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

Master 2 : SII

Année Universitaire : 2024/2025
Module Recherche d'Information

# Recherche d'Information

Projet

Le but de ce projet est d'appliquer les concepts abordés en cours concernant l'indexation des documents et l'appariement requête-document. La collection de test à utiliser dans le cadre de ce projet est le jeu de données "Medline".

**A. Collection de test :**

La collection "Medline" est un ensemble de données textuelles employée dans le domaine de la recherche d'information. Elle est accessible publiquement via le lien de l'Université de Glasgow :
http://ir.dcs.gla.ac.uk/resources/test_collections/medl/

Cette collection se compose de trois fichiers :

- **MED.ALL** : contient 1033 documents textuels.
- **MED.QRY** : contient 30 requêtes.
- **MED.REL** : contient une liste de correspondance entre requête et documents (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 permettant de :

. Extraire les termes en utilisant deux méthodes :

$$split()$$

$$\text{nltk.RegexpTokenizer}('(?:[A-Za-z]\.)+|[A-Za-z]+[\-@]\d+(?:\.\d+)?|\d+[A-Za-z]+|\d+(?:[\.\,\-]\d+)?\%?|\w+(?:[\-/]\w+)*').\text{tokenize}()$$

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

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

. Normaliser les termes extraits à l'aide de différentes méthodes :

- Normalisation simple sans aucun traitement supplémentaire (aucune modification des termes)
- Utiliser ces deux techniques de stemming pour réduire les mots à leur racine :

$$nltk.PorterStemmer().stem()$$

$$nltk.LancasterStemmer().stem()$$

. Pondérer les termes en utilisant la formule :

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

$\boldsymbol{poids}(\boldsymbol{t_i}, \boldsymbol{d_j})$ : le poids du terme $\boldsymbol{i}$ dans le document $\boldsymbol{j}$
$\boldsymbol{freq}(\boldsymbol{t_i}, \boldsymbol{d_j})$ : la fréquence du terme $\boldsymbol{i}$ dans le document $\boldsymbol{j}$
$\boldsymbol{MAX}\left(\boldsymbol{freq}((\boldsymbol{t}, \boldsymbol{d_j}))\right)$ : 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 :

$< N° \, document > \; < Terme > \; < Fréquence > \; < Poids >$

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

. Créer le fichier inverse, défini comme suit :

$< Terme > \; < N° \, document > \; < Fréquence > \; < Poids >$

. 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
$\boldsymbol{poids(t_i, Q)} = 1$, SI $\boldsymbol{t_i}$ appartient à $\boldsymbol{Q}$, 0 SINON

. Implémenter un SRI basé sur le modèle booléen, en utilisant les opérateurs logiques NOT, AND et OR.

. Implémenter un 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 $\boldsymbol{d}$
$\boldsymbol{avdl}$ : la taille moyenne des documents

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 pour chaque SRI implémenté.