MAAIN : programmer un moteur de recherche

M2 informatique – Université de Paris

Année 2020-2021

Le but de ce TP est de programmer un moteur de recherche simple mais fonctionnel. Il s'agit pour chaque groupe de 3 étudiants de créer un site permettant à l'utilisateur de faire des recherches sur les pages Wikipédia françaises. Le langage de programmation n'est pas imposé mais l'efficacité doit être au rendez-vous.

Vous devrez rendre un rapport au format `pdf` résumant le fonctionnement du moteur, expliquant les choix techniques que vous avez dû faire, et répondant aux principales questions des fiches de TP (longueur indicative : 10 pages). *Attention*, l'évaluation se fera principalement sur le produit fini : le site du moteur de recherche doit être fonctionnel à la fin du cours.

Un moteur de recherche basique fonctionne sur ce principe très simple :

— calculer « une fois pour toutes » le « score » des pages d'internet pour chaque mot ;
— lors d'une requête, renvoyer toutes les pages, triées par score décroissant, qui contiennent les mots cherchés.

Le score d'une page pour un mot va dépendre à la fois de l'importance de la page dans l'absolu, et de la fréquence d'apparition du mot dans la page. La pertinence du moteur de recherche dépend évidemment de ce score attribué à chaque page. Pour évaluer l'importance d'une page dans l'absolu, nous allons utiliser le *pagerank* qui a fait le succès de Google. Celui-ci attribue un score plus élevé aux pages qui reçoivent plus de liens, l'idée étant qu'une page reçoit d'autant plus de liens qu'elle est plus populaire.

Pour concevoir notre moteur de recherche, nous allons procéder en différentes étapes :

(TP1) *(durée indicative : 3 à 4 séances)* programmer un « collecteur » (en anglais : *web crawler*) qui indexera les pages Wikipédia en suivant les liens des pages.

Ce collecteur aura deux tâches :

— « apprendre » le graphe orienté G des pages visitées (un sommet = une page ; un arc = un lien),
— associer à chacun des mots français les plus fréquents la liste des pages dans lesquelles il apparaît, avec sa fréquence d'apparition ;

(TP2) *(durée indicative : 2 à 3 séances)* à partir du graphe G obtenu, calculer le *pagerank* de chaque sommet.

Pour chaque mot, calculer le score de chaque page dans lequel il apparaît ;

(TP3) *(durée indicative : 2 séances)* réaliser le site de sorte que, sur une requête de l'utilisateur, il affiche l'ensemble des pages contenant tous les mots de la requête, par score décroissant.

Conseil : à chaque étape, testez rigoureusement vos fonctions pour être certain de leur bon fonctionnement avant de passer à l'étape suivante. Concevez et testez vos algorithmes d'abord sur de courts extraits du fichier total.