TD Recherche d'information n°2
Cours de Bases de données avancées

1 Premiers pas vers la recherche plein texte

Voici quelques documents textuels très courts.

- A : "Le loup est dans la bergerie."
- B : "Les moutons sont dans la bergerie."
- C : "Un loup a mangé un mouton, les autres loups sont restés dans la bergerie."
- D : "Il y a trois moutons dans le pré, et un mouton dans la gueule du loup."

Prenons le vocabulaire suivant : "loup", "mouton", "bergerie", "pré", "gueule".

1.1) Construisez la fonction qui associe chaque document à un vecteur dans $\{0,1\}^5$. Vous pouvez représenter cette fonction sous forme d'une matrice d'incidence.

1.2) Calculer le score de chaque document pour les recherches suivantes, et en déduire le classsement :

q1 : "loup et pré"

q2 : "loup et mouton"

q3 : "bergerie"

q4 : "gueule du loup"

Correction : Les vecteurs des documents :

- $v(A) = [1,0,1,0,0]$
- $v(B) = [0,1,1,0,0]$
- $v(C) = [1,1,1,0,0]$
- $v(D) = [1,1,0,1,1]$

Les vecteurs des documents :

- $v(q_1) = [1,0,0,1,0]$
- $v(q_2) = [1,1,0,0,0]$
- $v(q_3) = [0,0,1,0,0]$
- $v(q_4) = [1,0,0,0,1]$

En ce qui concerne les requêtes :

- $E(q_1,A) = \sqrt{2}\,; E(q_1,B) = \sqrt{4}\,; E(q_1,C) = \sqrt{3}\,; E(q_1,D) = \sqrt{2}$
  A et D sont les plus pertinents, suivi de C et enfin B. Notez que A de contient pas le mot "pré". Pourquoi obtient-il le même score que D ?
- $E(q_2,A) = \sqrt{2}\,; E(q_2,B) = \sqrt{2}\,; E(q_2,C) = \sqrt{1}\,; E(q_2,D) = \sqrt{2}$
- $E(q_3,A) = \sqrt{1}\,; E(q_3,B) = \sqrt{1}\,; E(q_3,C) = \sqrt{2}\,; E(q_3,D) = \sqrt{5}$
- $E(q_4,A) = \sqrt{2}\,; E(q_4,B) = \sqrt{4}\,; E(q_4,C) = \sqrt{3}\,; E(q_4,D) = \sqrt{2}$
  C'est donc D qui l'emporte, mais à égalité avec A, ce ne correspond pas vraiment à l'intuition.

2 À propos de la fonction de distance

Supposons que l'on prenne comme distance non pas la distance Euclidienne mais le carré de cette distance. Est-ce que cela change le classement ? Qu'est-ce que cela vous inspire ?

Correction : On peut effectuer un calcul *simplifié* de la distance, tant que l'ordre est respecté. Ce qui nous intéresse en fait, ce n'est pas le score proprement dit, mais l'ordre des scores.

3 Critique de la distance euclidienne

La distance que nous avons utilisée mesure la **différence** entre la requête et un document, par comparaison des termes un à un. Cela induit des inconvénients qu'il est assez facile de mettre en évidence.

Supposons maintenant que le vocabulaire a une taille très grande. On fait une recherche avec 1 mot-clé.

Questions :

3.1) Quel est le score pour un document qui ne contient 99 termes et pas ce mot-clé ?

3.2) Quel est le score pour un document qui contient 101 termes **et** le mot-clé ?

Conclusion ? Le classement obtenu sera-t-il satisfaisant ? Trouvez un cas où un document est bien classé même s'il ne contient pas le mot-clé !

Correction : Distance de 100 dans le premier cas ; de 100 dans le second également. Ils seront classés au même niveau, ce qui ne va pas du tout ! Il suffit de prendre un document avec 50 termes : il sera mieux classé que n'importe quel document de 100 termes contenant ou non le mot-clé.

4 Critique de l'hypothèse d'uniformité des termes

Enfin, dans notre approche très simplifiée, tous les termes ont la même importance. Calculez le classement pour la requête :

q5 : "bergerie et gueule"

Tentez d'expliquer le résultat. Est-il satisfaisant ? Quel est le biais (pensez au raisonnement sur la longueur du document dans l'exercice précédent).

Correction : Vecteur de la requête : $v(q_5) = [0,0,1,0,1]$.
Calcul des classements : $E(q_5, A) = \sqrt{2}\,; E(q_5, B) = \sqrt{2}\,; E(q_5, C) = \sqrt{3}\,; E(q_2, D) = \sqrt{4}$.
Tous les documents sont mieux classés que D, alors que "gueule" est un terme plus discriminant.

5 Calculons des tf / idf et des classements

La table ci-dessous montre une matrice d'incidence avec une ligne par terme, une colonne par document, l'idf de chaque terme et le tf dans chaque cellule.

terme	**d1**	**d2**	**d3**
voiture (1,65)	27	4	24
marais (2,08)	3	33	0
serpent (1,62)	0	33	29
baleine (1,05)	14	0	17

Quelques calculs :

5.1) Normaliser les vecteurs des tf pour chaque document.

