Examen du cours Recherche d'information / web sémantique

ENSEEIHT 3A Info – année 2018-2019

Les documents sont autorisés.

Partie « recherche d'information »
(durée indicative : 1h – barême 10 pts)

Exercice 1 (2)

Répondre aux questions suivantes :

a) Un document ne comportant aucun terme de la requête ne peut pas être pertinent pour cette requête ? vrai/faux Justifier
b) La représentation en sac de mots permet de capturer (représenter) le sens des mots. Vrai/Faux
c) La représentation en Trigrammes permet de capturer la syntaxe. Vrai/Faux
d) La racinisation permet-elle d'améliorer le rappel ou la précision ou les deux ? justifier ?
e) A votre avis, quel modèle de RI (parmi ceux étudiés en cours) serait le plus approprié pour rechercher des *Tweets* ? Justifiez ? On suppose qu'un Tweet est formé d'un message où les mots se répètent rarement.

Exercice 2 (5 pts)

Considérons la collection suivante composée de 3 documents :
D1 = { 4 t1, 2 t5}, D2 = {1 t1, 1 t2, 4 t5, 4 t6}, D3 = {2t2, 3 t3, 5 t4, 6 t5}
Et la requête suivante : Q1 = {2 t1, 2 t2}
Un document (resp. requête) est représenté par une liste de termes pondérés ayant la forme suivante : Dj{w_{ij} t_i}, i=1..6. w_{ij} t_i signifie la fréquence du terme t_i dans D_j. Les poids nuls ne sont pas représentés.

Questions

Donner l'ordre dans lequel seront renvoyés les documents qui répondent à la requête pour les 3 modèles suivants :

1. Le modèle vectoriel utilisant la pondération de type *qqq.ddd=nnn.ltn*
2. Le modèle probabiliste (*Probabilistic Ranking Principle*) (BIR Model)
3. le modèle de langue basé sur l'interpolation de Jelinek Mercer (JM)

Exercice 3 (3pts) :

La table ci dessous présente les documents retrouvés dans une collection de 14 documents, par deux algorithmes de recherche différents Alg1 et Alg2 en réponse à une requête. La valeur **1** de la table indique que l'algorithme a effectivement retrouvé le document spécifié dans la colonne correspondante et la valeur **0** indique que le document n'a pas été pas retrouvé. La dernière ligne « Pert » indique si le document est très pertinent (noté 2), pertinent (noté 1) ou non pertinent (noté 0) pour la requête.

docs	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
Alg1	1	0	0	1	1	1	0	0	1	1	0	0	1	1
Alg2	0	1	1	0	1	1	1	0	0	1	1	0	1	1
Pert	1	1	2	1	0	0	2	1	1	0	0	2	1	1

Questions

1- Calculer la précision moyenne pour chaque algorithme (Alg1 et Alg2)
2- Calculer également la R-Précision (de chaque algo).