Programmer un moteur de recherche

M2 informatique – Université Paris Diderot

Année 2018-2019

Le but de ce TP est de programmer un moteur de recherche simple mais fonctionnel. Il s'agit pour chaque binôme 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 fonctionnant 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 :
— ordonner « une fois pour toutes » les pages d'internet par « score » décroissant ;
— lors d'une requête, renvoyer toutes les pages, triées par score décroissant, qui contiennent les mots cherchés.

Pour ce genre de moteur, donc, le résultat ne dépend pas de la fréquence des mots dans la page, mais seulement du score de la page. Toutefois, afin d'améliorer la qualité des résultats, on pourra ensuite prendre en compte d'une façon ou d'une autre la fréquence des mots dans les pages.

La pertinence du moteur de recherche dépend évidemment du score attribué à chaque page. 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 ;

(TP2) *(durée indicative : 2 à 3 séances)* à partir du graphe G obtenu, calculer le *pagerank* de chaque sommet.
Pour chaque mot, trier par ordre de *pagerank* décroissant la liste des pages contenant ce mot ;

(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 ordre de *pagerank* décroissant.

Conseil : à chaque étape, testez rigoureusement vos fonctions pour être certain de leur bon fonctionnement avant de passer à l'étape suivante.