HeadMind Partners

Recherche d'information sur le Web: Développement d'un moteur de recherche

ST4 – EI – Groupe 5

Théo Schneider
Jules Berard
Zenta Utagawa
Salma Maazoum
Joshua Noullier-Jacques

CentraleSupélec

Sommaire

1. Exploration et traitement des données
2. Indexation
3. Conception d ’un moteur de recherche
   a) Différents modèles conçus
   b) Amélioration et comparaison
4. Test sur des requêtes

Indexation

- Exploration des données
- Conception de l'index inversé

Indexation des données

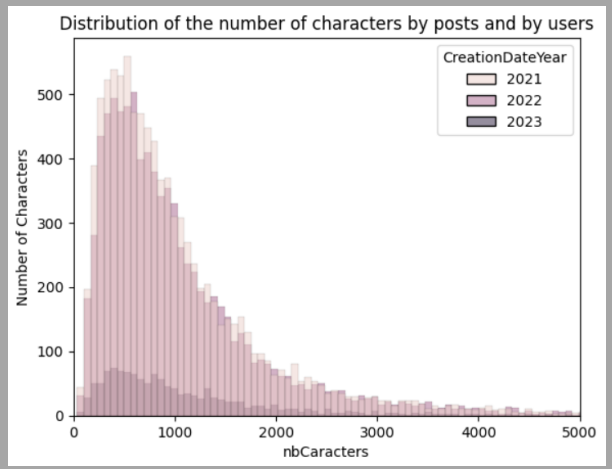
Prétraitement des données:

- Balises HTML
- Tokenisation
- Mots vides
- Ponctuation
- Lemmatisation

(la bibliothèque nltk)

Structure de données :

- Dictionnaire
- Index de fréquence (df et idf)
- Longueur du poste



Accès à l'index inversé:

Fichier Pickle

```
{'I': {'df': 49841,
  'inv_ind': [(5, 0.06481481481481481),
   (7, 0.034482758620689655),
   (14, 0.03896103896103896),
   (16, 0.0125),
   (21, 0.0019342359767891683),
   (22, 0.011627906976744186),
   (24, 0.013245033112582781),
   (29, 0.014492753623188406),
```

