TD - Recherche d'information Multimédia

Exercice 1 – modèle booléen pondéré

Considérons deux documents en utilisant un modèle booléen pondéré D1 et D2 tels que :

t	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
$W_{D1}(t)$	0.5	0	0.8	0	1	0	0.6	0.8	0	0.9	1	0	0	0

t	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
$W_{D2}(t)$	1	0.7	0	0	1	0	0	0	0	0.9	0	0.3	0	0.5

1. Représenter le fichier inverse relatif à un corpus composé uniquement de ces deux documents, puis représenter l'arbre de requête et les étapes de correspondance entre ce corpus de deux documents, pour les deux requêtes suivantes

Q1 : $t_1 \wedge t_5$

Q2 : $(t_1 \wedge t_5) \vee (t_8 \wedge t_{10})$

pour les deux versions de correspondance vue en cours (similarité basée sur la logique floue, puis la seconde similarité vue en cours).

Exercice 2 – modèle vectoriel

Considérons deux documents d1=(0.5, 0.5) et d2=(0.25, 1), et une requête q=(1, 0.5). Représenter ces vecteurs graphiquement, et déduire d'après vous l'ordre des réponses d'un système vectoriel. Valider votre intuition en utilisant une correspondance utilisant un cosinus.

Exercice 3 – modèle vectoriel

Considérons les documents suivants :

d1 = (3,0,3,0,0,0) d2 = (3,0,2,1,0,0) d3 = (1,2,3,2,1,0)

Considérons les requêtes q1 = (1,0,1,0,0,0) et q2 =(0,0,0,1,0,1).

1. Utiliser comme fonction de correspondance la méthode du cosinus théorique pour calculer la valeur de pertinence système de ces documents. Les ordonner par pertinence décroissante.
2. Refaire les mêmes calculs en se basant sur la représentation par fichier inverse comme vu en cours, et vérifier que les résultats sont exactement les mêmes.

Exercice 4 – correspondances dans le modèle vectoriel

Reprendre la question 1 de l'exercice 3 en utilisant les deux autres correspondances théoriques vues en cours : Dice et Jaccard. Commenter les résultats obtenus par ces 3 similarités, et en déduite pourquoi, en se basant sur ces exemples, la correspondance par cosinus est la plus susceptible de bien fonctionner sur des gros corpus de documents.

Exercice 5 – pondération dans le modèle vectoriel

Considérons les textes suivants :
Considérons les textes suivants :
Document 1 : L'opacité à droite du genou du corps calleux signale une tumeur.
Document 2 : Le corps calleux unit la partie droite et la partie gauche du cerveau.
Document 3 : Le genou du corps calleux est formé par des fibres.

1. En considérant un anti-dictionnaire composé des termes :
{à, au, d, de, du, des, elle, elles, est, et, je, il, ils, l, le, la, les, lui, par, qui, son, s, sa, ses, tu, un, une} représenter l'ensemble des termes d'indexation de chacun des documents ci-dessus.
2. Calculer les tf de chacun de ces termes pour chaque document.
3. Calculer l'idf de chacun des termes présents dans les documents
4. En déduire le tableau du fichier inverse pour ce corpus.
5. Calculer les normes des vecteurs documents.

Exercice 6 – pondération dans le modèle vectoriel

Fournir les résultats des requêtes suivantes pour le corpus de l'exercice 5 :
Q0 : pomme de terre
Q1 : corps calleux
Q2 : corps calleux gauche
Q3 : tumeur sur le genou du corps calleux
Commencer par analyser les requêtes comme les documents (anti-dictionnaire), et utiliser une pondération des requêtes par tf.idf. Vous devez faire utiliser les calculs théoriques ET PAS la version basée sur des fichiers inverses.

Exercice 7 – modification de requête par thesaurus existant

Considérons un thésaurus qui représente des liens de généricité/spécificité de la manière suivante :
tissu → cellule
tissu → épithélium
tissu → glande
tissu → membrane
tissu → muscle

D'autre part, considérons un vocabulaire tel que t1 = cellule, t2 = épithélium, t1 = glande, t4 = membrane, t5 = muscle, t6 = vitesse.
Prenons les vecteurs suivants :
d1 = (1,0,1,0,0,0) d2 = (3,0,2,1,0,0)
d3 = (1,2,3,0,1,0) d4 = (0,0,0,2,1,2)
d5 = (1,1,1,4,2,1) d6 = (1,1,0,0,3,2)

Si nous posons une requête "tissu" en le faisant comme habituellement avec un système basé sur un modèle vectoriel, quelle serait la réponse du système? (on suppose que la pondération de la requête est uniquement le tf, pour simplifier, et on utilise la formule théorique).
Etendre la requête en propageant à tous ses spécifiques le poids du terme de la requête initiale, "tissu", puis évaluer la réponse de la requête en utilisant le cosinus.

