Projet RI

Sorbonne Université

Implémentation d'un système de recommandation grâce à l'algorithme topical PageRank

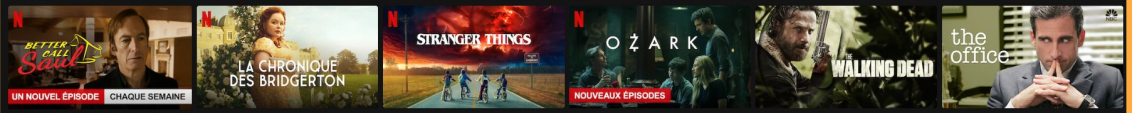
Sommaire

Introduction

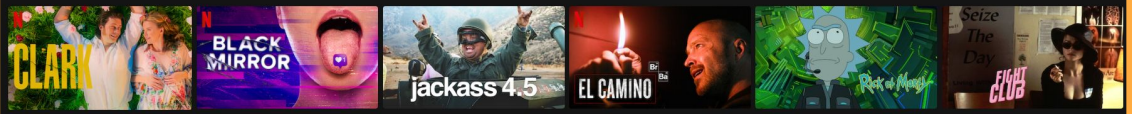
Comparer à des articles similaires

						
	Cet article Melitta 6762887 Cafetière Filtre avec Verseuse en Verre, Easy II, 1023-02, Noir	Russell Hobbs Machine à Café, Cafetière Filtre 625ml, Café Rapide en 7mn, Plaque Chaude, Porte Filtre Amovible - Inox Brossé 24210-56 Compact Home	Moulinex Subito Cafetière filtre électrique, 0,6 L soit 6 tasses, Machine à café, Système anti goutte, Porte-filtre pivotant, Auto off 30 min FG150813	MOULINEX Cafetières filtre SUBITO inox 10/15 Tasses Machine à café cafetière électrique Cafetière Capacité 1.25L Antigoutte Porte-filtre pivotant Auto off 30 minutes FG360811	Ufesa CG7114 Capriccio Cafetière Filtre Américaine, 6 Tasses, 0.6L, 600W, Filtre Réutilisable Pivotant, Système Anti Goutte, Plaque Chauffante Antiadhésive, Arrêt Automatique, Bleu	Brandt CAF815X – Cafetière Filtre – Capacité 15 Tasses – 800W – Maintien Au Chaud – Système Antigouttes - 1,5L - Inox
	#1 Meilleure vente					
	Ajouter au panier	Ajouter au panier	Ajouter au panier	Ajouter au panier	Ajouter au panier	Ajouter au panier
Évaluation des clients	★★★★½ (6455)	★★★★½ (1380)	★★★★½ (1389)	★★★★½ (658)	★★★½☆ (119)	★★★★☆ (62)
Prix	25,90 €	32,75 €	28,25 €	39,00 €	26,99 €	35,99 €

Séries primées



Tendances actuelles



Appréciés sur Netflix

Tendances

NOUVEAUTÉS | MUSIQUE | JEUX VIDÉO | FILMS

AVATAR 2 : LA VOIE DE L'EAU Bande Annonce (2022)
FilmsActu · 1,8 M de vues · il y a 12 jours
☆ Les Films à VOIR ? Ils sont ICI ► https://www.youtube.com/playlist?list=PL843D2ED8D80FA673 AVATAR 2 : LA VOIE DE L'EAU Bande Annonce (2022) Sam Worthington, Zoe Saldana, James Cameron...

CETTE SEMAINE DANS DEMAIN NOUS APPARTIENT | BANDE ANNONCE DU 16 AU 20 MAI 2022
Demain Nous Appartient Family · 75 k vues · il y a 4 jours
Merci à @dna_tf1.ba pour la bande annonce

SHE-HULK Bande Annonce VF (2022) Série Marvel
JEUXACTU · 235 k vues · il y a 4 jours
► Rendez-vous sur Disney+ : https://bit.ly/2K7XSXC ► Découvrez les collections de produits Disney, Marvel et Star Wars sur : https://www.shopdisney.fr/ ► Marvel FR : https://www.youtube.com/...

ARTHUR MALEDICTION - Bande-annonce
EUROPACORP · 88 k vues · il y a 9 jours
Alex est un fan des films Arthur et les Minimoys depuis qu'il est enfant. Pour son anniversaire, ses meilleurs amis lui font la surprise de l'emmener dans la maison abandonnée où le film...

