

TD – Recherche d'information – Recherche et évaluation

Exercice 1 – Recherche dans le modèle vectoriel

Fournir les résultats des requêtes suivantes pour le corpus de l'exercice 3 des TDs précédents :

Q1 : pomme de terre

Q2 : recherche d'information textuelle

Q3 : domaine du modèle vectoriel

Question 1

Construire les vecteurs des requêtes. Commencer par analyser les requêtes comme les documents (anti-dictionnaire), et utiliser une pondération des requêtes par le tf.idf.

Calculer les normes.

**** Tous nos vecteurs sont 11D, cf la partie sur l'indexation. ****

Pour Q1 :

Extraction tokens, antidictionnaire (le même que les TD précédent):

~~pomme de terre~~

// //

pas de terme du vocabulaire donc requête vide :

$$Q0 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$$

donc on n'exécute même pas a requête : pas de document qui répond !

Pour Q2 :

~~recherche d' information~~ textuelle

// // (t10) tf=1 idf = 1.10

donc le vecteur requête est $Q2 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1.10\ 0)$

$$\|Q2\| = \text{rac}(1.10) = 1.10$$

Pour Q3 :

~~domaine~~ ~~et~~ ~~modèle~~ vectoriel

(t2) tf=1 idf=1.10 (t4) tf=1 idf=1.10 (t11) tf=1 idf = 1.10

donc le vecteur requête est $Q3 = (0\ 1.10\ 0\ 1.10\ 0\ 0\ 0\ 0\ 0\ 0\ 1.10)$

$$\|Q3\| = \text{rac}(3.63) = 1.91$$

Question 2

Calculer les similarités RSV(d,q) en reprenant la formule du cours (transparent 10) sans utiliser l'index inversé, avec chacun des documents.

Q1 : requête vide : c'est fini. On sait qu'il n'y pas de réponse.

$$Q2 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1.10\ 0)$$

$$RSV(D1,Q2) = RSV((0\ 0\ 0\ 0\ 0\ 1.10\ 0\ 1.10\ 0\ 1.10\ 0), (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0))$$

On connaît $\|D1\|$ et $\|Q2\|$

$$RSV(D1, Q2) = (0*0 + 0*0 + \dots + 1.10*1.10 + 0*0) / (1.90 * 1.10) = 0.58$$

$$RSV(D2, Q2) = RSV((0 \ 1.10 \ 1.10 \ 0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0), (0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0))$$

$$= 0 / \|D2\| * \|Q2\| = 0$$

→ D2 ne répond pas à Q2

$$RSV(D3, Q2) = 0$$

→ D3 ne répond pas à Q2

Réponse du système à Q2 ?

On sait que D2 et D3 ne répondent pas → pas dans la réponse.

Il ne reste que D1 (0,58), donc la réponse pour Q2 est : D1.

$$Q3 = (0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1.10)$$

$$RSV(D1, Q3) = RSV((0 \ 0 \ 0 \ 0 \ 1.10 \ 0 \ 1.10 \ 0 \ 1.10 \ 0), (0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1.10))$$

$$RSV(D1, Q3) = 0$$

$$RSV(D2, Q3) = RSV((0 \ 1.10 \ 1.10 \ 0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0), (0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1.10))$$

$$RSV(D2, Q3) = (1.10*1.10) / (2.20 * 1.91) = 0.29$$

$$RSV(D3, Q3) = RSV((1.10 \ 0 \ 2.20 \ 0 \ 0 \ 0 \ 0 \ 1.10 \ 0 \ 1.10), (0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1.10))$$

$$RSV(D3, Q3) = (2.20*1.10 + 1.10*1.10) / (2.91 * 1.91) = 0.66$$

La réponse pour Q3 est : D3, D2.

D1 NE REpond PAS A Q3.

Question 3

Refaire les calculs avec l'utilisation de l'index inversé.

Pour la **question 3** avec l'index inversé: exactement les mêmes résultats que pour la question 2.

