Travaux Pratiques Web Scraping :
De la collecte manuelle au bot automatique

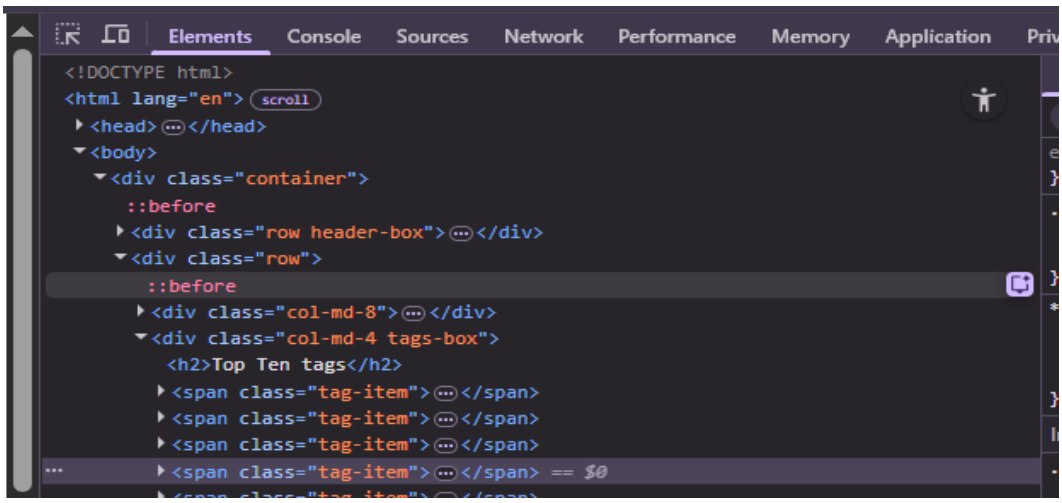
Partie 1 : Collecte Manuelle de Données

Étape 1 : Exploration Visuelle

1. http://quotes.toscrape.com/
2. Observe la page principale :
   - **Combien de citations par page ?**
     - Il y en a 10 par page
   - **Informations disponibles pour chaque citation ?**
     - Texte de la citation
     - Auteur
     - Tags
   - **Y a-t-il une pagination ?**
     - Oui des liens ‘Next’ et ‘Previous’

Étape 2 : Inspection du Code HTML

1. Inspecter l'élément



2. Analyse la structure :

Questions à répondre

a) Quelle est la classe CSS de la div contenant une citation ?
   - *quote*

b) Quel est le sélecteur pour le texte de la citation ?
   -

c) Comment est structuré l'auteur ?
   - <small class="author">

d) Comment sont organisés les tags ?
   - Dans une contenant des <a class="tag">

Étape 3 : Collecte Manuelle

Fichier Excel

	A	B	C	D	E	F
1	**N°**	**Citation (Texte)**	**Auteur**	**Tags**	**URL**	
2	1	“The world as we have create	Albert Einstein	change, deep-thoughts,	http://quotes.toscrape.com/	
3	2	“It is our choices, Harry, that	J.K. Rowling	abilities, choices	http://quotes.toscrape.com/	
4	3	“There are only two ways to l	Albert Einstein	inspirational, life, live, m	http://quotes.toscrape.com/	
5	4	“The person, be it gentleman	Jane Austen	aliteracy, books, classic	http://quotes.toscrape.com/	
6	5	“Imperfection is beauty, madr	Marilyn Monroe	be-yourself, inspiriona	http://quotes.toscrape.com/	
7	6	“Try not to become a man of	Albert Einstein	adulthood, success, val	http://quotes.toscrape.com/	
8	7	“It is better to be hated for wh	André Gide	life, love	http://quotes.toscrape.com/	
9	8	“I have not failed. I've just fou	Thomas A. Edisc	edison, failure, inspiratic	http://quotes.toscrape.com/	
10	9	“A woman is like a tea bag; y	Eleanor Rooseve	misattributed-eleanor-ro	http://quotes.toscrape.com/	
11	10	“A day without sunshine is lik	Steve Martin	humor, obvious, simile	http://quotes.toscrape.com/	

Chronomètre : environ 15 seconds pour chaque citation, presque 5 minutes en totale (4 min et 27 seconds)

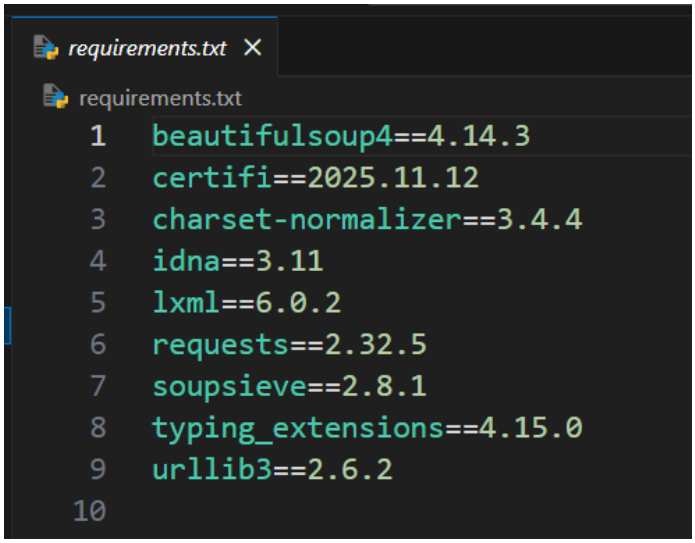
Questions de réflexion :

