Lic. Info. 2ème année — Université d'Orléans
SOL3IF02 - POO — Année 2013-2014

Feuille d'exercices n° ? + ...?

Exercice 1. On souhaite concevoir un moteur de recherche destiné à la recherche efficace de documents pertinents parmi un corpus. On s'intéresse plus particulièrement à la réalisation du coeur de l'application : l'index inversé. En effet, cette structure est particulièrement appropriée à la recherche d'information dans des textes.
Un index inversé est une structure de données dans laquelle on stocke les mots contenus dans les documents du corpus, auxquels on associe la liste de documents dans lesquels ces mots apparaissent ainsi que leur nombre d'apparitions dans chaque document (voir Figure 1). Cette structure permet de calculer rapidement la pertinence d'un document du corpus pour répondre à la requête d'un utilisateur.
Pour calculer la pertinence d'un document relativement à un mot-clé, on se sert du td-idf. "tf" signifie "term frequency" : fréquence du terme. C'est en fait le nombre d'apparitions du mot-clé dans le document. "idf" signifie "inverse document frequency" : fréquence inverse de document. L'idf d'un mot clé k parmi un ensemble D de documents d_i se calcule ainsi : $IDF_k = log\frac{|D|}{|d_i, k \in d_i|}$, avec $|D|$ le nombre de documents et $|d_i, k \in d_i|$ le nombre de document où k apparait. Le tf-idf d'un document pour un mot clé donné est la multiplication du tf par l'idf.

Réaliser l'index inversé ainsi que les méthodes permettant :
– d'ajouter un document ayant un titre (*String*) et un contenu (*String*) à cet index
– de supprimer un document de cet index en donnant le titre du document
– d'obtenir une description de cet index
– de connaître le nombre de mots contenus dans l'index
– de connaître le nombre de documents contenus dans l'index
– de calculer le tf-idf d'un document pour un mot-clé donné
– de renvoyer un ensemble de documents triés par tf-idf décroissant pour un mot clé donné

.

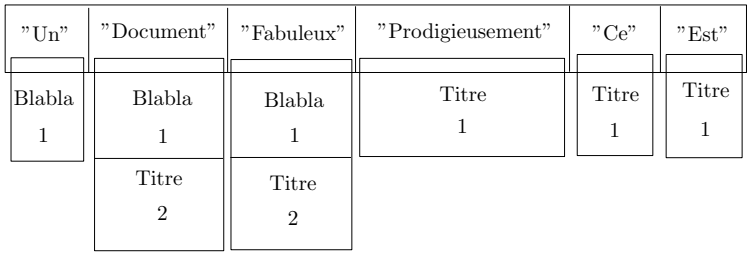

FIGURE 1 – L'index inversé d'un corpus contenant les deux documents suivants sous la forme (Titre, contenu) : ("Blabla", "Un document fabuleux") ; ("Titre", "Ce document est prodigiseument fabuleux. Document !")

Exercice 2. En utilisant uniquement les méthodes `add`, `size` et `iterator` des Set, réaliser :
– L'union de deux Set
– L'intersection de deux Set (vous pouvez utiliser un troisième Set)