



Exercice 1 :

Considérer un corpus de documents composé de trois documents :

- Document 1 : "Machine learning is an area of artificial intelligence"
- Document 2 : "Artificial intelligence is the study of algorithms"
- Document 3 : "Algorithms are used in various applications"

1) Donner la table des fréquences des termes dans les deux documents

Terme	Document 1	Document 2	Document 3
Machine	1	0	0
Learning	1	0	0
Artificial	0	1	0
Intelligence	1	1	0
Algorithms	0	0	1
Study	0	1	0
Area	1	0	0
Used	0	0	1
Various	0	0	1
Applications	0	0	1

2.1) Calculer le poids de chaque terme dans chaque document en utilisant la formule suivante :

$$w(t_i, d_j) = \frac{freq_{ij}}{\max\{\forall t_i \in d_j\} freq_{ij}} * \log\left(\frac{N}{n_i} + 1\right)$$

Où n_i représente le nombre de documents contenant le terme t_i , N est le nombre de documents et le logarithme est calculé en base 10.

Terme	Document 1	Document 2	Document 3
Machine	0.25	0	0
Learning	0.25	0	0
Artificial	0	1	0
Intelligence	0.5	0.5	0
Algorithms	0	0	1
Study	0	1	0
Area	0.25	0	0
Used	0	0	1
Various	0	0	1
Applications	0	0	1

2.2) Calculer le poids de chaque terme dans chaque document en utilisant la formule suivante :

$$w(t_i, d_j) = 1 + \log(freq_{ij}) * \log\left(\frac{N}{n_i} + 1\right)$$

Terme	Document 1	Document 2	Document 3
Machine	1	N.D	N.D
Learning	1	N.D	N.D
Artificial	Non déterminé (N.D)	1	N.D
Intelligence	1.693	1.693	N.D
Algorithms	N.D	N.D	1.693
Study	N.D	1.693	N.D
Area	1	N.D	N.D
Used	N.D	N.D	1.693
Various	N.D	N.D	1.693
Applications	N.D	N.D	1.693

2.3) Calculer le poids de chaque terme dans chaque document en utilisant la formule suivante :

$$w(t_i, d_j) = \frac{freq_{ij}}{1.5 * \left(\frac{Longueur Document d_j}{Longueur moyenne}\right) + freq_{ij} + 0.5} * \log\left(\frac{N}{n_i}\right)$$

Longueur moyenne = 21/3 = 7

Longueur Document $d_1 = 8$

Longueur Document $d_2 = 7$

Longueur Document $d_3 = 6$

Terme	Document 1	Document 2	Document 3
Machine	0.343	0	0
Learning	0.343	0	0
Artificial	0	0.299	0
Intelligence	0.343	0.343	0
Algorithms	0	0	0.299
Study	0	0.299	0
Area	0.343	0	0
Used	0	0	0.299
Various	0	0	0.299
Applications	0	0	0.299

Exercice 2 :

Soit la matrice de fréquence termes-documents suivante :

	t1	t2	t3	t4	t5
D1	6	2	3	6	2

D2	6	1	2	0	2
D3	6	5	1	0	0

Calculer la valeur de discrimination pour les deux termes t1 et t4 en utilisant la distance euclidienne normalisée sur la valeur maximale de la somme comme une mesure de similarité

$$\text{Disc}(t1) = \text{AvgSim}_2 - \text{AvgSim}_1$$

$$\text{Disc}(t4) = \text{AvgSim}_3 - \text{AvgSim}_1$$

1. Calculer la similarité moyenne AvgSim_i entre le document d_i et v (vecteur centroid du corpus)

$$\text{AvgSim}_i = \frac{1}{N} * \sum_{i=1}^N \text{sim}(d_i, v) \text{ tel que } N \text{ est le nombre de documents}$$

$$\text{sim}(d_i, v) = 1 - \sqrt{\frac{\sum_{j=1}^M |d_{ij} - v_j|^2}{\text{maxSomme}}}$$

$$\text{maxSomme} = \sum_{j=1}^M \sum_{i=1}^N \max(d_{ij})^2$$

$$d_{ij} = \sqrt{\sum_{k=1}^n (d_{ik} - d_{jk})^2}$$

La table de distance d_{ij} entre les documents est la suivante :

	D1	D2	D3
D1	0	$\sqrt{38}$	$\sqrt{53}$
D2	$\sqrt{38}$	0	$\sqrt{21}$
D3	$\sqrt{53}$	$\sqrt{21}$	0

Pour calculer $\text{maxSomme} = \sum_{j=1}^M \sum_{i=1}^N \max(d_{ij})^2$, il faut :

