TD - Stockage

Exercice 1 – Pondération

Un document d contient uniquement la phrase « deux un deux». Supposons que chaque mot est dans le vocabulaire d'indexation. Le corpus de documents contient 1 000 documents et le terme "deux" apparaît dans 150 documents et le terme "un" dans 50 documents. Si nous utilisons la pondération telle que $ptf_{ti,d}=tft_{i,d}$; pdf_{ti} est à base de logs népériens et la normalisation $n_d=1$, donner le poids de chacun des termes du document. Commenter les valeurs obtenues.
Valeurs possiblement utiles: ln(5)=1,61 ; ln(6.67)=1.90 ; ln(10)=2.30 ; ln(20)=3.00 .

Exercice 2 – Indexation vectorielle

Considérons les textes suivants :
Document 1 : « Le professeur parle de la recherche d'information textuelle »
Document 2 : « La recherche d'information est un domaine de recherche qui s'intéresse à des nombreux problèmes »
Document 3 : « Le modèle vectoriel de recherche d'information est un modèle simple à comprendre »

1. En considérant un anti-dictionnaire composé des termes :
{à, au, d', de, du, des, elle, elles, est, je, il, ils, le, la, les, lui, qui, son, s', sa, ses, tu, un, une}
représenter l'ensemble des termes d'indexation de chacun des documents ci-dessus. Attention aux majuscules !
2. Quel vocabulaire V est associé à cette collection dans le cas de l'utilisation de la troncature à base de 3 règles :
    1. s → /
    2. e → /
    3. ll → l
3. Calculer les pft à base de tf de chacun de ces termes pour chaque document.
4. Calculer les pdf à base d'idf de chacun des termes présents dans les documents
5. Donner les représentations vectorielles des documents sans normalisation.
6. En déduire le tableau de l'index inversé pour ce corpus.
7. Calculer les normes de chaque vecteur document.