TD Recherche d'information
Tri des données

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. 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.
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"

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 ?

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 :

1. Quel est le score pour un document qui ne contient 99 termes et pas ce mot-clé ?
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 !

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).

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 :

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

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.

7 Interrogeons et classons

Reprendre les requêtes de l'exercice 1.

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

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.

1. Classez les documents B, C, D par similarité cosinus décroissante avec A ;
2. Calculez la similarité cosinus entre chaque paire de document ; peut-on identifier 2 groupes évidents ?