



Web scraping Via Scrappy



Scrappy

Réalisé par :
HADOU Mohamed

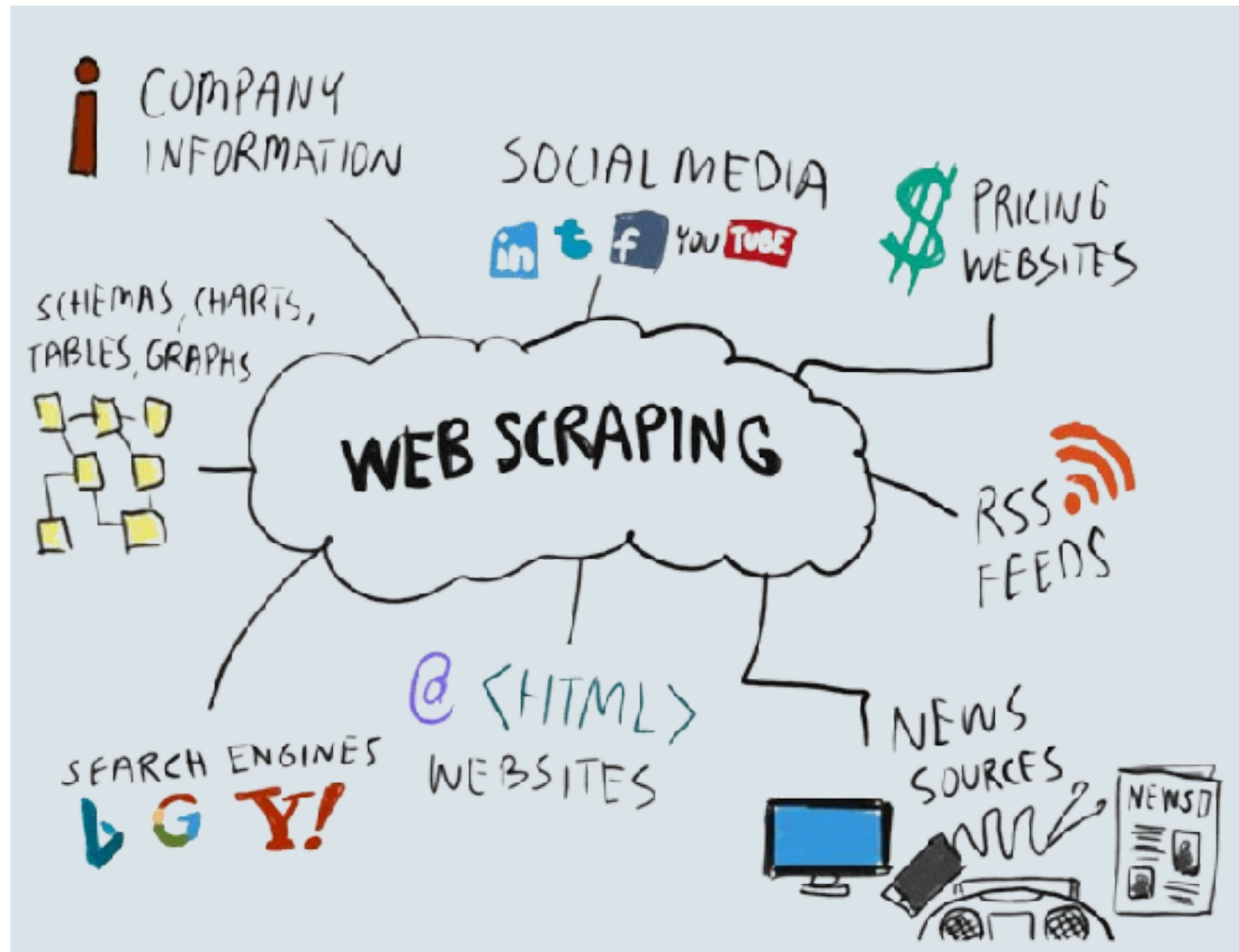
Sommaire



- 1 INTRODUCTION**
- 2 DEFINITION DE SCRAPY**
- 3 ARCHITECTURE DE SCRAPY**
- 4 APPLICATION PRATIQUE**
- 5 CONCLUSION**



Introduction



Le web scraping est une technique d'extraction de données depuis des sites web. Elle consiste à automatiser le processus de collecte d'informations disponibles sur des pages web, permettant de convertir les données souvent dispersées et formatées pour la lecture humaine en formats structurés et utilisables par des machines.

