Recherche d'Information sur le Web

Mise en place d'un moteur de recherche sur un index de questions Amazon avec Lucene

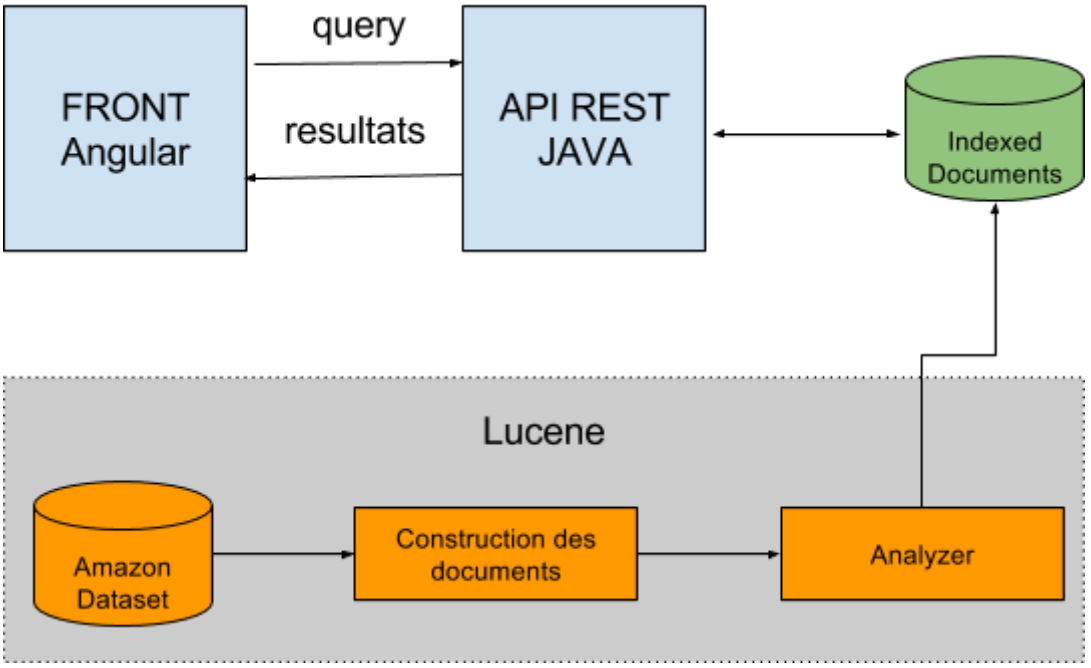
Par
Philippe Leboc
Rémi Bouguermouh

Présentation

Ce projet a pour objectif la mise en place d'un moteur de recherche. Le sujet du projet consiste à créer une version simplifiée d'un moteur de question réponse (slackoverflow, quora …). **Dans le cadre de ce projet, nous utiliserons un jeu de donnée de question / réponse provenant d'Amazon**[1].

Pour le moteur de recherche, nous utiliserons la bibliothèque Lucene[2].

Architecture du projet



Lucene

Lucene est une bibliothèque Java pour l'indexation et la recherche de contenu. Au coeur de l'architecture de Lucene pour l'indexation de contenu, il y a 2 notions importantes :

- Les *documents.* Un document correspondant à une unité dans la base indexé. Un document est composé de plusieurs fields.
- Les *fields*. Un *field* est un champ texte identifié par une clé et une valeur (textuel).

Dans le cadre du projet un document possède 2 *fields* :

- Field Question. Qui contient l'intitulé d'une question.
- Field Reponse. Qui contient l'intitulé d'une réponse à une question donné.

[1] http://jmcauley.ucsd.edu/data/amazon/qa/
[2] https://lucene.apache.org/

Cette implémentation permet d'indexer les intitulés de question et leurs réponse. L'intérêt d'indexer aussi les réponses c'est qu'elles apportent souvent plus d'information.

Example :
Si nous faisons une recherche avec le terme '*Motion Gaming*', aucun intitulé de question intègre ce terme. En revanche ce terme apparaît sur une réponse relative au Kinect de Microsoft. Ce document sera donc retourné avec un meilleur **score**.

Le score d'un document est calculé à partir du score combiné de ses *fields*.

Code source

Le code du projet est disponible sur la plateforme github :
https://github.com/Mathael/shlagoverflow