Master SII 2018/2019

Projet de TP en Recherche d'information

Ce TP vise à vous familiariser avec les procédures de base en recherche d'information (RI) à savoir : (1) l'indexation des documents, (2) la pondération des termes et (3) l'appariement requête-document. Dans ce TP nous allons réaliser le modèle Booléen, le modèle Vectoriel et le modèle Probabiliste.

1. La collection de documents

Dans ce TP, la collection de travail est un dossier contenant un ensemble de N documents textuels à créer vous-même.

2. Création d'un fichier inverse de fréquences

L'utilisation d'un fichier inverse peut grandement améliorer l'efficacité de la RI. On vous demande donc de convertir votre résultat d'indexation en un fichier inverse de fréquences. Ce fichier inverse correspond à la structure suivante :

[Mot, document] → fréquences

Remarque : Utilisez le fichier des mots vides donné en TP pour supprimer les mots vides.

3. Création d'un fichier inverse de poids

Ce fichier inverse correspond à la structure suivante :

[Mot, document] → poids

Utiliser la formule de pondération TF*IDF: $poids(ti, dj)=(freq(ti,dj)/Max(freq(t, dj))*Log((N/ni) +1)$

4. Fonctions d'accès

Pour connaître quels sont les mots qui apparaissent dans un document et avec quelles fréquences et quels poids, on doit programmer deux fonctions d'accès de base. La première accepte un numéro de document, et retourne sa liste des mots avec leurs fréquences et leurs poids. La deuxième accepte un mot, et retourne la liste des documents contenant ce mot avec sa fréquence et son poids dans chaque document.

5. Réalisation du modèle Booléen standard

Programmer une fonction qui accepte une requête au format booléen standard et renvoie tous les documents pertinents à cette requête (similarité=1).

6. Réalisation du modèle Vectoriel

Programmer une fonction qui accepte une requête au format vectoriel (un ensemble de mots) et renvoie tous les documents pertinents à cette requête avec leurs similarités par rapport à la requête, triés par ordre décroissant. Vous devez implémenter quatre fonctions d'appariements : 1- Produit Interne ; 2- Coef de Dice ; 3- Cosinus ; 4-Jaccard.

7. Réalisation du modèle probabiliste avec un échantillon d'apprentissage

Pour implémenter le modèle probabiliste, il nous faut un échantillon d'apprentissage. Dans notre cas, cet échantillon sera les documents jugés pertinents par l'utilisateur. Donc, dans ce TP, ce modèle probabiliste sera implémenté en deux étapes. La 1ère étape consiste à répondre à la requête utilisateur avec le modèle vectoriel déjà développé précédemment, puis l'utilisateur choisira parmi les documents trouvés ceux qui sont pertinents pour lui. La 2ème étape fait une deuxième recherche avec cette même requête en appliquant le modèle probabiliste, en utilisant l'ensemble des documents choisis par l'utilisateur comme un échantillon d'apprentissage.

8. IHM

Développer une IHM permettant d'afficher les différentes informations de ce projet.

9- Rapport à remettre

Vous devez remettre un rapport écrit, qui contiendra au minimum les points suivants :

- Présentation du projet.
- Explication de vos algorithmes
- Format (structure) de vos fichiers d'indexation, avec des exemples.
- Quelques interfaces avec les résultats des requêtes pour chacun des modèles.
- Analyse et discussion de vos résultats.