USTOMB -2025- Faculté MI Département Informatique

Recherche d'Informations L3 (ISIL)

TD2

Exercice 1

La loi de Zipf indique que dans un corpus de textes, la fréquence d'un mot est inversement proportionnelle à son rang dans la liste des mots triés par fréquence.
Soit le texte suivant composé de trois documents :

- D1 : "La recherche d'informations est un domaine fondamental de l'informatique."
- D2 : "Les moteurs de recherche utilisent des algorithmes complexes."
- D3 : "Un bon algorithme de recherche améliore la pertinence des résultats."

1. Construisez la liste des termes uniques présents dans ce corpus.
2. Calculez la fréquence d'apparition de chaque terme.
3. Vérifiez si ces fréquences suivent une distribution en accord avec la loi de Zipf en ordonnant les termes par fréquence.

   La formule de Zipf est :

   $$f(r) = \frac{C}{r^s}$$

   où :

   - C est une constante de normalisation (égale à la fréquence du premier mot dans notre cas, ici $C = 3$).
   - s est un coefficient (prenons $s = 1$, valeur classique).
   - r est le rang du mot.
4. Tracez un graphique de fréquence des termes en fonction du rang et commentez le résultat.

Exercice 2

Soit le corpus suivant où t_i représente le poids du terme *i* dans d_i.

	t_1	t_2	t_3	t_4	t_5
d1:	6	2	3	6	2
d2:	6	1	2	0	2
d3:	6	5	1	0	0

Appliquez « l'approche basée sur la discrimination » pour déterminer les termes les plus discriminants dans la phase 3 de l'indexation.

Exercice 3

Pondération des termes selon TF/IDF
Soit le tableau ci-dessous représentant le fichier inverse.

Terme	**D1**	**D2**	**D3**	**D4**	**D5**
Algo		1			1
Informa		1			1
Programm	3	2	2		1
lang	1		1	1	
fonct			1	1	1
const	1				

Calculer les poids $W(t_i, d_j)$ du terme t_i dans le document d_j selon *TF* ∗ *IDF* donnée selon les formules suivantes :

$$TF = \frac{freq(t_i, d_j)}{\sum_{\forall t' \in d_j} freq(t', d_j)}$$

$$IDF = \log(N/Nt)$$