Lycée de l'Hyrôme

24h du code : Le moteur de recherche

Introduction

Nous vous proposons pour cette première partie du concours de développer une plateforme web qui intégrera un moteur de recherche. Rassurez-vous, on ne vous demande pas de re-coder Google, votre plateforme se limitera simplement à la recherche d'articles Wikipédia. Nous vous conseillons pour cela, de diviser votre travail en deux parties bien distinctes mais qui peuvent être réalisées en parallèle (et avec des technologies différentes) : le peuplement de votre base de données grâce à un crawler et la recherche dans cette base de données.

Crawler ?

Un web crawler (de l'anglais, qui signifie robot d'exploration") est un outil développé et utilisé par les moteurs de recherche qui vise à sonder et lire les contenus sur internet pour être intégrés ensuite à une base de données. Cette dernière servira ensuite aux moteurs de recherche pour proposer aux utilisateurs les meilleurs résultats au vue de leur recherche.

Nous vous proposons ici de développer votre propre crawler qui devra partir d'un article Wikipédia au choix puis aller chercher pour chaque lien le contenu de la page ainsi que les articles auxquels elle fait référence. Toutes les données récupérées devront être stockées dans votre base de données de façon efficace (nous vous laissons le choix dans la structure de cette dernière). L'action devra être répétée jusqu'à ce que vous ayez assez de contenu pour utiliser votre moteur de recherche.

Moteur de recherche

Dans cette deuxième partie vous devez mettre en place une page web avec un champ de recherche, permettant à votre utilisateur de saisir sa demande. Votre base de données étant peuplée par le crawler, vous pouvez aller y lire les informations pour renvoyer les articles Wikipédia les plus cohérents à la recherche. Par exemple si je recherche "mésopotamie", je m'attends à ce que le moteur de recherche m'affiche des articles y faisant référence, voire même en premier résultat l'article "mésopotamie"de Wikipédia (si il a été récupéré par le crawler dans l'étape précédente). Vous ne pourrez bien sûr pas avoir tous les articles dans votre base de données, c'est pourquoi le jury sera indulgent, pragmatique dans les tests et privilégiera un code propre. L'aspect esthétique est important et sera évalué par les autres groupes.

Vous justifierez par un oral d'une minute environ votre charte graphique, lors de la présentation de votre site aux groupes concurrents.