

TD1 - Indexation

Exercice 1

$tf(t_i, d)$: nombre occurrences de t_i

$$idf(t_i) = \log\left(\frac{1+N}{1+df(t_i)}\right) \quad . \quad \text{simplifié} : \log\left(\frac{N}{df}\right)$$

$$|D| = 100$$

$$df(\text{maison}) = 20 ; \quad df(\text{belle}) = 35$$

$$tf(\text{maison}, d) = 2 ; \quad tf(\text{belle}, d) = 1$$

$$1.1. \quad tf \cdot idf(\text{maison}, d) = tf(\text{maison}, d) \times idf(\text{maison})$$

$$= 2 \times \log\left(\frac{100}{20}\right) = 2 \log 5 \approx 3,6 \quad \text{d'après la table}$$

$$tf \cdot idf(\text{belle}, d) = 1 \times \log\left(\frac{100}{35}\right) = \log(2) - \log(7) = \log(2 \times 10) - \log(7) \\ \approx 0,7 + 2,3 - 2 = 1$$

Exercice 2

— Doc 1 : the new home has been sold on top forecasts

— Doc 2 : the home sales rise in july

— Doc 3 : there is an increase in home sales in july

— Doc 4 : july encounter a new home sales rise

Ainsi qu'une liste de mots vides : the, a, an, have, be, on, behind, under, there, in,

2.1. Vocabulaire : $V = \{ \text{new, home, sold, top, forecasts, sale, rise, july, increase, encounter} \}$

$$|V| = 10$$

$$2.2. \quad tf(d_1, \text{new}) = 1$$

$$tf(d_1, \text{home}) = 1$$

$$tf(d_1, \text{sold}) = 1$$

$$tf(d_1, \text{top}) = 1$$

$$tf(d_1, \text{forecasts}) = 1$$

$$tf(d_2, \text{home}) = 1$$

$$tf(d_2, \text{sold}) = 1$$

$$tf(d_2, \text{rise}) = 1$$

$$tf(d_2, \text{july}) = 1$$

$$tf(d_3, \text{increase}) = 1$$

$$tf(d_3, \text{encounter}) = 1$$

$$tf(d_4, \text{new}) = 1$$

$$tf(d_4, \text{home}) = 1$$

$$tf(d_4, \text{sold}) = 1$$

$$tf(d_4, \text{rise}) = 1$$

$$tf(d_4, \text{july}) = 1$$

$$tf(d_4, \text{encounter}) = 1$$

$$tf(d_5, \text{new}) = 1$$

$$tf(d_5, \text{home}) = 1$$

$$tf(d_5, \text{sold}) = 1$$

$$tf(d_5, \text{rise}) = 1$$

$$2.3. \quad idf(\text{new}) = \log\left(\frac{4}{2}\right) \approx 0,7$$

$$idf(\text{home}) = \log\left(\frac{4}{2}\right) = 0$$

$$idf(\text{sold}) = \log\left(\frac{4}{1}\right) = 1,4$$

$$idf(\text{top}) = \log\left(\frac{4}{1}\right) = 1,4$$

$$idf(\text{forecasts}) = \log\left(\frac{4}{1}\right) = 1,4$$

$$idf(\text{sale}) = \log\left(\frac{4}{3}\right) = 0,3$$

$$idf(\text{rise}) = \log\left(\frac{4}{2}\right) = 0,7$$

$$idf(\text{july}) = \log\left(\frac{4}{3}\right) = \log(2 \times 2) - \log 3 = 0,7 + 0,7 - 1,1 = 0,3$$

$$idf(\text{increase}) = \log\left(\frac{4}{1}\right) = 1,4$$

$$idf(\text{encounter}) = \log\left(\frac{4}{1}\right) = 1,4$$

$$docs \text{ sur } tf = 4 \dots$$

donc il y

à eux appartient

doc.

$$df(\text{home}) = 4$$

$$df(\text{new}) = 2$$

$$df(\text{sold}) = 1$$

$$df(\text{top}) = 1$$

$$df(\text{forecasts}) = 1$$

$$df(\text{sale}) = 3$$

$$df(\text{rise}) = 2$$

$$df(\text{july}) = 3$$

$$df(\text{increase}) = 1$$

$$df(\text{encounter}) = 1$$

2.4. Index inverse :

new (1; 0,7), (4; 0,7)
 july (2; 0,3), (3; 0,3), (4; 0,3)
 home (1; 0), (2; 0), (3; 0), (4; 0)
 sold (1; 1,4)
 top (1; 1,4)
 forecasts (1; 1,4)
 sale (2; 0,3); (3; 0,3); (4; 0,3)
 rise (2; 0,7); (4; 0,7)
 increase (3; 1,4)
 encounter (4; 1,4)

Index : représentation simple des documents

d1	(t1, n11); (t2, n12);
d2	(t1, n21); (t2, n22);
...	
dk	(t1, nk1); (t2, nk2);

