LAMA-WeST Lab
Artificial Intelligence
Natural language processing
Semantic web

La Métrique BM25

INF8460 - Traitement automatique de la langue naturelle

Karou Diallo
Polytechnique Montréal

10 septembre 2024

Introduction

De TF-IDF à BM25

Introduction

De TF-IDF à BM25

Définition

La métrique BM25 (Best Matching 25) est une fonction de pondération utilisée dans la recherche d'information pour évaluer la pertinence des documents par rapport à une requête. Elle fait partie de la famille des modèles probabilistes de recherche.

Caractéristiques

- C'est un modèle de recherche d'information probabiliste
- BM25 calcule un score de pertinence en tenant compte de :
  - la longueur des documents
  - la fréquence des termes de la requête dans chaque document
  - la fréquence des termes de la requête dans le corpus global

Avantages

- offre un bon équilibre entre précision et rappel
- gère efficacement de grandes collections de documents
- permet des ajustements pour différents types de corpus
- est largement utilisé dans les moteurs de recherche et les systèmes de recommandation

Les limites de la métrique TF-IDF

La métrique BM25 règle les problèmes de TF-IDF, notamment :

- Le biais lié à la fréquence d'un terme (cf. 8)
- La non-considération de la longueur des documents (cf. 9)
- Le contrôle statique de la variation des longueurs de document dans la version améliorée de TF-IDF (cf. 12)

Introduction

De TF-IDF à BM25

Biais lié à la fréquence d'un terme

- Lorsqu'un terme est très fréquent dans un document le **TF** est élevé
- Les longs documents seront plébiscités par la métrique **TF-IDF** au détriment des plus courts
- Alors qu'on devrait plutôt favoriser les documents courts contenant un terme de la requête

Terme recherché "nlp"	**Fréquence terme**	**Longueur du document**
Document d_1	100	1000
Document d_2	1	10

- Est ce que d_1 est plus pertinent que d_2 ? pas forcément ...

TF-IDF normalisé (1)

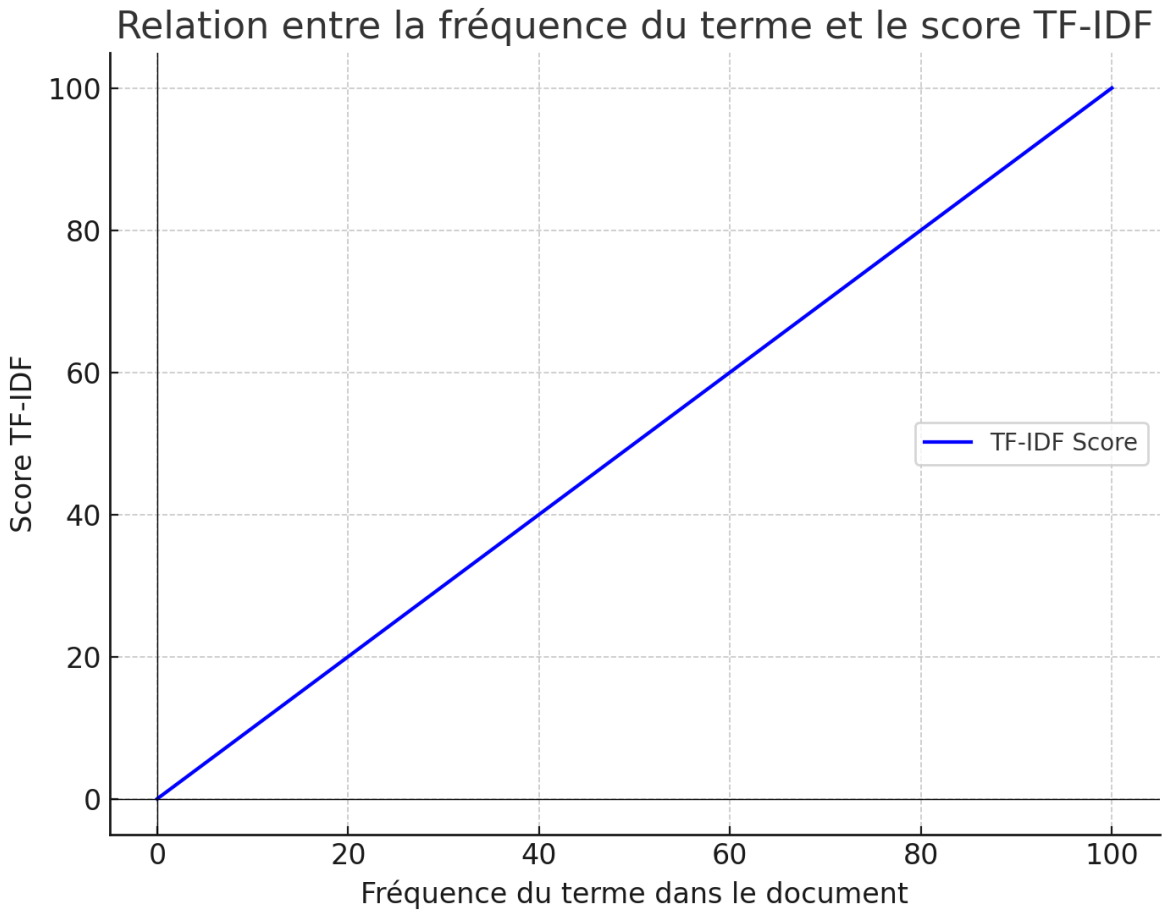
Pour pallier le problème du biais lié à la fréquence d'un terme, le **TF-IDF normalisé** inclut la longueur du document dans le calcul de la fréquence d'un terme t dans un document d :

