Examen du cours Recherche d'information

ENSEEIHT 3A Info – année 2021-2022

Les documents sont autorisés.

(durée indicative : 1h30)

Exercice 1 (5)

1- Donner l'impact des procédures ci-dessous, en termes de rappel et la précision :

a. Utilisation de la lemmatisation des mots ;
b. Utilisation des synonymes ;
c. Utilisation de la position des termes dans les documents;
d. Utilisation des expressions;
e. Utilisation des représentations distribuées (Word Embedding);

2- Répondre par Vrai ou Faux et justifiez votre réponse

a. Un document ne comportant aucun terme de la requête ne peut pas être pertinent pour cette requête?
b. La représentation en sac de mots permet de capturer (représenter) le sens des mots.
c. L'utilisation des représentations distribuées de mots (*Word embedding*) permet de sélectionner des documents pertinents même si ces derniers n'ont aucun terme en commun avec la requête
d. Un système de recherche d'information basé sur un modèle de langue de type bigrammes renvoie deux listes différentes pour les requêtes suivantes : « Information retrieval » et « Retrieval information »
e. L'analyse sémantique latente permet de récupérer des documents pertinents même si ces documents n'ont aucun terme en commun avec la requête

Exercice 2 (6)

Nous disposons d'une collection comportant les 3 documents suivants:
D1 = {2 t1, 3 t2, 4 t5}, D2 = {1 t2, 2 t3, 2 t5, 2 t6}, D3 = { 1t2, 1 t3, 5 t4}
Soit la requête suivante : Q1 = {2 t1, 1 t4}
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

1- Donner le fichier inversé permettant de représenter cette collection de documents
2- Donner l'ordre dans lequel seront renvoyés les documents qui répondent à la requête pour les 3 modèles suivants :
   a) Le modèle vectoriel utilisant la pondération de type *qqq.ddd=nnn.ltn*
   b) Le modèle probabiliste (*Probabilistic Ranking Principle*) (BIR Model)
   c) Le modèle de langue avec interoplation de direcjlt (avec λ=0.5)

Dirichlet

Exercice 3 (4 pts)

Soit $q = q_1, ..., q_m$ une requête, d un document et $P(q_i|d)$ la probabilité du mot q_i dans le modèle de langue de d. On suppose que nous disposons d'une collection de documents comportant au total 8 mots $w1, ..., w8$.

La Table ci-dessous liste pour chaque mot sa probabilité dans le modèle de langue de référence, $P_{ml}(w|REF)$, estimé sur la collection (2ème colonne), la fréquence du terme c(w; d) dans un document (3ème colonne). Les colonnes 4 et 5 représentent les probabilités du terme dans le modèle langue du document d, estimé respectivement selon le maximum de vraisemblance et Dirichlet avec le paramètre μ.

Mots	$P_{ml}(w\|REF)$	c(w, d)	$P_{ml}(w\|d)$	$P_{dir}(w\| d)$
w1	0.3	2		
w2	0.15	1		
w3	0.1	2		0.125
w4	0.1	4		
w5	0.05	1		
w6	0.1	0		
w7	0.1	0		
w8	0.1	0		

1- Remplir la colonne 4, ($P_{ml}(w|d)$), le modèle de langue du document.
2- La colonne 5 représente la probabilité du terme calculée après un lissage de Dirichlet effectuée sur la collection. Seule la probabilité de *w3* est donnée dans le tableau, déduire la valeur de μ? (posez l'équation puis déduire cette valeur)

Exercice 4 (5pts) :

La table ci-dessous montre les documents trouvés par un système de recherche d'information, **S**, en réponse à une requête, parmi les 10 documents d'une collection. Les documents retrouvés sont listés par ordre décroissant de leur pertinence (calculée par un modèle (BM25). La valeur **1** de la table indique que le système a effectivement sélectionné le document spécifié dans la colonne correspondante et la valeur **0** indique que le document n'a pas été pas retrouvé (l'ordre des documents est donc : D1, D4, D5, D6, D9, D10). La dernière ligne « Pert » indique si le document est pertinent (noté **1**) ou non pertinent (noté 0) pour la requête.

docs	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
S	1	0	0	1	1	1	0	0	1	1
Pert	1	0	1	0	1	0	1	0	1	0

Questions

1- Calculer les valeurs de précision et de rappel non interpolées et interpolées.
2- Calculer la précision moyenne de S
3- Calculer également sa R-Précision.