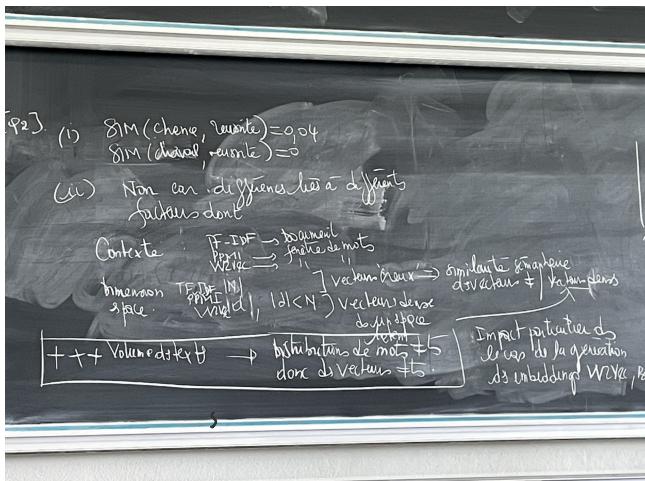


Sous-titre TD1:

Exercice 2 : Q2



Exercice 3 :

Dès Zipf $f(m) = C \times \frac{1}{m}$

fréquence du mot de rang m constante longueur du mot

$|V|$: taille du vocabulaire

$P(X_k=1) = \frac{f(k)}{\sum_{r=1}^{|V|} f(r)}$ X_k : le k ième mot le (+) fréquent rituel

et d'après la loi de Zipf: $f(k) = C \times \frac{1}{k^z}$

$P(X_k=1) = \frac{C/k}{\sum_{r=1}^{|V|} \frac{1}{r}} \rightarrow \sum_{i=1}^{|V|} \frac{1}{i} \approx \ln(|V|)$

$P(X_k=1) = \frac{1}{k \times \ln(|V|)}$

0) $|V| = 757476$
 Document 0 $\lg(0) = 4,16$

$X_{k,i}$, $i \in \{1, \dots, |V|\}$ qui apparaît
 $S_k = \sum X_{k,i} \rightarrow$ loi de Bernoulli

Nouvelle moyen d'apparitions du mot le plus fréquent dans un document de longueur: $E(S_k) = m \times p$

$$P(X_1=1) = \frac{1}{1 \times \ln(757476)} = \frac{1}{13,5}$$

$$E(S_k) = 4,16 \times \frac{1}{13,5}$$

Quelques onomatopées :

... ' ... ' ai ' in ' ne'

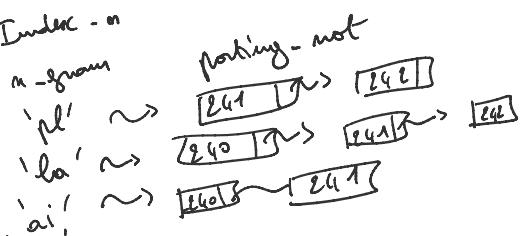
Question ouverte :

Bigramme : 'plain' → 'pl', 'la', 'ai', 'in', 'ne'
 'chain' → 'ch', 'ai', 'ai', 'in'
 'plate' → 'pl', 'la', 'at', 'ta', 'ne'

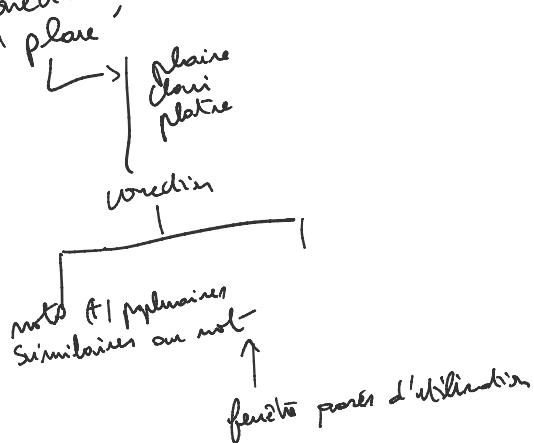
Index - m :

	<u>mot</u>	<u>DF</u>	<u>TF</u>	Posting-doc
240	chain	250		[245:60] ↗ []
241	plain	450		[120:15] ↗ []
242	plate	220		[120:10] ↗ []