- Trouver le maximum de chaque ligne de la table de distance
- Prendre le carrée de chaque maximum
- Somme des maximums de chaque ligne

Pour la ligne 1 (D1) : $\max(0, \sqrt{38}, \sqrt{53})^2 = 53$

Pour la ligne 2 (D2) : $\max(\sqrt{38}, 0, \sqrt{21})^2 = 38$

Pour la ligne 3 (D3) : $\max(\sqrt{53}, \sqrt{21}, 0)^2 = 53$

$$\text{MaxSomme} = 53 + 38 + 53 = 144$$

2) Calculer le vecteur centroid v du corpus :

Le poids du terme t_j dans v est la moyenne de ses poids dans les documents :

$$p_j = \frac{\sum_{i=1}^N p_{ij}}{N} \text{ tel que: } N \text{ est le nombre de documents}$$

v	6	2.667	2	2	1.333
v_1	0	2.667	2	2	1.333
v_2	6	0	2	2	1.333
v_3	6	2.667	0	2	1.333
v_4	6	2.667	2	0	1.333
v_5	6	2.667	2	2	0

	t1	t2	t3	t4	t5
D1	6	2	3	6	2
D2	6	1	2	0	2
D3	6	5	1	0	0

- Calculer $\text{sim}(d_1, v), \text{sim}(d_2, v), \text{sim}(d_3, v)$

$$\text{sim}(d_1, v) \approx 0.648$$

$$\text{sim}(d_2, v) \approx 0.777$$

$$\text{sim}(d_3, v) \approx 0.697$$

- Calculer $\text{sim}(d_1, v_1), \text{sim}(d_2, v_1), \text{sim}(d_3, v_1)$

$$\text{sim}(d_1, v_1) \approx 0.389$$

$$\text{sim}(d_2, v_1) \approx 0.453$$

$$\text{sim}(d_3, v_1) \approx 0.421$$

- Calculer $\text{sim}(d_1, v_2), \text{sim}(d_2, v_2), \text{sim}(d_3, v_2)$

$$\text{sim}(d_1, v_2) \approx 0.614$$

$$\text{sim}(d_2, v_2) \approx 0.805$$

$$\text{sim}(d_3, v_2) \approx 0.522$$

- Calculer $\text{sim}(d_1, v_3), \text{sim}(d_2, v_3), \text{sim}(d_3, v_3)$

$$\text{sim}(d_1, v_3) \approx 0.463$$

$$\text{sim}(d_2, v_3) \approx 0.519$$

$$\text{sim}(d_3, v_3) \approx 0.768$$

- Calculer $\text{sim}(d_1, v_4), \text{sim}(d_2, v_4), \text{sim}(d_3, v_4)$

$$\text{sim}(d_1, v_4) \approx 0.500$$

$$\text{sim}(d_2, v_4) \approx 0.860$$

$$\text{sim}(d_3, v_4) \approx 0.761$$

- Calculer $\text{sim}(d_1, v_5), \text{sim}(d_2, v_5), \text{sim}(d_3, v_5)$

$$\text{sim}(d_1, v_5) \approx 0.648$$

$$\text{sim}(d_2, v_5) \approx 0.774$$

$$\text{sim}(d_3, v_5) \approx 0.690$$

La valeur de discrimination pour les deux termes t4 et t1:

$$\text{AvgSim}_1 = \frac{1}{3}(\text{sim}(d_1, v) + \text{sim}(d_2, v) + \text{sim}(d_3, v)) \approx 0.707$$

$$\text{AvgSim}_2 = \frac{1}{3}(\text{sim}(d_1, v_1) + \text{sim}(d_2, v_1) + \text{sim}(d_3, v_1)) \text{ (pour t1)} \approx 0.421$$

$$\text{AvgSim}_3 = \frac{1}{3}(\text{sim}(d_1, v_4) + \text{sim}(d_2, v_4) + \text{sim}(d_3, v_4)) \text{ (pour t4)} \approx 0.707$$

$$\text{Disc}(t1) = \text{AvgSim}_2 - \text{AvgSim}_1 = -0.286 \rightarrow \text{Le terme t1 n'est pas discriminant}$$

$$\text{Disc}(t4) = \text{AvgSim}_3 - \text{AvgSim}_1 = 0 \rightarrow \text{Le terme t4 est indifférent}$$

Exercice 3

Soient les deux documents suivants :

Doc1: Now is the time for all good men to come to the aid of their country

Doc2: It was a dark and stormy night in the country manor. The time was past midnight