Cette méthode est largement utilisée dans plusieurs domaines pour diverses applications :

- **Analyse de données** : Le web scraping permet de recueillir de grandes quantités de données pour l'analyse statistique, le suivi des tendances du marché, ou encore l'évaluation des comportements des consommateurs.
- **Surveillance des prix** : Les entreprises utilisent le web scraping pour suivre les changements de prix de produits sur différents sites de commerce en temps réel, leur permettant d'ajuster leurs stratégies de prix compétitivement.
- **Collecte de données de recherche** : Dans le secteur académique, le scraping est utilisé pour agréger des données à partir de publications, de forums et d'autres sources pour des études et des recherches.
- **Génération de leads** : Les marketeurs utilisent le web scraping pour collecter des informations de contact ou des données démographiques à des fins de marketing ciblé.

scrapy

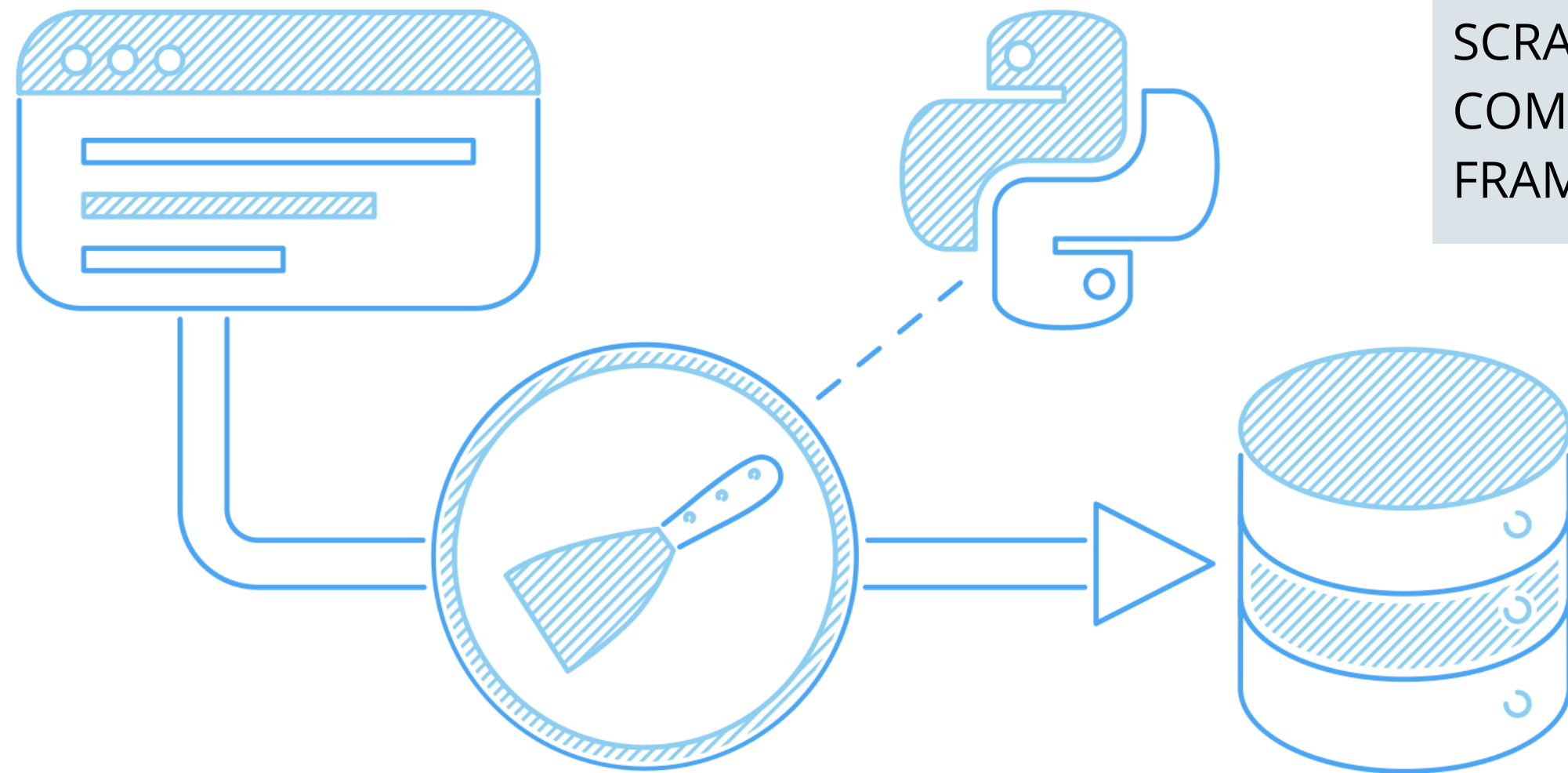
SCRAPY EST UN FRAMEWORK ROBUSTE SPÉCIALEMENT CONÇU POUR LE WEB SCRAPING ET L'EXTRACTION DE DONNÉES, QUI OFFRE PLUSIEURS AVANTAGES SIGNIFICATIFS :