$$\mathrm{TF(t, d)} = \frac{\mathrm{freq(t)}}{\left|d\right|} \tag{1}$$

où t représente le terme de la requête, d est le document, freq(t) est la fréquence du terme et $\left|d\right|$ représente la longueur du document.

TF-IDF normalisé (2)

Mais un problème demeure : le score TF-IDF est linéairement dépendant de la fréquence $\mathrm{TF}(t, d)$.



TF-IDF normalisé (3)

- Considérons le cas suivant :

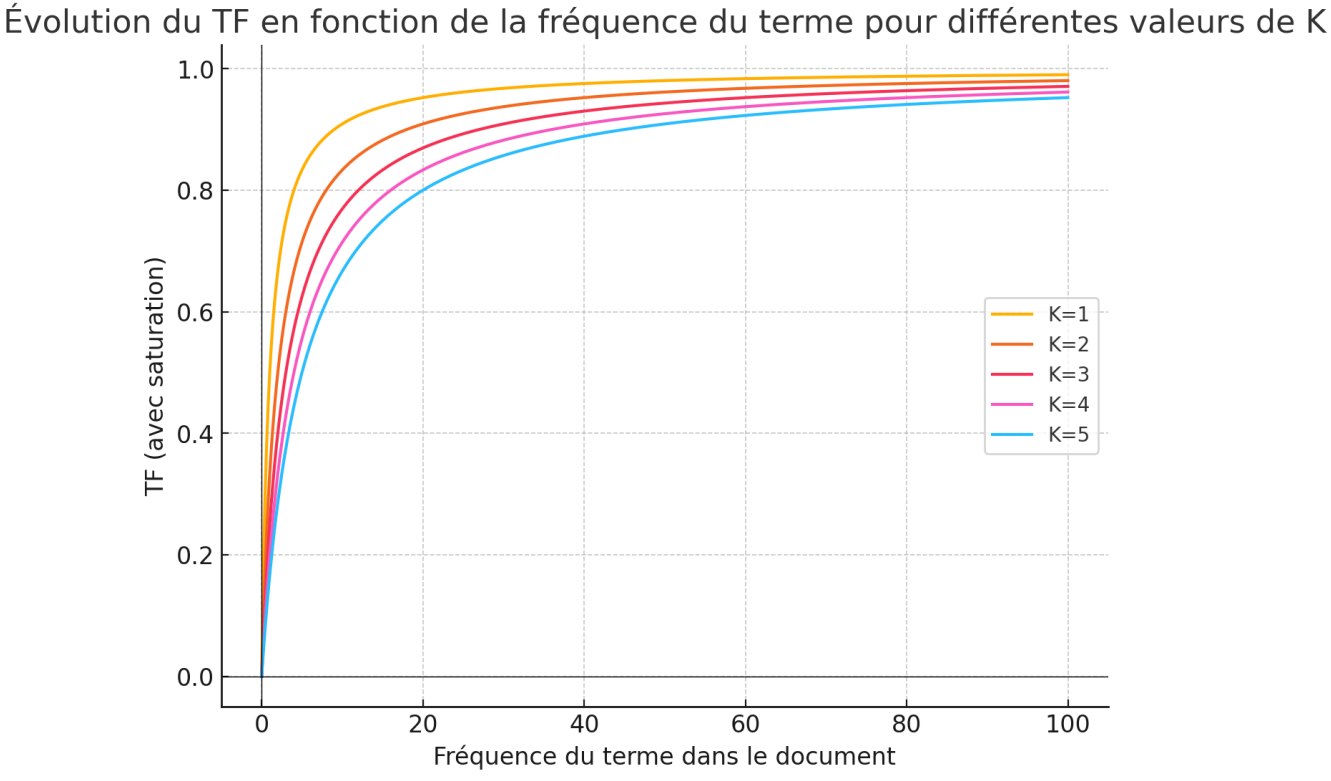
Terme recherché "nlp"	**Fréquence terme**
Document d_1	50
Document d_2	100

- Est-ce que d_2 est deux-fois plus pertinent que d_1 ? pas forcément ...
- On voudrait idéalement avoir une influence de la fréquence du terme qui diminue (sature) à partir d'un certain seuil.

