Examen du module : Recherche d'information

Exercice 1 (4 pts) : *Question de cours*

1. Est ce qu'il y a une différence entre un système de gestion de base de données et un système de recherche d'information ? OUI 1pt
2. Un terme qui apparaît dans un seul documents d'un corpus est-il discriminant ou non ? Justifier votre réponse. 1.5pt
   Un terme qui apparaît dans un document est discriminant car ce terme distingue ou faire la différence entre ce document des autres documents.
3. Dans le modèle vectoriel, à quoi correspondent les axes de l'espace vectoriel ? 1.5pt
   Les axes de l'espace vectoriel correspondent aux termes de la collection de documents.

Exercice 2 (4 pts) : *Indexation*

Soient les ensembles des termes obtenus de l'indexation des documents D1 et D2 suivants:

D1 = {efficacité, recherche, mesurée, précision, moyenne}

D2 = {modèles, recherche, efficaces, langage, vectoriel}

1. Donner la table des fréquences : terme, document; 2pts

termes	D1	D2
efficacité	1	0
recherche	1	1
mesurée	1	0
précision	1	0
moyenne	1	0
modèles	0	1
efficaces	0	1
langage	0	1
vectoriel	0	1

2. Calculer TF*IDF de chaque terme où TF = $\frac{freq(t_i,d_j)}{1.5*\left(\frac{Longueur\_doc\_dj}{Longueur\_moy\_doc}\right)+freq(t_i,d_j)+0.5}$; IDF=$Log(\frac{N}{Nt})$; où N_t est le nombre de documents contenant le terme t_i et N est le nombre de documents. 2pts

   Longueur_doc_D_1=5; Longueur_doc_D_1=5; Longueur_moy_doc=10/2=5

termes	D1	D2	TF*IDF (D1)	TF*IDF (D2)
efficacité	1	0	[1/(1.5+1+0.5)]*log(2) =0.1	0
recherche	1	1	0	0
mesurée	1	0	(1/3)*log(2) =0.1	0
précision	1	0	(1/3)*log(2) =0.1	0
moyenne	1	0	(1/3)*log(2) =0.1	0
modèles	0	1	0	(1/3)*log(2)=0.1
efficaces	0	1	0	(1/3)*log(2) =0.1
langage	0	1	0	(1/3)*log(2) =0.1
vectoriel	0	1	0	(1/3)*log(2) =0.1

Examen du module : Recherche d'information

Exercice 3 (6 pts) : *Modèles de recherche d'information*

Nous voulons mesurer la correspondance (la similarité) entre les documents d'un corpus qui ont été préalablement pondérés. Pour la correspondance entre un document **A** et un document **B** on utilisera la formule du cosinus du modèle vectoriel.
Soit la fonction **COR_COS( )** qui prend en argument deux tableaux de poids **WEIGHT_A** et **WEIGHT_B** contenant les poids des termes des deux documents, et qui renvoie la valeur de la correspondance entre ces deux documents.

- A quoi correspond la taille de **WEIGHT_A** et de **WEIGHT_B.**

La taille de WEIGHT_A et de **WEIGHT_B est égale à la taille de l'**espace vectoriel engendré par termes de la collection de documents donc **WEIGHT_A** et de **WEIGHT_B ont la même taille.** 0.5

- Écrivez l'algorithme de la fonction **COR_COS( ).**

La formule du cosinus de calcul de correspondance entre un document d_i et d_k est donnée comme suit: **0.5**

$$R(d_i, d_k) = \quad Cos(d_i, d_k) = \frac{\sum_{j-1}^{m} w_{ij} * w_{kj}}{\sqrt{\sum_{j-1}^{m} w_{kj}^2 * \sum_{j-1}^{m} w_{ij}^2}}$$

