USTOMB -2021- Faculté MI Département Informatique

Recherche d'Informations L3 (ISIL)

TD1

Q1 : Quelle est la condition pour qu'un corpus soit significatif ?

Pour qu'un corpus soit significatif, il doit contenir un nombre assez important de documents.

Q2 : Y'a-t-il une différence entre un système de base de données et un système de recherche d'information ?

OUI : un SGBD est utilisé dans une base totalement structurée, contrairement à un SRI qui est utilisé dans une partie seulement des spécifications des documents (également structurée)

Q3 :Un terme qui apparaît dans tous les documents d'un corpus est-il discriminant?**NON**

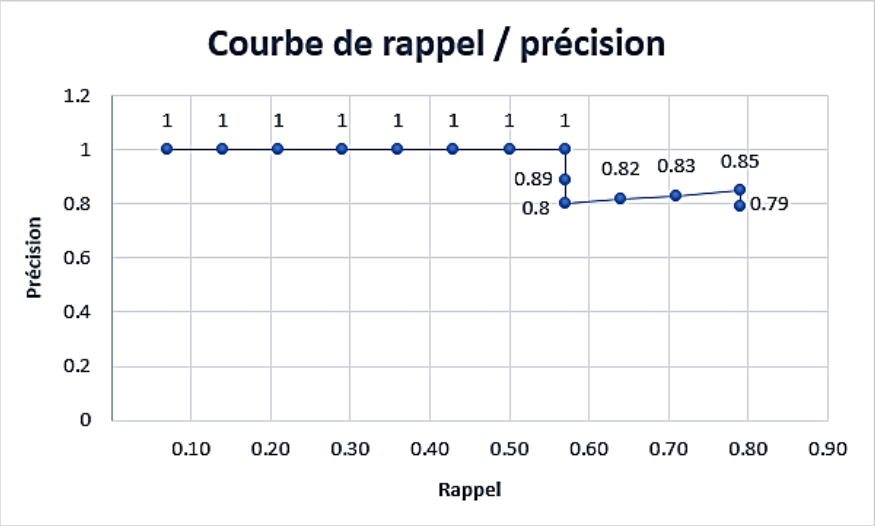
Q4:Soit les informations suivantes contenues dans le tableau ci-dessous concernant un corpus C pour les requêtes indiquées. Calculer les taux de précision et de rappel du système à chaque réponse et remplir le tableauci-dessus.

Précision= Documents pertinents rapportés/ Documents rapportés.

Rappel= Documents pertinents rapportés/ Documents pertinents.

Supposons quepour les requêtes indiquées, le nombre total de documents pertinents dans l'ensemble de tous les documents est 14.

Rang	n°doc	Pertinent	Rappel	Précision
1	324	X	1/14=0.07	1/1=1
2	589	X	2/14=0.14	2/2=1
3	528	X	3/14=0.21	3/3=1
4	590	X	4/14=0.29	4/4=1
5	986	X	5/14=0.36	5/5=1
6	592	X	6/14=0.43	6/6=1
7	899	X	7/14=0.50	7/7=1
8	988	X	8/14=0.57	8/8=1
9	578		8/14=0.57	8/9=0.89
10	985		8/14=0.57	8/10=0.8
11	537	X	9/14=0.64	9/11=0.82
12	591	X	10/14=0.71	10/12=0.83
13	772	X	11/14=0.79	11/13=0.85
14	990		11/14=0.79	11/14=0.79



Q5 : Soit le tableau ci-dessous résumant les résultats obtenus avec un système de recherche

- Donner la précision et le rappel.
- Quelle méthode donne toujours un rappel maximal ? une précision maximale?

	Pertinent	Non pertinent
Sélectionné	50	10
Non sélectionné	30	120

Précision= Documents pertinents rapportés/ Documents rapportés= 50/60=0.83

Rappel= Documents pertinents rapportés/ Documents pertinents=50/80=0.62

Afin d'augmenter le rappel, il faut augmenter le nombre de documents rapportés cependant cette méthode risque d'engendrer une précision faible et vice versa si on essaye d'augmenter la précision en diminuant le nombre de documents rapportés le rappel diminuera.

