Travaux dirigés sur l'indexation des documents

Exercice 1 Soit un document d contenant la phrase (et que la phrase) "Il fait beau ce soir.". La collection de documents associée contient 10 000 documents et chaque mot de d a une fréquence documentaire df égale à 1 000. Quel est le poids de chaque terme selon le critère $tf \times idf$ en utilisant la mesure de pondération globale de termes idf basée sur :

1. La mesure simple, et ensuite,
2. La mesure basée sur le logarithme.

Exercice 2 Supposons que nous nous intéressons aux trois termes : "domaine", "théoriques", "informatique" du document suivant :

L'informatique est un domaine vaste. Comme tout domaine, elle se base sur des fondements théoriques. Les études théoriques se fondent sur des preuves. Entre informatique opérationnelle et informatique décisionnelle, le débat est aujourd'hui d'actualité. Un domaine d'application de l'informatique est la recherche d'information.

Sachant que :

1. la taille du corpus est 64,
2. le nombre de document contenant le premier terme est 16,
3. le nombre de document contenant le second terme est 8,
4. le nombre de document contenant le troisième terme est 4,

déterminer les poids de ces termes selon la formule $tf \times idf$ en utilisant la mesure de pondération globale de termes (idf) basée sur :

1. La mesure simple, et ensuite,
2. La mesure basée sur le logarithme.

Exercice 3 Soit le corpus contenant les documents suivants :

- Doc_1 : La forêt est pleine d'animaux.
- Doc_2 : Les animaux sauvages sont dangereux.
- Doc_3 : La forêt est pleine d'arbres.
- Doc_4 : L'animal le plus puissant est le lion.
- Doc_5 : Beaucoup de naturalistes s'intéressent aux lions et aux tigres.
- Doc_6 : Il y a tout un processus permettant de préserver les espèces en voie de disparition.
- Doc_7 : Les humains ont généralement peur des animaux considérés dangereux.

Il est à noter que lors de l'étape de l'extraction les termes d'indexation sont extraits sous forme de racines. Par ailleurs, un thésaurus permet d'indiquer que le terme espèce et le terme animal sont considérés comme étant synonymes pour le cas de ce corpus. Il en est de même pour le terme humain et le terme naturaliste, d'une part, et le terme sauvage et le terme dangereux d'autre part.

1. Donner les matrices de présence, de fréquence et de poids associées aux termes retenus suivants : forêt, animal, lion, arbre, sauvage, humain. La formule de poids considérée est :

$$\text{poids}(t,d) = \text{tf}(t,d) \times log(\frac{|D|}{df(t,D)})$$

2. En déduire des matrices déterminées dans la question précédente, la matrice du poids normalisée sachant que l'on considère la fréquence maximale d'un terme dans un document comme moyen de normalisation.
3. Reprendre la question 1 en considérant parmi les termes qui y sont indiqués seulement ceux qui ne seront pas élagué puisque leur rang associé est compris entre :
   - Seuil minimal : 2
   - Seuil maximal : 3

   Il est à noter que les termes de même fréquence ont le même rang.

Exercice 4 Soit la collection de documents suivante (Extraite du site Evene)

- Doc_1 : Abandonnez celui qui s'abandonne.
- Doc_2 : Le bonheur est l'absence des peines, comme la santé est l'absence des maladies.
- Doc_3 : Une danse est un poème.
- Doc_4 : Il faut toujours avoir les mêmes égards pour ses amis, qu'ils soient présents ou absents.
- Doc_5 : Comme un pays laissé à l'abandon
- Doc_6 : Absent le chat, les souris dansent.

Lors de l'étape d'extraction, nous considérons seulement les mots suivants regroupés par catégories :

(Abandonnez ; abandon ; abandonne), (absence ;absent ;absents), (danse ;dansent)

1. En supposant qu'une étape de normalisation des termes moyennant la lemmatisation est effectuée, en déduire les termes retenus ainsi que leurs fréquences dans tout le corpus.
2. En supposant qu'une étape de normalisation des termes moyennant la troncature est effectuée, en déduire les termes retenus ainsi que leurs fréquences dans tout le corpus. La troncature sera effectuée de la manière suivante :
   - retenir les 7 premières lettres pour la première catégorie de termes
   - retenir les 5 premières lettres pour la deuxième et la troisième catégorie de termes
3. Discuter l'impact des deux techniques de normalisation citées précédemment sur la recherche d'information, en vous appuyant sur un exemple précis basé sur la collection. Votre réponse doit comparer le résultat obtenu selon les cas suivants :
   - Sans normalisation/ avec normalisation
   - Troncature/lemmatisation

Exercice 5 Soit une collection constituée de 3 documents : d_1, d_2 et d_3. La matrice de fréquence de 3 termes notés t_1, t_2 et t_3 dans cette collection est donnée comme suit :

$$\begin{pmatrix} & d_1 & d_2 & d_3 \\ t_1 & 2 & 2 & 0 \\ t_2 & 3 & 2 & 4 \\ t_3 & 0 & 7 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

1. En déduire la matrice de présence associée.