1) Construire le fichier inverse correspondant

Terme	Nombre de documents	Fréquence	Pointeur	
				Doc2, 1
a	1	1		Doc1,1
aid	1	1		Doc1,1
all	1	1		Doc2,1
and	1	1		Doc1, 1
come	1	1		Doc1,1
				Doc2, 1
contry	2	2		Doc2, 1
dark	1	1		Doc1,1
for	1	1		Doc1,1
good	1	1		Doc2,1
in	1	1		Doc1,1
is	1	1		Doc2,1
it	1	1		Doc2,1
manor	1	1		Doc1,1
men	1	1		Doc2,1
midnight	1	1		Doc2,1
night	1	1		Doc1,1
now	1	1		Doc1,1
of	1	1		
past	1	1		Doc2,1
stormy	1	1		Doc2,1
				Doc2,1
the	2	2		Doc1,1
their	1	1		Doc1,1
				Doc2,1
time	2	2		Doc2,1
to	2	2		Doc1,1
				Doc2,1
was	2	2		Doc1,1
				Doc2,1

Soit la requête booléenne q : time and past and the,

2) Construire la matrice d'incidence documents-termes comme requis par le modèle booléen

	a	aid	all	and	come	contry	dark	fot	good	in	is	it	manner	men	midnight	night	now	of	past	stormt	the	their	time	to	was
Doc1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1
Doc2	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	0	0	1	1	1	0	1	1	1

3) Calculer Rsv(q, Doc2) et Rsv(q, Doc1)

q : time and past and the

Rsv(q, Doc1)=0

Rsv(q, Doc2) =1

4) Evaluer la requête (résultats de la recherche) : time and past and the :

$\text{numDoc}_{\text{time}} \cap \text{numDoc}_{\text{past}} \cap \text{numDoc}_{\text{the}} = \{1, 2\} \cap \{2\} \cap \{1, 2\} = \{2\}$

Résutat retourné est le document 2

Exercice 4

Soient les trois documents suivants :

D 1 : {Le langage de programmation python est très utilisé pour le traitement de texte}

D2 : {Le langage JAVA est basé sur le langage C++}

D3 : {Un langage de programmation est un langage utilisé pour traduire un algorithme en un programme}

1) Indexer les 3 documents et donner le fichier inverse

- Segmentation du texte en termes
- Suppression des mots vides
- Normalisation des termes (Porter, troncature à x caractères, N-gram)
- Calcul des fréquences des termes

3.1) Construire le dictionnaire des termes

Terme	Numéro de document	Trie des termes	Terme	Numéro de document
langage	1		algorithme	3
programmation	1		basé	2
python	1		c++	2
utilisé	1		java	2
traitement	1		langage	1
texte	1		langage	3
langage	3		langage	3
programmation	3		langage	2
langage	3		langage	2
utilisé	3		programmation	1
traduire	3		programmation	3
algorithme	3		programme	3
programme	3		python	1
langage	2		texte	1
java	2		traduire	3
basé	2		traitement	1
langage	2		utilisé	1
c++	2		utilisé	3

3.2) Construire le fichier inverse :

Terme	Nombre de documents	Fréquence	Pointeur
algorithme (t1)	1	1	→ Doc3, 1
basé (t2)	1	1	→ Doc2, 1
c++(t3)	1	1	→ Doc2, 1
java(t4)	1	1	→ Doc2, 1
langage (t5)	3	5	→ Doc1, 1 → Doc2, 2 → Doc3, 2
programmation (t6)	2	2	→ Doc1, 1 → Doc3, 1
programme (t7)	1	1	→ Doc3, 1
python(t8)	1	1	→ Doc1, 1
texte(t9)	1	1	→ Doc3, 1
traduire(t10)	1	1	→ Doc1, 1
traitement(t11)	1	1	→ Doc1, 1
utilisé (t12)	2	2	→ Doc1, 1 → Doc3, 1

3.3) Indexer les documents :

D1 : {Le langage de programmation python est très utilisé pour le traitement de texte}

D1 : langage, 5, programmation, 2, python, 1, utilisé, 2, traitement, 1, texte, 1

D2 : {Le langage JAVA est basé sur le langage C++}

D2 : langage, 5, java, 1, basé, 1, c++, 1}

D3 : {Un langage de programmation est un langage utilisé pour traduire un algorithme en un programme}

D3 : langage, 5, programmation, 2, utilisé, 2, traduire, 1, algorithme, 1, programme, 1

Considérer la formule suivante pour la pondération :

$$w(t_i, d_j) = \frac{freq_{ij}}{\max\{\forall t_l \in d_j\} freq_{lj}} * \log\left(\frac{N}{n_i} + 1\right)$$

Où n_i représente le nombre de documents contenant le terme t_i , N est le nombre de documents et le logarithme est calculé en base 2.