TF modifié au niveau de BM25 (1)

- Pour éliminer cette constante linéarité, le paramètre K est introduit dans le calcul de la fréquence :

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \frac{\mathrm{TF}(\mathrm{t},\mathrm{d})}{\mathrm{TF}(\mathrm{t},\mathrm{d}) + K} \tag{2}$$



TF modifié au niveau de BM25 (2)

- En prenant en compte la longueur de tous les documents, l'équation est ajustée comme suit :

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \frac{\mathrm{TF}(\mathrm{t},\mathrm{d})}{\mathrm{TF}(\mathrm{t},\mathrm{d}) + \left(K \times \frac{\left|\mathrm{d}\right|}{\mathrm{moy}\left(\left|\mathrm{D}\right|\right)}\right)} \tag{3}$$

où $\left|\mathrm{d}\right|$ est la longueur du document d, $\mathrm{moy}\left(\left|\mathrm{D}\right|\right)$ est la longueur moyenne de l'ensemble des documents D.

- On pénalise les documents longs et favorise les plus courts
- Problème : Jusque là, on a un contrôle statique de l'impact de la longueur des documents alors qu'on voudrait une flexibilité sur le traitement de chacun d'eux

TF modifié au niveau de BM25 (3)

- Solution : On contrôle maintenant l'importance de la longueur pour chaque document avec le paramètre b

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \frac{\mathrm{TF}(\mathrm{t},\mathrm{d})}{\mathrm{TF}(\mathrm{t},\mathrm{d}) + K \times \left(1 - b + b \times \frac{\left|\mathrm{d}\right|}{\mathbf{moy}\left(\left|\mathrm{D}\right|\right)}\right)} \quad (4)$$

- $b = 0$: on ignore la longueur du document
- $b = 1$: on a un impact statique de la longueur
- $b \in]0, 1[$: on a une normalisation partielle par la longueur du document ie elle est prise en compte, mais l'effet est atténué par rapport à $b = 1$

IDF au niveau de BM25

- Avec la formule Eq.5, le IDF est négatif si le terme apparaît dans plus de la moitié des documents. On corrige cela en ajoutant 1 comme dans Eq.6

$$\mathrm{IDF(t)} = \log\left(\frac{(N - \mathrm{DF(t)} + 0.5)}{\mathrm{DF(t)} + 0.5}\right) \tag{5}$$

$$\mathrm{IDF(t)} = \log\left(\frac{1 + (N - \mathrm{DF(t)} + 0.5)}{\mathrm{DF(t)} + 0.5}\right) \tag{6}$$

où N est le nombre total de documents et $DF(\mathrm{t})$ est le nombre de documents contenant le terme t

BM25 au complet (1)

- Enfin en rassemblant les composants TF et IDF, nous avons l'équation complète de BM25 comme suit :

$$\mathrm{BM25(t,d)} = \frac{\mathrm{TF(t,d)} \times \left(K+1\right)}{\mathrm{TF(t,d)} + K \times \left(1 - b + b \times \frac{\left|\mathrm{d}\right|}{\mathbf{moy}\left(\left|\mathrm{D}\right|\right)}\right)} \times \log\left(\frac{1+(N-\mathrm{DF(t)}+0.5)}{\mathrm{DF(t)}+0.5}\right) \quad (7)$$

BM25 au complet (2)

Considération pour le choix des valeurs des paramètres.

- Dans la majorité des cas $K = 1.2$ et $b = 0.75$ sont de bonnes valeurs pour ces deux paramètres.
- Le paramètre K doit être choisi en considérant la longueur moyenne des documents.
- Pour les longs documents K peut doit être choisi suffisamment grand afin de ne pas atteindre le point de saturation très rapidement.
- Le paramètre b est choisi en prenant en compte le type de document et l'impact de la longueur de document sur la pertinence d'un terme de recherche.
- Pour des documents scientifiques une petite valeur de b est mieux alors que pour un document contenant beaucoup de bruits une grande valeur sera meilleure.

Ces diapos ont été adaptés à partir de la référence suivante : ”BM25 Algorithm : Overcoming the Limitations of TF-IDF” MLWorks, YouTube, `https://youtu.be/YL-3G5-xVYU?si=VNT2Vshf2nuxngPU` à la date du 19 août 2024.