Avantages :

- **Rapidité** : Scrapy est construit sur Twisted, une bibliothèque réseau asynchrone, qui lui permet de gérer un grand nombre de requêtes simultanément. Cela le rend particulièrement rapide pour télécharger et traiter des données provenant de plusieurs sources en même temps.
- **Extensibilité** : Le framework est hautement configurable et peut être étendu par des plugins et des middlewares personnalisés. Cela permet aux développeurs de rajouter des fonctionnalités spécifiques nécessaires pour leurs scrapers.
- **Gestion efficace des requêtes asynchrones** : L'architecture asynchrone de Scrapy permet une gestion plus efficace des temps d'attente et de la bande passante, réduisant ainsi les temps de réponse et optimisant les ressources du système.

Architecture de Scrapy



POUR BIEN COMPRENDRE L'ARCHITECTURE DE SCRAPY, IL EST ESSENTIEL DE SAISIR CERTAINS COMPOSANTS CLÉS QUI FORMENT LE CŒUR DE CE FRAMEWORK DE CRAWLING ET DE SCRAPING.

Architecture de Scrapy

Spiders

Scripts définissant la logique de navigation et d'extraction des données depuis les sites web ciblés.

Engine

Composant central orchestrant les requêtes, les réponses et le passage des items entre les différentes parties du framework.

Scheduler

Gestionnaire des requêtes qui les organise en fonction de la priorité, décidant de l'ordre dans lequel elles seront traitées par le downloader.

Downloader

Module chargé d'envoyer les requêtes aux sites web et de collecter les réponses à traiter.