Q6 : Soit le corpus de documents ci-dessous, calculer les valeurs de discrimination de tous les termes suivant l'approche basée sur la valeur de discrimination. Quels sont les termes les plus discriminants du corpus ?

d1: 6 2 3 6 2
d2: 6 1 2 0 2
d3: 6 5 1 0 0

Le vecteur centroide V : 6 2.667 2 2 1.333

MaxS=5*36=180

Calcul des similarités:
Sim(d1,V)=0,685 = sqrt (((6-6)^2 + (2-2.667)^2 + (3-2)^2 +(6-2)^2 +(2-1.333)^2) /180)
Sim(d2,V)=0,800 = sqrt (((6-6)^2 + (1-2.667)^2 + (2-2)^2 +(0-2)^2 +(2-1.333)^2) /180)
Sim(d3,V)=0,739 = sqrt (((6-6)^2 + (5-2.667)^2 + (1-2)^2 +(0-2)^2 +(0-1.333)^2) /180)

Calcul de l'uniformité du corpus U

U=1/3*(0,685+ 0,800+ 0,739)= 0.741

Normaliser t_1 à 0, le vecteur V devient alors
V1: 0 2.667 2 2 1.333
Les vecteurs D1 D2 et D3 deviennent alors
D1 : 0 2 3 6 2
D2 : 0 1 2 0 2
D3 : 0 5 1 0 0

Calcul des similarités:
Sim(d1,V1)=0,685 = 1- sqrt (((0-0)^2 + (2-2.667)^2 + (3-2)^2 +(6-2)^2 +(2-1.333)^2) /180)
Sim(d2,V1)=0,800 = 1-sqrt (((0-0)^2 + (1-2.667)^2 + (2-2)^2 +(0-2)^2 +(2-1.333)^2) /180)
Sim(d3,V1)=0,739 = 1-sqrt (((0-0)^2 + (5-2.667)^2 + (1-2)^2 +(0-2)^2 +(0-1.333)^2) /180)

Calculer l'uniformité du corpus U1
U1=1/3*(0,685+ 0,800+ 0,739)= 0.741
Calcul de la valeur de discrimination: v1=U1-U=0

Normaliser t_2 à 0, le vecteur V devient alors
V2: 6 0 2 2 1.333
Les vecteurs D1 D2 et D3 deviennent alors
D1 : 6 0 3 6 2
D2 : 6 0 2 0 2
D3 : 6 0 1 0 0

Calcul des similarités:
Sim(d1,V2)=0,689 = 1- sqrt (((6-6)^2 + (0-0)^2 + (3-2)^2 +(6-2)^2 +(2-1.333)^2) /180)
Sim(d2,V2)=0,843 = 1-sqrt (((6-6)^2 + (0-0)^2 + (2-2)^2 +(2-0)^2 +(2-1.333)^2) /180)
Sim(d3,V2)=0,806 = 1-sqrt (((0-0)^2 + (0-0)^2 + (2-1)^2 +(0-2)^2 +(0-1.333)^2) /180)
Calculer l'uniformité du corpus U2
U2=1/3*(0,689+ 0,843+ 0,806)= 0.779
Calcul de la valeur de discrimination: v2=U2-U=0.039

Normaliser t_3 à 0, le vecteur V devient alors
V3: 6 2.667 0 2 1.333
Les vecteurs D1 D2 et D3 deviennent alors
D1 : 6 2 0 6 2
D2 : 6 1 0 0 2
D3 : 6 5 0 0 0

Calcul des similarités:
Sim(d1,V3)=0,694 = 1- sqrt (((6-6)^2 + (2-2.667)^2 + (0-0)^2 +(6-2)^2 +(2-1.333)^2) /180)
Sim(d2,V3)=0,800
Sim(d3,V3)=0,750
Calculer l'uniformité du corpus U3
U3=1/3*(0,694+ 0,800+ 0,750)= 0.748
Calcul de la valeur de discrimination :v3=U3-U=0.007

Normaliser t_4 à 0, le vecteur V devient alors
V4: 6 2.667 2 0 1.333
Les vecteurs D1 D2 et D3 deviennent alors
D1 : 6 2 3 0 2
D2 : 6 1 2 0 2
D3 : 6 5 1 0 0

Calcul des similarités:
Sim(d1,V4)=0,898
Sim(d2,V4)=0,866
Sim(d3,V4)=0,786
Calculer l'uniformité du corpus U4
U4=1/3*(0,898+ 0,866+ 0,786)= 0.850
Calcul de la valeur de discrimination :v4=U4-U=0.109

Normaliser t_5 à 0, le vecteur V devient alors
V5: 6 2.667 2 2 0
Les vecteurs D1 D2 et D3 deviennent alors
D1 : 6 2 3 6 0
D2 : 6 1 2 0 0
D3 : 6 5 1 0 0

Calcul des similarités:
Sim(d1,V5)=0,689
Sim(d2,V5)=0,806
Sim(d3,V5)=0,759
Calculer l'uniformité du corpus U5
U5=1/3*(0,689+ 0,806+ 0,759)= 0.751
Calcul de la valeur de discrimination : v5=U5-U=0.010

Voici l'ordre de pertinence des termes :

v4 0.106

v2 0.039

v5 0.010

v3 0.007

v1 0

v1 n'est pas un terme pertinent par contre v4 est considéré comme le pertinent .