2 heures - Tous documents autorisés – *le barème est indicatif*

MIASHS
Recherche d'information 2018-2019 – Session 1
--- Calculatrices autorisées - Ordinateurs interdits ---

Exercice 1 – Généralités sur les modèles et systèmes de recherche d'information (3 pts)

Question 1 : Indiquer si, selon vous, le fait d'utiliser des troncatures sur les mots (comme l'algorithme de Porter) privilégie davantage le rappel ou la précision. (10 lignes maximum)

Question 2 : Décrire l'impact de l'utilisation d'un anti-dictionnaire : sur la taille du vocabulaire, sur la taille du fichier inverse, sur les poids (tf.idf) des termes qui ne sont pas dans l'anti-dictionnaire. (10 lignes maximum)

Question 3 : Supposons que D est un document d'un corpus de documents C. Supposons que l'on retire D du corpus C, et appelons C' ce nouveau corpus. Expliquer quel serait l'impact de ce retrait sur les valeurs d'idf pour : a) les termes qui apparaissaient dans D et dans d'autres documents de C' et; b) les termes qui n'apparaissent pas dans D mais dans un document de C'. (10 lignes maximum)

Exercice 2 : Modèle vectoriel et pondération idf (3 pts)

Question 1 : Selon vous, avec le modèle vectoriel, est-il possible pour des utilisateurs de rechercher des documents qui NE CONTIENNENT PAS un mot-clé donné ? Expliquer votre réponse.

Question 2 : Il a été proposé des moyens de "normaliser" les tf, par exemple avec la formule suivante qui définit une nouvelle valeur de *term frequency* appelé $tf'_{i,j}$ pour un terme j dans un document i, basée sur le $tf_{i,j}$ vu en cours :

$$tf'_{i,j} = 0{,}5 + \frac{0{,}5 * tf_{i,j}}{\max(tf_{i,\circ})}$$

Avec $\max(tf_{i,\circ})$ indiquant la valeur de tf maximale pour les termes du document i. La valeur $tf'_{i,j}$ n'est calculée que pour les termes dans le document. Pour un terme non-présent sa valeur est nulle.

a- Donner et expliquer les bornes (max et min) des valeurs de $tf'_{i,j}$. Quelles sont les différences avec celles de $tf_{i,j}$ classiques ?

Considérons un document D_{10} pour lequel $\max(tf_{10,\circ}) = 20$.
Dans D_{10} le terme t_4 apparaît 10 fois, et le terme t_7 apparaît 15 fois.

b- Donner les valeur de $tf_{10,4}$, $tf'_{10,4}$, $tf_{10,7}$ et $tf'_{10,7}$.
c- Discuter brièvement les différences relatives entre les tf des termes 4 et 7, et les tf ' de ces même termes dans le document D_{10}.

Exercice 3 : Modèle vectoriel (5 pts)

Chacun des **3 documents** suivants a été indexé et est représenté dans le modèle vectoriel de recherche d'information par un vecteur à 5 dimensions, uniquement basé sur les tf des termes :

D_1 = (1 2 1 2 1) D_2 = (4 1 1 1 1) D_3 = (2 4 2 4 2)

2 heures - Tous documents autorisés – *le barème est indicatif*

Question 1 : Considérons une requête Q contenant trois termes qui, une fois traitée, est représentée par le vecteur suivant :

$$\overrightarrow{Q_1} = (1\ 2\ 0\ 0\ 1)$$

Si cette requête est posée sur le corpus de 3 documents dont les vecteurs sont décrits plus haut, indiquer a) les valeurs de correspondance de ces documents en utilisant la fonction de correspondance cosinus vue en cours sans modifier les vecteurs documents et requête, et b) l'ordre des réponses.

Question 2 : Commenter les valeurs obtenues dans la question 1 : expliquer l'ordre des 3 documents, et expliquer particulièrement l'ordre entre D1 et D2, en fonction des poids de termes dans ces deux documents.

Question 3 : a- En reprenant la requête Q de la question 1, et en supposant que le corpus est toujours le même, donner le vecteur requête $\overrightarrow{Q_{idf}}$ si l'on **considère uniquement** une pondération idf comme poids des termes **de la requête** et non plus le tf comme en question 1. Nous rappelons que, les termes qui n'apparaissent pas dans la requête ont toujours un poids nul dans Q_{idf}.
Utiliser la formule d'idf du cours avec un népérien $\log_e$, noté classiquement *ln* (**en utilisant les valeurs (*) fournies à la fin de l'exercice**). Afin de pouvoir faire les calculs, on pose que : le corpus total contient 100 documents (incluant les 3 décrits plus haut), et les df des termes, dans l'ordre des dimensions, sont les suivants : 10 50 100 50 95.

b- Expliquer **en détail** la modification des poids entre le vecteur requête de la question 1 de l'exercice 3 et ce que vous obtenez ici.

