Lexicologie

4 février 2020

—Examen—

Tous les supports existants (papier, internet, ordinateur) sont autorisés, à l'exception des téléphones, messageries et demander de l'aide sur des forums. Envoyez les solutions en email à kata.gabor@inalco.fr avant 12h30. Les réponses à 1) peuvent être écrites directement dans le texte de l'email ou affichées par un programme. Pour les autres exercises, un programme python est demandé.

1. Calculez le produit scalaire.

$$\begin{bmatrix}1 & 3 & 11 & 8\end{bmatrix} \times \begin{bmatrix}7 & 7 & 2 & 0\end{bmatrix} =$$

$$\begin{bmatrix}3 & 1 & 4 & 0\\ 2 & 8 & 5 & 0\\ 8 & 1 & 9 & 1\\ 1 & 0 & 0 & 6\end{bmatrix} \times \begin{bmatrix}1 & 7 & 0 & 0\end{bmatrix} =$$

$$\begin{bmatrix}3 & 1 & 4 & 0\\ 2 & 8 & 5 & 0\\ 8 & 1 & 9 & 1\end{bmatrix} \times \begin{bmatrix}0 & 2 & 0\\ 7 & 3 & 3\\ 3 & 2 & 5\\ 1 & 6 & 2\end{bmatrix} =$$

2. Soit une matrice terme-documents avec 4 documents ($d1, d2...d4$), un vocabulaire qui correspond à la liste ordonnée {*sport, match, rugby, football, foot, handball, tennis, tournoi, équipe*}, et les valeurs d'occurrence des termes dans les documents :

$$\begin{bmatrix}1411 & 239 & 1 & 9264 & 524 & 0 & 3 & 8754 & 3242\\ 9343 & 1746 & 0 & 7351 & 9467 & 45 & 99 & 7253 & 8263\\ 7593 & 7362 & 6782 & 5 & 2 & 0 & 99673 & 4766 & 23\\ 78746 & 13864 & 8732 & 924 & 51 & 34 & 24 & 12442 & 18322\end{bmatrix}$$

3. Représentez la matrice terme-document en tant que matrice numpy.

4. Ecrivez un programme qui
   - demande à l'utilisateur de rentrer une requête d'un ou plusieurs mots,
   - transforme la requête en vecteur one-hot, c'est-à-dire un vecteur dont les dimensions correspondent aux colonnes de la matrice terme-doc et qui prend une valeur 1 si la dimension correspond à un mot de la requête et 0 ailleurs,

— si un ou plusieurs mots de la requête ne sont pas dans la matrice, redemande à l'utilisateur de rentrer une requête.

5. Complétez le programme de façon à ce qu'il retourne, pour une requête utilisateur, la liste des documents (nommés d1, d2...) ordonnés selon leur pertinence, où la pertinence est la somme des occurrences des mots de la requête. Utilisez le produit scalaire.

6. Quel document sera retourné en premier pour la requête {*match football*} ?
7. Appliquez la pondération $PMI(t, d)$ à cette matrice selon les instructions. La probabilité d'observer un terme $P(t)$ est estimée comme

$$P(t) = \frac{freq(t)}{\sum_t freq(t)} \tag{1}$$

La probabilité d'observer un document $P(d)$ est estimée comme

$$P(d) = \frac{nombre - des - mots - dans(d)}{\sum_d nombre - des - mots - dans(d)} \tag{2}$$

On considère $P(t, d)$ la probabilité d'observer le terme t dans le document d, avec $freq(t, d)$ le nombre d'occurrences de t dans d :

$$P(t, d) = \frac{freq(t, d)}{\sum_d nombre - des - mots - dans(d)} \tag{3}$$

Sauvegardez le résultats dans une deuxième matrice, avec les valeurs PMI à la place des valeurs d'occurrence.

8. Si la pertinence des documents est donnée par le produit scalaire du vecteur requête one-hot avec la matrice *pondérée*, quel document sera retourné en premier pour la requête {*match football*}, et avec quelle valeur de pertinence ?