Session 2 Lesson 2- Crawl and scrape

Sometimes, you may want a little bit of information - a movie rating, stock price, or product availability - but the information is available only in HTML pages, surrounded by ads and extraneous content.

To do this we build an automated web fetcher called a crawler or spider. After teh HTML contents have been retrived from the remote web servers, a scraper parses it to find the needle in the haystack.

BeautifulSoup Module

First, the document is converted to Unicode, and HTML entities are converted to Unicode characters. Beautiful Soup then parses (analyses) the document using the best available parser. It will use an HTML parser unless you specifically tell it to use an XML parser.

Beautiful Soup transforms a complex HTML document into a complex tree of Python objects. But you'll only ever have to deal with about four kinds of objects: Tag, NavigableString, BeautifulSoup, and Comment.

- A ***Tag*** object corresponds to an XML or HTML tag in the original document
- The ***BeautifulSoup*** object itself represents the document as a whole
- The ***NavigableString*** class to contains the bit of text within a tag

https://www.crummy.com/software/BeautifulSoup/bs4/doc/
(https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

In [ ]:

```
<h1 id="HEADING" property="name" class="heading_name   ">
    <div class="heading_height"></div>
     "
     Le Jardin Napolitain
     "
</h1>
```

Step 1 : Making the soup

First we need to use the BeautifulSoup module to parse the HTML data into Python readable Unicode Text format.

Let us write the code to parse a html page. We will use the trip advisor URL for Le Jardin Napolitain - https://www.tripadvisor.fr/Restaurant_Review-g227613-d3531819-Reviews-Le_Jardin_Napolitain-Jouy_en_Josas_Versailles_Yvelines_Ile_de_France.html (https://www.tripadvisor.fr/Restaurant_Review-g227613-d3531819-Reviews-Le_Jardin_Napolitain-Jouy_en_Josas_Versailles_Yvelines_Ile_de_France.html)

In [3]:

```
from bs4 import BeautifulSoup
scrape_url = 'https://www.tripadvisor.fr/Restaurant_Review-g227613-d3531819
-Reviews-Le_Jardin_Napolitain-Jouy_en_Josas_Versailles_Yvelines_Ile_de_Fran
ce.html'
response = requests.get(scrape_url)
print(response.status_code)

if response.status_code == 200:
    soup = BeautifulSoup(response.text, 'html.parser') # Soup
#    print(soup)
```

```
200
```

Step 2: Inspect the element you want to scrape

In this step we will inspect the HTML data of the website to understand the tags and attributes that matches the element.

Let us inspect the HTML data of the URL and understand where (under which tag) the review data is located.

In [ ]:

```
<p class="partial_entry">“En plus du renommé super gentil accueil,
les pizzas sont toujours excellentes... On adore! Petit bémol sur la déco.
A rafraichir peut-être? Mais vraiment pas un frein à la fréquentation de
cet établissement!</p>
```

Step 3: Searching the soup for the data

Beautiful Soup defines a lot of methods for searching the parse tree (soup), the two most popular methods are: find() and find_all().

The simplest filter is a tag. Pass a tag to a search method and Beautiful Soup will perform a match against that exact string.

Let us try and find all the < p > (paragraph) tags in the soup:

In [44]:

```
from bs4 import BeautifulSoup
scrape_url = 'https://www.tripadvisor.fr/Restaurant_Review-g227613-d3531819
-Reviews-Le_Jardin_Napolitain-Jouy_en_Josas_Versailles_Yvelines_Ile_de_Fran
ce.html'
response = requests.get(scrape_url)
print(response.status_code)

if response.status_code == 200:
    soup = BeautifulSoup(response.text, 'html.parser')
    for review in soup.find_all('p'):
        print(review.text) #We are interested only in the text data, since
the reviews are stored as text
```

200
Ce restaurant propose-t-il principalement des pizzas ?
Ce restaurant convient-il pour un repas d'affaires ?
Ce restaurant convient-il pour le déjeuner ?
Ce restaurant convient-il pour le dîner ?
Ce restaurant convient-il pour un brunch ?
Ce restaurant propose-t-il des plats végétariens satisfaisant
s ?
Ce restaurant propose-t-il des plats sans gluten satisfaisant
s ?
Ce restaurant convient-il pour le petit déjeuner ?
Ce restaurant propose-t-il des plats végétaliens satisfaisant
s ?
Ce restaurant propose-t-il des plats à emporter ?
Merci pour votre aide !

Rapport qualité/prix très bon.
Les pizzas sont très bonnes, bien garnies et copieuses.
L'ambiance toujours très sympa.

Un endroit fort sympathique, on y mange de délicieuses pizzas e
t pâtes, l'accueil est chaleureux et le cadre convivial.

En plus du renommé super gentil accueil, les pizzas sont toujou
rs excellentes... On adore! Petit bémol sur la déco. A rafraich
ir peut-être? Mais vraiment pas un frein à la fréquentation de
cet établissement!

