Série TD- n°2
Appariement – Evaluation d'un SRI

Exercice 1

Soit une collection composée de :

D1 : {indexation, tokenisation, ponderation, stoplist}

D2: {informatique, 3logiciel, 2materiel, machine, 2programme, 3instruction, executer}

D3 : {indexation, 2recherche, information, 2ri, appariement}

D4 : {3recherche, 2operationnel, 2ro, mathematique}

L'indexation des documents repose sur les traitements : lemmatisation, tokenisation, élimination des mots vides et pondération avec $tf_{(Termej, Di)}$=*fréquence* T_j *normalisée par la fréquence maximale dans Di.*

1) Formuler les besoins suivants sous forme de requêtes booléennes (expressions logiques) :

Besoin1 : « recherche d'information ou indexation » ;

Besoin2 : « indexation sans appariement ni recherche » ;

Besoin3 : « En informatique, un logiciel est un programme ou séquences d'instructions interprétables par une machine ????»

2) Lister dans l'ordre les documents retournés aux requêtes Q1 et Q2 (associées aux besoins 1 et 2 respectivement) avec :

2.1) Le modèle booléen standard. Justifier.

2.2) Le modèle logique floue. Justifier.

2.3) La mesure produit scalaire du modèle vectoriel. Justifier.

3) Evaluer et commenter les performances des trois modèles en termes de : Silence, Bruit, Precision, Rappel et P@1.

Exercice 2

Etant donnée une collection du contexte biomédicale, constituée de 100000 mots (avec redondance). Soit un échantillon de documents de la collection composé de :

D1 : {hiv, 2virus, infect, blood, aids} ; *D2 :* {covid-19, virus, sars-cov-2} ; *D3 :* {strong, muscle, masseter}

SRI1 est un système de recherche qui utilise la technique de lemmatisation dans l'indexation des documents. Le processus d'appariement repose sur le modèle vectoriel avec la mesure produit scalaire. La pondération des termes est $tf_{(Termej, Di)}$=*fréquence* T_j *normalisée par la somme des fréquences des termes dans Di.*

1) Lister les documents restitués aux requêtes :

Q1 : sars-cov-2 virus ; Q2 : strongest ligament ?

2) Evaluer la précision des résultats retournés en réponse aux requêtes Q1 et Q2.

3) **SRI2** est un système de recherche similaire au **SRI1**, la différence entre eux est dans le choix des termes d'indexation. Le SRI2 utilise le filtrage en considérant les termes d'indexation dont leur fréquence ∈ [3000, 50000]

3.1) Donner les index obtenus par le SRI2, sachant que :

- Le taux fréquentiel du mot «hiv » dans la collection est : 15%
- Le taux fréquentiel du mot «virus » dans la collection est : 9%
- Le taux fréquentiel du mot «infect » dans la collection est : 55%
- Le taux fréquentiel du mot «blood » dans la collection est : 35%
- Le taux fréquentiel du mot «covid-19 » dans la collection est : 10%

- Le taux fréquentiel du mot «sars-cov-2» dans la collection est : 7%
- Le taux fréquentiel du mot «strong» dans la collection est : 1 %
- Le taux fréquentiel du mot «muscle» dans la collection est : 40%
- Le taux fréquentiel du mot «masseter» dans la collection est : 20%

3.2) Lister les documents restitués aux requêtes Q1 et Q2 avec le SRI2. Déduire la précision des résultats obtenus.

3.3) Commenter les performances des deux SRIs en réponse à Q1 et Q2.

4) Lister les documents retournés pour Q3 : virus. Que peut-on remarquer ?

Exercice 3

Soit un fond documentaire composé de : $d1 = (T_1, T_3)$; $d2 = (2T_1, T_3, 2T_4)$; $d3 = (3T_1, 2T_2, T_4)$; $d4 = (T_5)$
Considérons : $q1 = (T_1, 2T_2, T_3)$.
Les termes d'indexation sont pondérés avec : *tf*Idf.*

1) Donner la liste ordonnée des documents retournés à *q1* avec (justifier):

1.1) Mesure Cosinus du modèle vectoriel ;

1.2) Mesure de Dice du modèle vectoriel ;

1.3) Modèle booléen étendu ;

1.4) Modèle unilangue avec lissage de Laplace.

2) Soit *q2* une requête dont le contenu est : T1T4. La séquence T1T4 possède 2 occurrences dans *d2* et une seule apparition dans *d3*. Lister dans l'ordre les documents retournés en utilisant le modèle bi-langue.

Exercice 4

Soient les listes suivantes des réponses retournés par un SRI aux requêtes Q1 et Q2 respectivement.

Rang	IdDoc (Docs Sélectionnés pour *Q1*)
1	D588
2	D589
3	D25
4	D30
5	D250
6	D11
7	D15
8	D22
9	D35
10	D40

Rang	IdDoc (Docs Sélectionnés pour *Q2*)
1	D1
2	D5
3	D25
4	D40

Les jugements de pertinence en RI est l'ensemble de documents jugés sémantiquement pertinents à une requête donnée, par des assesseurs (documentalistes). Soient les jugements de pertinence de :

Q1 :{D588, D30, D15, D40} Q2 : {D5, D40, D250, D111, D15, D8}

1) Calculer la precision, le rappel, le silence, le bruit et la F-mesure (F-Score) du SRI.

2) Donner : la R-Precision et les précision P@x (x=1, 2 et 5).

3) Calculer les taux de précision et de rappel du système dans chaque réponse. Déduire la MAP.

4) Tracer la courbe rappel-précision interpolée du système.