

Exercice N°1

A. On considère la collection de documents suivants qui correspondent à l'acte I, scène 1 de la pièce « Romeo et Juliette » de Shakespeare

D1 : Do you quarrel, sir ?

D2 : Quarrel sir ! no, sir

D3 : If you do sir, I am for you: I serve as good a man as you

D4 : No better

D5 : Well, sir

Questions

[Q1] Construire un extrait de fichiers inverses classiques de cette collection en considérant les mots *a*, *better*, *sir*, *you*, *do*

[Q2] Construire un extrait de l'index étendu basé sur : a) les positions, b) bigrammes caractères (mettre en avant les bigrammes qui marquent la différence avec l'index classique basé mots).

[Q3] Construire la matrice document-mot basé sur le schéma de pondération TF-IDF de type TF-Max

1) mots/dico

	Df	TTF
<i>a</i>	1	1 → [3:1]
<i>better</i>	1	1 → [4:1]
<i>sir</i>	4	5 → [2:1] → [2:2] → [3:1] → [5:1]
<i>you</i>	2	4 → [1:1] → [3:3]
<i>do</i>	2	2 → [1:1] → [3:1]

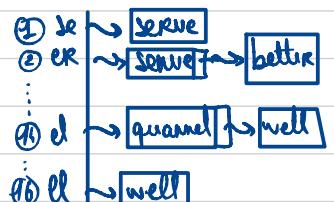
2) a) les positions:

mot	Df	TTF
<i>a</i>	1	1 → [3:<1:8>]
<i>sir</i>	4	5 → [1:<1:4>] → [2:<2:2;4>] → [3:<1:4>] → [5:<1:2>]

b) bigrammes caractères: Soient les mots: *serve*, *better*, *quarrel*, *well*

bigrammes ("serve") : { 'se', 'er', 'rv', 've', 've' }
 bigrammes ("better") : { 'be', 'er', 'tt', 'te', 'er' }
 bigrammes ("quarrel") : { 'qu', 'ua', 'ar', 're', 'el' }
 bigrammes ("well") : { 'we', 'el', 'll' }

mot	Df	TTF
<i>serve</i>	1	1
<i>better</i>	1	1
<i>quarrel</i>	2	2
<i>well</i>	1	1



3) Tf-max:

$$TF_{max}(t, d) = \frac{n(t, d)}{n(t_{max}, d)}$$

$$IDF(w) = \log \left(\frac{N}{D_f(w)} \right)$$

D1 : Do you quarrel, sir ?

D2 : Quarrel sir ! no, sir

D3 : If you do sir, I am for you: I serve as good a man as you

D4 : No better

D5 : Well, sir

N=5

On suppose:

- élimination mots vides

- Seuls les mots soulignés sont dans l'index

mot	Idf
<i>better</i>	$\log(5/1)$
<i>good</i>	$\log(5/1)$
<i>quarrel</i>	$\log(5/2)$

Dico	Tf max
D1	2
D2	2
D3	1
D4	1
D5	1

	better	good	quand ^{TF}	mar ^{IDF}	mar	six	venue
D ₁	0	0	$\frac{1}{2} \log(5/2)$	0	$\frac{1}{2} \log(5/4)$	0	0
D ₂	0	0	$\frac{1}{2} \log(5/2)$	0	$\frac{2}{2} \log(5/4)$	0	0
D ₃	0	$\frac{1}{2} \log(5)$	0	$\frac{1}{2} \log(5)$	$\frac{1}{2} \log(5/4)$	$\frac{1}{2} \log(5)$	0
D ₄	$\frac{1}{2} \log(5)$	0	0	0	0	0	0
D ₅	0	0	0	0	$\frac{1}{2} \log(5/4)$	0	0

One-hot. Vecteurs documents:

	better	good	quand	mar	six	venue
D ₁	(0, 0, log(5/2), 0, log(5/4), 0)					
D ₂	(0, 0, $\frac{1}{2} \log(5/2)$, 0, log(5/4), 0)					
D ₃	(0, log(5), 0, log(5), log(5/4), log(5))					
D ₄	(log(5), 0, 0, 0, 0, 0)					
D ₅	(0, 0, 0, 0, log(5/4), 0)					

Exercice N°=2

Soit le corpus de documents suivants :

D₁ : 'Réussite et résultat du travail et de la volonté'

D₂ : 'Pas de réussite sans la chance dans la vie'

D₃ : 'La réussite est le résultat du travail, la chance est un plus'

D₄ : 'Réussite et chance font bon ménage'

Questions

[Q1] Calculer les vecteurs des mots 'chance', 'réussite' et 'travail' en utilisant la mesure PPMI basée sur une fenêtre de 3 mots. Précision : ne considérer que les mots soulignés.