Moteur de recherche

- Différents modèles conçus
- Avantages
- Inconvénient

Moteur de recherche

Approche naïve

Recherche booléenne

MIB

BM25

Tf-Idf

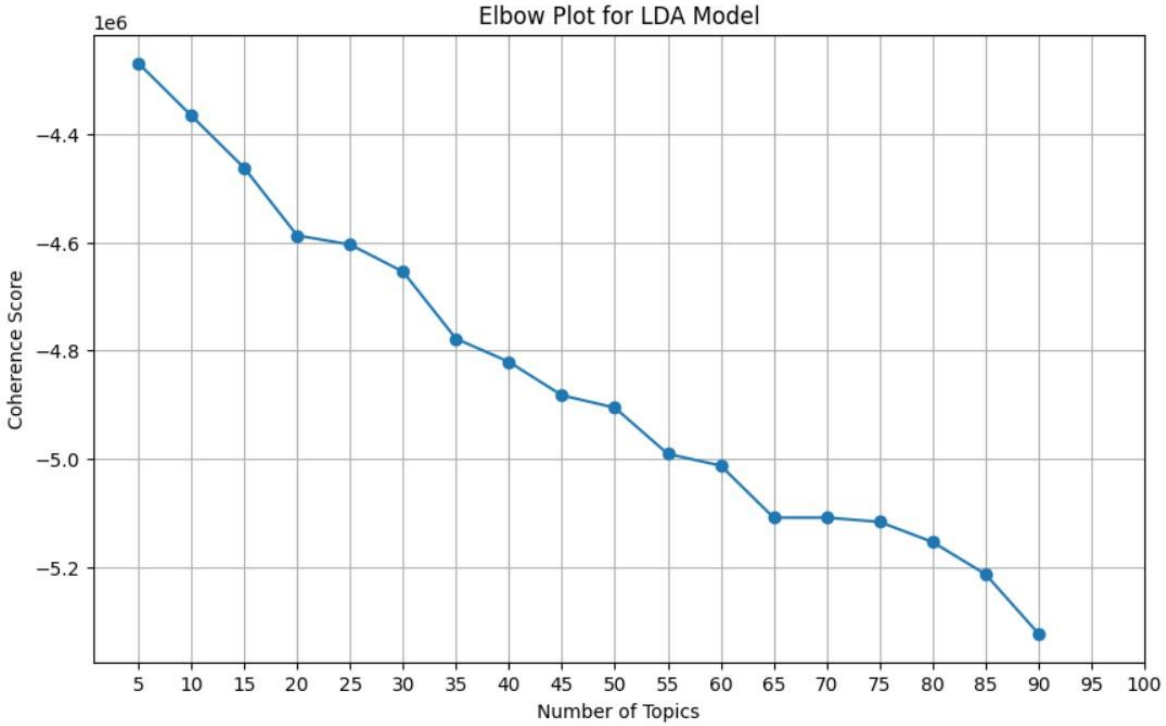
Similarité Sémantique

Améliorations

- Clustering
- Utilisation des métadonnées
- Fusion des méthodes

Clustering

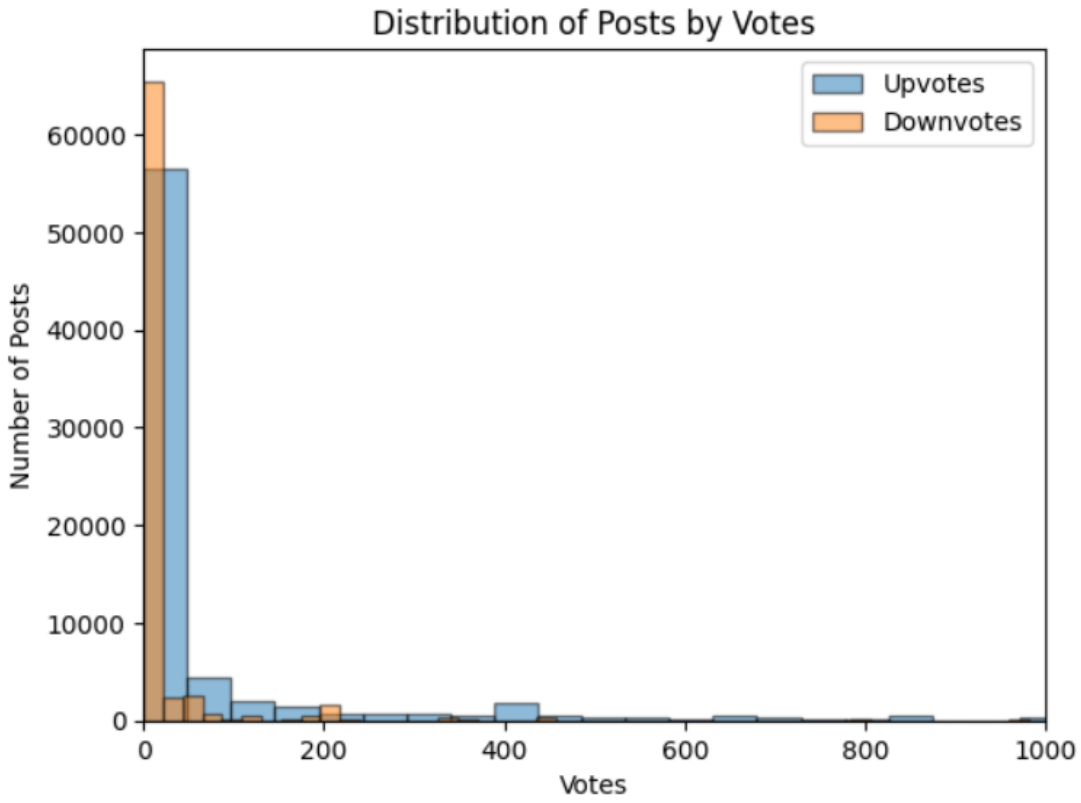
- Latent Dirichlet Allocation
- Associer un thème à chaque document
- Restreindre la recherche au document de même thème que la requête



Choix du nombre de thème

Métadonnées

- Pour la recherche (*tags, titre, …*)
- Pour une nouvelle méthode *Hub/Authorities*
- Pour le filtrage (*score , flags,…*)



La distribution des scores est normale, on discrimine les extrêmes.

Fusion des méthodes

	Query 1		Query 2		Query 3		Query 4		Query 5	
	ɒrmance for multiclassifi		draw neural network		neural network layers		how sklearn working		treat categorical data	
Post ID	Judgement	Search Algo	Judgement	Search Algo	Judgement	Search Algo	Judgement	Search Algo	Judgement	Search Algo
22		10		7		7	2		3	3
694		7	3	3	4	3	4	4		8
5706		6	2	4	5	4		10		6
6107				6	6	1		6		
9302		5	1	5	1	5	3	9	1	
9443		3		9		9	1	8	4	2
12321		4					6	1	2	1
12851			5	2	3	2		2		9
13490	1	9		10		10		7		7
14899		2	4	1	2	6				10
15135		8		8		8	5	3		5
15989	2	1						5		4

Merci pour votre attention

EI – Groupe 5