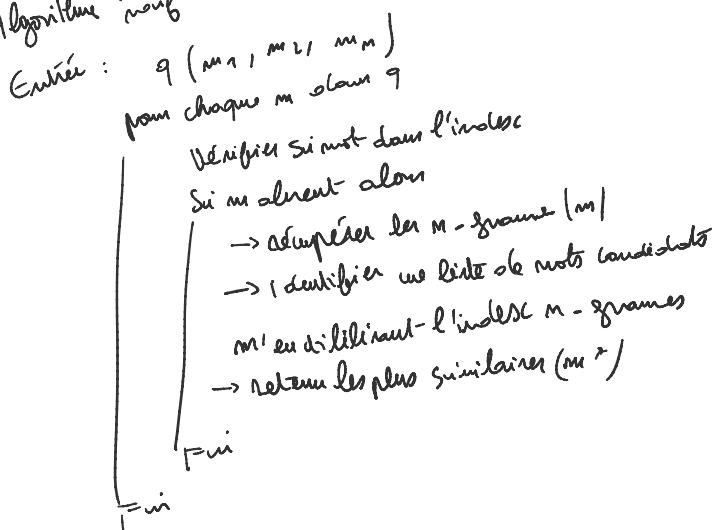
Index - n :

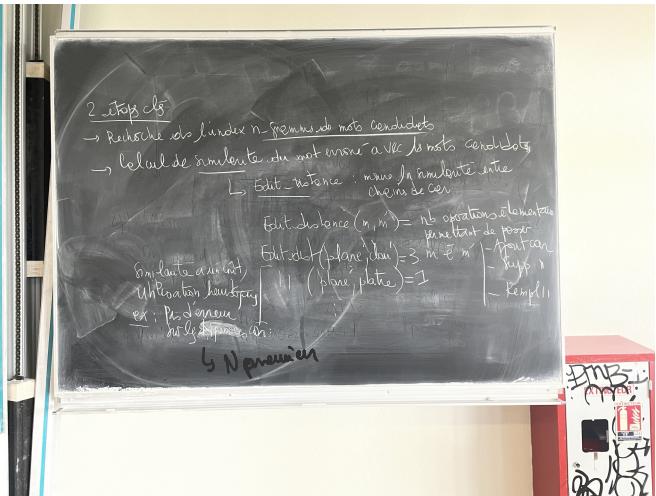


Connexion automatique:



Algorithmus 'naïf'





TD 2: modèles de recherche d'information:

Exercice 1:

Q1:

$$TF(t, d) = \log(1 + tf(t, d))$$

$$IDF(t) = \log \frac{N}{m_t}$$

similitude

$$RSV(q, D) = SIM(\vec{q}, \vec{D})$$

$$= \arccos(\vec{q}, \vec{D}) = \sum \frac{w_i q_i}{\|\vec{q}\| \|\vec{D}\|}$$

$$= \text{prod}(\vec{q}, \vec{D}) = \sum w_i q_i$$

RF → slide ? (ex, BN25)

	cord	vacuum	virus	sigmatone	respiration	
D1	0	log 5 * log 1/2	0	log 7 * log 4/3	0	monne
D2	0	0	log 5 * log 1/4	0	log 4 * log 4	
D3	0	log 4 * log 1/2	0	log 5 * log 4/3	0	
D4	0	0	0	log 3 * log 4/3	0	
\vec{q}	1	0	0	1	0	

← à calculer
pour calculer
vairne ($\vec{P}^T \vec{D}$)

En utilisant le produit intérieur:

$$RSV(q, D_1) = \log 7 * \log 3/4$$

$$RSV(q, D_2) = 0$$

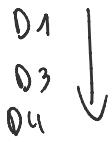
$$RSV(q, D_3) = \log 5 * \log 3/4$$

$$RSV(q, D_4) = \log 3 * \log 3/4$$

Donc: la liste ordonnée de documents retournée en réponse à la requête en question:

D1 1

Tutoriel
réponse à la question en cours



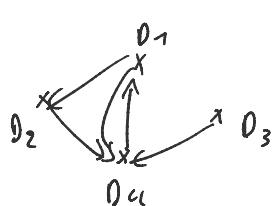
Q2] On considère le graphe $G(D, E)$ des itérations entre éléments :

$$D = \{D_1, D_2, D_3, D_4\}$$

$$E = \{e_{ij} | D_i \text{ cite } D_j (D_i \rightarrow D_j)\}$$

La matrice d'adjacence du G :

$$\begin{matrix} & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 1 & 0 & 0 & 0 \end{matrix}$$