5pts

```
float COR _COS(WEIGHT_A[],WEIGHT _B[],   int taille) {
        float sommeProduits = 0.0;
        float sommePAcarre = 0.0;
        float sommePBcarre = 0.0;
for (int i=0; i<taille; i++){
        sommePAcarre += WEIGHT _A[i]* WEIGHT _A[i];
        sommePBcarre += WEIGHT _B[i]* WEIGHT _B[i];
        sommeProduits += WEIGHT_A[i]* WEIGHT_B[i];
        }
return sommeProduits/(sqrt(sommePAcarre* sommePBcarre));
}
```

Exercice 4 (6 pts) : *Évaluation des SRI*

Nous souhaitons évaluer un système de recherche d'information Sys1. Supposons que pour une requête Q1 le système S1 testé renvoie les réponses suivantes:

rang	n° doc	pertinent	précision	rappel
1	588	X	1/1=1	1/10=0.1
2	589	X	2/2=1	2/10=0.2
3	576	X	3/3=1	3/10=0.3
4	590			
5	986			
6	592	X	4/6=0.67	4/10=0.4
7	884			
8	988			
9	578	X	5/9=0,56	5/10=0.5
10	985			

Examen du module : Recherche d'information

Supposons qu'il y a dans l'ensemble de tous les documents, 10 documents jugés pertinents pour la requête Q1.

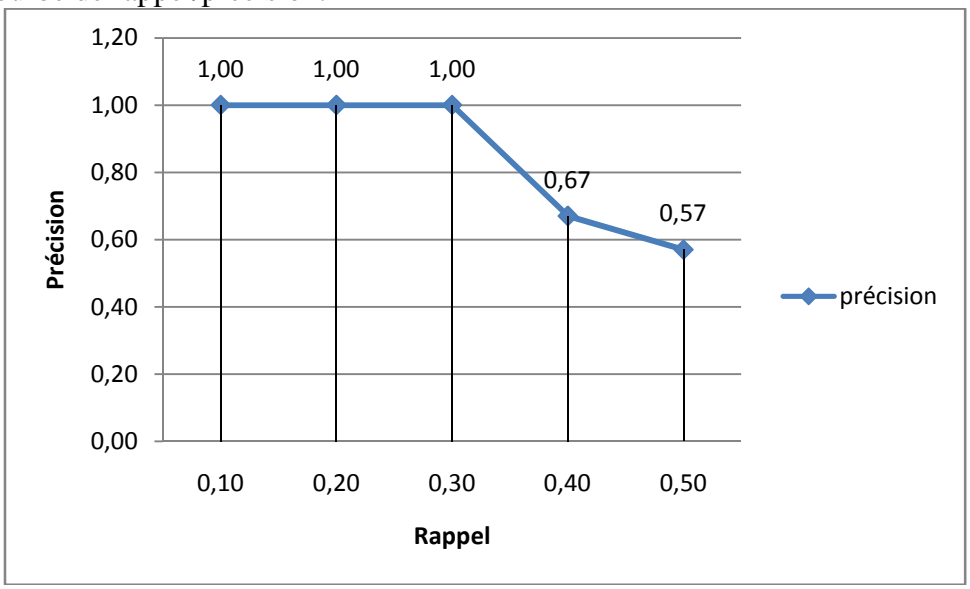
1. Calculer les taux de précision et de rappel du système et remplir le tableau ci-dessus. 2pts

$$rappel = \frac{Nombre\ de\ documents\ pertinents\ séléctionnés}{Nombre\ total\ de\ documents\ pertinents}$$

$$précision = \frac{Nombre\ de\ documents\ pertinents\ \ sélectionnés}{Nombre\ total\ de\ documents\ sélectionnés}$$

précision	rappel
1,00	0,10
1,00	0,20
1,00	0,30
0,67	0,40
0,56	0,50

2. Dessiner la courbe de rappel/précision. 2pts



3. Calculer les taux de précision "interpolés" pour les taux de rappels 0, 0.1, … 1.0. 2pts

rappels	précisions
0	1
0,1	1
0,2	1
0,3	1
0,4	0.67
0,5	0.56
0,6	0
0,7	0
0,8	0
0,9	0
1	0