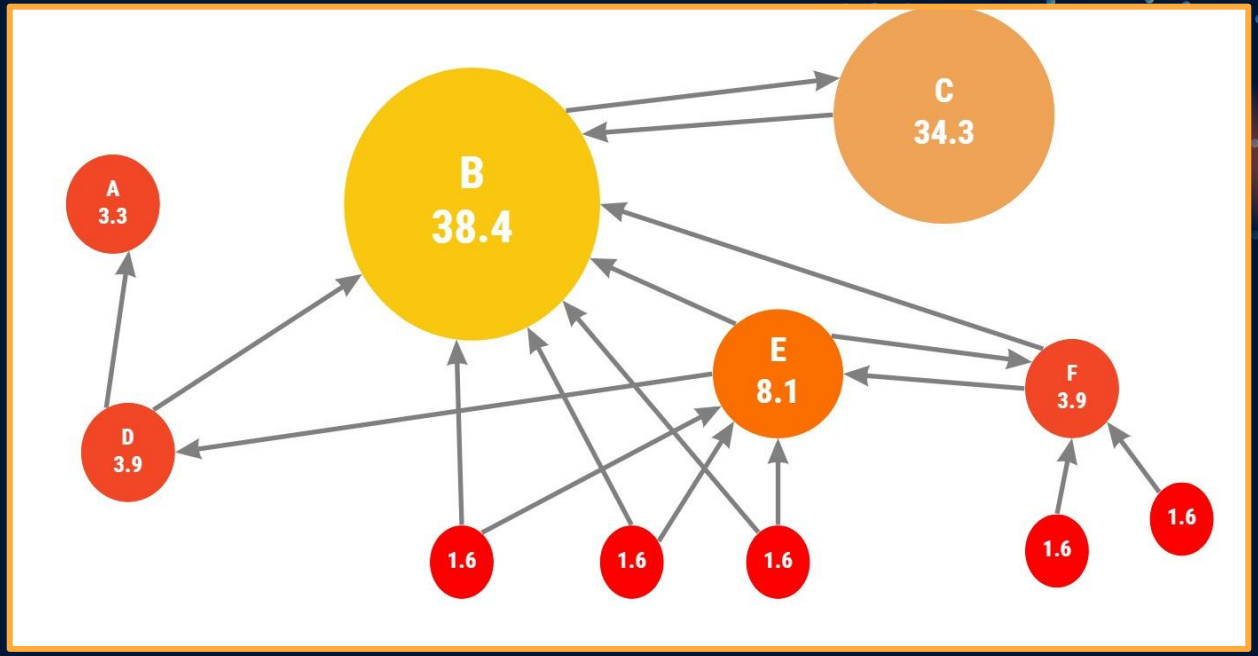
VIKRAM - Official Trailer | Kamal Haasan | VijaySethupathi, FahadhFaasil | LokeshKanagaraj | Anirudh
Sony Music South · 27 M de vues · il y a 6 jours
#VikramTrailer #VikramAudioLaunch #VikramInAction #VIKRAM WORLDWIDE THEATRICAL RELEASE ON JUNE 3RD. Movie : VIKRAM Starring: Kamal Haasan, Vijay Sethupathi, Fahadh Faasil. Banner: Raaj...

COUPEZ ! Bande Annonce Teaser (2022) Romain Duris
Bandes Annonces Cinéma · 62 k vues · il y a 2 semaines
COUPEZ ! Bande Annonce Teaser (2022) Romain Duris © 2022 - PAN Distribution