1. Combien de temps avez-vous mis pour 10 citations ?
   - Presque 5 minutes en totale (4 min et 27 seconds)

2. Combien de temps faudrait-il pour 1000 citations ?
    - 1000 = 500 min ≈ 8 heures.
3. Quels problèmes avez-vous rencontrés ?
    - Erreurs de copier-coller, ennuyeux …
4. Quelles erreurs pourriez-vous faire en copiant manuellement ?
    - Oubli les guillemets, Fautes de frappe …

Partie 2 : Introduction à BeautifulSoup

Étape 1 : Installation des Bibliothèques



Étape 2 : Premier script - Structure de base

```
(env)
HP@Joulz MINGW64 /d/master IAID/Py Data Science/TP_WebScraping/TP_WEBSCRAPING (main)
$ python scraper_basic.py

Extrait du HTML:
<!DOCTYPE html>
<html lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   Quotes to Scrape
  </title>
  <link href="/static/bootstrap.min.css" rel="stylesheet"/>
  <link href="/static/main.css" rel="stylesheet"/>
 </head>
 <body>
  <div class="container">
   <div class="row header-box">
    <div class="col-md-8">
     <h1>
      <a href="/" style="text-decoration: none">
       Quotes to Scrape
      </a>
     </h1>
    </div>
    <div class="col-md-4">
     <p>
      <a href="/login">
```

Étape 3 : Comprendre les Sélecteurs BeautifulSoup

Exécution de la version modifiée de fichier **scraper_basic.py**

```
Les 3 premières citations :

Citation 1 :
  Texte : “The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”
  Auteur : Albert Einstein
  Premier tag : change
Citation 2 :
  Texte : “It is our choices, Harry, that show what we truly are, far more than our abilities.”
  Auteur : J.K. Rowling
  Premier tag : abilities
Citation 3 :
  Texte : “There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”
  Auteur : Albert Einstein
  Premier tag : inspirational
```

Partie 3 : Création du Bot de Scraping

1. Exécution de script

```
$ python scraper_v1.py
Nombre de citations scrapées: 10

Première citation:
{'text': '“The world as we have created it is a process of our t
in', 'tags': 'change, deep-thoughts, thinking, world'}

==================================================
Temps d'exécution: 0.48 secondes
Temps d'exécution manuel: 4 min 27 sec (267 secondes)
Différence: 266.52 secondes de gain
Rapport de performance: 560.5x plus rapide
```

2. Comparer temps
    - Le fichier scraper_v1.py réduit un travail de 4 minutes 27 secondes à moins de 0.5 seconde avec un facteur de performance de 560 fois.

Ajoute la fonction sauvegarder en CSV:

```python
if __name__ == "__main__":
    # Début du chronomètre
    start_time = time.time()

    url = "http://quotes.toscrape.com/"
    quotes = scrape_single_page(url)

    print(f"Nombre de citations scrapées: {len(quotes)}")
    print("\nPremière citation:")
    print(quotes[0])

    # Sauvegarde dans un fichier CSV
    save_to_csv(quotes)
```

Étape 4 : Scraper plusieurs pages (Pagination)

1. Exécution de script

```
================================================
Scraping terminé !
Total de citations: 30
Temps d'exécution: 2.98 secondes
Total de citations: 30
Temps d'exécution: 2.98 secondes
Temps d'exécution: 2.98 secondes
Temps d'exécution manuel (1 page): 4 min 27 sec (267 secondes)
Temps manuel estimé pour 30 citations: 801 secondes
Différence: 798.02 secondes de gain
Rapport de performance: 268.8x plus rapide
================================================

Données sauvegardées dans quotes_all.csv
```

2. Observer le temps d'exécution
    - En scrappant 3 pages avec 30 citations en **3 secondes seulement**.
3. Comparer temps
    - Le code automatise ce qui prendrait manuellement 13 minutes, donc un facteur de 269 fois en performance.

Partie 4 : Améliorations et Bonnes Pratiques

1. Exécution de script **python scraper_v3.py**

```
================================================ …
  BOT DE WEB SCRAPING - QUOTES TO SCRAPE
============================================================
📄 Page 1... ✓ 10 citations
📄 Page 2... ✓ 10 citations
📄 Page 3... ✓ 10 citations
📄 Page 4... ✓ 10 citations
📄 Page 5... ✓ 10 citations

============================================================
  RAPPORT FINAL
============================================================
✓ Citations scrapées: 50
✗ Erreurs rencontrées: 0
⏱️ Temps d'exécution: 5.84 secondes
⚡ Vitesse: 8.56 citations/seconde
============================================================
✓ Données sauvegardées dans quotes_final.csv
```

2. Exécute le script **analyze_data.py**

```
============================================================
  ANALYSE DES DONNÉES SCRAPÉES
============================================================

 Statistiques générales:
  • Nombre total de citations: 50
  • Nombre d'auteurs uniques: 28
  • Nombre de tags uniques: 79

 Top 5 auteurs les plus cités:
   8x Albert Einstein
   6x J.K. Rowling
   6x Marilyn Monroe
   3x Bob Marley
   3x Dr. Seuss

 Top 10 tags les plus utilisés:
   9x inspirational
   9x love
   8x life
   5x humor
   4x books
   4x reading
   3x friends
   3x friendship
   2x paraphrased
   2x simile

 Citation la plus longue:
  “This life is what you make it. No matter what, you're going to mess up
  (1084 caractères)

 Citation la plus courte:
  “We read to know we're not alone.”
  (34 caractères)
```