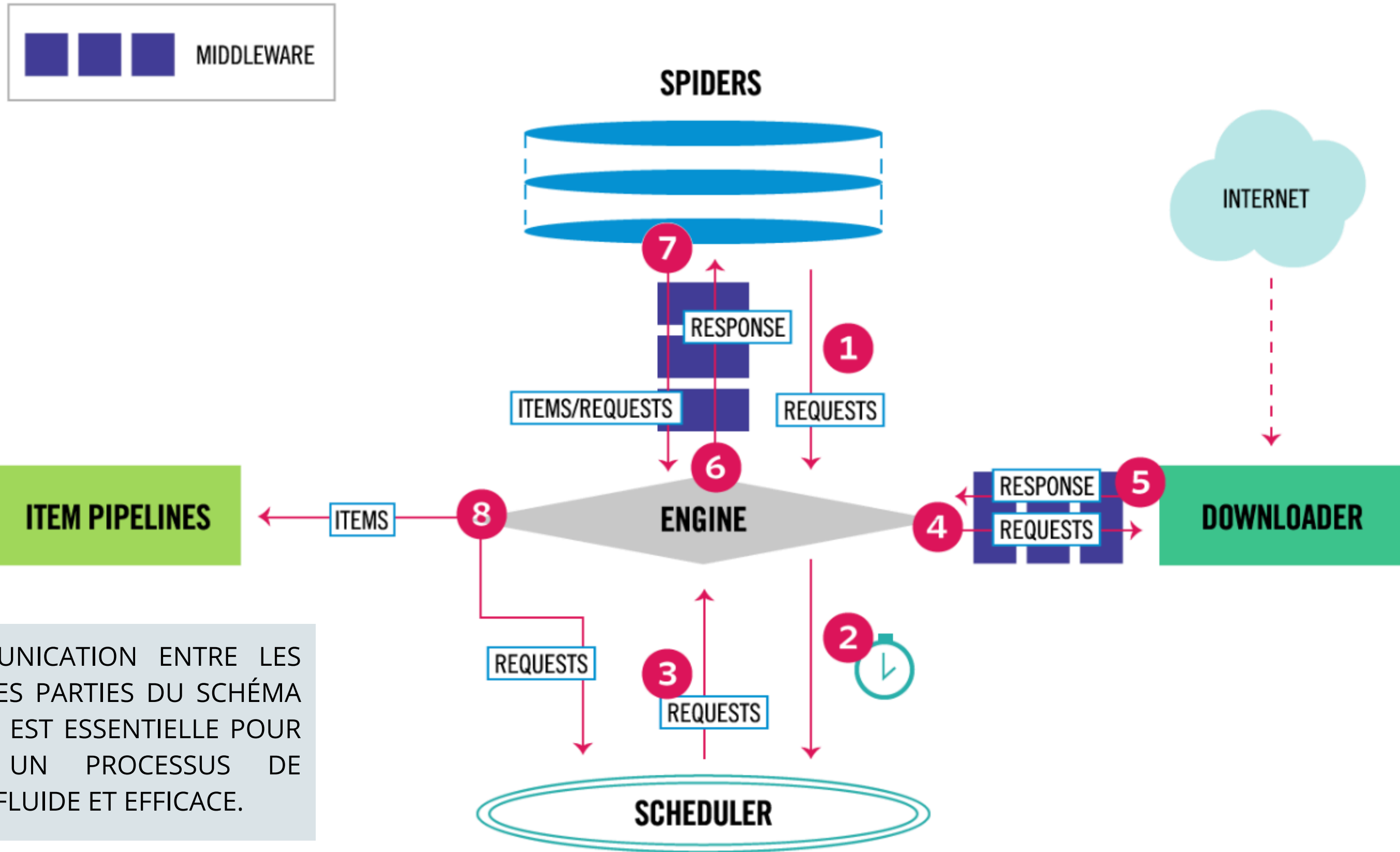
Pipelines

Série de processus permettant de nettoyer, valider, transformer et stocker les items (données extraites).

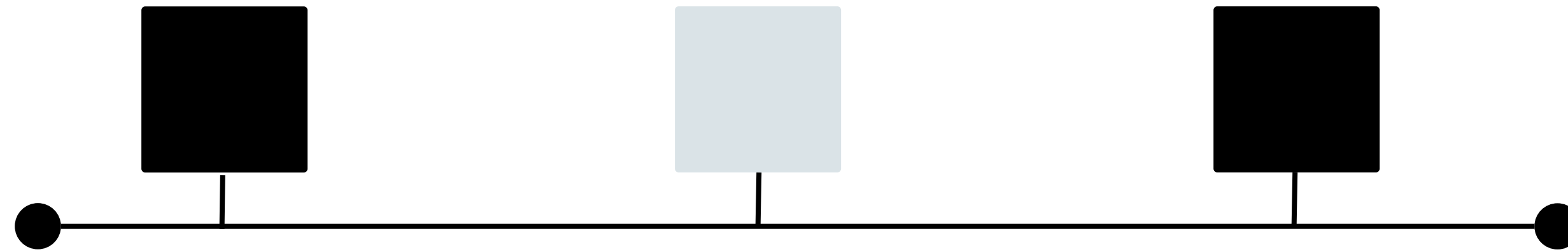
Middlewares

Composants intermédiaires qui modifient ou filtrent les requêtes et réponses, permettant de personnaliser le processus de web scraping

La Dynamique des Composants dans l'Architecture de Scrapy



La Dynamique des Composants dans l'Architecture de Scrapy



Tout d'abord, le moteur Scrapy est au cœur de cette communication. Il initie le processus en lançant les spiders appropriées pour chaque tâche de scraping. Le moteur communique avec le Scheduler pour obtenir les URLs à visiter, en tenant compte des priorités définies et des politiques de filtrage. Une fois que le Scheduler fournit une URL à visiter, le moteur passe cette information à la spider correspondante.

Spider, en tant qu'agent de récupération de données, communique étroitement avec le moteur pour recevoir les URLs à visiter et envoyer les résultats de son travail. Elle utilise également les middlewares pour manipuler les requêtes et les réponses HTTP en fonction des besoins spécifiques du scraping.

Lorsque spider extrait des données des pages web, elle les renvoie sous forme d'items au moteur. Le moteur transmet ensuite ces items aux pipelines pour un traitement ultérieur. Les pipelines sont des éléments clés de la communication, car ils permettent de nettoyer, valider et stocker les données extraites dans divers formats ou bases de données.

Application Pratique

Dans cet exemple de création de projet, nous avons extrait les informations de cette page Amazon

amazon.fr/gp/bestsellers/lawn-garden/

Les meilleures ventes








Nos produits les plus populaires en vertu du nombre de ventes. Mis à jour fréquemment.

