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 $\boldsymbol{RSV}$ d'un document $\boldsymbol{d}$ par rapport à une requête $\boldsymbol{Q}$ est calculé à l'aide de **Scalar Product** comme suit :

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

$\boldsymbol{Q} =< v_1, v_2, v_3, \dots, v_n >$
$\boldsymbol{d} =< w_1, w_2, v_3, \dots, v_n >$

$\boldsymbol{n}$: la taille du vocabulaire
$\boldsymbol{v_i}$: poids du terme $\boldsymbol{t_i}$ dans la requête $\boldsymbol{Q}$ (par défaut $\boldsymbol{v_i = 1}$ si la requête $\boldsymbol{Q}$ contient le terme $\boldsymbol{t_i}$, **0** sinon)
$\boldsymbol{w_i}$: poids du terme $\boldsymbol{t_i}$ dans le document $\boldsymbol{d}$

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 $\boldsymbol{RSV}$ d'un document $\boldsymbol{d}$ par rapport à une requête $\boldsymbol{Q}$ est calculé à l'aide de **Cosine Measure** comme suit :

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

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 $\boldsymbol{RSV}$ d'un document $\boldsymbol{d}$ par rapport à une requête $\boldsymbol{Q}$ est calculé à l'aide de **Jaccard Measure** comme suit :

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

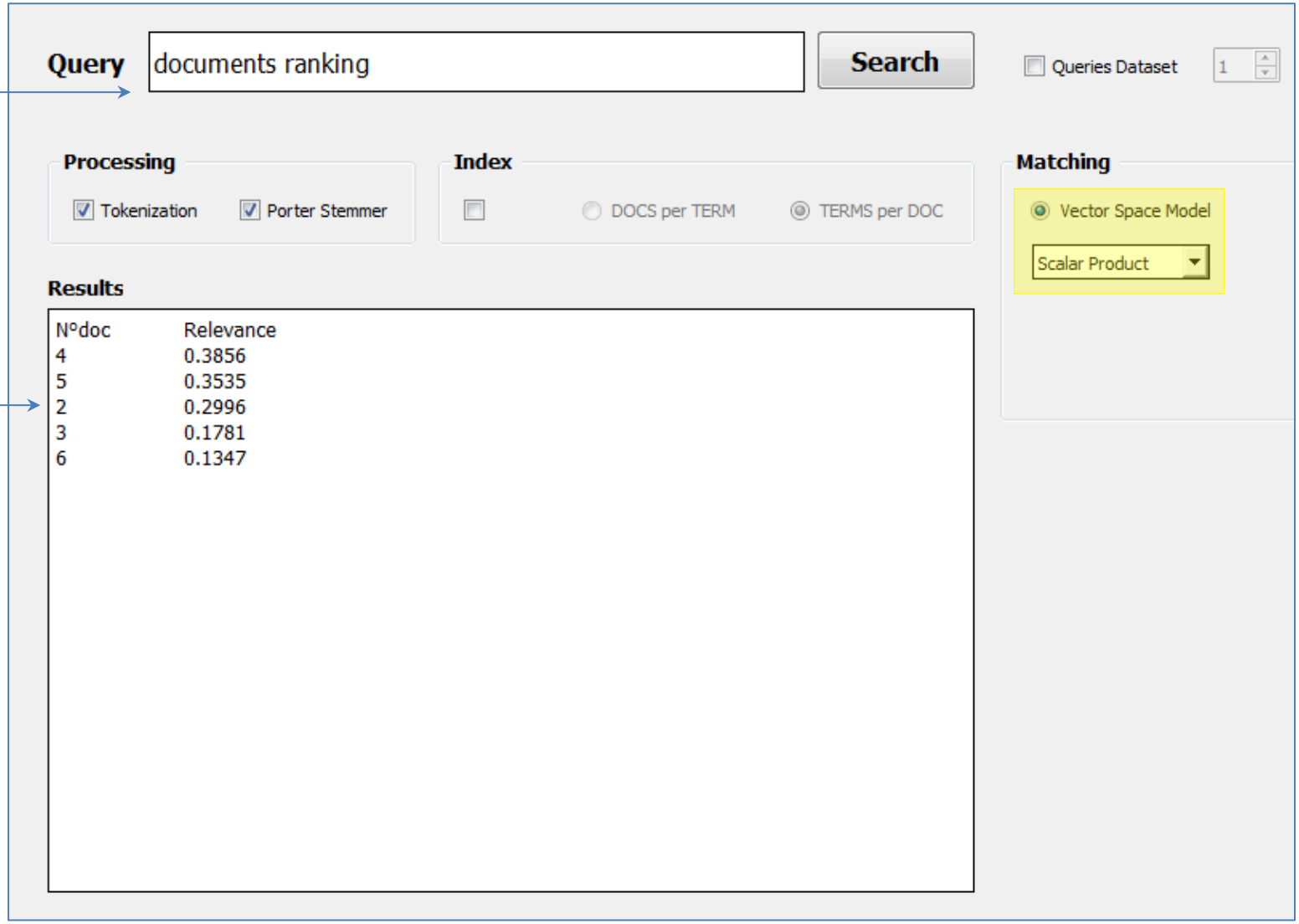
Exercice :