Index inversé : point d'entrée par les mots

t1	(d1, n11); (d3, n13);
t2	(d4, n24); (d5, n25);
...	
tj	(d1, dj1); (d7, dj2);

Index :

d_1 (new; 0,7), (home; 0), (sold; 1,4), (top; 1,4), (forecasts; 1,4)
 d_2 (home; 0), (sale; 0,3), (rise; 0,7), (july; 0,3)
 d_3 (increase; 1,4), (home; 0), (sale; 0,3), (july; 0,3)
 d_4 (july; 0,3), (encounter; 1,4), (new; 0,7), (home; 0), (sale; 0,3), (rise; 0,7)

2.5. But : $\|q\| = 1$

$$d_1 = \begin{pmatrix} 0,7 \\ 0 \\ 1,4 \\ 1,4 \\ 1,4 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\tilde{d}_1 = \frac{d_1}{\|d_1\|} = \frac{d_1}{\sqrt{0,7^2 + 1,4^2 + 1,4^2 + 1,4^2}}$$

$$\frac{q}{\|q\|} \cdot \frac{d}{\|d\|} = \cos(q, d)$$

tf-idf avec produit scalaire : $q \cdot d = \sum_{i=1}^{|V|} q_i \cdot d_i$

tf-idf avec angle

2.6. Q2 : "july new"

Q1 : "sale home"

on regarde l'index inversé.

→ réponse : d_4 score 1
 d_1 score 0,7
 d_2 score 0,3
 d_3 score 0,3

en regardant l'index normal :

d_1 : 0,7
 d_2 : 0,3
 d_3 : 0,3
 d_4 : 1

d_1 : 0
 d_2 : 0,3
 d_3 : 0,3
 d_4 : 0,3

suite l.5. donc si j'ai compris

$$d_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0,3 \\ 0,3 \\ 0,7 \\ 0 \\ 0 \end{pmatrix} \quad \tilde{d}_2 = \frac{d_2}{\sqrt{0,3^2 + 0,3^2 + 0,7^2}}$$

$$d_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0,3 \\ 0,3 \\ 0 \\ 0 \\ 1,1 \\ 0 \end{pmatrix} \quad \tilde{d}_3 = \frac{d_3}{\sqrt{0,3^2 + 0,3^2 + 1,1^2}}$$

$$d_4 = \begin{pmatrix} 0,7 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0,3 \\ 0,3 \\ 0,7 \\ 0 \\ 0 \\ 1,1 \end{pmatrix} \quad \tilde{d}_4 = \frac{d_4}{\sqrt{0,7^2 + 0,3^2 + 0,3^2 + 0,7^2 + 1,1^2}}$$

Exercice 4

4.1. Soit k le rang du mot.

$$f(k) = \frac{\lambda}{k^s}, \quad \lambda, s \in \mathbb{R}^{+*}$$

$$\begin{aligned} f(1) &= 5000000 = 5 \cdot 10^6 = \lambda \\ f(10) &= 500000 = 5 \cdot 10^3 = \frac{\lambda}{10^s} \end{aligned} \Rightarrow 5 \cdot 10^{5+s} = 5 \cdot 10^6 \Leftrightarrow s = 1$$

• On a $P(X_k=1) \propto \frac{1}{k}$ et X_k forment une partition de l'univers de proba

$$\text{d'où } \sum P(X_k=1) = 1 \propto \sum_{k=1}^M \frac{1}{k} \sim \log M$$

$$\text{et donc } P(X_k=1) \sim \frac{1}{k \log M}$$

4.2. Quel est le nombre moyen d'apparition du mot le plus fréquent dans un doc. de taille 416?

$$\rightarrow \text{loi binomiale } Y \sim B\left(\frac{1}{\log M}, 416\right)$$

$$E(Y) = \frac{416}{\log M} \approx 30,8$$

4.3. $Y_k = \# \text{ d'apparition du } k^{\text{e}} \text{ mot}$

$$E(Y_k) = \frac{416}{k \log M} \text{ décroissant}$$

On cherche K tel que $E(Y_K) \geq 1$.

$$\Leftrightarrow K \leq \frac{416}{\log M}$$

$$\Leftrightarrow K^* = 30$$

Exercice 3

a	(1,1)	(2,2)	(3,2)	(4,3)	(5,2)
b	(2,7)	(10,5)			

3.1. index TAAT : $w_{t,i} \geq w_{t,i+1}$

a	(4,3)	(2,2)	(3,2)	(5,2)	(1,1)
b	(2,7)	(10,5)			

Etape	(4,3)	(2,2)	(3,2)	(5,2)	(1,1)	(2,7)	(10,5)	
Ebat	{4:3}	{2:2}	{3:2}	{5:2}	{1:1}	{2:7}	{10:5}	→ Heap (tri)

\Rightarrow Etap 1 : (d_2, g)

3.2. $\max(a) : 3$
 $\max(b) : 7$

on traite les scores les plus grands en premier.

