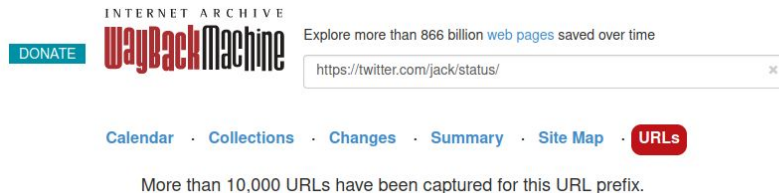
# Les archives du web

- Les archives du web peuvent contenir des pages difficilement accessibles (ou même n'existant plus)
- <https://web.archive.org/>
- Site très lent



# Les archives du web

- Possibilité de rechercher des URLs par préfixe, ou même d'avoir accès à un sitemap



INTERNET ARCHIVE  
**WaybackMachine** Explore more than 866 billion web pages saved over time

DONATE

Calendar · Collections · Changes · Summary · Site Map · **URLs**

More than 10,000 URLs have been captured for this URL prefix.

## URL ↑

[http://twitter.com/jack/status/https://o.twimg.com/1/proxy.jpg?t=FQQVBBIMAWH0dHA6Ly9YWNrLjAubXNoY2RuLmNvbS9iZWRpYS9aZ2t5TURFekx6QXhMekkwTHpVMUwvUnBZMnREYjNOMGlyeHZMbVpoWkRBeUxtcHdad3B3Q1hSb2RXMWDVFUyTUhmM05UQUtaUWwxY0djLzAyMwblaabhu2N9i5Ztov8VgLu-\\_4](http://twitter.com/jack/status/https://o.twimg.com/1/proxy.jpg?t=FQQVBBIMAWH0dHA6Ly9YWNrLjAubXNoY2RuLmNvbS9iZWRpYS9aZ2t5TURFekx6QXhMekkwTHpVMUwvUnBZMnREYjNOMGlyeHZMbVpoWkRBeUxtcHdad3B3Q1hSb2RXMWDVFUyTUhmM05UQUtaUWwxY0djLzAyMwblaabhu2N9i5Ztov8VgLu-_4)

[http://twitter.com/jack/status/https://si0.twimg.com/profile\\_images/2852410605/6e6da28a06cfd7aea20a3cc393ef1182\\_normal.png](http://twitter.com/jack/status/https://si0.twimg.com/profile_images/2852410605/6e6da28a06cfd7aea20a3cc393ef1182_normal.png)

<http://twitter.com/jack/status/10126447293239296>

<http://twitter.com/jack/status/10211421040156672>

<http://twitter.com/jack/status/10211486303522817>

<http://twitter.com/jack/status/10327785578>



# Aspects légaux

- Scrapping légal en général
  - Mais peut être contraire aux conditions générales d'utilisation du site
- Par contre, l'utilisation des données peut être contrôlée ou illégale
  - Le droit d'auteur est toujours présent
  - Les données personnelles très encadrées par la RGPD en Europe
- CNIL :  
<https://www.cnil.fr/fr/la-reutilisation-des-donnees-publicquement-accessibles-en-ligne-des-fins-de-demarchage-commercial>
- En cas de doute, demander à un avocat

# En route vers le TP