N'importe quelle catégorie

Jardin

- Barbecue et repas en extérieur
- Bassins d'agrément
- Chauffage et refroidissement extérieur
- Décoration d'extérieur
- Déneigement
- Jardinage
- Luminaires extérieur
- Matériels d'arrosage et outils pour jardins
- Mobilier de jardin
- Oiseaux et animaux sauvages
- Piscines, baignoires et accessoires
- Plantes, graines et bulbes
- Rangement et stockage extérieurs
- Thermomètres et instruments météorologiques
- Tondeuses et outillage de jardin motorisé

Les meilleures ventes en Jardin

#1  KB KFOUB4 - Anti-Fourmis Boîtes Appât x4 - Détruit durablement et en profondeur les fourmilières - Actif à faible dose - Anti-four... ★★★★☆ 3 034 3 offres à partir de 14,90 €	#2  ALGOFLASH Terreau Semis, Bouturage et Repiquage, UAB, Prêt à l'Emploi, Fabriqué en France, 6 L, TSEB6 ★★★★☆ 3 097 1 offre à partir de 16,90 €	#3  Neudorff - Anti-Limaces Ferramol Protection Contre Les limaces et Les escargots dans Jardin et Potager. Approprié pour l'agriculture... ★★★★☆ 186 3 offres à partir de 13,80 €	#4  PIC - Pièges à Mites Alimentaires PIC - Paquet Triple = 6 pièges Anti Mites - Piège à phéromones pour la Cuisine et Les magasins... ★★★★☆ 12 698 1 offre à partir de 19,99 €
#5  EDIESI, Sardines de Fixation, Sardines Camping, 50 Pcs + 8 Rondelles, en Acier (150x40x3mm), pour Le Jardin, Le Paillis et L... ★★★★☆ 900 10,99 €	#6  Pulvérisateur à pression TUKAN 5 litres, Pulvérisateur de jardin pour protection des plantes, Réservoir de 5 L, Flacon pulvérisateu... ★★★★☆ 22 547 15,51 €	#7  CAUSSADE Anti-nuisibles Rats & Souris Efficacité Radicale 6 Blocs pour Garage et Cave Lieux Humides CARSBL180 ★★★★☆ 2 004 5 offres à partir de 6,79 €	#8  vounot Toile de Paillage avec 30pcs de Piquets Contre Les Mauvais Herbes 10Mx2M en Fibres de Polypropylène Tissées Anti-UV 100g/m2... ★★★★☆ 3 938 24,99 €



Application Pratique



L'installation de Scrapy

Utilisez la ligne de commande suivante pour l'installer : **pip install scrapy**

Ou avec conda : **conda install -c conda-forge scrapy**

La creation d'un projet de Scrapy

Pour générer le projet on utilise la ligne de commande suivante :