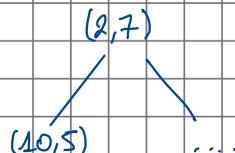
- On commence par la liste b :

(2,7) (10,5)

$D(2,7)_b$ on ajoute le document 2

$$2 : 7 \quad (10) \quad \Leftrightarrow 7 < d_2 \leq 10$$

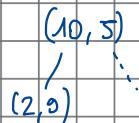
$D(10,5)_b \rightarrow s(d_{10}) \in [5,8]$ potentiel candidat, on l'ajoute



$D(4,3)_a \rightarrow s(d_4) \in [3,3]$

$\hookleftarrow s \otimes$

$D(2,2)_a \rightarrow s(d_2) \in [9,9]$



on pourrait s'arrêter là en fin de l'algorithme

$D(3,2)_a \rightarrow s(d_3) \in [2,2]$

$D(1,1)_a \rightarrow s(d_1) \in [1,1]$

$D(5,2)_a \rightarrow s(d_5) \in [2,2]$

on le délimite.

3.3. ($k=3$) ($\text{top } 3$)

$$\begin{array}{c} a \quad (1,1) \quad (2,2) \quad (3,2) \quad (4,3) \quad (5,2) \\ b \quad (2,7) \quad (10,5) \end{array}$$

- $(\underline{1},1)_a \quad (2,7)_b \quad \text{candidate} = 1 \quad s(d_1) = 1$

$$\begin{matrix} 1,1 \\ / \quad \backslash \\ - \quad - \end{matrix}$$

- On avance les cursus dont le doc = 1

$$(\underline{2},2)_a \quad (\underline{2},7)_b \quad s(d_2) = 9$$

$$\begin{matrix} 1,1 \\ / \quad \backslash \\ 2,9 \quad \dots \end{matrix}$$

- On avance les cursus dont le doc = 2

$$(\underline{3},2)_a \quad (10,5)_b \quad \text{candidate} = 3 \quad s(d_3) = 2$$

$$\begin{matrix} 1,1 \\ / \quad \backslash \\ 2,9 \quad 3,2 \end{matrix}$$

- $(\underline{4},3)_a \quad (10,5)_b \quad s(d_4) = 3$

$$\begin{matrix} 3,2 \\ / \quad \backslash \\ 2,9 \quad 4,3 \end{matrix}$$

- $(5,2)_a \quad (10,5)_b \quad s(d_5) = 2$

- $(10,5)_b \quad s(d_{10}) = 5$

$$\begin{matrix} 4,3 \\ / \quad \backslash \\ 2,9 \quad 10,5 \end{matrix}$$

$$\begin{matrix} \text{top 3 :} \\ (\text{fini}) \end{matrix} \quad \begin{matrix} 2,9 \\ 10,5 \\ 4,3 \end{matrix}$$

34. WAND ($k=1$) . On connaît le max des lists comme dans TAAFT.

- $(1,1)_a \quad (2,7)_b \quad \theta = 0$

$3 > \theta$ on regarde le score maximum qu'un doc peut avoir

$$\Rightarrow s(d_1) = 1$$

Maj Heap : $1,1 \rightsquigarrow \theta = 1$

- $(0,2)_a \quad (2,7)_b$

$$3 > \theta$$

$$\Rightarrow s(d_2) = 9 \quad 2,9 \rightsquigarrow \theta = 9$$

- $(3,2)_a \quad (0,5)_b$

$$3 < \theta$$

$$10 > \theta$$

pivot = 10



on considère le 2^e document

- $\otimes \quad (10,5)_b$

on s'arrête.

$$7 < \theta$$

on ne calcule pas les scores des documents 3, 4, 5

et si il y avait plus de documents dans L, on se serait arrêté à d₁₀.

TD2 - Modèles d'ordonnancement

Exercice 1

$$1.1. \text{ RSV}(q_1, d_1) = 1 \wedge 0 \wedge 1 = 0$$

$$\text{RSV}(q_1, d_2) = 1 \wedge 1 \wedge 2 = 1$$

$$\text{RSV}(q_1, d_3) = 0 \wedge 1 \wedge 2 = 0$$

$$\text{RSV}(q_2, d_1) = 1 \wedge (1 \vee 2) = 1 \wedge 1 = 1$$

$$\text{RSV}(q_2, d_2) = 2 \wedge (1 \vee 0) = 2 \wedge 1 = 1$$

$$\text{RSV}(q_2, d_3) = 2 \wedge (0 \vee 0) = 2 \wedge 0 = 0$$

$$\text{RSV}(q_3, d_1) = 0$$

$$\text{RSV}(q_3, d_2) = 3 \wedge ((1 \vee 0) \vee 0) = 3 \wedge (1 \vee 0) = 1$$

$$\text{RSV}(q_3, d_3) = 2 \wedge ((0 \vee 0) \vee 0) = 0$$

1.2.

