Création d'un web crawler à recherche d'offres d'emploi sur le web

Genèse

Actuellement pour un jeune diplômé, un demandeur d'emploi, les sites web spécialisés en recrutement et les réseaux sociaux sont des outils indispensables pour les aider à (re)trouver une offre d'emploi ou un stage. Donc habituellement la personne concernée visite régulièrement ces sites web et peut-être créer des comptes sur chacune de ces plateformes pour mettre toutes les chances de son coté afin d'être rapidement contacté par un recruteur.
Jusque là c'est bien mais seulement en Afrique, il existe des centaines de sites web spécialisés en recrutement !
Il est donc très difficile (pour un humain) de consulter ces portails numériques toutes les semaines.
Par contre il faut absolument rester connecté pour ne rien louper…

Comment faire alors ? Il suffit de… créer un bot internet!

Nous allons donc créer un robot qui va faciliter la tâche à ces utilisateurs.

Un bot qui explore des pages web est appelé web crawler (ou web scraper, web spider, etc.). Nous allons appeler ce bot "**JobTracker**"

JobTracker va continuellement explorer des pages web spécialisés en recrutement (avec l'accord des éditeurs) et rassembler toutes ces informations sur un seul endroit !
Et l'utilisateur final n'aura plus besoin visiter tous ces sites là, un seul lui suffit.

Pour rassembler toutes ces offres dans un seul endroit il faudra donc créer un nouveau site et y stocker ces offres (références).
On parle de référence car ce nouveau site ne copie pas bêtement les données des autres pages web mais uniquement l'essentiel et garde un lien vers la page d'origine. De ce fait, si un utilisateur clique sur les détails d'une offre d'emploi, il est directement redirigé vers le site en question.
Pour une lecture plus simple par la suite, ce nouveau site, nous allons l'appeler **Kelemploi**.

En résumé:

- **JobTracker** explore des sites d'offres d'emploi tous les jours et stocke les informations collectées dans une base de données.
- **Kelemploi** affiches un petit résumer de chaque offre d'emploi présent dans cette base de données que les utilisateur pourront consulter.

On peut aller encore plus loin en proposant aux utilisateurs de créer des alertes par mail. Et par la suite ils n'auront plus besoin de visiter aucun site.
Si une offre correspondante à une alerte d'un utilisateur est créée, le système envoie automatiquement un mail à cet utilisateur. Tout fonctionne de façon presque magique!

Implémentation

Pour mener à bien ce projet nous allons le découper en plusieurs parties:

1. Exploration des pages web
2. Stockage des informations collectées dans une base de données
3. Exploitation par des utilisateurs

Exploration des pages web

C'est surement la partie la plus difficile.
Cette partie consiste à créer une base de connaissance à partir de laquelle le robot pourra savoir quel site visiter, quel page contenu de quelle page sauvegarder, à quelle fréquence et à quelle vitesse faut-il exploiter chacune des pages désirées, etc.
Pour chaque site web, on a un fichier robots.txt unique, en plus la disposition des balises HTML des différentes, des noms de classes HTML différentes, des layout différents et peut-être le code HTML est régulièrement mis à jour. Donc il faudra créer des sous explorateurs de pages pour chaque site internet.

Stockage des informations

Le stockage des informations se fera au niveau de la base de données du site web. Nous allons donc créer une **API REST** pour faciliter l'échange de données entre le bot et la plateforme web.

Exploitation par des utilisateurs

Cette partie consiste à créer un site web traditionnel mais en plus qui affiche des offres d'emploi provenants d'autres sites web grâce aux données collectées par le bot.
Un module utilisateur (connexion/déconnexion) est requis pour que chaque utilisateur puisse créer une alerte (type de contrat, domaine d'activité, etc.) et donc recevoir un mail lorsque une offre qui match ces critères est créée.

NB: Pour pouvoir crawler le réseaux LinkedIn, il faut leur approbation. Nous allons donc leur envoyer un mail à l'adresse whitelist-crawl@linkedin.com pour qu'il whitelist le bot.