scrapy startproject amazon

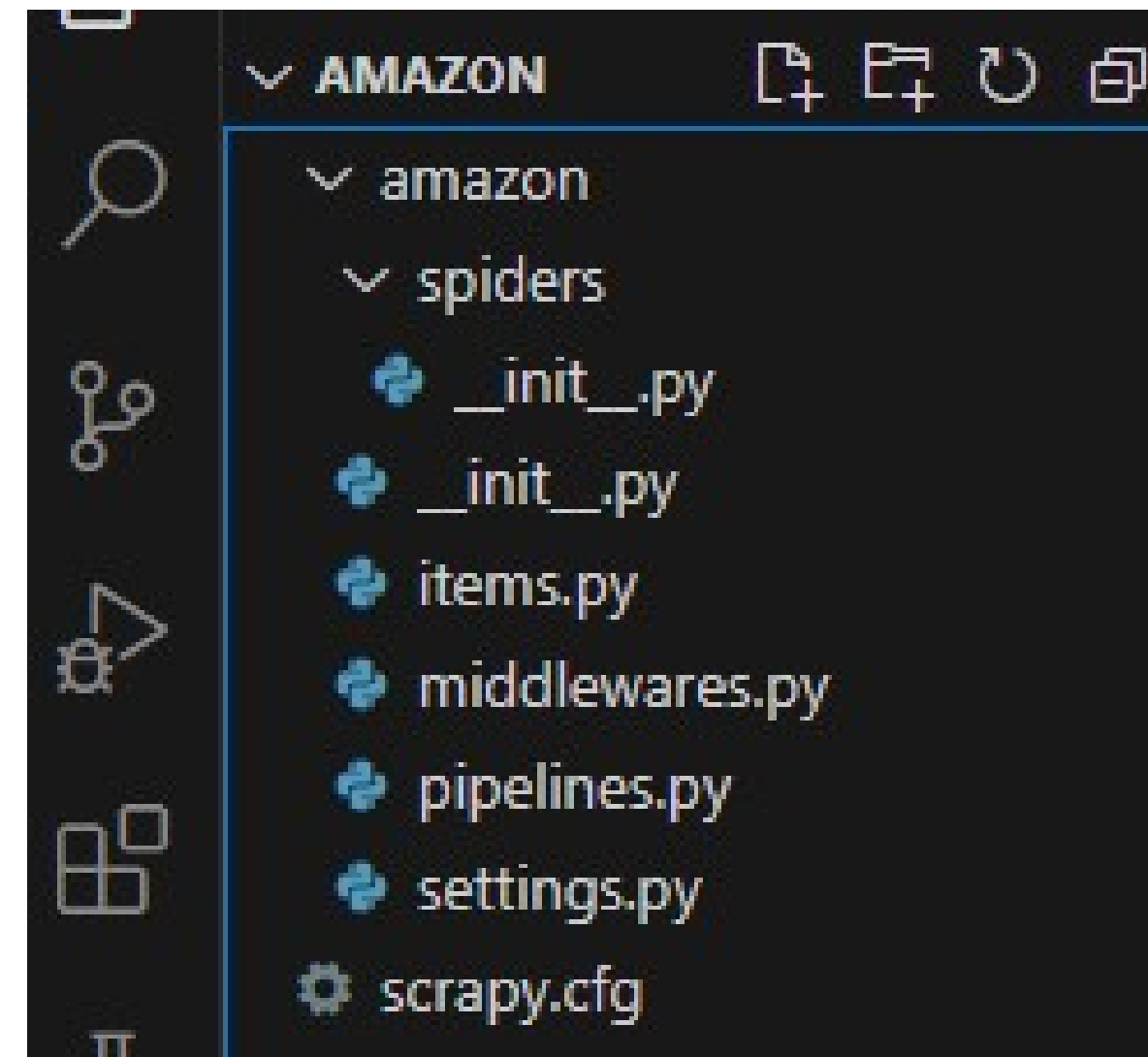
Application Pratique

Après la creation on trouve plusieurs fichiers dans notre projet

La racine du projet contient le fichier **scrapy.cfg** qui est un fichier de configuration qui contient des variables telles que le nom du module qui contient les paramètres du projet et d'autres variables de déploiement.

La racine contient un autre dossier datasets qui contient le projet en lui même. Ce dossier qui est un package Python (d'où le **init.py**) contient le package spiders (qui pour l'heure est vide) ainsi que les modules : **items**, **middlewares** et **pipelines**.

Le fichier **settings.py** contient des variables qui sont utilisées par l'engine et le reste du projet.



Application Pratique

Le Fichier items.py

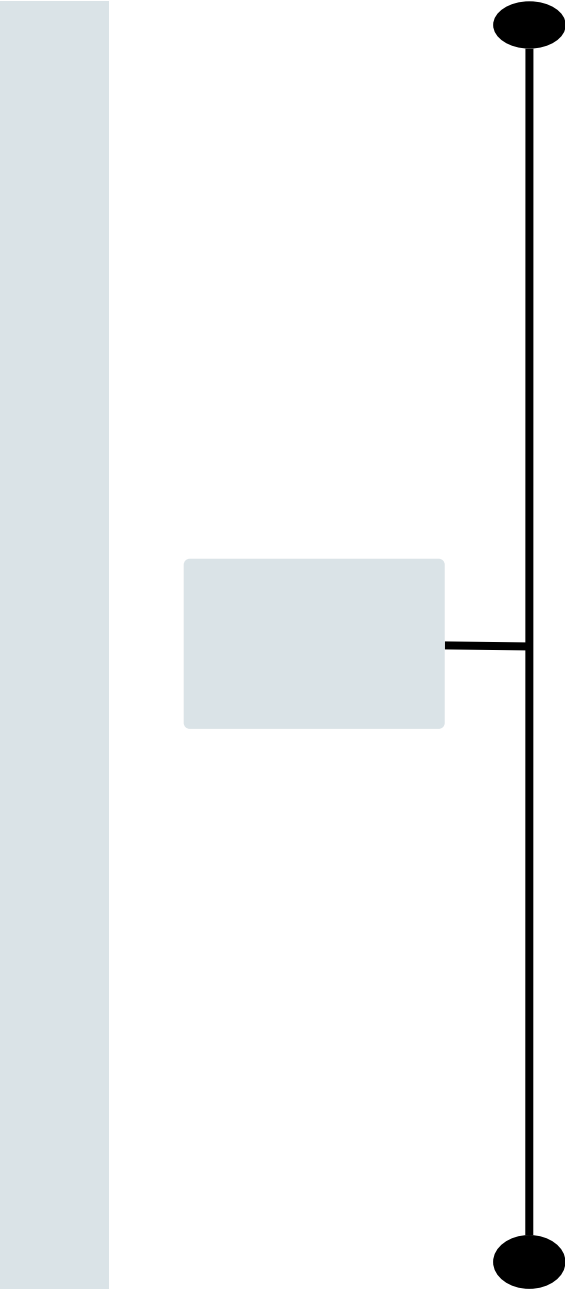
Définissez les éléments (items) qui représentent les données à extraire. Par exemple :

```
items.py  article.py

amazon > items.py > AmazonItem
1  # Define here the models for your scraped items
2  #
3  # See documentation in:
4  # https://docs.scrapy.org/en/latest/topics/items.html
5
6  import scrapy
7
8
9  class AmazonItem(scrapy.Item):
10
11     titre = scrapy.Field()
12     image = scrapy.Field()
13     prix = scrapy.Field()
14
15     # Le mot-clé 'pass' indique la fin de la définition de la classe
16     pass
```

