Examen du cours Recherche d'information

ENSEEIHT 3A Info – année 2019-2020

Les documents sont autorisés.

(durée indicative : 1h30)

Exercice 1 (6)

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

a. 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 de mots à la place des mots simples ;
f. utilisation de l'opérateur AND dans le modèle booléen

2- Répondre par Vrai ou Faux et justifiez votre réponse

a. Dans un système de de recherche d'information qui utilise des n-grammes de caractères plutôt que des mots, la racinisation n'est pas nécessaire.
b. L'analyse sémantique latente permet de récupérer des documents pertinents même si ces documents n'ont aucun terme en commun avec la requête
c. La normalisation par la longueur des documents vise à éviter que les documents courts ne soient classés trop haut
d. En général, comme les documents d'un ensemble de résultats sont listés par ordre décroissant de pertinence estimée, la précision diminue à mesure que le rappel augmente

Exercice 2 (6 pts)

On considère un système de recherche d'information basé le modèle de langue mixte (JM) combinant le modèle de document et le modèle de collection. Les probabilités sont estimées en se basant sur la fréquence des termes dans le document (pour le modèle de document) et la fréquence des termes dans la collection (pour le modèle de collection).

Considérons la requête suivante : q (XML, document), dont les ICF (Inverse Collection Frequency, fréquence relative des termes dans la collection) sont : 10/10 000 000 pour le terme "XML" et 10 000/10 000 000 pour le terme "document".

Soit une collection de 4 documents : d1, d2, d3 et d4. On suppose que le document d1 contient 1000 termes dont le terme « XML » qui apparaît une seule fois, et le document d2 contient également 1000 termes dont le terme « document » qui apparaît 1 fois. Les documents d3 et d4 ne contiennent pas ces deux termes.

Questions :

Donner l'ordre dans lequel les 4 documents seront renvoyés en réponse à la requête q

1. dans le cas d'un modèle de langue mixte JM avec lambda=0.5
2. dans le cas du modèle probabiliste BIR
3. dans le cas du modèle vectoriel utilisant la pondération de type *qqq.ddd=nnn.ltn*
4. On suppose que la requête q est pondérée q (XML 3, document 1), donner l'ordre dans lequel les documents seront renvoyés dans le cas du modèle (JM).

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 (4pts) :

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. 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- Représenter graphiquement la courbe de rappel et de précision interpolée de S.
2- Calculer la précision moyenne de S
3- Calculer également sa R-Précision.