TD – Recherche d'information

Exercice 1 – modèle booléen pondéré

Considérons un document D1 représenté sur un vocabulaire T={t_1, …, t_{10}}.

- W_{D1} est défini par :

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

Calculer le score de similarité entre D1 et chacune des deux requêtes suivantes :
Q1 : ($t_1 \wedge t_5$)
Q2 : (($t_1 \wedge t_5$) $\vee$ ($t_8 \wedge t_{10}$))

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 dans un espace à deux dimensions.
En déduire visuellement l'ordre des réponses d'un système vectoriel avec la similarité cosinus.
Valider votre intuition en calculant la correspondance par la formule du cosinus, et en ordonnant les réponses suivant les scores de similarité décroissants.

Exercice 3 – modèle vectoriel

Considérons les documents suivants :
d1 = (1,0,1,0,0,0) d2 = (3,0,2,1,0,0) d3 = (1,2,3,0,1,0)
Considérons les requêtes q1 = (2,0,2,0,0,0) et q2 =(0,0,0,2,0,2).
Avec un tel espace à 6 dimensions, on ne peut plus faire de représentation graphique.

Utiliser comme fonction de correspondance cosinus pour calculer la valeur de pertinence système de ces documents pour chacune des deux requêtes. Les ordonner par pertinence décroissante et donner la liste de réponse pour chaque requête.

Exercice 4 – pondération dans le modèle vectoriel

Un document contient uniquement la phrase « deux un deux». Supposons que chaque mot est dans le vocabulaire d'indexation. Le corpus de documents contient 1 000 documents et le terme "deux" apparaît dans 150 documents et le terme "un" dans 50 documents. Si nous utilisons la pondération tf.idf vue en cours, donner le poids de chacun des termes du document. Faire les calculs avec les deux manières de calculer l'idf vues en cours. Commenter les valeurs obtenues.
Valeurs possiblement utiles: ln(5)=1.609 ; ln(6.67)=1.898 ; ln(10)=2.303 ; ln(20)=2.996; 1/100 = 0.01 ; 1/150 = 0.0067 ; 1/200 = 0.005 ; 1/50 = 0.02 .

Exercice 5 – indexation dans le modèle vectoriel

Considérons les textes suivants :
Document 1 : « Le professeur parle de la recherche d'information textuelle. »
Document 2 : « La recherche d'information est un domaine de recherche qui s'intéresse à des nombreux problèmes. »

Document 3 : « Le modèle vectoriel de recherche d'information est un modèle simple à comprendre. »

1. En considérant un anti-dictionnaire composé des termes :
{à, au, d, de, du, des, elle, elles, est, je, il, ils, le, la, les, lui, qui, son, s, sa, ses, tu, un, une}
représenter l'ensemble des termes d'indexation de chacun des documents ci-dessus.
1' Appliquer une troncature avec les règles suivantes
    1. s → /
    2. 1 → /
    3. e → /

2. Calculer les tf de chacun des termes pour chaque document.
3. Calculer l'idf de chacun des termes présents dans les documents
4. En déduire les vecteurs de chaque document.
5. Calculer les normes de chaque vecteur document.
6. Déduire de la question 4 le tableau du fichier inverse pour ce corpus.

Exercice 6 – pondération dans le modèle vectoriel

Fournir les résultats des requêtes suivantes pour le corpus de l'exercice 5 :
Q1 : recherche d'information textuelle
Q2 : domaine du modèle vectoriel
Commencer par analyser les requêtes comme les documents (muniscule, anti-dictionnaire, troncature), et utiliser une pondération des requêtes tf.idf.

Exercice 7 - bouclage de pertinence

Reprenons les documents de l'exercice 3 : en considérant que les documents d1 et d2 sont pertinents et que le document d3 est non pertinent pour la requête q1, utiliser la formule de Rocchio reformuler la requête et pour l'évaluer, avec alpha=1, beta=0.4 et gamma=0.2

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.
Les règles de réécriture que nous utilisons sont celles vues en cours :
1. s → /
2. ed → /
3. ing → /
4. er →
5. e → /
6. ment → /
7. double consonne et non (*l ou *s ou *z) → la consonne

Prenons 4 documents :
D1 = "computing programs written software development"
D2 = "programming language softwares"
D3 = "computer software program"

D4 = "information retrieval"

Question 1
Considérons initialement le cas sans utilisation de l'algorithme de Porter (donc sans troncature) :

- le vocabulaire T = {computer, computing, development, information, language, program, programming, programs, retrieval, software, softwares, written};
- des pondérations uniquement basées sur le tf;
- la similarité basée sur le cosinus;

donner le résultat d'une requête "programs" de vecteur requête :
Q0 = (0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0).

Question 2
Donner pour ces 4 documents les termes tronqués avec Porter qui les indexent. En déduire le vocabulaire d'indexation de ce corpus.

Question 3
Reprendre la même requête "programs" qu'à la question 1, avec les mêmes pondérations (tf), lui appliquer la troncature, calculer le vecteur requête et la réévaluer.

Exercice 9 – Evaluation de SRI

Nous réalisons ici une analyse d'un système de recherche d'information.
Question 1
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	X		
2	589	X		
3	576			
4	590	X		
5	986			
6	592	X		
7	884			
8	988			
9	578			
10	985			
11	103			
12	591			
13	572	X		
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.
Donner le tableau de résultats normalisé pour cette requête, et en déduire la courbe de rappel/précision.

Question 2
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.

Question 3
En regardant les courbes, que pouvez-vous déduire de la qualité relative du système pour ces deux requêtes?

Question 4
Donner le tableau global 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.