Application Pratique

Créez une spider pour scraper un site web spécifique. Par exemple :



```
items.py  article.py  X
amazon > spiders > article.py > SpiderArticle > parse
1  #Importation des modules nécessaires
2  from scrapy import Request, Spider
3  from ..items import AmazonItem
4
5  #la classe SpiderArticle : Cette classe hérite de la classe de base Spider de Scrapy.
6  class SpiderArticle(Spider):
7
8      name = "article"
9      url = "https://www.amazon.fr/gp/bestsellers/lawn-garden/"
10
11      """
12      Méthode start_requests() : Cette méthode génère les requêtes initiales à envoyer au serveur.
13      Dans ce cas, elle envoie une requête à l'URL définie et spécifie la méthode 'parse' comme callback.
14      """
15      def start_requests(self):
16          yield Request(url=self.url, callback=self.parse)
17
18      #Méthode parse() : Cette méthode est appelée pour traiter la réponse reçue du serveur.
19      def parse(self, response):
```

Application Pratique

Créez une spider pour scraper un site web spécifique. Par exemple : (suite)

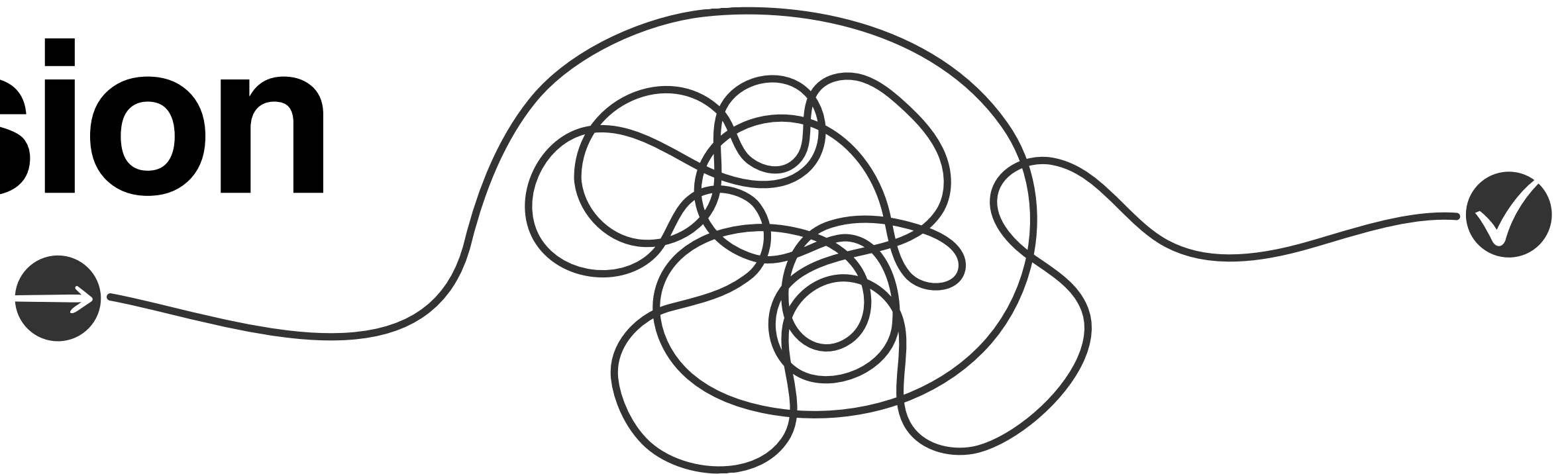
```
18 #Méthode parse() : Cette méthode est appelée pour traiter la réponse reçue du serveur.
19 def parse(self, response):
20     #On utilise des sélecteurs CSS pour extraire les informations pertinentes de la réponse HTML.
21     listArticle = response.css("div.p13n-sc-uncoverable-faceout")
22     #On peut utiliser des sélecteurs XPath avec ce format. Le résultat est le même.
23     #listArticle = response.xpath("//div[contains(@class, 'p13n-sc-uncoverable-faceout')]")
24
25     for article in listArticle:
26         titre = article.css("div._cDEzb_p13n-sc-css-line-clamp-3_g3dy1::text").extract_first()
27         image = article.css("img.a-dynamic-image::attr(src)").extract_first()
28         prix = article.css("span.p13n-sc-price::text").extract_first()
29         #extract_first() : extrait le premier élément
30
31     #Pour chaque article extrait, un objet AmazonItem est créé
32     #les données sont assignées à ses champs (titre, image, prix) et l'objet est renvoyé.
33     item = AmazonItem()
34
35     item['titre'] = titre
36     item['image'] = image
37     item['prix'] = prix
38
39     #"yield" est envoyé l'item pour être traité plus loin par Scrapy,
40     #généralement via le pipeline de traitement, où il peut être stocké dans une base de données
41     yield item
```