Exemple pour Q3 (cf. transparent 12):

$$Q3 = (0 \ 1.10 \ 0 \ 1.10 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1.10)$$

1. On garde les $t_i < 0 \Rightarrow t2, t4, t11$

(Note : ligne_res est un vecteur avec comme indice D1, D2, D3, initialisé à [0, 0, 0])

2. Boucle sur t2, t4, t11

pour t2 : (note : la ligne t2 de l'index inversé est [0, 1.10, 0])

$$\text{ligne_res}[D1] += 0 * 1.10 \rightarrow = 0$$

$$\text{ligne_res}[D2] += 1.10 * 1.10 \rightarrow = 1.21$$

$$\text{ligne_res}[D3] += 0 * 1.10 \rightarrow = 0$$

(note : pour le moment, calcul partiel, uniquement en comptant t2, avec tous les docs
 $\Rightarrow [0, 1.21, 0]$)

pour t4 : (note, la ligne t4 de l'index inversé est [0, 0, 2.20])

$$\text{ligne_res}[D1] += 0 * 1.10 \rightarrow = 0$$

$$\text{ligne_res}[D2] += 0 * 1.10 \rightarrow = 1.21$$

$$\text{ligne_res}[D3] += 2.20 * 1.10 \rightarrow 2.42$$

(note : pour le moment, calcul partiel, en comptant maintenant t2 et t4, avec tous les docs => [0, 1.21, 2.20])

Sur t11 : (note, la ligne t11 de l'index inversé est [0, 0, 1.10])

$$\text{ligne_res [D1]} += 0 * 1.10 \rightarrow = 0$$

$$\text{ligne_res[D2]} += 0 * 1.10 \rightarrow = 1.21$$

$$\text{ligne_res[D3]} += 1.10 * 1.10 \rightarrow 3.63$$

(note : pour le moment, ligne_res ([0, 1.21, 3.63]) contient le produit scalaire des docs et de tout Q3 (car on a passé les 3 termes non-nuls de la requête)

3. Calcul final (obtention du cosinus pour tous les docs): (rappel: $\|d1\| = 1.90$, $\|d2\| = 2.20$; $\|d3\| = 2,91$; $\|Q1\| = 1.91$)

$$\text{ligne_res [D1]} = \text{ligne_res [D1]} / (\|D1\| * \|Q3\|) \rightarrow = 0$$

$$\text{ligne_res[D2]} = \text{ligne_res [D2]} / (\|D2\| * \|Q3\|) \rightarrow = 0.29$$

$$\text{ligne_res[D3]} = \text{ligne_res [D3]} / (\|D3\| * \|Q3\|) \rightarrow = 0.66$$

(Note : on a maintenant les cosinus pour les 3 docs, en ne regardant QUE les éléments intéressants pour les calculs, c'est beaucoup plus rapide !)

La suite est la même que pour la question 2.

Exercice 2 – Evaluation de SRI

Nous réalisons ici une évaluation d'un système de recherche d'information.

Question 1

Supposons que pour une requête Q1 le système de recherche d'information testé renvoie les réponses suivantes:

rang	n° doc	pertinent	rappel	précision
1	588	1	1 / 5 = 0.2	1 / 1 = 1
2	589	1	2 / 5 = 0.4	2 / 2 = 1
3	576	0	2 / 5 = 0.4	2 / 3 = 0.67
4	590	1	3 / 5 = 0.6	3 / 4 = 0.75
5	986	0	3 / 5 = 0.6	3 / 5 = 0.6
6	592	1	4 / 5 = 0.8	4 / 6 = 0.67
7	884	0	4 / 5 = 0.8	4 / 7 = 0.57
8	988	0	4 / 5 = 0.8	4 / 8 = 0.50
9	578	0	4 / 5 = 0.8	4 / 9 = 0.44
10	985	0	4 / 5 = 0.8	4 / 10 = 0.4
11	103	0	4 / 5 = 0.8	4 / 11 = 0.36
12	591	0	4 / 5 = 0.8	4 / 12 = 0.33
13	572	1	5 / 5 = 1	5 / 13 = 0.38
14	990	0	5 / 5 = 1	5 / 14 = 0.36

Les documents pertinents pour Q1 sont : 572, 588, 589, 590, 592.

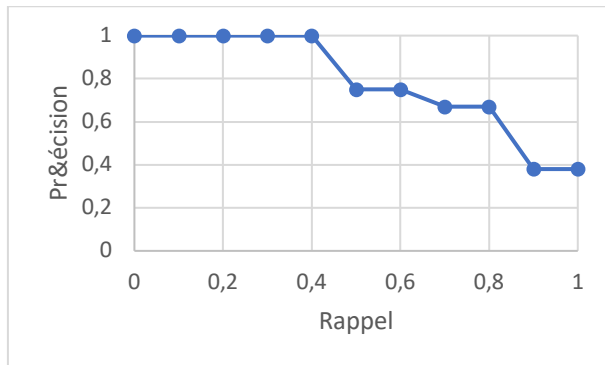
Calculer les taux de précision et de rappel du système à chaque réponse et remplir le tableau ci-dessus.