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.**

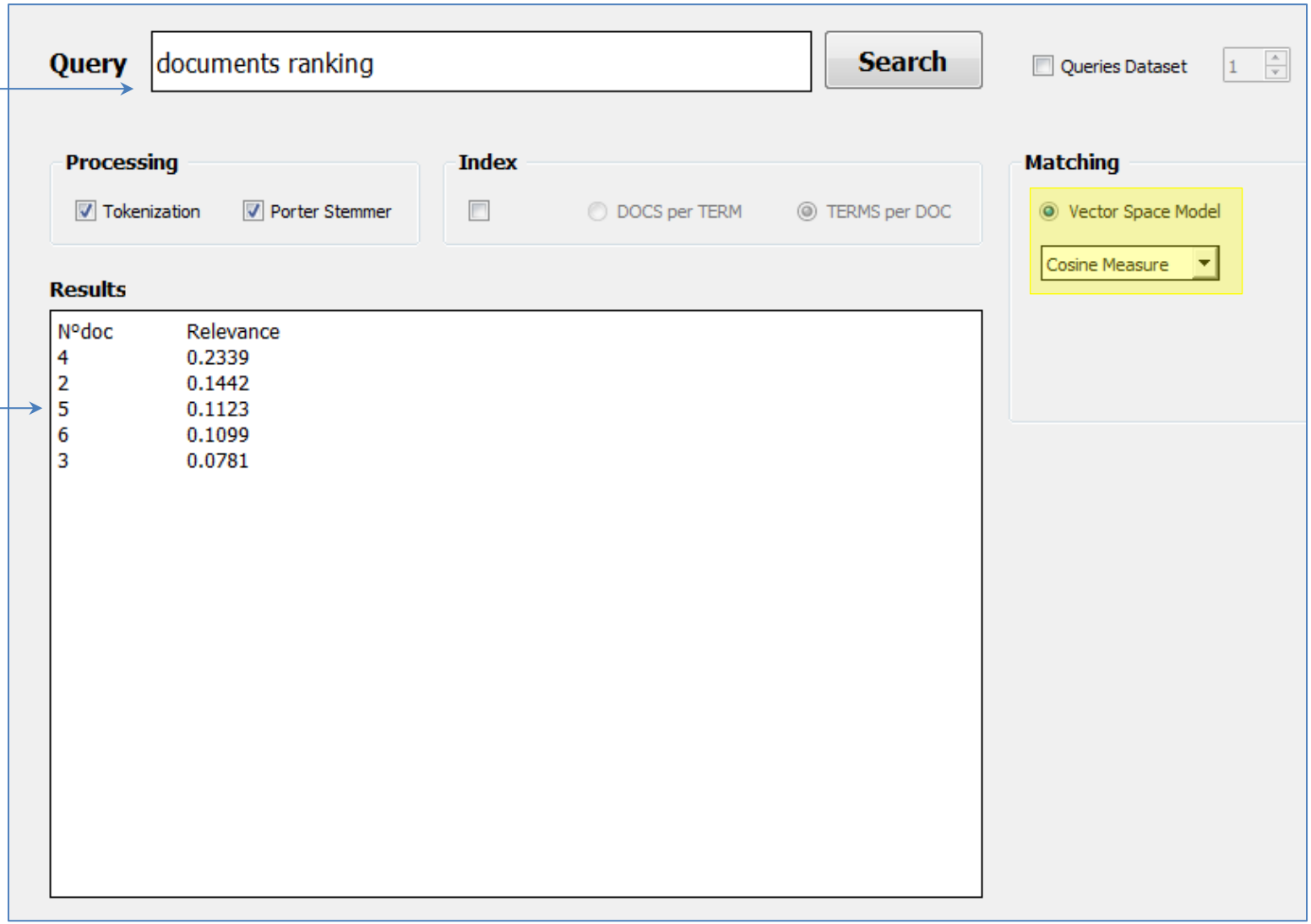
Introduire la requête

Résultats retournés par le modèle vectoriel basé sur le **Produit Scalaire**



Introduire la requête

Résultats retournés par le modèle vectoriel basé sur la **Similarité Cosinus**



Introduire la requête

Résultats retournés par le modèle vectoriel basé sur **Indice de Jaccard**