Sommaire

PageRank

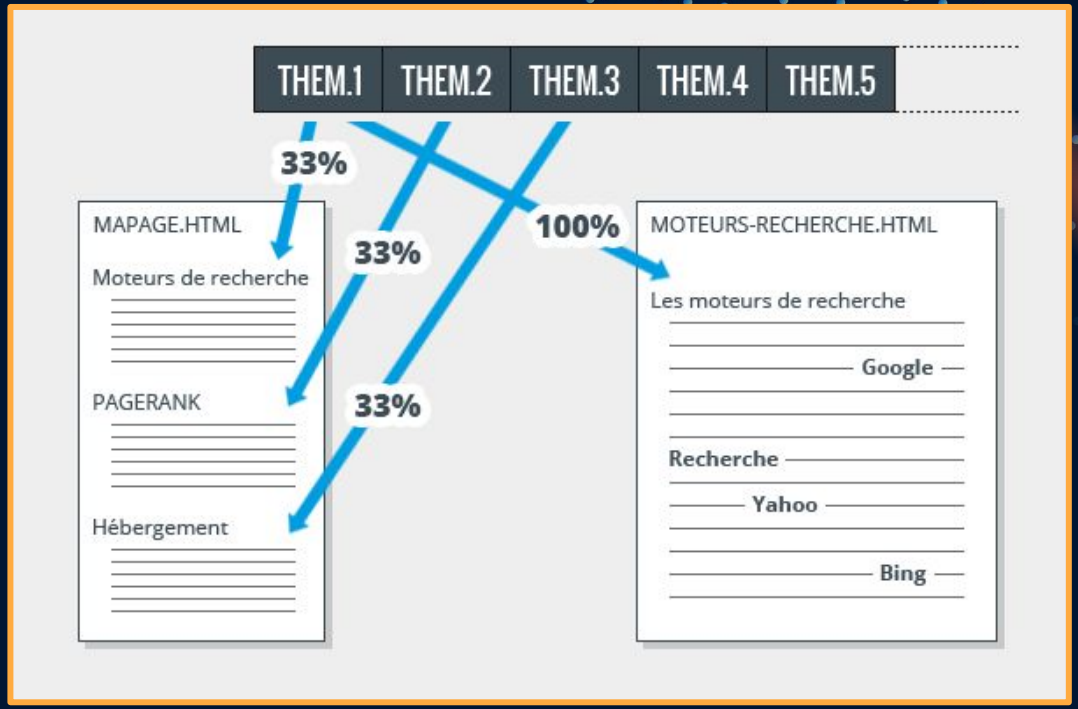
- PR(p_i) donne la valeur pour un noeud p_i
- Constante d = 0.85
- N représente le nombre total de pages
- M(p_i) représente le nombre de page qui pointe vers p_i
- L(p_j) donne le nombre d'arc sortant pour un noeud p_j



$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Topical PageRank

- d et α sont des constantes
- N est le nombre total de page
- A_z(p) représente le score de la page p du thème z
- q représente les pages pointant vers p
- O(p) représente le nombre d'arc sortant du noeud p
- C_z(p) est la probabilité que la page p appartienne au thème z



$$A_z(p) = d \sum_{q:q\to p} \frac{\alpha A_z(q) + (1-\alpha) C_z(q) A(q)}{O(q)} + \frac{1-d}{N} C_z(p)$$

Système de recommandation

$$A_z(p) = d \sum_{q:q\to p} \frac{\alpha A_z(q) + (1-\alpha)C_z(q)A(q)}{O(q)} + \frac{1-d}{N} C_z(p)$$

Équation du topical PageRank

$$A = d\alpha G \bullet A + d(1-\alpha)G \bullet F_{CA} + (1-d)\frac{C}{N}$$

Équivalent matriciel de cette équation

$$R = d\alpha \mathcal{M} \bullet R + d(1-\alpha)\mathcal{M} \bullet F + (1-d)I$$

Modification de la formulation pour notre cas

$$\begin{cases} R_{ig}^{u_k}(0) = \dfrac{1}{|M| * n} (1 \le i \le |M|, 1 \le g \le n) \\ F_{ig}(t) = (\sum_{g=1}^{n} R_{ig}^{u_k}(t-1)) * P_{ig} \\ R^{u_k}(t) = d\alpha \mathcal{M} R^{u_k}(t-1) + d(1-\alpha)\mathcal{M}F(t) + (1-d)I^{u_k} \end{cases}$$

Formulation finale de l'algorithme itératif

$$TR_{ki} = \sum_{g=1}^{n} (R_{ig}^{u_k} * P_{ig})$$

Estimation du score des films pour chaque utilisateur

Sommaire

Données brutes

- 100 000 Notes
- 9000 films
- 600 utilisateurs

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy
...	...	...	...
9737	193581	Black Butler: Book of the Atlantic (2017)	Action|Animation|Comedy|Fantasy
9738	193583	No Game No Life: Zero (2017)	Animation|Comedy|Fantasy
9739	193585	Flint (2017)	Drama
9740	193587	Bungo Stray Dogs: Dead Apple (2018)	Action|Animation
9741	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy

