Réalisation d'un premier robot web

(sources : scrapy.org , IUT Aix en Provence)

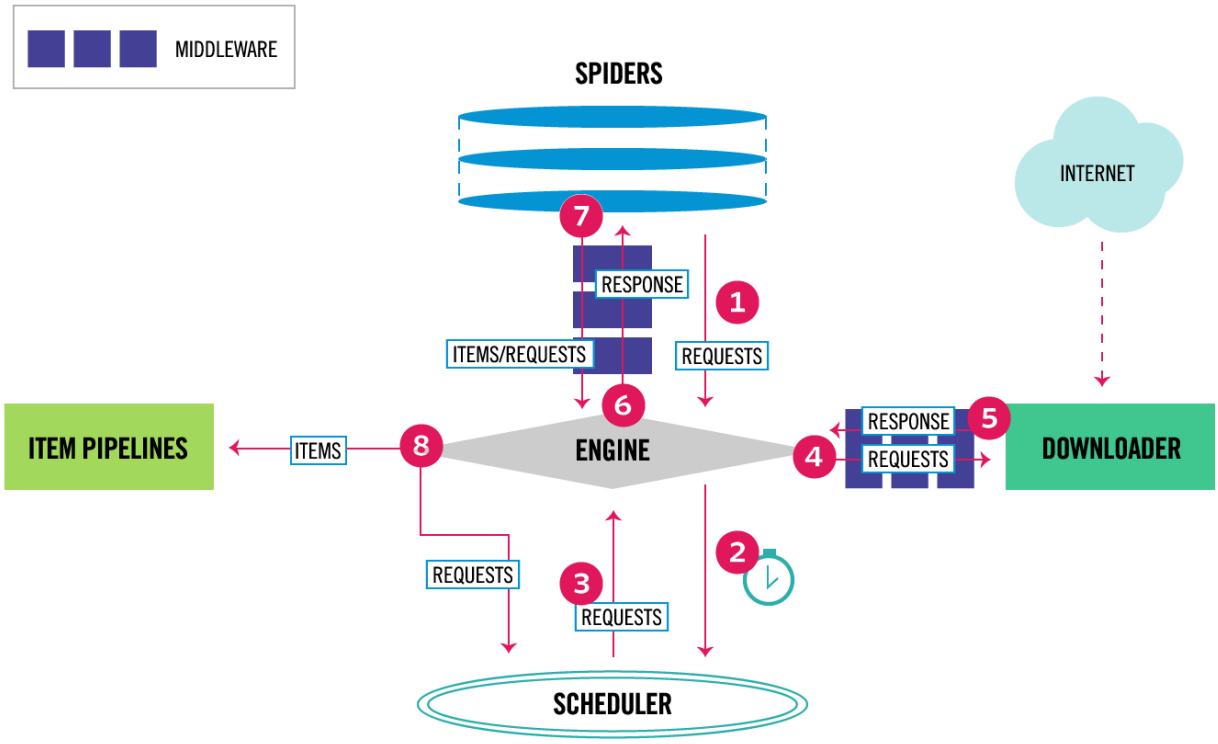
Le web est une collection immense de pages -- noeuds contenant de l'information -- reliés entre elles avec les hyperliens. On appelle 'web crawling' le balayage (de tout ou partie) de ces noeuds, à l'aide d'un programme qui télécharge les pages web, noeud par noeud, et qui découvre les liens qu'elles contiennent afin de les explorer et les télécharger à leurs tours. L'opération s'appelle aussi le data mining.

Le but en est donc de récupérer ces informations, pour les filtrer, classifier, indexer, mettre dans des bases de données, et les exploiter par la suite, par exemple avec des moteurs de recherche.

Un web crawler complet est composé d'au moins quatre éléments :

- Un groupe de robots qui explorent le web
- Un aspirateur de site web
- Un planificateur des requêtes
- Un serveur de données

Il en existe des libres et gratuits, comme **Scrapy** pour Python, dont le détail du fonctionnement est exposé ci-dessous.



Le flux de données dans Scrapy est contrôlé par un programme appelé moteur et qui exécute les différentes tâches dans l'ordre ci-dessous.

1. Le moteur récupère la requête initiale (adresse web à télécharger) d'un des robots (spider).
2. Le moteur envoies la requête au planificateur et demande la prochaine requête à traiter.

3. Le planificateur renvoies la requête à traiter en priorité au moteur.
4. Le moteur envoies la requête à l'aspirateur de site web.
5. Une fois le site aspiré, l'aspirateur envoies sa réponse contenant la page web au moteur.
6. Le moteur envoies la page à un des robots.
7. Le robot traite la page, en extrait les éléments d'intérêt et les nouvelles requêtes (adresses web à télécharger), et les envoies au moteur.
8. Le moteur envoies les éléments d'intérêt au serveur et les nouvelles requêtes au planificateur.
9. Le processus est répété depuis l'étape 3 et jusqu'à ce qu'il n'y ai plus de requêtes à traiter.

Programmation des fonctions principales d'un robot web

Un web crawler dispose d'une liste d'URLs qu'il doit visiter. Au démarrage, on donne au web crawler une URL de départ. Cette liste est une queue, dont les URLs sont extraites une par une et visitées. La liste est re-remplie ensuite avec les nouvelles URLs découvertes dans les pages téléchargées et filtrées, et le processus continue jusqu'à atteindre la profondeur maximale.

Les étapes principales du processus de web crawling sont les suivantes :

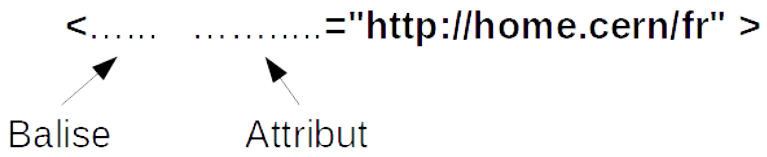
initialisation de la queue des URLs avec l'URL de départ
tant qu'on peut encore visiter
    prendre l'URL disponible suivant depuis la queue
    télécharger le contenu et marquer l'URL comme visitée
    extraire les hyperliens depuis le document nouvellement téléchargé et les rajouter
        dans la queue s'ils satisfont les critères nécessaires
    réévaluer les conditions pour continuer à visiter des sites
    attendre "un peu" avant de continuer (pour ne pas "assommer" le serveur)

A l'aide du module python **module_browser_snt.py**, coder les premiers aspects du crawler exposés ci-dessous.

1. Téléchargement de la page HTML.

2. Récupération de toutes les balises contenant les hyperliens. De quel type sont-t-elles ? Compléter l'exemple de balise ci-dessous.

.....
.....



3. Récupération des hyperliens. Constat. Ne conserver que les liens contenant le terme 'http'.

.....
.....

4. Élimination des liens contenant l'adresse du site traité.

Que reste-t-il à faire pour obtenir un robot web somaire ?

.....
.....
.....
.....