La matrice de transition W^T

$$\begin{matrix} & 1 & 2 & 3 & 4 \\ 1 & 0 & 0 & 0 & 1 \\ 2 & \frac{1}{2} & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & \frac{1}{2} & 1 & 1 & 0 \end{matrix}$$

$$PR^{t+1} = (1 - d) * \frac{1}{N} + d * W^T PR^t, \quad d = 0.85, \quad N = 4$$

$$PR^0 = (0.25, 0.25, 0.25, 0.25)$$

$$PR^1 = 0.15 * \frac{1}{4} + 0.85 * \begin{bmatrix} 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$= D_1 \left[\frac{0.15}{4} + 0.85 * \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix} \right] \frac{0.15}{4} + 0.85 * 0.25$$

A compléter avec une feuille excel

Autoposition et de convergence.

→ Recherche $T(q, d)$ dans les documents de la collection.

→ Cela donne $\begin{bmatrix} \text{PR} \\ \text{D}_1 \\ \text{D}_2 \\ \text{D}_3 \\ \text{D}_4 \end{bmatrix}$

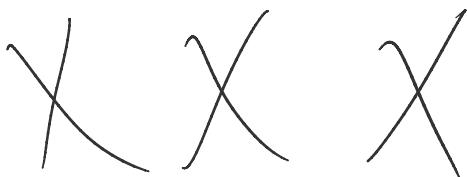
$RSV^c(q, D) = \alpha \text{PR}(D) + (1-\alpha) RSV(q, D)$

Ex: $RSV^c(q, D) = 0,5 + 0,5 * \log 7 * \log 3 / 4$

$RSV^c(q, D_2) = 0,5 + 0,5 * \log 5 * \log 3 / 4$

$RSV(q, D_3) = 0,5 * 0,5 + 0,5 * \log 5 * \log 3 / 4$

$RSV(q, D_4) = 0,5 * 0,5 + 0,5 * \log 3 * \log 3 / 4$



Q: Word segmentation

$RSV(Q, D)$

$\alpha = 0,5$

		$\log 7 * \log 4/3$
D_1		$0,5 * 0,5$
D_2		0
D_3		$\log 5 * \log 4/3$
		$\log 3 * \log 4/3$

$RSV^{(1)}(Q, A_1) = 0,5 * 0,5 + \frac{0,5}{2} * (\log 7 * \log 4/3 + \log 3 * \log 4/3)$

$RSV^{(1)}(Q, A_2) = 0,5 * 0,3$

$RSV^{(1)}(Q, A_3) = 0,5 * 0,1 + \frac{0,5}{1} * (\log 5 * \log 4/3)$

16/03/2023: Continuez T02:

@exercice:

$$RSV(Q, d_1) = \log 4/3$$

$$RSV(Q, d_2) = \log 4/3 + \log 2$$

$$RSV(Q, d_3) = \log 4/3 + \log 2$$

$$RSV(Q, d_4) = \log 2$$

Technique de Rocchio

$$q' = \vec{q} + \frac{0,75}{2} (\vec{d}_2 + \vec{d}_3)$$

Q6 - abc

$$q' = (0, 0, 1, 0, 1, 1, 0, 0) + \frac{0,75}{2} \left[(\log 4/3, 0, 0, 0, \log 2, \log 4/3, \log 2, 0) + (0, 0, 0, 0, \log 2, \log 4/3, \log 2, \log 2) \right]$$

$$= (0, 0, 1, 0, 1, 1, 0, 0) + \frac{0,75}{2} \left[(\log 4/3, 0, 0, 0, 2 \times \log 2, 2 \times \log 4/3, 2 \times \log 2, \log 2) \right]$$

→ Application de la requête ('Dany', 'love', 'give')

$$= (0,0,1,0,1,1,0,0) + \overbrace{(\text{word})}^{\text{expression et reformulation de la requête}} (1,0,0,0,0,0,0,0)$$

(expression et reformulation de la requête ('Dany', 'love', 'give'))

- Ajout des mots : 'book', 'read', 'think'
- Ajout des mots dérivés : 'give', 'love', 'mang'
- Augmentation des poids des mots