$$d_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}, \quad d_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 2 \\ 0 \\ 3 \end{pmatrix}, \quad d_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 2 \\ 0 \\ 2 \end{pmatrix}$$

$$q_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow$$

$$\cos(q_1, d_1) = \frac{q_1 \cdot d_1}{\|q_1\| \cdot \|d_1\|} = \frac{q_1}{\sqrt{2}} \cdot \frac{d_1}{\sqrt{10}}$$

$$= \frac{0 \times 0 + 1 \times 1 + 0 \times 0 + 2 \times 0 + \dots + 1 \times 2 + 2 \times 0 + 0}{\sqrt{2} \sqrt{10}}$$

$$= \frac{2}{\sqrt{2} \sqrt{10}} \approx 0,145$$

$$\cos(q_1, d_2) = \frac{q_1}{\sqrt{2}} \cdot \frac{d_2}{\sqrt{10}} = \frac{1+2}{4\sqrt{2}} = \frac{3}{4\sqrt{2}} \approx 0,153$$

$$\cos(q_1, d_3) = \frac{q_1}{\sqrt{2}} \cdot \frac{d_3}{\sqrt{10}} = \frac{2}{\sqrt{2} \sqrt{10}} \approx 0,13$$

$$q_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \cos(q_2, d_1) = \frac{q_2}{\sqrt{5}} \cdot \frac{d_1}{\sqrt{10}} = \frac{2+1}{\sqrt{5} \sqrt{10}} = \frac{3}{\sqrt{5} \sqrt{10}} \approx 0,142$$

$$\cos(q_2, d_2) = \frac{q_2}{\sqrt{5}} \cdot \frac{d_2}{\sqrt{10}} = \frac{2+2}{4\sqrt{5}} = \frac{1}{\sqrt{5}} \approx 0,15$$

$$\cos(q_2, d_3) = \frac{q_2}{\sqrt{5}} \cdot \frac{d_3}{\sqrt{10}} = \frac{1}{\sqrt{5} \sqrt{10}} = 0,127$$

$$q_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

 \Rightarrow

$$\cos(q_3, d_1) = \frac{q_3}{\sqrt{3}} \cdot \frac{d_1}{\sqrt{10}} = \frac{1}{\sqrt{3}\sqrt{10}} \approx 0,18$$

$$\cos(q_3, d_2) = \frac{q_3}{\sqrt{3}} \cdot \frac{d_2}{4} = \frac{1+1+3}{4\sqrt{3}} = \frac{5}{4\sqrt{3}} \approx 0,32$$

$$\cos(q_3, d_3) = \frac{q_3}{\sqrt{3}} \cdot \frac{d_3}{\sqrt{11}} = \frac{3}{\sqrt{3}\sqrt{11}} \approx 0,52$$

$$1.3. \text{ BIM}(t) = \log \frac{P(t|R)}{P(t|\neg R)}$$

d_1 et d_2 sont pertinents

$$q_t = \{t_2, t_8\}$$

$$s(d, q) = \sum_{t \in q \cap d} \text{BIM}(t)$$

$$s_{\text{BIM}}(d_1) = \text{BIM}(t_2) + \text{BIM}(t_8) = s_{\text{BIM}}(q_2)$$

$$s_{\text{BIM}}(d_3) = \text{BIM}(t_8)$$

t_2	R	$\neg R$
$t_2 E d$	$1+\varepsilon$	$1+\varepsilon$
$t_2 \notin d$	$1+\varepsilon$	$0+\varepsilon$
	2	$1-N-m_t$
	$\frac{2}{N}$	$\frac{1}{N-R}$

$$P(t_2|R) \approx \frac{1}{2} \approx \frac{1}{2}$$

$$P(t_2|\neg R) \approx 1 \approx \frac{3}{4}$$

(La priorité de Dirichlet)
avec lissage, $\varepsilon = 0,5$

	t_2	t_8
d_1	1	1
d_2	1	2
d_3	0	2

$$\text{BIM}(t_2) = \log \left(\frac{n_t + \varepsilon}{R - n_t + \varepsilon} \times \frac{N - R - m_t + n_t + 0,5}{m_t - n_t + 0,5} \right)$$

$$= \log \left(\frac{1+0,5}{2-1+0,5} \times \frac{3-2-2+1+0,5}{2-1+0,5} \right) = \log \left(\frac{0,5}{1,5} \right) = \log \left(\frac{1}{3} \right)$$

= $-\log(3)$?? régatif ???

