Cours Technologies Web : M1INFO 2023

Conception et développement d'un projet Web/Informatique :
Outil de recherche d'informations pour un corpus de documents

1. Planning des séances

09 mars : Présentation du projet, contexte et développement à faire
16 mars : Suivi et présentation de l'avancement des développements
23 mars : Suivi et présentation de l'avancement des développements
30 mars : Suivi et présentation de l'avancement des développements
06 avril : Présentation finale et validation

2. Présentation du travail à réaliser : Un outil de recherche d'information

L'objectif est de concevoir et développer un outil de recherche d'informations, cela consiste à permettre **une recherche rapide** et **pertinente** avec des mots-clés dans un ensemble de documents TEXTES.

Les notions de **recherche rapide** et **pertinente** évoquées doivent être prises en compte dans la conception et le développement de l'outil pour permettre de réduire le temps de recherche et mettre en ordre la pertinence des documents trouvés pour l'utilisateur.

Pour faire, le travail doit être structuré autour des entités obligatoires suivantes :

1. Module d'indexation de documents TEXTEs
2. Base de données : mots ↔ documents
3. Interface de recherche simple mots-clés → liste de documents pertinents
4. La lecture récursive de dossier

I. Module d'indexation de document (Séance 1 et 2)

indexer n'importe quel texte brute (fichiers .txt) : : qui permet de générer une représentation sous forme d'une liste (vecteur) de mots la plus réduite possible de chaque texte.

1- opération de tokenisation : on découpe le texte sur la base d'un ensemble de caractères de séparation et d'espacement et autres… Suppression avec une liste (fichier) de mots dits vides (contrairement à des mots pleins), la liste se nomme une stopwords.
Une liste stopwords c'est quoi : Les experts en sémantique parlent de liste stopwords pour **les mots qui, en théorie, n'ont aucune influence sur le référencement naturel d'un texte**. Un mot vide est un mot non significatif figurant dans un texte. On l'oppose à "mot plein", qui dégage de la sémantique. Cette liste est fournie, donc vous n'avez pas à vous en occuper de sa constitution mais juste de son usage algorithmique.

2- on filtre et on calcule, **la fréquence des mots**, quand on calcule la fréquence cela supprime automatiquement les doublons et donne aussi des statistiques précises sur le poids (l'importance par présence) des mots dans le texte.

À la fin, ce module (script) donnera à partir d'un fichier, une liste de mots avec la fréquence de chacun dans le texte (fichier).

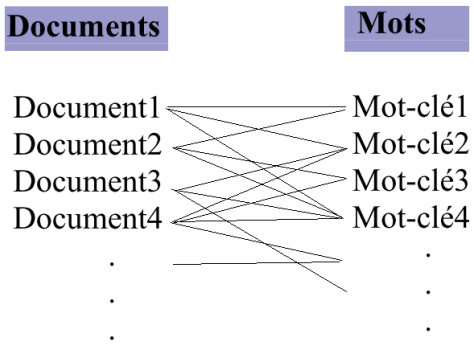
3- affichage et présentation des traces du traitement des textes dans le module indexation :

A l'exécution du module, le module doit présenter le déroulement du traitement de l'indexation par des statistiques pour chaque document traité et NON par une suite de listes de mots longues et des messages difficiles à comprendre et à suivre. Les statistiques sous forme de synthèse résumeraient mieux les résultats pour chaque document traité et du coup pour l'ensemble des textes à indexer.

II. Base de données : enregistrement des résultats de l'indexation (relation Document ↔ Mot (+ Occurrence) (séance 2 et 3)

La base de données est considérée ICI comme un support intermédiaire entre l'utilisateur et les documents pour faciliter et rendre l'accès à l'information rapide, elle doit contenir les résultats du traitement des documents du module indexation. La base de données doit mettre en évidence dans une base de données la relation :

documents ↔ mots



+ la propriété fréquence, cela traduit la présence de mots dans un document et sa fréquence (nombre de fois d'apparition dans ce même document).

III. Module de recherche et visualisation (séance 3 et 4)

1- rechercher par mots-clés : interface avec simple champ de recherche

Cette interface se présente comme celle de google, simple mais efficace car elle cache le gros du travail derrière ! Elle doit chercher dans la base de données

-dans la table des mots, si un mot est trouvé alors on trouve le document et le nombre d'occurrences associé

-le classement se fera uniquement par occurrences du mot dans les documents trouvés
-chaque résultat se présente comme sur google ou proche pour le moment :

MyEngine

Rechercher Mot

Nombre de réponses pour (mot) : 2

1. fichier1.txt (5)

2. source2.txt (4)

V. Lecture récursive du contenu d'un dossier (séance 4 et pour la validation séance 5)

Cette lecture comme nous l'avions vu dans le cours du 1er semestre, permettra de lire et d'explorer facilement n'importe quelle structure arborescente de dossiers et de fichier sans se soucier de la taille et de la complexité de arborescence.
C'est ce module qui est lancé au démarrage de chaque traitement d'indexation sur un ensemble de documents dans un dossier, il fera appel au 1er module de traitement (indexation) pour chaque type de document souhaité ou ciblé par l'indexation (dans ce travail seulement les .TXT). L'importance d'utiliser cette lecture automatique sera d'éviter d'aller chercher chaque document à indexer manuellement et de rendre la tâche impossible dans la cas d'un nombre important de documents ou dans le cas d'une augmentation régulière de documents à indexer.