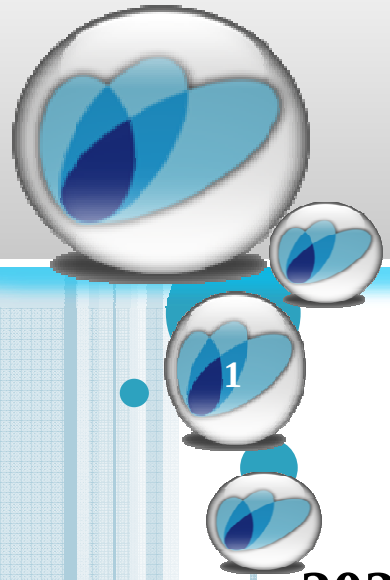
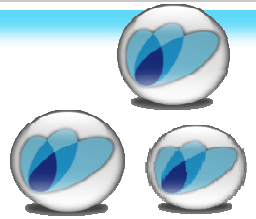
Cours Recherche D'information (Information Retrieval)

3ème Année Licence

Option : ISIL





2023-2024

Mme Z.LAAREDJ

Le modèle vectoriel
VSM: Vector Space Model

Plan

- Définition du modèle VSM
- Les concepts de base du modèle VSM.
- Représentation d’un document dans le VSM
- Représentation d’une requête dans le VSM
- Calcul de la similarité dans VSM.
- Avantages du modèle vectoriel
- Limites du modèle vectoriel
- Extensions du modèle vectoriel

Définition de modèle vectoriel

- Le Modèle VSM introduit par Gerard Salton dans le système SMART (System for the Mechanical Analysis and Retrieval of Text) dans les années 1971 [1].
- Il se base sur une formalisation vectorielle. A l'origine; il concernait les documents textuels puis il est étendu à d'autres types de documents.
- Idée de base : représenter les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents : T (un terme = une dimension) autrement dit :
  - L'ensemble des termes d'indexation forment un espace vectoriel.
  - Les documents sont des vecteurs dans cet espace.
  - Les requêtes sont des vecteurs dans cet espace.

Les concepts de base du modèle VSM

- T= {t1,t2,…,tn}: est l'ensemble de termes d'indexation (termes de la collection des documents).
- D={d1,d2,…,dm} : est l'ensemble des documents du corpus
- Un document di est représenté par un vecteur dont les composantes sont les poids des termes tj dans le document di:

di=(wi1,wi2,…,win} avec i=1…,m

- Q={q1,q2,…,qk} : est l'ensemble des requêtes du corpus

Les concepts de base du modèle VSM

- Une requête q est représentée par un vecteur dont les composantes sont les poids des termes tj dans la requête Q