dans un document pertinent (ici d_1 et d_3)
dans un document non pertinent (ici d_2)

terme présent
terme absent

t_8	R	$\neg R$
$t_8 E d$	2	1
$t_8 \notin d$	0	0
	2	1
	$\frac{2}{R}$	

$$p_t = \frac{2+0,5}{2} = 1,25$$

$$u_t = \frac{2-2+0,5}{1} = 0,5$$

$$\text{BIM}(t_8) = \log \left(\frac{n_t + \varepsilon}{R - n_t + \varepsilon} \times \frac{N - R - m_t + n_t + 0,5}{m_t - n_t + 0,5} \right) = \log \left(\frac{2+0,5}{2-2+0,5} \times \frac{8-2-3+2+0,5}{3-2+0,5} \right)$$

$$= \log \left(\frac{2,5}{0,5} \times \frac{0,5}{1,5} \right) = \log \left(\frac{5}{3} \right) = \log(5) - \log(3)$$

$$s_{\text{BIM}}(d_1, q) = s_{\text{BIM}}(d_2, q) = \log(5) - 2\log(3) \approx -0,26 \quad ??$$

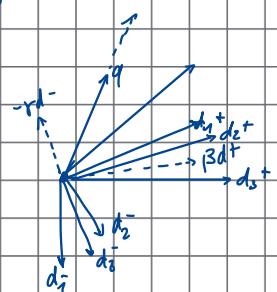
$$s_{\text{BIM}}(d_3, q) = \log(5) - \log(3) \approx 0,22$$

Exercise 2

$$2.1 \quad q_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$q_{\text{nocchio}} = \alpha q + \beta d^+ - \gamma d^-$$

$$\frac{1}{10+1} \sum_{i=1}^{10+1} d_i^+$$



$$d^+ = (d_1, d_3) = \frac{1}{10+1} \sum_{i=1}^2 d_i^+ = \frac{1}{2} (d_1 + d_3)$$

$$= \left(\frac{1}{2} \ 1 \ 0 \ 1 \ \frac{1}{2} \ 0 \ \frac{1}{2} \ \frac{3}{2} \ 1 \ 1 \right)$$

$$d^- = d_2 = (1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 2 \ 0 \ 3)$$

$$\Rightarrow q_{\text{nocchio}} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 0,5 \begin{pmatrix} 1/2 \\ 1/2 \\ 0 \\ 1 \\ 1/2 \\ 0 \\ 1/2 \\ 3/2 \\ 1 \\ 1 \end{pmatrix} - 0,3 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 2 \\ 0 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 1/4 - 0,3 \\ 1 + 1/4 - 0,3 \\ 0 \\ 0,5 \\ 1/4 \\ 0 \\ 1/4 - 0,3 \\ 1 + 3/4 - 0,3 \cdot 2 \\ 1/2 \\ 1/2 - 0,9 \end{pmatrix} = \begin{pmatrix} -0,05 \\ 0,95 \\ 0 \\ 0,5 \\ 0,25 \\ 0 \\ -0,05 \\ 1,15 \\ 0,5 \\ -0,4 \end{pmatrix}$$

$$\|q_n\| = \sqrt{-0,05^2 + 0,95^2 + \dots + 0,4^2} = 1,718$$

$$\text{sc}(q_{\text{nocchio}}, d_1) = \cos(q_n, d_1) = \frac{q_n \cdot d_1}{\|q_n\| \|d_1\|}$$

$$= 0,95 + 2 \cdot 0,5 + 1,15 \cdot 2 \cdot 0,5 = \frac{4,1}{1,718 \times 3,16} \approx 0,7545$$

$$s(q_1, d_2) \approx 0,2837$$

$$s(q_n, d_3) \approx 0,2895$$

(python + faible flemme de détailler)

$$\Rightarrow d_1 > d_3 > d_2$$

d_1 bat tout le monde.

2.2 ???

On a reformuler notre requête mais il faut la binariser et on fixe un seuil, par exemple 0,5.

$$\Rightarrow q_R = t_2, t_4, t_8$$

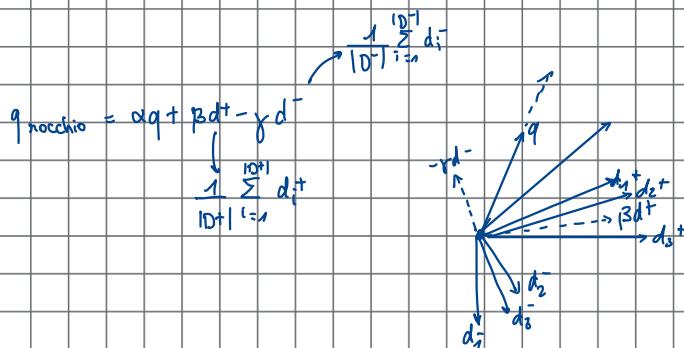
	t_4
d_1	2
d_2	0
d_3	0

ici, seul le score de d_1 changera.

il suffit ensuite de calculer $BIN(t_4) \dots$

t_4	R	γR	
$t_2 Ed$	1+ε	1+ε	$2 = n_t$
$t_2 Bd$	1+ε	0+ε	$1 = N - n_t$
	R	1	etc...
	0	$\frac{n}{N} - R$	

