Série d'exercices de TP
Module : RI

Collection

Créez un dossier appelé « TPRI », créez dans ce dossier N documents de type texte. Chaque document i est nommé « Di ». Ce dossier sera notre collection de travail pour cette série d'exercices. Réalisez les exercices suivants avec le langage Python.

Exercice 1

Ecrire un programme python qui permet la création des structures suivantes : (en ignorant les mots vides et les ponctuations (. , ! ? ' …etc). pour les mots vides voir le fichier : stopwords_fr , qui vous sera donné en TP) :

1- Créer pour chaque document une structure, contenant les termes (les mots) du document avec leurs fréquences. Dans cette structure la clé sera le terme, la valeur sera sa fréquence d'apparition dans le document.

2- Créer une seule structure, contenant tous les termes des documents avec leurs fréquences (chaque terme est associé avec sa fréquence d'apparition dans son document). Dans cette structure la clé sera une paire (terme, N° du document), la valeur sera la fréquence d'apparition de ce terme dans ce document. Si le terme n'existe pas dans ce document, sa valeur sera 0. Nous appelons cette structure le Fichier Inverse de fréquences de cette collection.

Exercice 2

En utilisant la structure du Fichier Inverse de fréquences obtenue dans l'exercice 1, programmer deux fonctions d'accès de base. La première reçoit un numéro de document, et retourne la liste des termes avec leurs fréquences dans ce document. La deuxième reçoit un terme, et retourne la liste des documents et les fréquences de ce terme dans chaque document.

Exercice 3

En utilisant la structure du Fichier Inverse de fréquences obtenue dans l'exercice 1, créer une autre structure contenant tous les termes des documents avec leurs poids (chaque terme est associé à son poids). Ce poids est calculé par la formule de pondération suivante :

$$poids(ti, dj) = (freq(ti,dj)/Max(freq(t, dj))*Log((N/ni) +1)$$

avec :

- poids(ti, dj) : le poids du terme i dans le document j
- freq(ti,dj) : la fréquence du terme i dans le document j, qui est déjà calculée dans le Fichier Inverse de fréquences.
- Max(freq(t, dj) : La fréquence max dans le document j
- N : le nombre de documents dans la collection
- ni : le nombre de documents contenant le terme i
- Log : c'est le Log de 10.

Dans cette structure la clé sera une paire (terme, N° du document), la valeur sera son poids. Si le terme n'existe pas dans ce document, son poids sera 0. Nous appelons cette structure le Fichier Inverse pondéré de cette collection.

Exercice 4

En utilisant la structure du Fichier Inverse pondéré obtenue dans l'exercice 3, programmer deux fonctions d'accès de base. La première reçoit un numéro de document, et retourne la liste de ses termes avec leurs poids. La deuxième reçoit un terme, et retourne la liste des documents et les poids de ce terme dans chaque document.