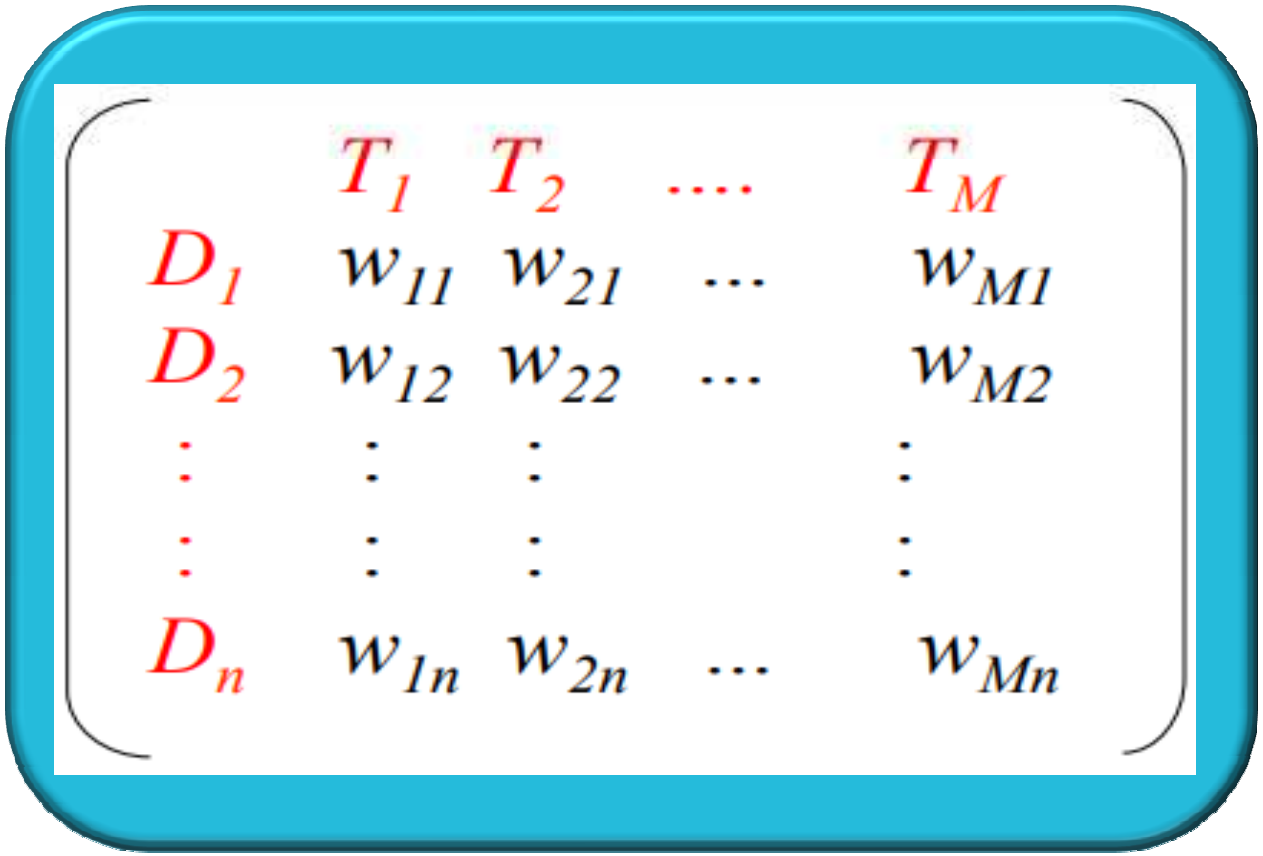
  q=(wq1,wq2,…,wqn)

Oû: Wij est le poids du terme tj dans le document di, i=1…m, j=1…n

Wqj est le poids du terme tj dans la la requete q, j=1….n

Les concepts de base du modèle VSM

- Une collection de n documents et M termes distincts peut être représentée sous forme de matrice.
- La requête est également représentée par un vecteur[1]



Matrice documents-termes

➢ Une collection de documents D={d1,d2,…,dm} et un ensemble T={t1,t2…,tn} de termes distincts sont représentés par la matrice documents-termes suivante :

	t_1	t_2	$\cdots$	t_n
d_1	w_{11}	w_{12}	$\cdots$	w_{1n}
d_2	w_{21}	w_{22}	$\cdots$	w_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_m	w_{m1}	w_{m2}	$\cdots$	w_{mn}

Matrice requêtes-termes

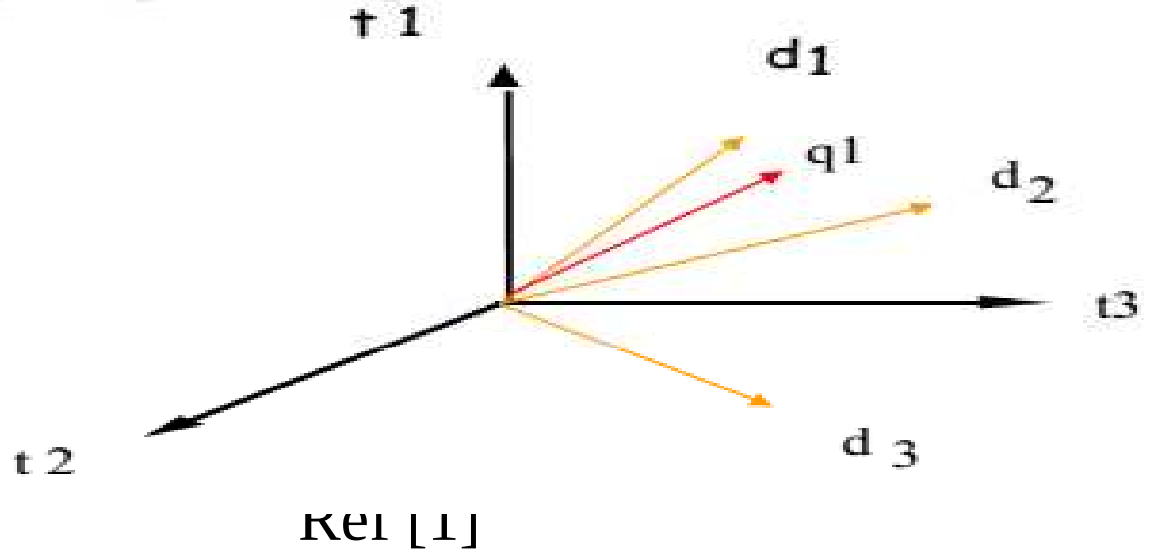
➢ Un ensemble de requêtes Q={q1,q2,…,qk} et un ensemble T={t1,t2…,tn} de termes distincts sont représentés par la matrice requêtes termes suivante :