Notes



$$\cosine(q, d_i) = \frac{q \cdot d_i}{\|q\| \|d_i\|}$$

Calcul BIR dinmo :

notes :

$$s(q, d) = \sum_{t \in q \cap d} \log \frac{p_t(1 - u_t)}{u_t(1 - p_t)}$$

$$p_t = \frac{n_t}{R}, \quad u_t = \frac{m_t - n_t}{N - R}$$

$$s(q, d) = \sum_{t \in q \cap d} \log \frac{r_t + 0.5}{R - r_t + 0.5} \times \frac{N - n_t - R + r_t + 0.5}{n_t - r_t + 0.5}$$

$$\frac{\pi_L}{R} \times \frac{N-R}{m_L - \pi_L} \times \frac{N-R-m_L+\pi_L}{N-R} \times \frac{R}{R-\pi_L}$$

avec lissage

$$= \frac{n_L + \epsilon}{R - \pi_L + \epsilon} \times \frac{N - R - m_L + \pi_L + 0.5}{m_L - \pi_L + 0.5} \quad \text{ok}$$

TD3 - Evaluation d'un SRI

Exercice 1

$$\text{1.1. Precision} = \frac{tp}{tp + fp}$$

$$\text{Rappel} = \frac{tp}{tp + fn}$$

18 documents pertinents
20 documents retrieved
1000 documents au total

	Relevant	Not relevant
Retrieved	TP	FP
~Retrieved	FN	TN

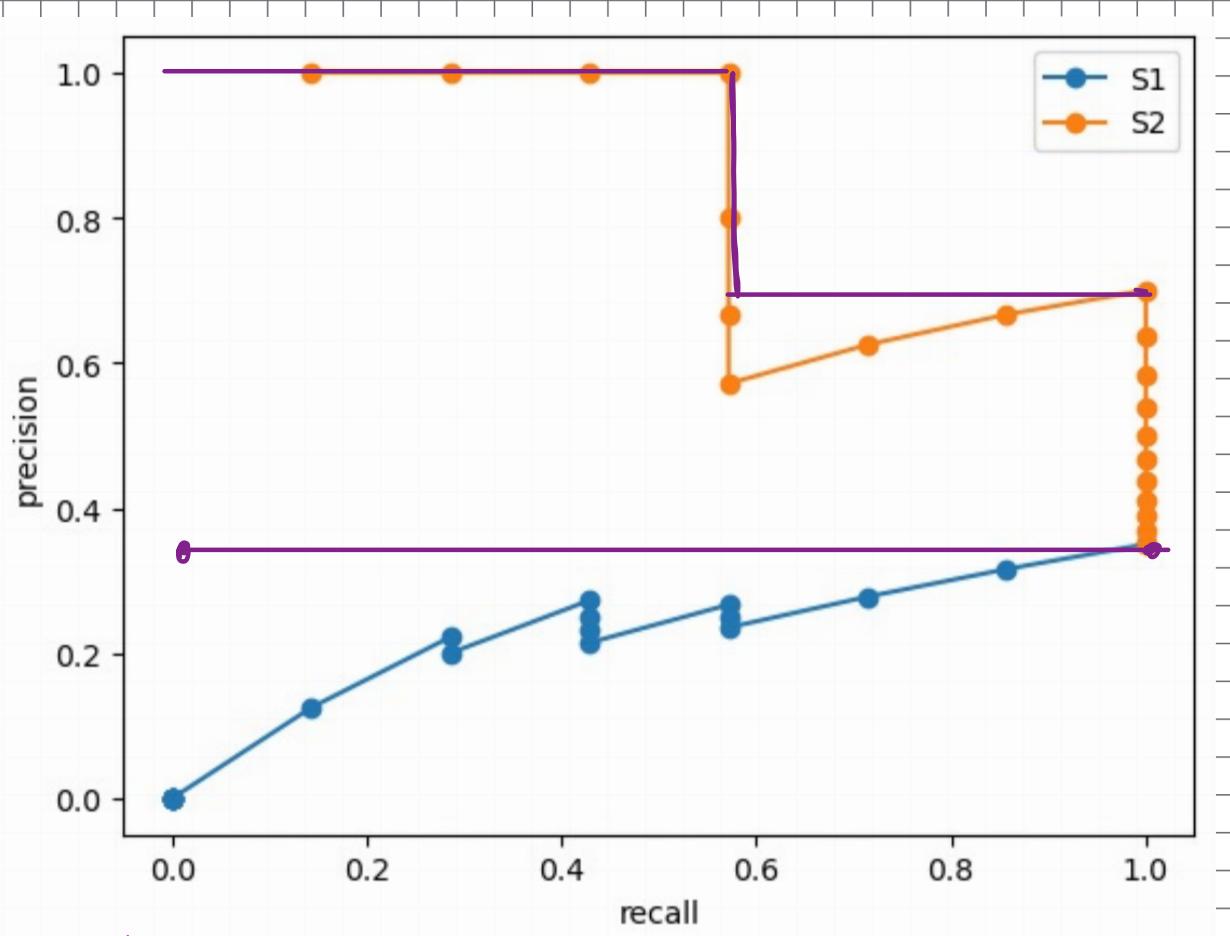
$$S_1 : 7 \text{ documents pertinents} \rightarrow \text{Rappel}_{S_1} = \frac{7}{12} \approx 0,583$$