[Q2] (i) Calculer la similarité sémantique de la paire de mots ('chance', 'réussite') et ('travail', 'réussite') en utilisant une mesure de similarité vectorielle ;

(ii) D'après vous, obtiendrait-on les mêmes tendances avec des similarités basées sur des vecteurs TF-IDF ou Word2Vec ? Préciser et expliquer brièvement 2 facteurs qui impacteraient ces tendances.

1) D₁: w₁: 'réussite, résultat, travail';
w₂: 'résultat, travail, volonté'

D₂: w₃: 'réussite, chance, vie'

D₃: w₄: 'réussite, résultat, travail'
w₅: 'résultat, travail, chance'

D₄: w₆: 'réussite, chance, ménage'

Méthode PPMI:

étape 1:

	chance	réussite	travail	ménage	réultat	volonté	rie	P(w)
chance	0/21	2/21	1/21	1/21	1/21	0	1/21	6/21
réussite	2/21	0	2/21	1/21	2/21	0	1/21	8/21
travail	1/21	2/21	0	0	3/21	1/21	0	7/21
P(c)	3/21	4/21	3/21	2/21	6/21	1/21	2/21	4/21

étape 2: $\frac{P_{ij}}{P_i \times P_j}$

	chance	réussite	travail	ménage	réussit volante	vie
chance	0	$\frac{2/21}{6/21 \times 4/21}$	$\frac{1/21}{6/21 \times 3/21}$		0	0
réussite	$\frac{2/21}{8/21 \times 3/21}$	0			0	0
travail			0	0		0

étape 3: $\log\left(\frac{P_{ij}}{P_i \times P_j}\right)$

étape 4: \rightarrow ppmi $\log() < 0 \rightarrow 0$

étapes:						
	chance	réussite	travail	ménage	réussit volante	vie
chance	0	$\log\left(\frac{2/21}{6/21 \times 4/21}\right)$	$\log\left(\frac{1/21}{6/21 \times 3/21}\right)$	$\log\left(\frac{1/21}{6/21 \times 2/21}\right)$	$\log\left(\frac{1/21}{6/21 \times 6/21}\right) = 0$	$\log\left(\frac{1/21}{6/21 \times 2/21}\right)$
réussite	$\log\left(\frac{2/21}{8/21 \times 3/21}\right)$	0	$\log\left(\frac{2/21}{8/21 \times 3/21}\right)$	$\log\left(\frac{1/21}{8/21 \times 2/21}\right)$	$\log\left(\frac{2/21}{8/21 \times 6/21}\right) = 0$	$\log\left(\frac{1/21}{8/21 \times 2/21}\right)$
travail	$\log\left(\frac{1/21}{4/21 \times 2/21}\right)$	$\log\left(\frac{2/21}{2/21 \times 6/21}\right)$	0	0	$\log\left(\frac{3/21}{2/21 \times 6/21}\right) = 0$	$\log\left(\frac{1/21}{2/21 \times 6/21}\right) = 0$

→ chance réussite travail ménage réussit volante vie
 chance (0; 0,24; 0,15; 0,56; 0; 0; 0,56)

réussite (0,56; 0; 0,56; 0,27; 0; 0; 0,27)

travail (0; 0,18; 0; 0; 0,18; 0,48; 0)

$$\cos(\overline{\text{chance}}, \overline{\text{réussite}}) = 0,86$$

$$\cos(\overline{\text{travail}}, \overline{\text{réussite}}) = 0$$

rappels

$$\cos(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

q2 -

$$\text{i)} \text{cossim}(\text{chance}, \text{réussite}) = 0,86$$

$$\text{cossim}(\text{travail}, \text{réussite}) = 0$$

ii) Non car différences liées à différents futurs dont

Contexte : TF-IDF → document
 PPMI → fenêtre de mots
 W2V4C → / / /

Dimension : TF-IDF, INI] vecteurs 'creux' ⇒ similitude
 PPMI] sémantique des vecteurs ≠ vecteurs denses

space INI W2V4C, INI < N] vecteurs denses
 dans un space
 + + + volumes des textos → distribution de mots ≠ donc des vecteurs ≠

⇒ Impact particulier dans le cas de génération des embeddings W2V4C, BERT