CY Cergy Paris Université
Master 2 Recherche SIC

Intégration de données

Travaux dirigés : recherche textuelle

A. Soit un corpus de documents texte indexés par un moteur de recherche utilisant la formule de similarité vue en cours. Pour une requête *q* composée de termes *t* et un document *d*:

$$similarité(q,d) = coord(q,d) \times queryNorm(q) \times norm(d) \times \sum_{t \in q} (tf(t,d) \times idf(t)^2)$$

- $tf(t,d)$ = $\sqrt{(freq(t,d))}$ (où *freq(t,d)* = nb apparitions *t* dans *d*)
- $idf(t)$ = 1+ log ($numDocs$/($docFreq(t)$+1)) (où *numDocs* = nb documents du corpus et *docFreq(t)* = nb documents où *t* apparaît)
- $norm(d)$ = 1 / $\sqrt{(numTerms(d))}$ (où *numTerms(d)* = nb termes dans *d*)
- $coord(q,d)$ = ($numTerms(q \cap d)$) / ($numTerms(q)$)
- $queryNorm(q)$ = 1 / $\sqrt{(\sum_{t \in q} idf(t)^2)}$

On considère que le corpus contient 4000 documents. La requête *q* contient les termes 'crise', qui apparaît dans 39 documents et 'dette', dans 3 documents. On considère les documents suivants du corpus :

- *d1* : contient 100 termes, dont 'crise' apparait 4 fois et 'dette' une fois.
- *d2* : contient 400 termes, dont 'crise' apparait une fois et 'dette' une fois.
- *d3* : contient 36 termes, dont 'crise' n'apparait pas et 'dette' apparait 4 fois.
- *d4* : contient 64 termes, dont 'crise' apparait 9 fois et 'dette' n'apparait pas.

Calculez les similarités des quatre documents par rapport à la requête.

B. Soit le graphe de liens entre documents suivant. En considérant *d*=0,2, la probabilité de sauter aléatoirement sur une page, calculez les deux premières itérations de l'algorithme PageRank pour le calcul de l'importance des documents A, B, C et D, en utilisant la formule : $PR^{(k+1)} = d\ U + (1-d)\ L^T * PR^{(k)}$