9742 rows × 3 columns

Fichier CSV des films

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931
...	...	...	...	...
100831	610	166534	4.0	1493848402
100832	610	168248	5.0	1493850091
100833	610	168250	5.0	1494273047
100834	610	168252	5.0	1493846352
100835	610	170875	3.0	1493846415

100836 rows × 4 columns

Fichier CSV des notes

Données pré-traitées

	movieId	title	(no genres listed)	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	...	Film-Noir	Horror
0	0	Toy Story (1995)	0	0	1	1	1	1	0	0	...	0	0
1	1	Jumanji (1995)	0	0	1	0	1	0	0	0	...	0	0
2	2	Grumpier Old Men (1995)	0	0	0	0	0	1	0	0	...	0	0
3	3	Waiting to Exhale (1995)	0	0	0	0	0	1	0	0	...	0	0
4	4	Father of the Bride Part II (1995)	0	0	0	0	0	1	0	0	...	0	0

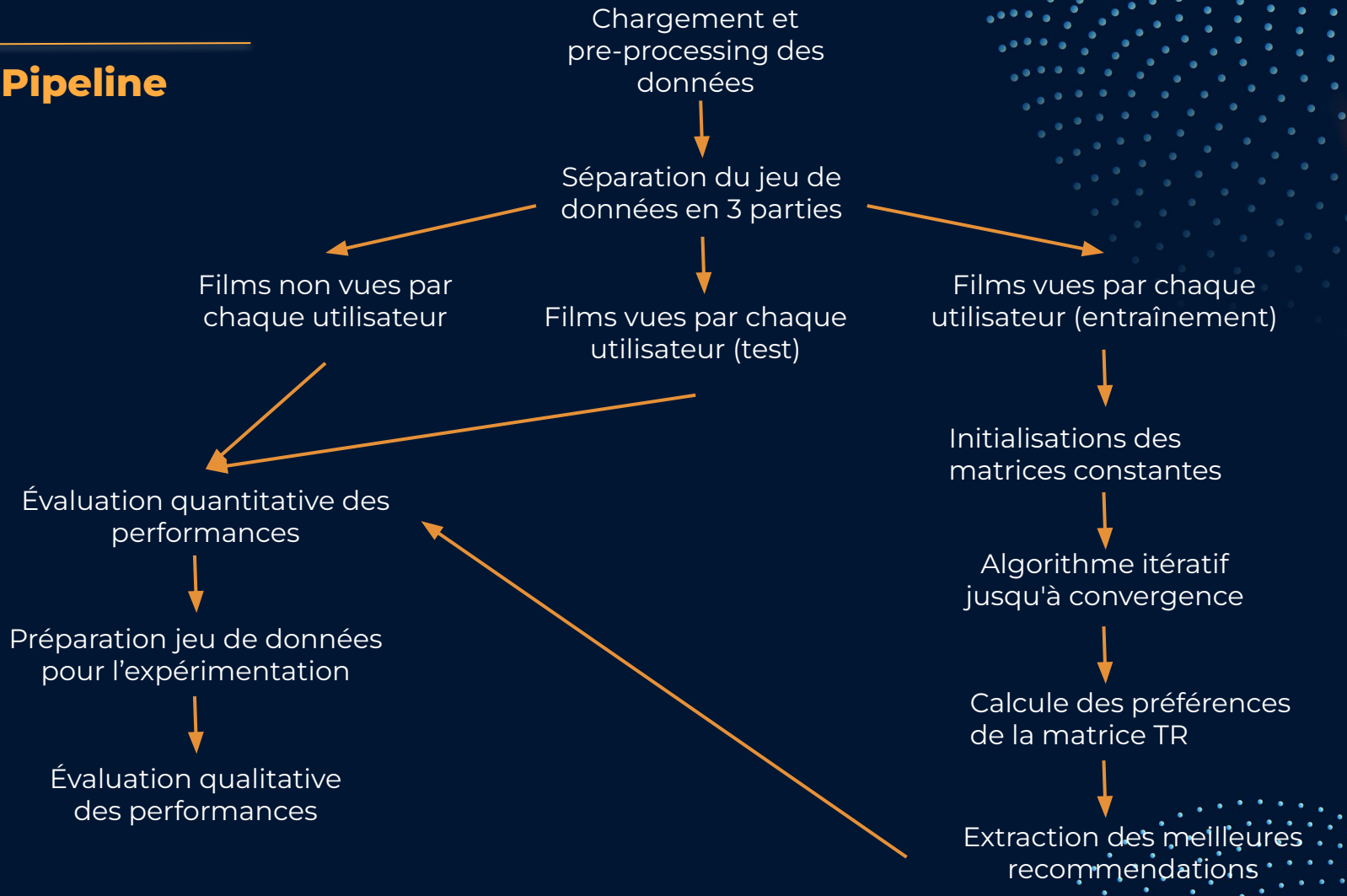
Fichier CSV des films

	userId	movieId	rating
0	0	0	4.0
1	0	2	4.0
2	0	5	4.0
3	0	43	5.0
4	0	46	5.0
...	...	...	...
100831	609	9416	4.0
100832	609	9443	5.0
100833	609	9444	5.0
100834	609	9445	5.0
100835	609	9485	3.0

100836 rows × 3 columns

Fichier CSV des notes

Pipeline


Chargement et pre-processing des données
Séparation du jeu de données en 3 parties
Films non vues par chaque utilisateur
Films vues par chaque utilisateur (test)
Films vues par chaque utilisateur (entraînement)
Initialisations des matrices constantes
Algorithme itératif jusqu'à convergence
Calcule des préférences de la matrice TR
Extraction des meilleures recommendations
Évaluation quantitative des performances
Préparation jeu de données pour l'expérimentation
Évaluation qualitative des performances

Sommaire

Évaluation quantitative

$$DOA_{u_k} = \frac{\sum_{(m_i \in Te_{u_k}, m_j \in Nw_{u_k})} f_{u_k}(m_i, m_j)}{|Te_{u_k}| \bullet |Nw_{u_k}|}$$

Table 1: DOA comparison of our algorithm (TR) with PaperRank (PR) on the 5 splits.

	Split1	Split2	Split3	Split4	Split5	**Mean**
PR	87.69	87.71	87.65	87.51	88.14	87.74