Donner le tableau de résultats normalisé pour cette requête, et en déduire la courbe de rappel/précision.

Tableau normalisé

Rappel	Précision
0	1
0.1	1
0.2	1
0.3	1
0.4	1
0.5	0.75
0.6	0.75
0.7	0.67
0.8	0.67
0.9	0.38
1.0	0.38

Dessin :



Question 2

Réaliser le même travail pour la requête Q2, avec les réponses suivantes :

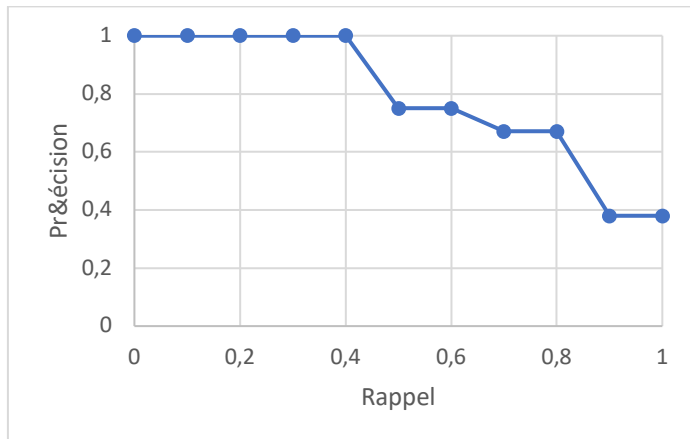
Rang	n° doc	pertinent	Rappel	précision
1	324	X	1/11 = 0.09	1
2	589	X	2/11 = 0.18	1
3	528	X	3/11 = 0.27	1
4	590	X	4/11 = 0.36	1
5	986	X	5/11 = 0.45	1
6	592	X	6/11 = 0.55	1
7	899	X	7/11 = 0.64	1
8	988	X	8/11 = 0.73	1
9	578		8/11 = 0.73	8/9 = 0.89
10	985		8/11 = 0.73	8/10 = 0.80
11	537	X	9/11 = 0.82	9/11 = 0.82
12	591	X	10/11 = 0.91	10/12 = 0.83
13	772	X	11/11 = 1	11/13 = 0.85
14	990		11/11 = 1	11/14 = 0.79

La liste des tous les documents pertinents pour la requête Q2 est : 324, 528, 537, 589, 590, 591, 592, 772, 899, 986, 988.

Tableau normalisé

Rappel	Précision
0	1
0.1	1

0.2	1
0.3	1
0.4	1
0.5	1
0.6	1
0.7	1
0.8	0.85
0.9	0.85
1.0	0.85



Question 3

En regardant les courbes, que pouvez-vous déduire de la qualité relative du système pour ces deux requêtes?

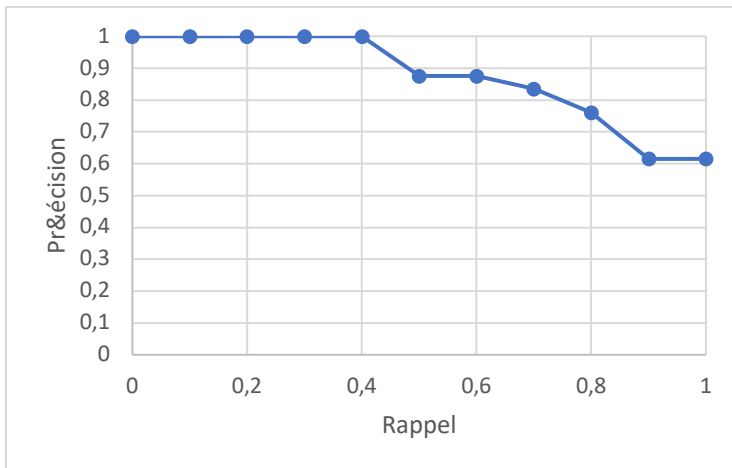
Il répond bien aux deux requêtes. Il répond mieux à Q2 qu'à Q1.

Dans les 2 cas il est capable de retourner tous les docs pertinents (car on atteint le point de rappel 1 avec une précision non-nulle).