Application Pratique

Pour lancer notre spider et avoir les données scrapées dans un fichier CSV, on fait la commande suivante : **scrapy crawl article -o article.csv**

	A ^B _C titre	A ^B _C image	A ^B _C prix
1	KB KFOUB4 - Anti-Fourmis Boîtes Appât x4 - Détruit durablement et en...	https://images-eu.ssl-images-amazon.com/images/I/91tK2duUvxL._A...	13,90 €
2	ALGOFLASH Terreau Semis, Bouturage et Repiquage, UAB, Prêt à l'Em...	https://images-eu.ssl-images-amazon.com/images/I/61FMUoowK8L._...	5,39 €
3	Neudorff - Anti-Limaces Ferramol Protection Contre Les limaces et L...	https://images-eu.ssl-images-amazon.com/images/I/91ihJekgONL._AC...	13,80 €
4	EDIESI, Sardines de Fixation, Sardines Camping, 50 Pcs + 8 Rondelles, e...	https://images-eu.ssl-images-amazon.com/images/I/71QM9SHjrIL._A...	10,99 €
5	Pulvérisateur à pression TUKAN 5 litres, Pulvérisateur de jardin pour p...	https://images-eu.ssl-images-amazon.com/images/I/61tqDnIPo3L._AC...	15,51 €
6	PIC – Pièges à Mites Alimentaires PIC - Paquet Triple = 6 pièges Anti Mi...	https://images-eu.ssl-images-amazon.com/images/I/81BGjVaRQSL._A...	10,99 €
7	vounot Toile de Paillage avec 30pcs de Piquets Contre Les Mauvais Her...	https://images-eu.ssl-images-amazon.com/images/I/7151T2EeHpL._A...	24,99 €
8	Super Ninja Anti Mouches Interieur – 10 Pièges – Attrape Mouche ...	https://images-eu.ssl-images-amazon.com/images/I/71pZ67qP9rL._A...	5,99 €
9	La cordeline Ficelle Jute Naturel Ø1.5mm ±100m pour Le Jardinage, Bri...	https://images-eu.ssl-images-amazon.com/images/I/81FahJH9JYL._AC...	3,49 €
10	Weber Barbecue à Charbon Compact Kettle 47cm - Barbecue à Couver...	https://images-eu.ssl-images-amazon.com/images/I/71F-MmjQqmL._...	69,53 €
11	Kärcher Nettoyeur Haute Pression K 2 Universal Edition, Pression : ma...	https://images-eu.ssl-images-amazon.com/images/I/61Gu5Ac6V8L._A...	62,67 €
12	BARRIERE A INSECTES Spécial Fourmis et autres rampants, Prêt à l'em...	https://images-eu.ssl-images-amazon.com/images/I/615f8wAjVOL._A...	8,69 €
13	FERTILIGENE FTUBFOU - Anti-Fourmis Tube Appât Gel 20 g - Détruit du...	https://images-eu.ssl-images-amazon.com/images/I/81mau5njd1L._A...	7,50 €
14	KETER Abri Horizontal Range Poubelle SIO MIDI - 880 litres	https://images-eu.ssl-images-amazon.com/images/I/71Fkp3UK4OL._A...	116,62 €
15	CAUSSADE Anti-nuisibles Rats & Souris Efficacité Radicale 6 Blocs pour ...	https://images-eu.ssl-images-amazon.com/images/I/91KALLNnt7L._A...	6,79 €

conclusion



En conclusion, le scraping, en tant que technique de collecte de données à partir de sources en ligne, est devenu un outil indispensable dans de nombreux domaines, de la recherche académique à l'analyse commerciale en passant par le développement de produits. Cette pratique offre la possibilité d'extraire des informations précieuses et pertinentes à partir de sites web variés, ce qui peut permettre une prise de décision plus éclairée, une analyse approfondie des tendances du marché et une automatisation de nombreux processus.



Merci.