TR	89.05	89.12	89.26	88.97	89.01	**89.08**

Table 2: Overall DOA comparisons of our algorithm (TR) with other scoring algorithms.

	L+	PR	Katz	Dijkstra	**TR**
DOA	87.18	87.74	85.26	49.65	**89.08**

- La métrique DOA permet d'avoir une idée de la performance de notre système de recommandation
- Les films vus par un utilisateur doivent arriver avant les films non vus dans les recommandations
- Pour se faire, nous devons séparer notre jeu de données en jeu d'entraînement, jeu de test et films non vues
- Nous obtenons un DOA de 0.75 en test sur le petit jeu de données

Évaluation qualitative

```
PS C:\Users\karna\Desktop\Projects\PythonProject\PageRankRecommandation\src> python3 .\experiment.py
------------------------------------------ User 1 / 10 ------------------------------------------

--> List of film seen by user
['Groundhog Day (1993)',
 'Sense and Sensibility (1995)',
 'Terminator, The (1984)',
 'Princess Bride, The (1987)',
 'Cable Guy, The (1996)']

--> Choose 5 films you would recommand for that user in the following list :
{'257': 'Pulp Fiction (1994)',
 '314': 'Forrest Gump (1994)',
 '1938': 'Matrix, The (1999)',
 '6840': 'Duchess, The (2008)',
 '7339': 'Ricky Gervais Live 3: Fame (2007)',
 '6530': 'Nanny Diaries, The (2007)',
 '8009': 'Man with the Iron Fists, The (2012)',
 '224': 'Star Wars: Episode IV - A New Hope (1977)',
 '277': 'Shawshank Redemption, The (1994)',
 '8690': 'Kenny & Company (1976)'}

--> Your choice must be this format : 145/323/447/565/23
257/314/6840/224/277
```

- On montre une liste de film vus par un utilisateur à notre expérimentateur
- On montre une liste de film comprenant des films aléatoires et des films recommandés par notre système
- On effectue cela pour plusieurs utilisateurs
- On calcule un score de précision classique
- On fait une moyenne des scores reçu grâce à plusieurs expérimentateurs
- On obtient une précision de 87%