Toujours ouvert en toutes occasions
Carte agréable
Bonnes pizza et autres plats italiens
Personnel serviable
À revenir sans modération

Restaurant dans une ambiance sympathique
La terrasse est agréable
Les serveuses adorables
Les pizzas aussi bonnes qu'en Italie !!! Elles sont bien garnie
s et juste parfaitement cuites !!!

Un des rares endroits ouverts le soir à JOUY de plus le personn
el est sympa et les pizzas plutôt bonnes.

Toujours un accueil personnalisé par Mehdi le patron du lieu.
Les pizzas sont impeccables : mention spéciale pour la pizza 'V
iagra' qui ne manque pas de piquant.
Les enfants adorent.
Des plats simples mais de qualité.

Les pizzas sont mangeable. Mais dommage pour les viandes et les

sauces trop de plats à la carte pour avoir du frais les pâtes s
ont limite pasta box

Les pizzas sont très bonnes et toujours bien garnies. Le servic
e est agréable et le repas commence toujours par un kir offert.
.
Très bien à emporté aussi.

Pizzeria à connaître. Accueil du patron toujours chaleureux. Le
s pizzas sont un peu salées à mon gout. Service rapide mais un
peu impersonnel

Vous possédez ou gérez cet établissement ? Prenez le contrôle d
e votre page pour répondre gratuitement aux avis, mettre à jour
votre page et bien plus encore.
Bonjour, le Jardin Napolitain de Jouy en Josas est il climatis
é? merci Cdlt J.B.

Voir toutes les questions
(1)

Expanding this further

To add additional details we can inspect the tags further and add the reviewer rating and reviwer details.

In [68]:

```
from bs4 import BeautifulSoup
scrape_url = 'https://www.tripadvisor.fr/Restaurant_Review-g227613-d3531819
-Reviews-Le_Jardin_Napolitain-Jouy_en_Josas_Versailles_Yvelines_Ile_de_Fran
ce.html'
response = requests.get(scrape_url)
print(response.status_code)

if response.status_code == 200:
    soup = BeautifulSoup(response.text, 'html.parser')

    for rev_data in soup.find_all('span', class_= 'noQuotes'):
        print(rev_data.text)

    for rev_data in soup.find_all('p', class_= 'partial_entry'):
        print(rev_data.text)

    for rev_data in soup.find_all('div', class_='rating reviewItemInline'):
        for img_data in rev_data.find_all('img'):
            print(img_data.get('alt'))
```

200
Pizza très bonnes
très sympa
De bonnes pizzas
Une constance
Quelles sont bonnes les pizzas !!!
Bon accueil, pizzas très correctes
Toujours sympa !
Quel domage
Très bien
A connaotre
Bonjour, le Jardin Napolitain de Jouy en Josas est il climatis
é? merci Cdlt J.B.

Rapport qualité/prix très bon.
Les pizzas sont très bonnes, bien garnies et copieuses.
L'ambiance toujours très sympa.

Un endroit fort sympathique, on y mange de délicieuses pizzas e
t pâtes, l'accueil est chaleureux et le cadre convivial.

En plus du renommé super gentil accueil, les pizzas sont toujou
rs excellentes... On adore! Petit bémol sur la déco. A rafraich
ir peut-être? Mais vraiment pas un frein à la fréquentation de
cet établissement!

Toujours ouvert en toutes occasions
Carte agréable
Bonnes pizza et autres plats italiens
Personnel serviable
À revenir sans modération

Restaurant dans une ambiance sympathique
La terrasse est agréable
Les serveuses adorables
Les pizzas aussi bonnes qu'en Italie !!! Elles sont bien garnie
s et juste parfaitement cuites !!!

Un des rares endroits ouverts le soir à JOUY de plus le personn
el est sympa et les pizzas plutôt bonnes.

Toujours un accueil personnalisé par Mehdi le patron du lieu.
Les pizzas sont impeccables : mention spéciale pour la pizza 'V
iagra' qui ne manque pas de piquant.
Les enfants adorent.
Des plats simples mais de qualité.

Les pizzas sont mangeable. Mais dommage pour les viandes et les
sauces trop de plats à la carte pour avoir du frais les pâtes s
ont limite pasta box

```
Les pizzas sont très bonnes et toujours bien garnies. Le servic
e est agréable et le repas commence toujours par un kir offert.
.
Très bien à emporté aussi.


Pizzeria à connaître. Accueil du patron toujours chaleureux. Le
s pizzas sont un peu salées à mon gout. Service rapide mais un
peu impersonnel

4 sur 5 bulles
5 sur 5 bulles
4 sur 5 bulles
4 sur 5 bulles
4 sur 5 bulles
3 sur 5 bulles
4 sur 5 bulles
3 sur 5 bulles
4 sur 5 bulles
3 sur 5 bulles
```

Iterate through pages

Let us try and iterate the code so that it picks review from the other pages of the search results

In [ ]:

```
```