Projet Moteur de Recherche

Pierre Senellart
pierre.senellart@telecom-paristech.fr

23 avril 2026

Il s'agit de réaliser, en équipe et sur une période de deux semaines, un moteur de recherche sur les documents du Web. Le but est – en se restreignant éventuellement à un domaine précis, à un corpus de documents précis, en se concentrant sur une fonctionnalité précise, etc. – de « faire mieux que Google ».

La première étape (lundi 3 octobre) est d'établir un cahier des charges détaillé des fonctionnalités attendues, et de répartir les tâches entre les membre de l'équipe. Un debriefing aura lieu le mardi 4 octobre pour valider les choix avec l'enseignant et orienter le développement. Il y aura ensuite une réunion quotidienne pour faire le point sur l'avancement du projet.

Le cahier des charges doit fournir des choix pour chacun des points suivants.

1 Domaine

On pourra décider de se restreindre, au choix, à :

— Une application donné : sites d'actualités français, réseaux sociaux, annonces immobilières, sites en esperanto, etc.
— Une ou plusieurs hiérarchies de noms de domaines (p. ex., les sites des écoles de ParisTech, `lemonde.fr`, `inria.fr`, `legifrance.gouv.fr`, etc.)
— Des corpus ne nécessitant pas de crawl : Wikipedia (`http://dumps.wikimedia.org/`), débats parlementaires (fournis), corpus Le Monde (fourni), crawls d'Internet Archive (p. ex., Wikileaks) disponibles sur `http://www.archive.org/details/web`, etc.

On peut aussi partir sur un crawl du Web dans son ensemble (évidemment restreint à quelques millions de pages vues les contraintes).

Quel que soit le choix retenu, il doit y avoir au minimum une centaine de milliers de documents indexés. Tenir compte des contraintes de politesse du crawl pour réfléchir à ce qu'il est raisonnable de crawler en deux semaines, si crawl il y a.

2 Fonctionnalités

Tout ou partie des fonctionnalités suivantes pourront être intégrées au système :

— Crawler
— Détection de la langue
— Prétraitements linguistiques pour une langue donnée
— Indexation plein-texte
— Ranking (PageRank, etc.)
— Clustering
— Extraction d'informations

— Système de recommandation
— etc.

Chacune de ces fonctionnalités pourra soit être développée à partir de zéro, soit s'appuyer sur des outils existants, au choix. Dans le cas d'un développement à partir de zéro, prévoir [1] 1 jour-homme par fonctionnalité pour un comportement de base, quelques jours-homme pour quelque chose de plus abouti, de l'ordre d'une semaine-homme pour un produit fini pouvant s'intégrer avec d'autres composants, plus pour des fonctionnalités supplémentaires.

Le moteur devra également être accessible via une interface Web : une attention particulière sera portée à rendre cette interface dynamique et ergonomique ; la sécurité de l'interface Web sera également une priorité.

3 Architecture

Une réflexion sera portée sur l'architecture, en fonction de la taille des données à traiter et des intérêts : architecture centralisée en utilisant un SGBD ou un moteur d'indexation existant, architecture centralisée avec stockage disque brut, architecture distribué avec HDFS et MapReduce, etc.

Un choix sera également fait sur les méthodes de travail collaboratif utilisées : utilisation d'un système de contrôle de version pour le code (Git, SVN, etc.), espace de partage pour la documentation, mailing liste, etc.

4 Ressources disponibles

Une machine puissante et connectée en permanence est mise à disposition ; il est également possible d'utiliser les machines des salles de TP, et d'organiser ces dernières en cluster. Dans la mesure du raisonnable, d'autres ressources matérielles et logicielles peuvent également être mises à disposition.

5 Répartition des tâches

Chaque membre de l'équipe a un rôle bien défini en fonction de ses capacités et intérêts personnels. Le choix de ses rôles se fait conjointement et devra être inscrit, nominativement, dans le cahier des charges.

Les rôles suivant devront être représentés :

— 1 chef de projet intégrateur ; son rôle est d'assurer la bonne répartition des tâches et le progrès du projet dans son ensemble, et de s'assurer que l'ensemble des composants développés peuvent interagir. Il est de sa responsabilité de décider de réorienter le développement de l'ensemble du projet en cas de besoin.
— 1 développeur pour chaque module logiciel du moteur de recherche (plusieurs modules logiciels simples peuvent être confiés au même développeur, un module logiciel complexe pouvant être partagé entre deux développeurs).
— 1 concepteur d'interface pour la partie « côté client » de l'interface Web.
— 1 programmeur Web pour la partie « côté serveur » de l'interface Web.
— 1 responsable qualité, dont le rôle est d'une part de tester chacun des modules développés et d'autre part de mettre en place un protocole de comparaison des résultats du moteur avec ceux de moteurs existants (au moins Bing et Google). Il réalisera lui-même ces tests et rendra compte au chef du projet et aux développeurs des améliorations à apporter.

1. Estimation grossière, certaines fonctionnalités étant plus simples que d'autres à implémenter.

— (éventuellement) 1 rédacteur technique, qui rédigera la documentation interne (APIs des différents modules) et externe (rapport final sur les fonctionnalités et comment les utiliser) du projet.
— (éventuellement) tout autre poste qui semblera nécessaire à l'exécution du projet