Université des Sciences et de la Technologie Houari Boumediene
Faculté d'Informatique
Département d'Informatique

Master 2 : SII

Année Universitaire : 2022/2023
Module Recherche d'Information

# Recherche d'Information

Projet

Le but de ce projet est de mettre en application les concepts vus en cours quant à l'indexation des documents et l'appariement requête-document. La collection de test à considérer pour ce projet est le dataset : "CISI".

**A. Collection de test :**

La collection CISI est un ensemble de données textuelles utilisée pour la recherche d'information. Elle est accessible au public auprès de l'Université de Glasgow : http://ir.dcs.gla.ac.uk/resources/test_collections/
Cette collection est constituée de trois fichiers : CISI.ALL, CISI.QRY et CISI.REL.

- Le fichier CISI.ALL contient 1460 documents textuels. Chaque document est représenté par cinq champs : un identifiant unique (.I), un titre (.T), un auteur (.A), un résumé (.W) et une liste de références croisées (.X). Pour ce projet, nous intéressons qu'aux champs (.I), (.T) et (.W).
- Le fichier CISI.QRY contient 112 requêtes. Chaque requête est représentée par deux champs : un identifiant unique (.I) et le contenu de la requête (.W). Une requête peut être représentée par trois autres champs : un auteur (.A), un titre (.T) et une référence (.B).
- Le fichier CISI.REL contient une liste correcte de correspondance requête-document (jugements de pertinence).

**B. Actions à réaliser :**

Concevoir et développer une application avec une IHM permettant de réaliser les actions I. II. et III.

I. Indexation :

Implémenter les algorithmes qui permettent de :

. Extraire les termes à l'aide des deux méthodes :

$split()$

$nltk.RegexpTokenizer('(?:[A-Za-z]\.)+|\d+(?:\.\d+)?%?|\w+(?:\-\w+)*').tokenize()$

. Supprimer les mots vides à l'aide de la méthode :

$nltk.corpus.stopwords.words('english')$

. Normaliser les termes extraits à l'aide des deux méthodes :

$nltk.PorterStemmer().stem()$

$nltk.LancasterStemmer().stem()$

. Pondérer les termes à l'aide de la formule :

$$poids(t_i, d_j) = \frac{freq(t_i, d_j)}{MAX\left(freq\left((t, d_j)\right)\right)} * log\left(\left(\frac{N}{n_i}\right) + 1\right)$$

$\boldsymbol{poids(t_i, d_j)}$ : le poids du terme $\boldsymbol{i}$ dans le document $\boldsymbol{j}$
$\boldsymbol{freq(t_i, d_j)}$ : la fréquence du terme $\boldsymbol{i}$ dans le document $\boldsymbol{j}$
$\boldsymbol{MAX\left(freq\left((t, d_j)\right)\right)}$ : la fréquence max dans le document $\boldsymbol{j}$
$\boldsymbol{N}$ : le nombre de documents dans la collection
$\boldsymbol{n_i}$ : le nombre de documents contenant le terme $\boldsymbol{i}$
$\boldsymbol{log}$ : c'est le log de 10.

. Créer le fichier descripteurs, défini comme suit :

$$< Num\ document > \ < Terme > \ < Fréquence > \ < Poids >$$

. Retourner la liste des termes d'un document donné (avec fréquences et poids).

. Créer le fichier inverse, défini comme suit :

$$< Terme > \ < Num\ document > \ < Fréquence > \ < Poids >$$

. Retourner la liste des documents contenant un terme donné (avec fréquence et poids).

II. Appariement :

. Implémenter un système de recherche d'information (SRI) basé sur le modèle vectoriel en utilisant les fonctions d'appariement suivantes :

**Scalar Product :**

$$RSV(Q,d) = \sum_{i=1}^{n} poids(t_i, Q) * poids(t_i, d)$$

**Cosine Measure :**

$$RSV(Q,d) = \frac{\sum_{i=1}^{n} poids(t_i, Q) * poids(t_i, d)}{\sqrt{\sum_{i=1}^{n} poids(t_i, Q)^2} * \sqrt{\sum_{i=1}^{n} poids(t_i, d)^2}}$$

**Jaccard Measure :**

$$RSV(Q,d) = \frac{\sum_{i=1}^{n} poids(t_i, Q) * poids(t_i, d)}{\sum_{i=1}^{n} poids(t_i, Q)^2 + \sum_{i=1}^{n} poids(t_i, d)^2 - \sum_{i=1}^{n} poids(t_i, Q) * poids(t_i, d)}$$

$\boldsymbol{n}$ : la taille du vocabulaire
$\boldsymbol{poids}(\boldsymbol{t_i}, \boldsymbol{Q}) = 1$, SI $\boldsymbol{t_i}$ appartient à $\boldsymbol{Q}$, 0 SINON

. Implémenter un système de recherche d'information (SRI) basé sur le modèle booléen en utilisant les opérateurs logiques NOT, AND et OR.

. Implémenter un système de recherche d'information (SRI) basé sur le modèle probabiliste en utilisant la fonction BM25 suivante :

$$RSV(Q,d) = \sum_{t_i \in Q} \frac{freq(t_i, d)}{K\left((1-B) + B * \frac{dl}{avdl}\right) + freq(t_i, d)} * log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

$1.20 \leq \boldsymbol{K} \leq 2.00$ ; $0.50 \leq \boldsymbol{B} \leq 0.75$ : sont des constantes
$\boldsymbol{dl}$ : la taille du document $\boldsymbol{d}$ (nombre de termes)
$\boldsymbol{avdl}$ : la taille moyenne des documents (nombre de termes)

. Implémenter un système de recherche d'information (SRI) basé sur les techniques de Data Mining :

- Utiliser la technique de Clustering DBSCAN pour regrouper les documents de la collection.
- Utiliser la technique de Classification naïve bayésienne pour classifier la requête.

III. Evaluation :

. Comparer les SRI ci-dessus en termes de Précision (P@5 & P@10), Rappel et de F-measure.
. Tracer la courbe rappel-précision de chaque SRI implémenté.

**C. Rapport à remettre :**

Remettre un rapport du travail effectué au plus tard le **20 Décembre 2022**. Le rapport doit contenir, au minimum, les points suivants :

. Introduction
. Présentation du projet
. Explication de vos algorithmes et de vos SRIs.
. Montrez quelques résultats pour chacun des modèles, avec des captures d'écran
. Analyse et discussion des résultats des SRIs implémentés.
. Conclusion