$$\text{Precision}_{S_1} = \frac{7}{20} \approx 0,35$$

$$S_2 : 7 \text{ documents pertinents} \rightarrow \text{Rappel}_{S_2} = \frac{7}{12}$$

$$\text{Precision}_{S_2} = \frac{7}{20}$$

	S ₁	Rappel	Precision	Precision interpolée	S ₂	Rappel	Precision	Precision interpolée
1	N	0	0 = 0/1	7/20	P	1/7	1	1
2	N	0	0 = 0/2		P	2/7	1 = 2/2	1
3	N	0	0 = 0/3		P	3/7	1 = 3/3	1
4	N	0	0 = 0/4		P	4/7	1 = 4/4	1
5	N	0	0		N	4/7	4/5	4/5
6	N	0	0		N	4/7	4/6	7/10
7	N	0	0		N	4/7	4/7	
8	P	1/7	1/8		P	5/7	5/8	
9	P	2/7	2/9		P	6/7	6/9	
10	N	2/7	2/10		P	1	7/10	
11	P	3/7	3/11		N	1	7/11	
12	N	3/7	3/12		N	1	7/12	
13	N	3/7	3/13		N	1	7/13	
14	N	3/7	3/14		N	1	7/14	
15	P	4/7	4/15		N	1	7/15	
16	N	4/7	4/16		N	1	7/16	
17	N	4/7	4/17		N	1	7/17	
18	P	5/7	5/18		N	1	7/18	
19	P	6/7	6/19		N	1	7/19	
20	P	1	7/20		N	1	7/20	



en violet la précision interpolée.

⇒ S_2 a une meilleure performance que S_1 pour tous les points de rappel - Cela signifie que S_2 récupère plus de documents pertinents que S_1 à chaque niveau de rappel, tout en maintenant une précision globalement plus élevée.

$$1.3. \quad NDCG_p = \frac{DCG_p}{IDCG_p} \quad DCG = \text{rel}_1 + \sum_{i=2}^P \frac{\text{rel}_i}{\log_2(i)} = \sum_{i=1}^P \frac{\text{rel}_i}{\log_2(i+1)}$$

Pour exemple, on assigne 1 aux documents pertinents et 0 aux documents très pertinents. et 0 aux documents non pertinents.

$$\begin{aligned} & (d_1, \dots, d_4) \\ & \downarrow \\ & (n_1, \dots, n_4) \in \{0, 1\} \end{aligned}$$

non pertinent
pertinent

$$\begin{cases} P@k = \sum_{i=1}^k n_i / k \\ R@k = \sum_{i=1}^k n_i / R_{\text{total}} \end{cases}$$

$$P_{\text{interpolé}}@R = \max_{k/R \geq R} P@k$$

	S_1	reli:	$\text{reli:}/\log_2(i+1)$		S_2	reli:	$\text{reli:}/\log_2(i+1)$		$\log_2(i+1)$
1	N	0	0		TP	2	2		1
2	N	0	0		TP	2	1,26		1,58
3	N	0	0		TP	2	1		2
4	N	0	0		TP	2	0,86		2,32
5	N	0	0		N	0	0		2,58
6	N	0	0		N	0	0		2,81
7	N	0	0		N	0	0		3
8	TP	2	0,63		P	1	0,32		3,17
9	P	1	0,3		TP	2	0,6		3,32
10	N	0	0		P	1	0,29		3,46
11	TP	2	0,56		N	0	0		3,58
12	N	0	0		N	0			3,7
13	N	0	0		N	0			3,81
14	N	0	0		N	0			3,91
15	P	1	0,25		N	0			4,0
16	N	0	0		N	0			4,09
17	N	0	0		N	0			4,17
18	TP	2	0,47		N	0			4,38
19	P	1	0,23		N	0			4,32
20	P	1	0,23		N	0	0		4,39

$$\text{DCG}_{20}(S_1) = 2,67$$

$$\text{DCG}_{20}(S_2) = 6,33$$

Ideal (IDCG):

$$\text{pour } S_1: 2, 2, 2, 1, 1, 1, 1, 0 \dots 0 \rightarrow \text{IDCG}_{20}(S_1) = 5,77$$

$$\text{pour } S_2: 2, 2, 2, 2, 2, 1, 1, 0 \dots 0 \rightarrow \text{IDCG}_{20}(S_2) = 6,59$$

$$\text{donc: IDCG}_{20}(S_2) \sim \frac{2,67}{5,77} = 0,46$$

$$\text{IDCG}_{20}(S_2) \sim 0,96$$

1.4. S_2 a une performance nettement supérieure à S_1 . Cela renforce notre interprétation de la courbe de rappel/précision interpolé, qui montre également que S_2 avait de meilleures performances que S_1 .