Exercice 8 – algorithme de Porter

Rappel : L'algorithme de Porter sur la langue anglaise tente de définir des troncatures de mots pour améliorer la réponse des systèmes de recherche d'information. L'hypothèse est que des mots proches sémantiquement auront une troncature identique, cela amenant à améliorer la qualité des réponses du SRI.
L'algorithme de Porter est décrit en détail en http://snowball.tartarus.org/porter/stemmer.html
Nous ne faisons qu'en reprendre des parties dans cet exercice.
Les règles de réécriture que nous utilisons sont différentes de celles vues en cours :
1. S → /
2. ED → /
3. Y → /
4. ION →
5. E → /
6. R → /
7. AL → /
8. double consonne → la consonne

Prenons 4 documents :
D1 = "white blood cells abcesses"
D2 = "central fibrovascular cell"
D3 = "evidence centered invasion"
D4 = "abcesses diagnosed clinically"

1. En considérant le vocabulaire T = {abcesses, blood, cell, cells, central, centered, clinically, diagnosed, evidence, fibrovascular, invasion, stromal, white}, des pondérations uniquement basées sur le tf et la similarité basé sur le cosinus, donner le résultat d'une requête "cells" : Q0 = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0). Utiliser la formule du cours sans passer par les fichiers inverses.
2. Donner pour ces 4 documents les termes tronqués qui les indexent. En déduire le vocabulaire d'indexation de ce corpus.
Reprendre la même requête Q0, lui appliquer la troncature et la réévaluer.

Exercice 9 – Evaluation de SRI

Nous réalisons ici une évaluation d'un système de recherche d'information.
Supposons que pour une requête Q1 le système de recherche d'information testé renvoie les réponses suivantes:

rang	n° doc	pertinent	rappel	précision

1	588			
2	589			
3	576			
4	590			
5	986			
6	592			
7	884			
8	988			
9	578			
10	985			
11	103			
12	591			
13	572			
14	990			

Les documents pertinents pour Q1 sont : 572, 588, 589, 590 ,592.
Calculer les taux de précision et de rappel du système à chaque réponse et remplir le tableau ci-dessus. Dessiner la courbe de rappel/précision. Calculer le taux de précision moyen.
Normaliser les résultats de façon à obtenir les valeurs de précisions maximales par taux de rappel, puis obtenir les taux de précision "interpolés" pour les taux de rappels 0, 0.1, … 1.0 . Calculer le taux de précision moyen sur les valeurs de rappels 0, 0.1, …, 1.0 .

Réaliser le même travail pour la requête Q2, avec les réponses suivantes :

Rang	n° doc	pertinent	Rappel	précision
1	324	X		
2	589	X		
3	528	X		
4	590	X		
5	986	X		
6	592	X		
7	899	X		
8	988	X		
9	578			
10	985			
11	537	X		
12	591	X		
13	772	X		
14	990			

La liste des tous les documents pertinents pour la requête Q2 est : 324, 528, 537, 589, 590, 591, 592, 772, 899, 986, 988. En regardant les courbes, que pouvez-vous déduire de la qualité relative du système pour ces deux requêtes?
Réaliser l'intégration des résultats du système pour les deux requêtes et dessiner le schéma résultant.

Exercice 10 – Comparaison de SRI

Nous voulons comparer deux systèmes de recherche d'information. Le premier système S1 est celui de l'exercice 9. Le second système, S2, a pour tableau de rappel/précision pour les deux requêtes Q1 et Q2:

Rappel	Précision
0	0..92
0.1	0.88
0.2	0.86
0.3	0.84
0.4	0.80
0.5	0.75
0.6	0.72
0.7	0.70

0.8	0.65
0.9	0.63
1.0	0.61

Tracer les courbes de S1 et S2 sur la même figure. Analyser les courbes pour en déduire lequel des deux systèmes semble le meilleur.