2) Calculer la similarité entre chaque document et la requête Q : {langage python java} en utilisant les quatre formules du modèle vectoriel (produit scalaire, coefficient de Dice, Cosine, et indice de Jaccard)

Produit scalaire :

$$RSV(q, d_j) = \sum_{i=1}^n w_{iq} w_{ij}$$

Coefficient de Dice :

$$RSV(q, d_j) = \frac{2 \sum_{i=1}^n w_{iq} w_{ij}}{\sum_{i=1}^n (w_{iq})^2 + \sum_{i=1}^n (w_{ij})^2}$$

Coefficient de Cosine :

$$RSV(q, d_j) = \frac{\sum_{i=1}^n w_{iq} w_{ij}}{\sqrt{\sum_{i=1}^n (w_{iq})^2} \sqrt{\sum_{i=1}^n (w_{ij})^2}}$$

Indice de Jaccard :

$$RSV(q, d_j) = \frac{\sum_{i=1}^n w_{iq} w_{ij}}{\sum_{i=1}^n (w_{iq})^2 + \sum_{i=1}^n (w_{ij})^2 - \sum_{i=1}^n w_{iq} w_{ij}}$$

Calculer la matrice de fréquences des termes-documents

	D1	D2	D3
algorithme (t1)	0	0	1
basé (t2)	0	1	0
c++(t3)	0	1	0
java(t4)	0	1	0
langage (t5)	1	2	2
programmation (t6)	1	0	1
programme (t7)	0	0	1
python(t8)	1	0	1
texte(t9)	1	0	1
traduire(t10)	0	0	1
traitement(t11)	1	0	0
utilisé (t12)	1	0	1

Produit scalaire :

$$RSV(q, d_1) = \sum_{i=1}^n w_{iq} w_{ij=1} = 1*1 + 1*2 + 0*1$$

$$RSV(q, d_1) = 3$$

$$RSV(q, d_2) = \sum_{i=1}^n w_{iq} w_{ij=2}$$

$$SV(q, d_3) = \sum_{i=1}^n w_{iq} w_{ij=1}$$

	w(D1)	w(D2)	w(D3)	w(q) (=freq)
algorithmme (t1)	0	0	1	/
basé (t2)	0	1	0	/
c++(t3)	0	1	0	/
java(t4)	0	1	0	1
langage (t5)	1	1	1	1
programmation (t6)	1.32	0	1	/
programme (t7)	0	0	0.66	/
python(t8)	2	0	0	1
texte(t9)	2	0	0	/
traduire(t10)	0	0	1	/
traitement(t11)	2	0	0	/
utilisé (t12)	1.32	0	0.66	/

Coefficient de Dice

$$RSV(q, d_1)=0.4$$

$$RSV(q, d_2)=0.85$$

$$RSV(q, d_3)=0.76$$

Coefficient de Cosine (devoir)**Indice de Jaccard (devoir)**

Exercice 6

Soit un ensemble des termes d'indexation $T = (\text{document}, \text{web}, \text{information}, \text{recherche}, \text{image}, \text{contenu})$

Avec : $d1 = (\text{document } 1,0, \text{ web } 0,5, \text{ information } 0,3)$

$q1 = (\text{document} \vee \text{web})$

$q2 = (\text{web} \wedge \text{document})$

$q3 = ((\text{web} \vee \text{document}) \wedge \text{image})$

Calculer la similarité entre $d1$ et chaque requête en considérant :

- Le modèle booléen basé sur les ensembles flous

$$RSV(q_1, d_1) = \max(1, 0.5) = 1$$

$$RSV(q_2, d_1) = \min(1, 0.5) = 0.5$$

$$RSV(q_3, d_1) ?$$

$$S1q_3 = \text{web} \vee \text{document}$$

$$S2q_3 = S1q_3 \wedge \text{image}$$

$$RSV(S1q_3, d_1) = \max(1, 0.5) = 1$$

$$RSV(S2q_3, d_1) = \min(1, 0) = 0$$

- Le modèle p_norme avec $p = 2$ (requête non pondérée de deux termes)

$$RSV(q_1, d_1) = \frac{\sqrt{1^2 + 0.5^2}}{\sqrt{2}} = 0.787$$

$$RSV(q_2, d_1) = 1 - \frac{\sqrt{(1-1)^2 + (1-0.5)^2}}{\sqrt{2}} = 0.354$$

$$RSV(q_3, d_1) = 1 - \frac{\sqrt{(1-0.787)^2 + (1-0)^2}}{\sqrt{2}} = 0.27$$