

Année Universitaire : 2023/2024 Master 2 : SII Module : Recherche d'Information	Université des Sciences et de la Technologie Houari Boumediene Faculté d'Informatique Département d'Intelligence Artificielle et Sciences des Données	TP N°4 Recherche de l'Information : Appariement Partie 1
---	---	---

Support :

1. Extraction de termes (Tokens) à l'aide de l'expression régulière suivante :

```
nlTK.RegexpTokenizer('(?:[A-Za-z]\.)+|[A-Za-z]+[\-@]\d+(?:\.\d+)?|\d+[A-Za-z]+|\d+(?:[\.\,\,]\d+)?%?|\w+(?:[\-\/]\w+)*')
```

2. Appariement :

2.1. Modèle vectoriel (Vector space model)

2.1.1. Modèle basé sur le Produit Scalaire (Scalar Product)

Entrée (requête) :

Un ensemble de termes normalisés

Sortie :

Une liste de documents ordonnés selon leurs degrés de pertinences. Le degré de pertinence **RSV** d'un document **d** par rapport à une requête **Q** est calculé à l'aide de **Scalar Product** comme suit :

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

$$Q = \langle v_1, v_2, v_3, \dots, v_n \rangle$$

$$d = \langle w_1, w_2, w_3, \dots, w_n \rangle$$

n : la taille du vocabulaire

v_i : poids du terme **t_i** dans la requête **Q** (par défaut **v_i = 1** si la requête **Q** contient le terme **t_i**, **0** sinon)

w_i : poids du terme **t_i** dans le document **d**

2.1.2. Modèle basé sur la Similarité Cosinus (Cosine Measure)

Entrée (requête) :

Un ensemble de termes normalisés

Sortie :

Une liste de documents ordonnés selon leurs degrés de pertinences. Le degré de pertinence *RSV* d'un document *d* par rapport à une requête *Q* est calculé à l'aide de **Cosine Measure** comme suit :

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

2.1.3. Modèle basé sur l'Indice de Jaccard (Jaccard Measure)

Entrée (requête) :

Un ensemble de termes normalisés

Sortie :

Une liste de documents ordonnés selon leurs degrés de pertinences. Le degré de pertinence *RSV* d'un document *d* par rapport à une requête *Q* est calculé à l'aide de **Jaccard Measure** comme suit :

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

Exercice :

I. Implémenter les trois méthodes de recherche du modèle vectoriel:

- . Produit Scalaire
- . Similarité Cosinus
- . Indice de Jaccard

II. Visualiser les résultats retournés par chaque méthode de recherche.

Introduire la requête

Query

documents ranking

Search

☐ Queries Dataset

1

Processing

☒ Tokenization

☒ Porter Stemmer

Index

☐

☐ DOCS per TERM

☒ TERMS per DOC

Matching

☒ Vector Space Model

Scalar Product

Results

N°doc	Relevance
4	0.3856
5	0.3535
2	0.2996
3	0.1781
6	0.1347

Résultats retournés par le
modèle vectoriel basé sur le
Produit Scalaire

Introduire la requête

Query

documents ranking

Search

☐ Queries Dataset

1

Processing

☒ Tokenization

☒ Porter Stemmer

Index



☐ DOCS per TERM

☒ TERMS per DOC

Matching

☒ Vector Space Model

Cosine Measure

Results

N°doc	Relevance
4	0.2339
2	0.1442
5	0.1123
6	0.1099
3	0.0781

Résultats retournés par le
modèle vectoriel basé sur la
Similarité Cosinus

Introduire la requête

Query

documents ranking

Search

☐ Queries Dataset

1

Processing

☒ Tokenization

☒ Porter Stemmer

Index



☐ DOCS per TERM

☒ TERMS per DOC

Matching

☒ Vector Space Model

Jaccard Measure

Results

N°doc	Relevance
4	0.1297
2	0.0776
5	0.0535
6	0.0515
3	0.0403

Résultats retournés par le
modèle vectoriel basé sur

Indice de Jaccard