Question 4 : Reprendre la question 1 de l'exercice 3 avec le vecteur requête $\overrightarrow{Q_{idf}}$ que vous venez de calculer en gardant les vecteurs documents initiaux.

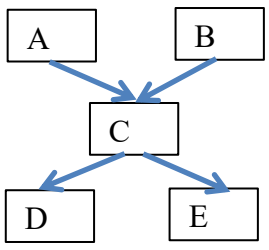
(*) Valeurs fournies, à utiliser dans l'exercice 3 :

$\log_e(0,2) = -1,609$	$\log_e(0,4) = -0,916$	$\log_e(0,5) = -0,693$	$\log_e(0,7) = -0,357$	$\log_e(1) = 0$
$\log_e(1,053) = 0,051$	$\log_e(1,25) = 0,223$	$\log_e(1,429) = 0,357$	$\log_e(1,67) = 0,511$	$\log_e(2) = 0,693$
$\log_e(2,25) = 0,801$	$\log_e(2,5) = 0,916$	$\log_e(2,75) = 1,012$	$\log_e(3) = 1,099$	$\log_e(4) = 1,386$
$\log_e(5) = 1,609$	$\log_e(6) = 1,792$	$\log_e(7) = 1,946$	$\log_e(10) = 2,303$	$\log_e(20) = 2,996$

Exercice 4 : Pagerank (6 pts)

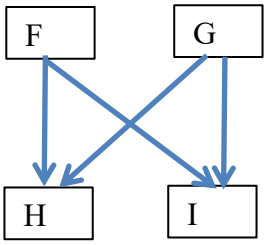
Dans cet exercice, on dira qu'il **y a convergence si entre deux boucles successives il n'y a pas de différence de valeur supérieure à 0.02 pour une même page entre deux itérations.** Dans tous les cas, **on ne fera pas plus de 6 boucles** en plus de l'initialisation**.**

Question 1 : Donner les formules de PR pour chaque page (en utilisant la valeur de p du cours), puis calculer les valeurs de PR jusqu'à convergence de la configuration suivante (en utilisant le valeur d'initialisation classique du cours) :



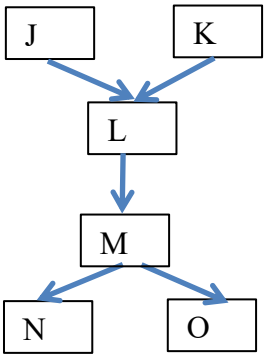
Question 2 : Discuter les valeurs obtenues en question 1 (les valeurs de PR des pages, l'ordre des valeurs entre les pages).

Question 3 : Reprendre la question 1 avec la configuration suivante :



Question 4 : Discuter les valeurs obtenues en question 3 (les valeurs de PR des pages, l'ordre des valeurs entre les pages). Comparer aussi les trois valeurs PR(C), PR(H) et PR(I).

Question 5 : Reprendre la question 1 avec la configuration suivante :



Question 6 : Discuter les valeurs obtenues en question 5 (les valeurs de PR des pages, l'ordre des valeurs entre les pages). Comparer en particulier les valeurs PR(L) et PR(M), et comparer PR(H) de la question précédente et PR(N)

Question 7 : Tentez d'expliquer brièvement pourquoi les valeurs des 3 configurations des questions 1, 3 et 5 ne convergent pas après le même nombre d'itération.

Question 8 : Reprendre la question 1 en utilisant changeant la valeur de d, maintenant égale à 0,2 (en gardant la même valeur d'initialisation qu'en cours). Commentez les différences obtenues avec la réponse à la question 1 (par exemple : le nombre d'itérations avant convergence, l'ordre dans les valeurs de PR pour une même configuration, les différences entre les valeurs pour une même page pour les 2 configurations après convergence, et les différences relatives entre des pages pour chaque configuration).

Exercice 5 : Recherche d'information réseaux sociaux (3 pt)

Considérons un réseau social : un réseau social est composé d'utilisateurs et de messages (par exemple des tweets dans le cas du réseau tweeter) écrits par les utilisateurs.
- Décrire comment intégrer des éléments de popularité (par exemple Pagerank) entre des utilisateurs ou entre des messages lors de la recherche de tweets ?

********** Fin **********