IDCG prend en compte la pertinence graduelle des documents ce qui signifie qu'il tient en compte de la différence entre le niveau de relevance des documents (P, TP, ...).

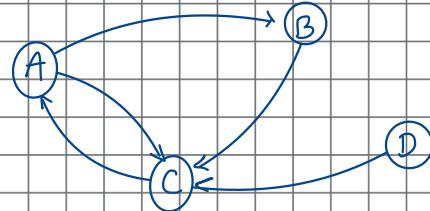
Exercice 2

On cherche le SR1 qui a une meilleure performance en terme de rappel = un SR1 qui récupère autant d'articles pertinents que possible.

En observant les courbes de précision-rappel, on observe que pour un rappel élevé, la précision interpolée du modèle 2 est supérieure à celle du 1. Il semble être donc le plus pertinent car plus susceptible de récupérer une plus grande proportion d'articles pertinents sans en rater.

TD4 - Page Rank

Exercice 1



Matrice d'adjacence :

$$A = \begin{pmatrix} & A & B & C & D \\ A & 0 & 1 & 1 & 0 \\ B & 0 & 0 & 1 & 0 \\ C & 1 & 0 & 0 & 0 \\ D & 0 & 0 & 1 & 0 \end{pmatrix}$$

soit

$$\begin{aligned} d_A &= 2 \\ d_B &= 1 \\ d_C &= 1 \\ d_D &= 1 \end{aligned}$$

(nombre de liens entrants)

d'où la matrice de transition

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

version matricielle

$$s_j = 0,85 \sum_i p_{ij} s_i + 0,15 = 0,85 sP + 0,15 \quad \text{nechif: c'est } \frac{1-d}{N} = \frac{0,15}{4}$$

 s_j correspond à la probabilité que la page j soit importante

pour avoir une sortie qui somme à 1

(0) On initialise : $s_A = s_B = s_C = s_D = \frac{1}{4}$

(1) $s_A = 0,85 \left(p_{AA} s_A + p_{BA} s_B + p_{CA} s_C + p_{DA} s_D \right) + \frac{(1-d)}{N} = 0,25$

$$s_B = 0,85 \left(\frac{1}{2} \times \frac{1}{4} \right) + \frac{0,15}{4} = 0,14375$$

$$s_C = 0,85 \left(\frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{4} + \frac{1}{4} \right) + \frac{0,15}{4} = 0,56875$$

$$s_D = 0,0375 = \frac{0,15}{4}$$

(2) on matricielle (think Numpy) $d \times s @ P + (1-d)/N$

$$0,85 \times \begin{pmatrix} 0,25 & 0,14375 & 0,56875 & 0,0375 \end{pmatrix} P + 0,15$$

$$= \begin{pmatrix} 0,15209 \\ 0,14375 \\ 0,2970125 \\ 0,0375 \end{pmatrix} \begin{matrix} s_A \\ s_B \\ s_C \\ s_D \end{matrix}$$

$$\begin{pmatrix} 0,3725 \\ 0,1958 \\ 0,3842 \\ 0,0375 \end{pmatrix}$$

 $C > A > B > D$

Après une 20aine d'itérations, on converge vers :

Exercice 2

Pour adapter l'algorithme PageRank à un graphe hétérogène comprenant de articles scientifiques et des auteurs, vous pouvez créer un modèle qui prend en compte les relations entre les deux types d'entités.

Approche possible :

1. créer une matrice d'adjacence combinée qui représente les relations entre les auteurs et les articles, ainsi que les relations de citation entre les articles et les auteurs.
Dans cette matrice, chaque ligne et chaque colonne représente une entité.
Avec 5 articles, 4 auteurs \rightarrow matrice taille 9×9 .

2. Normaliser les colonnes de la matrice d'adjacence pour créer la matrice de transition

3. Initialiser les scores PageRank et mettre à jour jusqu'à convergence.

4. Ordonner les auteurs et articles en réponse à un besoin en information.

Cette approche attribue un score PR à la fois aux auteurs et articles.

jsp si c'est ce qui est demandé mdr.

On souhaite gérer 2 types de liens

\rightarrow un score par noeud mais on modifie la formule

$$s_j = \frac{1-d}{N} + d \left(\sum_j p_{ij} s_j + \sum_k p_{ki} s_k \right)$$

où j parcourt les articles qui citent l'article i
 k auteurs

Somme de gauche : prend en compte les liens de citation
 Somme de droite : d'auteur

donc on attribue un score à chaque noeud en prenant en compte les 2 types de lien

