Recherche Documentaire

Jean-Baptiste Dalle & Romain Gaborieau

23 mars 2015

Corpus

- Un document par fichier
- Document au format XML
- Prise en compte uniquement de la partie "corps du texte"

- Plusieurs modèles possibles
- Choix du modèle vectoriel

"caesar" $\rightarrow$ D1 $\{3, 56\} \rightarrow$ D3 $\{7\}$

"world" $\rightarrow$ D6 $\{1, 5, 6\} \rightarrow$ D3 $\{8\} \rightarrow$ D8 $\{4\}$

- Suppression des caractères spéciaux
- Suppression des mots non spécifiques
- Réduction des mots à leur racine

Performances

L'indexation peut prendre un certain temps. Dans le cadre du projet, il n'est lancé qu'au démarrage de l'application ou sur demande.

Exemple

caesar is the king *and* dog *not* world

- Une requête est considérée comme un document à part entière
- Découpage en sous-requêtes
- Le même procédé de racinisation et de suppression des stopwords est réalisé sur chaque sous-requête

Exemple

caesar is the king *and* dog *not* world

Découpage

caesar ~~is the~~ king
and dog
not world

Exemple

caesar is the king *not* rom*

- L'indexation associe toutes les variantes non stemmées à leur racine
- Utilise les *regex* pour matcher le pattern
- *rom** peut correspondre à *roma* comme à *roman*

- Calcul du TF d'un mot dans un document : $\log_{10}(\frac{nbOccurences}{nbMot})$
- Calcul de l'IDF d'un mot dans le corpus : $\log_{10}(\frac{tailleDuCorpus}{nbOccurences})$
- Le TF-IDF d'un mot dans un document est alors calculé à partir de ces données

- Représentation des documents sous forme de vecteur
- Calcul de la proximité grâce au cosinus de Salton
- Plus la réponse est proche de 0, plus le document est pertinent

Les extensions

- Une correction orthographique est proposée
- Affichage dans la console
- Amélioration en sélectionnant automatiquement une correction ?