	t_1	t_2	⋯	t_n
q1	w_{11}	w_{12}	⋯	w_{1n}
q2	w_{21}	w_{22}	⋯	w_{2n}
⋮	⋮	⋮	⋱	⋮
qk	w_{11}	w_{22}	⋯	w_{kn}

Représentation du VSM dans l'espace vectoriel

$dj = (w_{1j}, w_{2j}, \ldots, w_{Mj})$

$q = (w_{1q}, w_{2q}, \ldots, w_{Mq})$



Rel [1]

- Plus les vecteurs représentant les documents sont « proches », plus les documents sont similaires :
- La pertinence est traduite en une similarité vectorielle : un document est d1 est d'autant plus pertinent à une requête que le vecteur associé est similaire à celui de la requête.

Représentation du VSM dans l'espace vectoriel

- La représentation vectorielle ne tient pas compte de l'ordre d'apparition des termes dans les documents.
- **Exemple** :

« Un garçon mange une pomme »

est représenté par le même vecteur que

« Une pomme mange un garçon »

- Pour cette raison, cette représentation est aussi connue sous le nom « sac de mots » ou « Bag of words » en anglais.

Le poids d’un terme dans un vecteur document

- Wij est le poids du terme document di et il représente :
  - L’importance du terme dans le document
  - La discrimination du document dans le corpus.
  - Le poids du terme dénote la capacité du terme à discriminer les documents[2].

Calcul du poids (fonction de pondération)

- Il existe plusieurs schémas de pondération pour la prise en compte du poids du terme dans le document.
- La majorité des schémas utilisent la pondération locale et la pondération globale.
- Le modèle vectoriel utilise tf-idf pour mesurer le poids d'un terme dans le document.
- La pondération locale TF: permet de mesurer l'importance du terme dans le document
- La pondération globale IDF: permet de mesurer l'importance du terme dans la collection[2].

La pondération locale TF [2]

Facteur binaire: présence/absence du terme dans le document	$tf = \begin{cases} 1 \ si \ t \in D \\ 0 \ si \ t \notin D \end{cases}$
Facteur fréquentiel: fréquence du terme dans le document (nombre d'occurrences)	$tf = \text{freq}(t,d)$
Facteur fréquentiel normalisé	$tf = \frac{\text{freq}(t,d)}{\max_i \text{freq}(t_i,d)}$
Facteur logarithmique	$tf = 1 + \log(\text{freq}(t,d)) \ si \ \text{freq}(t,d) \neq 0$
Facteur logarithmique (normalisé)	$tf = \frac{1+\log(\text{freq}(t,d))}{1+\log(\text{avg}_i \text{freq}(t_i,d))} \ si \ \text{freq}(t,d) \neq 0$
Facteur augmenté	$tf = 0.5 + 0.5 \times \frac{\text{freq}(t,d)}{max_i \text{freq}(t_i,d)}$

La pondération globale IDF

- Le facteur idf (inverse document frequency): importance du terme dans la collection: donner un poids plus important pour les termes moins fréquents (plus discriminants).
- Soit un corpus ou collection de documents D et un terme d'indexation tj, on mesure la rareté du terme tj par l'inverse de sa fréquence dans D:

$$\frac{N}{df_j}$$

Ou :

- N=le nombre des documents dans le corpus (taille de la collection)
- dfj=est le nombre de documents du corpus ou le terme apparaît.