Question 4

Donner le tableau global des résultats du système pour les deux requêtes et dessiner le schéma résultant.

Rappel	Précision Q1	Précision Q2	Précision
0	1	1	1
0,1	1	1	1
0,2	1	1	1
0,3	1	1	1
0,4	1	1	1
0,5	0,75	1	0,875
0,6	0,75	1	0,875
0,7	0,67	1	0,835
0,8	0,67	0,85	0,76
0,9	0,38	0,85	0,615
1	0,38	0,85	0,615



Exercice 3 – Comparaison de SRI

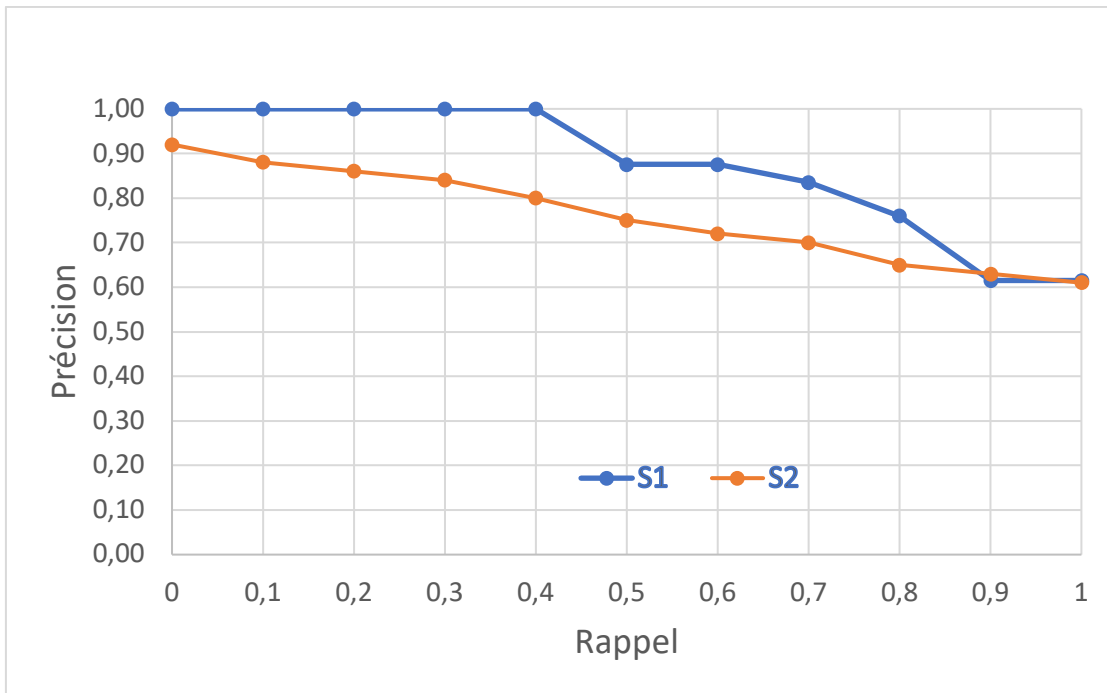
Nous voulons comparer deux systèmes de recherche d'information.

Le premier système S1 est celui de l'exercice 9. Le second système, S2, a pour tableau de rappel/précision pour les deux requêtes Q1 et Q2:

Rappel	Précision
0	0.92
0.1	0.88
0.2	0.86
0.3	0.84
0.4	0.80
0.5	0.75
0.6	0.72
0.7	0.70
0.8	0.65
0.9	0.63
1.0	0.61

Si on compare ces deux systèmes sur le même tableau :

Rappel	Précision S1	Précision 2
0	1	0.92
0,1	1	0.88
0,2	1	0.86
0,3	1	0.84
0,4	1	0.80
0,5	0,875	0.75
0,6	0,875	0.72
0,7	0,835	0.70
0,8	0,76	0.65
0,9	0,615	0.63
1	0,615	0.61



On voit que le système S1 est meilleur (courbe plus haute) quasiment pour toutes les valeurs de rappel. Il est très bon au début des réponses (entre 0 et 0,4 de rappel). Les 2 systèmes ont une bonne précision quand le rappel est à 1, ceci veut dire qu'ils sont capables tous les deux de trouver tous les documents pertinents, avec dans ce cas une précision d'environ 0,6 .