Projet: Webscraping of books.toscrape.com

De manière général le but de ce projet est de développer nos compétences en matière de collecte de donnée en utilisant la méthode de « websrapping » sur le logiciel python. Plus précisément celui-ci consiste à récolter les différentes informations concernant différentes catégories de livres. Le site sur lequel se trouve les informations à récupérer est : http://books.toscrape.com.

Afin de réaliser ce projet nous sommes passés par 4 étapes qui sont :

Etape_1-Récupération du fichier html du site

Etape_2-Récupération des liens menant aux différentes catégories de livres et sélection des catégories qui nous intéressent

Etape_3-Pour chaque catégorie de livre, récupérer les différents liens menant à chaque livre

Etape_4 : Récupérer à partir des liens de chaque livre les informations suivantes (le nom du livre, la description du livre, le prix, le nombre d'étoile, la disponibilité et nombre de livre disponible) et les stockés dans une base

Description et commentaire de chaque étape

Etape1 : Récupération du fichier html du site

Cette étape est la plus facile. En effet à l'aide de «BeautifulSoup » nous avons pu facilement récupéré les fichier Html du site où était stocké toutes les informations du site. Ces à partir de cet html que nous avons commencé à extraire les différents éléments qui nous conduiront vers notre objectif

Etape2 : Récupération des liens menant aux différentes catégories de livres et sélection des catégories qui nous intéressent

Cette étape consiste à chercher où sont stocké les liens de chaque catégorie de livre à partir du fichier html récupérer dans l'étape 1. Nous avons donc remarquer que les liens se trouvaient dans des balises « ***href*** » qui eux-mêmes se trouvaient dans des balises «***a*** » .Une fois cette remarque effectuée ,nous avons donc écris un code adéquat qui dans un premier temps nous récupère les balises **a** et enfin nous donne les **href** à partir duquel on extrait les différents urls de chaque catégories de livre. A partir de cela il est facile de sélectionner les catégories qui nous intéressent.

```
#selectionner mes les 5 catégories de livres à scrapper
#les catégories selectionner sont: Travel, classic, mystery, historical,sequential
interest=links[3:8]
interest
```

```
['http://books.toscrape.com/catalogue/category/books/travel_2/index.html',
 'http://books.toscrape.com/catalogue/category/books/mystery_3/index.html',
 'http://books.toscrape.com/catalogue/category/books/historical-fiction_4/index.html',
 'http://books.toscrape.com/catalogue/category/books/sequential-art_5/index.html',
 'http://books.toscrape.com/catalogue/category/books/classics_6/index.html']
```

Etape_3-Pour chaque catégorie de livre, récupérer les différents liens menant à chaque livre

Les catégories sélectionnées sont : Travel, mystery, historical, sequential, classic

Ici il s'agit donc de reprendre la procédure de l'étape 1 et de l'étape 2 mais pour les 5 url de chaque catégorie de livres choisies. Afin de mieux se retrouver nous avons créer les listes pour stockés les différentes url des livres de chaque catégorie. Les liens se trouvent également dans les « href » qui sont eux-mêmes dans des balises « a ». La nouveauté ici c'est que les balises se trouvent dans les balises « h3 ». Pour donner un exemple nous allons montrer les différents liens de la catégories **classics:**

```
classic
```

```
['https://books.toscrape.com/catalogue/the-secret-garden_413/index.html',
 'https://books.toscrape.com/catalogue/the-metamorphosis_409/index.html',
 'https://books.toscrape.com/catalogue/the-pilgrims-progress_353/index.html',
 'https://books.toscrape.com/catalogue/the-hound-of-the-baskervilles-sherlock-holmes-5_348/index.html',
 'https://books.toscrape.com/catalogue/little-women-little-women-1_331/index.html',
 'https://books.toscrape.com/catalogue/gone-with-the-wind_324/index.html',
 'https://books.toscrape.com/catalogue/candide_316/index.html',
 'https://books.toscrape.com/catalogue/animal-farm_313/index.html',
 'https://books.toscrape.com/catalogue/wuthering-heights_307/index.html',
 'https://books.toscrape.com/catalogue/the-picture-of-dorian-gray_270/index.html',
 'https://books.toscrape.com/catalogue/the-complete-stories-and-poems-the-works-of-edgar-allan-poe-cameo-edition_238/index.h
tml',
 'https://books.toscrape.com/catalogue/beowulf_126/index.html',
 'https://books.toscrape.com/catalogue/and-then-there-were-none_119/index.html',
 'https://books.toscrape.com/catalogue/the-story-of-hong-gildong_84/index.html',
 'https://books.toscrape.com/catalogue/the-little-prince_72/index.html',
 'https://books.toscrape.com/catalogue/sense-and-sensibility_49/index.html',
 'https://books.toscrape.com/catalogue/of-mice-and-men_37/index.html',
 'https://books.toscrape.com/catalogue/emma_17/index.html']
```

Etape_4 : Récupérer les différentes informations pour chaque livre

On recommence la même procédure en considérant cette fois si chaque lien de chaque catégorie de livre. Ici les informations concernant les livres se trouvent plutôt dans des balises « classes » à l'exception du nom du livre qui se trouve dans une balise h1. Cette étape est la moins facile de toutes car il faut être sûr de trouver la bonne balise. Une fois les différentes informations trouvés et stockés dans des listes, nous avons créé notre data base

	types	books_name	price	star_rating	avaibility	url_image	description
0	mystery	Sharp Objects	Â£47.82	One	In stock (20 available)	https://books.toscrape.com/media/cache/c0/59/c...	WICKED above her hipbone, GIRL across her hear...
1	mystery	In a Dark, Dark Wood	Â£19.63	Five	In stock (18 available)	https://books.toscrape.com/media/cache/95/84/9...	In a dark, dark wood Nora hasn't seen Clare fo...
2	mystery	The Past Never Ends	Â£56.50	One	In stock (16 available)	https://books.toscrape.com/media/cache/9d/f2/9...	A simple task, Attorney Chester Morgan thinks....
3	mystery	A Murder in Time	Â£16.64	Four	In stock (16 available)	https://books.toscrape.com/media/cache/cc/bd/c...	Beautiful and brilliant, Kendra Donovan is a r...
4	mystery	The Murder of Roger Ackroyd (Hercule Poirot #4)	Â£44.10	One	In stock (15 available)	https://books.toscrape.com/media/cache/86/38/8...	In the village of King's Abbot, a widow's sudd...
...	...	...	...	...	...	...	...
157	classic	The Story of Hong Gildong	Â£43.19	Two	In stock (1 available)	https://books.toscrape.com/media/cache/b5/5b/b...	The Story of Hong Gildong is arguably the sing...
158	classic	The Little Prince	Â£45.42	Four	In stock (1 available)	https://books.toscrape.com/media/cache/c0/d6/c...	Moral allegory and spiritual autobiography, Th...
159	classic	Sense and Sensibility	Â£37.46	Two	In stock (1 available)	https://books.toscrape.com/media/cache/a8/57/a...	'The more I know of the world, the more am I c...
160	classic	Of Mice and Men	Â£47.11	One	In stock (1 available)	https://books.toscrape.com/media/cache/a0/bb/a...	The compelling story of two outsiders striving...
161	classic	Emma	Â£32.93	Two	In stock (1 available)	https://books.toscrape.com/media/cache/ae/98/a...	'I never have been in love; it is not my way, ...