La pondération globale IDF

- La valeur obtenue croît très vite avec la taille N du corpus. On ajuste en prenant le logarithme de l'inverse normalisé.

$$idf_j = \log\left(\frac{N}{df_j}\right)$$

- Deux autres définitions sont utilisées:

$$idf_j = 1 + \log\left(\frac{N}{df_j}\right)$$

$$idf_j = \log\left(\frac{N - df_j}{df_j}\right)$$

Mesure de la pertinence pour un VSM

- La pertinence est traduite comme une similarité de vecteurs:
  - Un document est pertinent à une requête si le vecteur associé au document est similaire au vecteur associé à la requête.
  - Le traitement d'une requête est basée sur la comparaison des vecteurs documents et requête.
  - Dans l'espace vectoriel, plus les vecteurs représentants les documents sont proches et plus le document sont similaires.

Calcul de similarité dans le VSM

- L’appariement document-requête dans un modèle vectoriel consiste à trouver les vecteurs documents les plus proches du vecteur de la requête.
- Cet appariement est obtenu par l’évaluation de la distance entre les 2 vecteurs en utilisant la mesure RSV (Relevance Status Value).

Calcul de similarité dans le VSM

- Il y'a plusieurs façons de calculer la similarité (le RSV), le tableau ci donne les plus utilisées:

Le produit scalaire :	$RSV(d_i,q) = \sum_{j=1}^{n} w_{ij} \times w_{qj}$
Distance euclidienne :	$RSV(d_i,q) = \sqrt{\sum_{j=1}^{n} \left(w_{ij} - w_{qj}\right)^2}$
La mesure cosinus :	$RSV(d_i,q) = \frac{\vec{d_i}}{\lvert\vec{d_i}\rvert} \cdot \frac{\vec{q}}{\lvert\vec{q}\rvert} = \frac{\sum_{j=1}^{n} w_{ij} \times w_{qj}}{\sqrt{\sum_{j=1}^{n} w_{ij}^2} \times \sqrt{\sum_{j=1}^{n} w_{qj}^2}}$
La mesure de Dice :	$RSV(d_i,q) = \frac{2 \times \sum_{j=1}^{n} w_{ij} \times w_{qj}}{\sum_{j=1}^{n} w_{ij}^2 + \sum_{j=1}^{n} w_{qj}^2}$
La mesure de Jacard :	$RSV(d_i,q) = \frac{\sum_{j=1}^{n} w_{ij} \times w_{qj}}{\sum_{j=1}^{n} w_{ij}^2 + \sum_{j=1}^{n} w_{qj}^2 - \sum_{j=1}^{n} w_{ij} \times w_{qj}}$
Le coefficient de superposition (overlap) :	$RSV(d_i,q) = \frac{\sum_{j=1}^{n} w_{ij} \times w_{qj}}{min\left(\sum_{j=1}^{n} w_{ij}^2, \sum_{j=1}^{n} w_{qj}^2\right)}$

Avantages du modèle vectoriel

- La simplicité conceptuelle et de mise en œuvre du modèle vectoriel.
- Offre des moyens simples pour classer les résultats d'une recherche c'est-à-dire que la pondération améliore les résultats de recherche et la mesure de similarité permet d'ordonner les documents selon leur pertinence vis à vis de la requête.
- Le langage de requête est simple: liste de termes. Son efficacité dépend de la qualité de représentation : vocabulaire et le schéma de pondération.
- Les documents sont retournés dans un ordre décroissant de leur degré ou score de similarité avec la requête.
- Il est robuste et performant dans les tests.

Inconvénients du modèle vectoriel

- Approche vectorielle considère chaque terme comme étant indépendant des autres (pas de liens entre termes). Alors que peut avoir des termes qui sont reliée sémantiquement tel que le mot Véhicule et Automobile.

Extensions du modèle vectoriel

- Retour de pertinence (*relevance feedback)*
- Expansion de requêtes
  - Thesaurus
  - Intégration de co-occurrences
- Modèle d'analyse sémantique latente (*LSI -- Latent Semantic Indexing)*

Références bibliographiques

[1] https://www.irit.fr/~Mohand.Boughanem/slides/RI/chap4-mod-bool-vect.pdf

[2] https://miashs-www.u-ga.fr/prevert/SpecialiteIHS/RI/cours_RI.pdf

[3] https://lipn.univ-paris13.fr/~rozenknop/Cours/MICR_REI/Seance6/modeles-RI-1.pdf

[4]https://tel.archivesouvertes.fr/file/index/docid/785143/filename/2006_these_A_mercier_417I.pd

[5] https://www.youtube.com/watch?v=BDi3drDPibY