Université d'Alger 1

Faculté des Sciences
Département Informatique

Examen L3- Recherche D'Information

20 Mai 2024

Nom : Prénom : Groupe : Matricule :		/ 20

Partie cours (QCM) (12 points)

Cocher la ou les bonnes réponses : (0.5* 21 pts)

1. Quel est l'objectif principal d'un système de recherche d'information ?
   A. Indexer toutes les pages web disponibles
   B. Sélectionner des informations pertinentes répondant aux besoins des utilisateurs
   C. Analyser les habitudes de navigation des utilisateurs
   D. Proposer des recommandations de produits
2. Quel type de pertinence décrit la relation entre le sujet exprimé dans la requête et le sujet couvert dans le document ?
   A. Pertinence cognitive
   B. Pertinence situationnelle
   C. Pertinence au sujet (topique)
   D. Pertinence structurelle
3. Quelle tâche de la recherche d'information se concentre sur la recommandation de contenus pertinents à un utilisateur ?
   A. Recherche adhoc
   B. Filtrage d'information
   C. Fouille de textes
   D. Clustering
4. Quel est l'objectif principal de l'indexation dans un système de recherche d'information(SRI) ?
   A. Améliorer la vitesse de recherche
   B. Réduire la taille des documents
   C. Définir un ensemble d'éléments clés pour caractériser le contenu d'un document
   D. Classer les documents par ordre alphabétique
5. Quels sont les deux éléments qui doivent être stockés pour chaque terme dans un index inversé ?
   A. Fréquence locale du terme et longueur du document
   B. Longueur du document et position du terme
   C. Pointeur vers la liste de postings et fréquence globale du terme

D. Fréquence du document et pointeur vers la liste de postings

6. Quels sont les trois principaux composants du processus d'indexation ?
   A. Prétraiter les documents, créer une base de données, analyser les résultats
   B. Prétraiter les documents, constituer un dictionnaire de termes, construire une structure de données associant termes et documents
   C. Scanner les documents, convertir en texte, sauvegarder
   D. Analyser les documents, créer des métadonnées, archiver les documents
7. Qu'est-ce qu'une stop-liste ?
   A. Une liste de termes à utiliser pour indexer les documents
   B. Une liste de documents fréquemment consultés
   C. Une liste de mots vides à ignorer lors de l'indexation
   D. Une liste de requêtes fréquentes des utilisateurs
8. La normalisation linguistique peut inclure :
   A. L'ajout de termes aléatoires aux documents
   B. La suppression des ponctuations
   C. La lemmatisation ou le stemming des mots
   D. La conversion des mots en miniscules
9. Pourquoi ne peut-on pas garder tous les mots les plus fréquents selon la loi de Zipf ?
   A. Parce qu'ils sont trop longs
   B. Parce qu'ils n'aident pas à discriminer les documents
   C. Parce qu'ils sont en plusieurs langues
   D. Parce qu'ils sont trop complexes à analyser
10. Quelle méthode permet de convertir un texte en un ensemble de termes pour l'indexation ?
    A. Normalisation linguistique
    B. Segmentation
    C. Prétraitement
    D. Constitution d'un dictionnaire
11. Quelle formule est couramment utilisée pour pondérer les termes dans les SRI ?
    A. Le produit scalaire
    B. La distance euclidienne
    C. La distance Cosine
    D. Le TF-IDF
12. Dans la pondération TF-IDF, quel est l'effet de la partie "IDF"?
    A. Elle augmente le poids des termes fréquents dans le document
    B. Elle diminue le poids des termes fréquents dans l'ensemble des documents
    C. Elle n'a aucun effet sur le poids des termes
    D. Elle augmente le poids des termes rares dans le document
13. Quels types de termes ont généralement un poids élevé avec la pondération TF-IDF ?
    A. Les termes très fréquents dans tous les documents
    B. Les termes rares dans la collection de documents
    C. Les termes de longueur moyenne
    D. Les termes les plus courts
14. Quel est un inconvénient majeur du modèle booléen ?
    A. Il est difficile à comprendre
    B. Il nécessite une grande puissance de calcul
    C. Il ne prend pas en compte le poids des termes
    D. Il ne peut pas gérer les requêtes complexes
15. Avec le modèle booléen, une solution pour augmenter le rappel est-elle de :
    A. Utiliser l'opérateur ET
    B. Utiliser l'opérateur OU

C. utiliser l'opérateur NON entre les termes de la requête ?

Expliquer brièvement votre réponse sur un exemple.

- Utilisation de l'opérateur et : chat et chien

Cette requête retourne uniquement les documents qui contiennent à la fois "chat" et "chien". Cela restreint les résultats et peut réduire le rappel, car il y a moins de documents qui contiennent les deux termes simultanément.