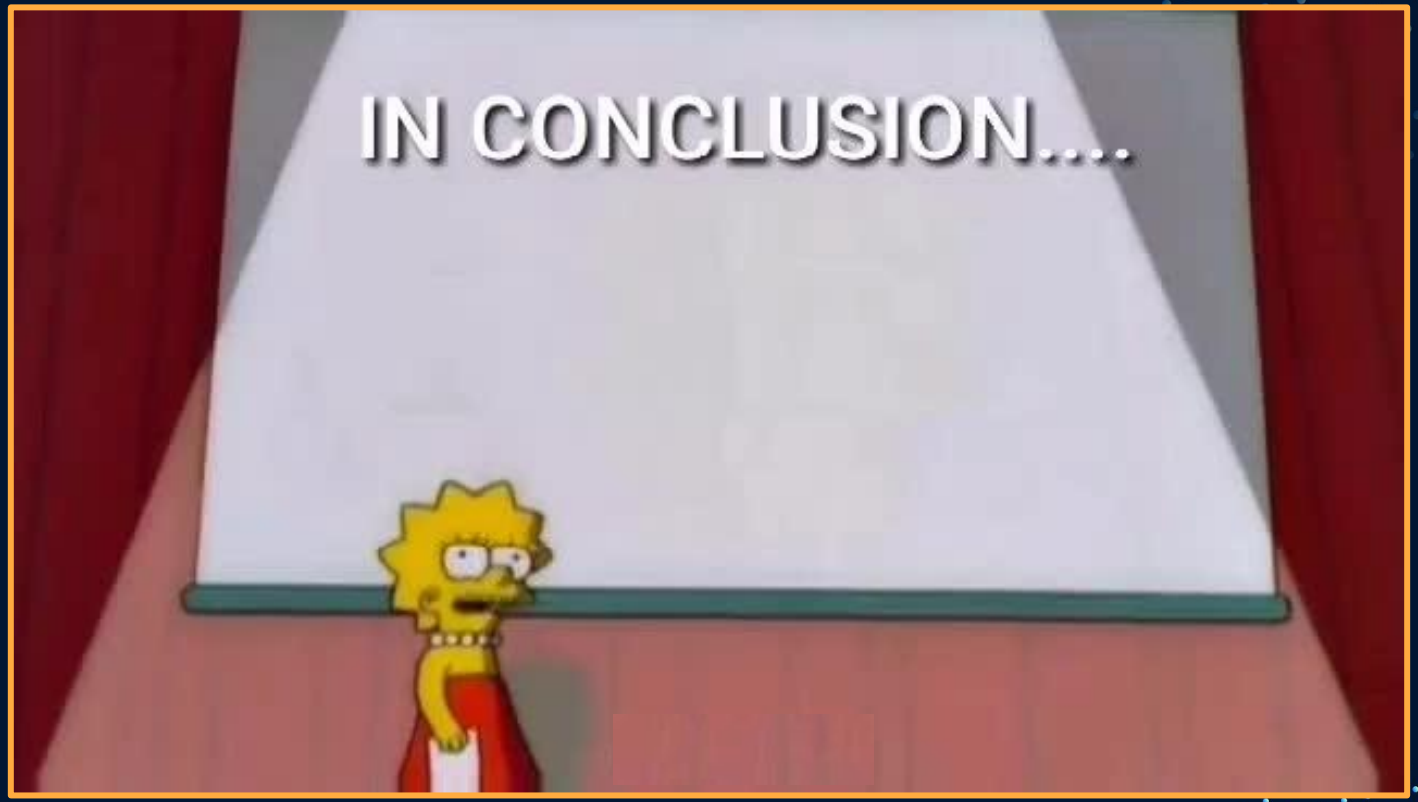
Sommaire

Améliorations

- Utiliser un score plus adapté que la précision pour notre évaluation qualitative (MRR, NDCG...)
- Utiliser des structures de données plus adaptés pour optimiser notre algorithme (sparse matrix, dictionnaires...)
- Paralléliser certains calculs pour pouvoir utiliser des jeux de données plus importants
- Ajouter des caractéristiques décrivant le film ou l'utilisateur (date, âge...)
- Optimisation grâce aux produits tensoriels

Sommaire

Conclusion



Merci pour votre attention !

“What I cannot create, I do not understand”

— Richard Feynman, physicien américain.

“When one teaches, two learn”

— Robert Heinlein, écrivain.