Cahier des Charges du Projet TIW

Partie théorique :

Titre :

"ExploreMots" - Moteur de Recherche Textuel**

Résumé :

Le projet vise à concevoir un moteur de recherche capable d'indexer et de récupérer des informations à partir de fichiers de différents formats tels que HTML, PDF et texte brut. Ce moteur de recherche facilitera la lecture, la sauvegarde, et l'indexation des contenus de ces fichiers dans une base de données. L'interaction avec le moteur se fera via une interface utilisateur permettant de rechercher des mots-clés dans les documents indexés.

Mots-clés :

1. Indexation

2. Occurrences : Le comptage des occurrences consiste à déterminer combien de fois chaque mot apparaît dans un ensemble de documents.

3. Lemmatisation : La lemmatisation est le processus de réduction des mots à leur forme de base

4.Tokenisation : La tokenisation est le processus de découpage d'un texte en unités linguistiques de base, appelées tokens.

5.La recherche dans la base de données : permet de trouver les documents contenant un mot spécifique et de récupérer des informations associées, telles que le nombre d'occurrences.

Contexte :

Le projet s'inscrit dans le contexte de l'analyse et de la gestion de contenu textuel, c’est pour faciliter l’acces au information a travers different type de fichiers.

Domaine d'Utilisation du Projet :

Le moteur de recherche sera utile dans divers domaines tels que l'éducation, la recherche documentaire, et toute autre activité nécessitant la récupération rapide d'informations à partir de fichiers textuels comme le moteur de google..

Usages :

1. Sauvegarde et Indexation : Les utilisateurs peuvent sauvegarder des fichiers textuels à travers l'interface et permettre au moteur de recherche de les indexer.

2. Recherche de Mots-clés : Les utilisateurs peuvent effectuer des recherches de mots-clés pour récupérer des informations à partir des fichiers indexés.

3.Visualisation de Statistiques : Le système affiche le nombre d'occurrences d'un mot dans les documents, ainsi que les documents associés.

Webographie :

https://www.comprendre-internet.com/Qu-est-ce-que-le-HTML.html

https://www.lebigdata.fr/sql-tout-savoir-guide

Partie Technique :

Langage Utilisé :

Technologies Utilisées :

PHP : Langage principal pour la logique serveur.

HTML, CSS, Bootstrap : Utilisés pour la création d'une interface utilisateur interactive et esthétique.

MySQL : Base de données pour le stockage des mots, occurrences et informations sur les fichiers.

JavaScript (Ajax) : Pour les requêtes asynchrones entre le frontend et le backend.

Fonctionnalités :

Exploration de Dossier :

La fonction file_get_contents parcourt tous le fichier d'un dossier donné,

$contenuFichier = file_get_contents("toto.txt");

-Tokenisation :

Exemple de tokenisation dans le projet:

$tok = strtok($texteSansMotsVides, " \n\t\r");

Dans cet exemple, la fonction `strtok` découpe le texte en mots en utilisant les espaces, les sauts de ligne et les tabulations comme délimiteurs.

- Nettoyage du Texte :

Exemple de nettoyage du texte dans le projet:

```
function nettoyerTexte($texte) {
   $texte = strtolower($texte); // Mettre en minuscules
   $texte = preg_replace("/[^a-zA-Z\s]/", "", $texte); // Supprimer la ponctuation et les chiffres
   return $texte;
}
```

Cette fonction prend en entrée le texte brut et renvoie le texte nettoyé en minuscules, sans ponctuation ni chiffres.

Comptage des Occurrences :

Exemple de comptage des occurrences dans le projet:

```
function compterOccurrences($mot) {
}
```

Cette fonction serait utilisée pour maintenir un compte des occurrences de chaque mot dans la base de données.

Recherche dans la Base de Données :

```
function rechercherMot($mot) {
   // les multiple requête qu'on peut utiliser ( select * from document …)
}
```

Cette fonction serait utilisée pour effectuer une requêteS SQL afin de récupérer les informations sur le mot recherché.

Recherche de Mots-clés :

Interface utilisateur permettant aux utilisateurs de rechercher un mot clé. Envoi d'une requête au backend, exécution d'une requête MySQL, et affichage du résultat.

Nuage de Mots-clés :

Fonctionnalité graphique permettant à l'utilisateur de visualiser la fréquence de chaque mot dans un cadre à travers un nuage de mots clés.