(1.5 pts)

- Utilisation de l'opérateur ou : chat ou chien

Cette requête retourne tous les documents qui contiennent soit "chat", soit "chien", soit les deux. Cela augmente le rappel car il inclut tous les documents contenant l'un des termes ou les deux, couvrant ainsi une plus grande part de la collection de documents.

- Utilisation de l'opérateur non : chat non chien

Cette requête retourne les documents qui contiennent "chat" mais excluent ceux qui contiennent "chien". Cela n'aide pas à augmenter le rappel et est plutôt utilisé pour affiner les résultats en excluant des documents.

16. Comment les documents sont-ils représentés dans le modèle vectoriel ?
    A. Par des listes de termes
    B. Par des graphes de concepts
    C. Par des arbres de décision
    D. Par des vecteurs de pondération de termes
17. Dans le modèle booléen flou, quel concept est introduit pour évaluer la pertinence des documents ?
    A. Les degrés d'appartenance
    B. La fréquence des termes
    C. La similarité euclidienne
    D. Les opérateurs conditionnels
18. Dans le modèle P-norm, que représente le paramètre "p" ?
    A. La probabilité d'apparition des termes
    B. La puissance de normalisation des termes
    C. Le degré de flou dans la recherche
    D. Un paramètre pour ajuster le compromis entre les requêtes booléennes strictes et les requêtes vectorielles
19. Quel indicateur mesure la proportion de documents pertinents parmi ceux qui ont été récupérés ?
    A. Précision
    B. Rappel
    C. F-mesure
    D. Taux de faux positifs
20. Quelle mesure évalue la capacité du système à récupérer tous les documents pertinents dans la collection ?
    A. Précision
    B. Rappel
    C. F-mesure
    D. F-mesure
21. Quelle mesure est utilisée pour évaluer la pertinence d'un classement de documents en prenant en compte leur position dans les résultats ?
    A. Précision
    B. Rappel
    C. Taux d'erreur
    D. DCG (Discounted Cumulative Gain)

Exercice (8 points)

Considérez la collection de documents suivante :

d1 = " Big cats are nice and funny "
d2 = " Small dogs are better than big dogs "
d3 = " Small cats are afraid of small dogs "
d4 = " Big cats are not afraid of small dogs "
d5 = " Funny cats are not afraid of small dogs "

1) Calculez les termes d'indexation pour chaque document.

*d*1 = " Big|cats|are|nice|and|funny"
*d*2 = " Small|dogs|are|better|than|big|dogs"
*d*3 = " Small|cats|are|afraid|of|small|dogs"
*d*4 = " Big|cats|are|not|afraid|of|small|dogs"
*d*5 = " Funny|cats|are|not|afraid|of|small|dogs"

(1.25 pts)

2) Normalisez les termes d'indexation par rapport aux pluriels et à la casse.

*d*1 = " big|cat|is|nice|and|funny"
*d*2 = " small|dog|is|better|than|big|dog"
*d*3 = " small|cat|is|afraid|of|small|dog"
*d*4 = " big|cat|is|not|afraid|of|small|dog"
*d*5 = " funny|cat|is|not|afraid|of|small|dog"

(1.25 pts)

3) Calculez le dictionnaire relatif à la collection de documents. (0.5 pts)

Dictionary = {big,cat,is,nice,and,funny,small,dog,better,than,afraid,of,not}

4) À partir de la collection de documents ci-dessus, construisez la table d'incidence documents -termes comme requis par le mode booléen.

	big	cat	is	nice	and	funny	small	dog	better	than	afraid	of	not
d_1	1	1	1	1	1	1	0	0	0	0	0	0	0
d_2	1	0	1	0	0	0	1	1	1	1	0	0	0
d_3	0	1	1	0	0	0	1	1	0	0	1	1	0
d_4	1	1	1	0	0	0	1	1	0	0	1	1	1
d_5	0	1	1	0	0	1	1	1	0	0	1	1	1

(2 pts)

5) À partir de la collection de documents de l'exercice et en considérant un modèle booléen.

(a) Répondez à la requête q1 = funny AND dog

R_{funny} ={d1, d5}, R_{dog} ={d2,d3,d4, d5}
q1→ $R_{funny} \cap R_{dog}$={d5}

(1 pts)

(b) Répondez à la requête q2 = nice OR dog

R_{nice} ={d1}, R_{dog} ={d2,d3,d4, d5}
q2→ $R_{funny} \cup R_{dog}$={ d, d2,d3,d4, d5}

(1 pts)

(c) Répondez à la requête q3 = big AND dog AND NOT funny

R_{big} ={d1, d2, d4}, R_{dog} ={d2,d3,d4, d5}, R_{funny} ={d1, d5}
q3→ ($R_{big} \cap R_{dog}$) ∩ R_{funny} = { d2 ,d4 } ∩ { d2 ,d3, d4 } ={ d2 ,d4 }

(1 pts)