5.2) Normaliser les vecteurs des tf pour chaque document, mais sur le sous-espace ("voiture", "baleine").

5.3) Normaliser les vecteurs des tf.idf pour chaque document.

Calculer le classement des requêtes suivantes, **sans tenir compte de l'idf** (donc, seul le tf entre en compte). Interprétez le résultat.

- voiture
- baleine
- voiture et baleine.
- voiture et baleine et marais et serpent

Correction : Dans les deux premier cas, on pourrait croire qu'il suffit de prendre le classement des tf du terme concerné, sans se lancer dans des calculs compliqués. Erreur ! Ce qui compte ce n'est pas la fréquence d'un terme, mais sa proportion par rapport aux autres.

— “voiture” : la requête est (1, 0, 0, 0) qui est un vecteur normalisé.

  — Pour d1, le cosinus vaut : $\frac{27}{30,56} = 0,88$
  — Pour d2, le cosinus vaut : $\frac{4}{46,84} = 0,085$
  — Pour d3, le cosinus vaut : $\frac{24}{41,30} = 0,58$

  Le classement est d1, d3, d2. Interprétation : on note que d1 et d3 parlent de voiture et de baleine en proportions à peu près équivalentes. Mais d1 ne parle **que** de voiture et de baleine, alors que d3 parle **aussi** de serpent, d’où la différence de classement.

— “baleine” : calcul identique, que vous devriez savoir faire.

— “voiture et baleine’‘ : là il faut se lancer dans le calcul du cosinus. Remarquons d’abord que les coefficients de la requête sont (1, 1) et sa norme $\sqrt{1+1} = 1,41$.

  — Pour d1, le cosinus vaut : $\frac{27+14}{1,41 \times 30,56} = 0,95$
  — Pour d2, le cosinus vaut : $\frac{4}{1,41 \times 46,84} = 0,06$
  — Pour d3, le cosinus vaut : $\frac{24+17}{1,41 \times 41,30} = 0,70$

  L’ordre est donc d1, d3, d2. Le document d3 présente un meilleur équilibre entre les composantes “voiture” et “baleine”, mais, contrairement à d1, il a une autre composante forte pour “serpent” ce qui diminue sa similarité.

— “voiture et baleine et marais et serpent”

  On reprend les calculs de la même manière, en partant des normes calculés précédemment. La norme de la requête est 2 (mais on remarque que l’on pourrait l’ignorer pour le classement).

  — Pour d1, le cosinus vaut : $\frac{27+3+14}{2 \times 30,56} = 0,72$
  — Pour d2, le cosinus vaut : $\frac{4+33+33}{2 \times 46,84} = 0,74$
  — Pour d3, le cosinus vaut : $\frac{24+29+17}{2 \times 41,30} = 0,84$

  d3 l’emporte car (intuitivement) il présente un meilleur équilibre entre les termes que les autres documents (la requête elle-même a la caractéristique d’être parfaitement équilibrée sur les termes du vocabulaire).
  Et si au lieu de prendre la norme complète on ne prenait que celle du sous-espace (“voiture“, “baleine”) ? À quoi correspondrait un tel calcul ?

6 Pesons le loup, le mouton et la bergerie

Nous reprenons nos documents de l’exercice 1.

— Donnez, pour chaque document, le tf de chaque terme.
— Donnez les idf des termes (ne pas prendre le logarithme, pour simplifier).
— En déduire la matrice d’incidence montrant l’idf pour chaque terme, le nombre de termes pour chaque document, et le tf pour chaque cellule.

Correction : Fréquence des termes par document (tf) :

— Document A : loup (1), bergerie (1)
— Document B : mouton (1), bergerie (1)
— Document C : loup (2), mouton (1), bergerie (1)
— Document D : loup (1), mouton (2), pré (1), gueule (1)

Fréquence inverse des termes dans les documents (idf) : loup (4/3), mouton (4/3), bergerie (4/3), pré (4), gueule (4).
Matrice d’incidence

	loup (4/3)	mouton (4/3)	bergerie (4/3)	pré (4)	gueule (4)
A	1	0	1	0	0
B	0	1	1	0	0
C	2	1	1	0	0
D	1	2	0	1	1

7 Interrogeons et classons

Reprendre les requêtes de l'exercice 1.

- "loup et pré"
- "loup et mouton"
- "bergerie"
- "gueule du loup"

7.1) Calculer le classement avec la distance cosinus, en ne prenant en compte que le vecteur des tf, comme dans l'exercice 5.

8 Comparons les loups et les moutons

Reprenez une nouvelle fois les documents de l'exercice 1. Vous devriez avoir la matrice des tf.idf calculée dans l'exercice 6.

8.1) Classez les documents B, C, D par similarité cosinus décroissante avec A ;

8.2) Calculez la similarité cosinus entre chaque paire de document ; peut-on identifier 2 groupes évidents ?