SORBONNE UNIVERSITÉ

PROJET DE RECHERCHE

Moteur de recommandation de séries par analyse de sous-titres

RIDA TALEB N°3804264

ACHRAF JDAY N°3802410

DAN SERRAF N°3971120

SOMMAIRE

- ❑ Nature des données
- ❑ Traitement des données
- ❑ Approche basée sur le contenu
- ❑ Approche basée sur le filtrage collaboratif
- ❑ Difficultés
- ❑ Approche hybride

NATURE DES DONNÉES

Sous-titres

- ❑ Tokenisation
- ❑ Normalisation
- ❑ Lemmatisation
- ❑ Stemming
- ❑ Stop-words

Notes utilisateurs

- ❑ Nom utilisateur
- ❑ Nom de la série
- ❑ Note personnelle
- ❑ Note globale
- ❑ Genre

TRAITEMENT DES DONNÉES

$f_{t,d}$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

$f_{t,d}$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0
n_t	**1**	**3**	**1**	**1**	**1**	**2**	**3**	**1**

$N = 4$

$$\mathrm{idf}(t,D) = \log_{10} \frac{N}{n_t}$$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

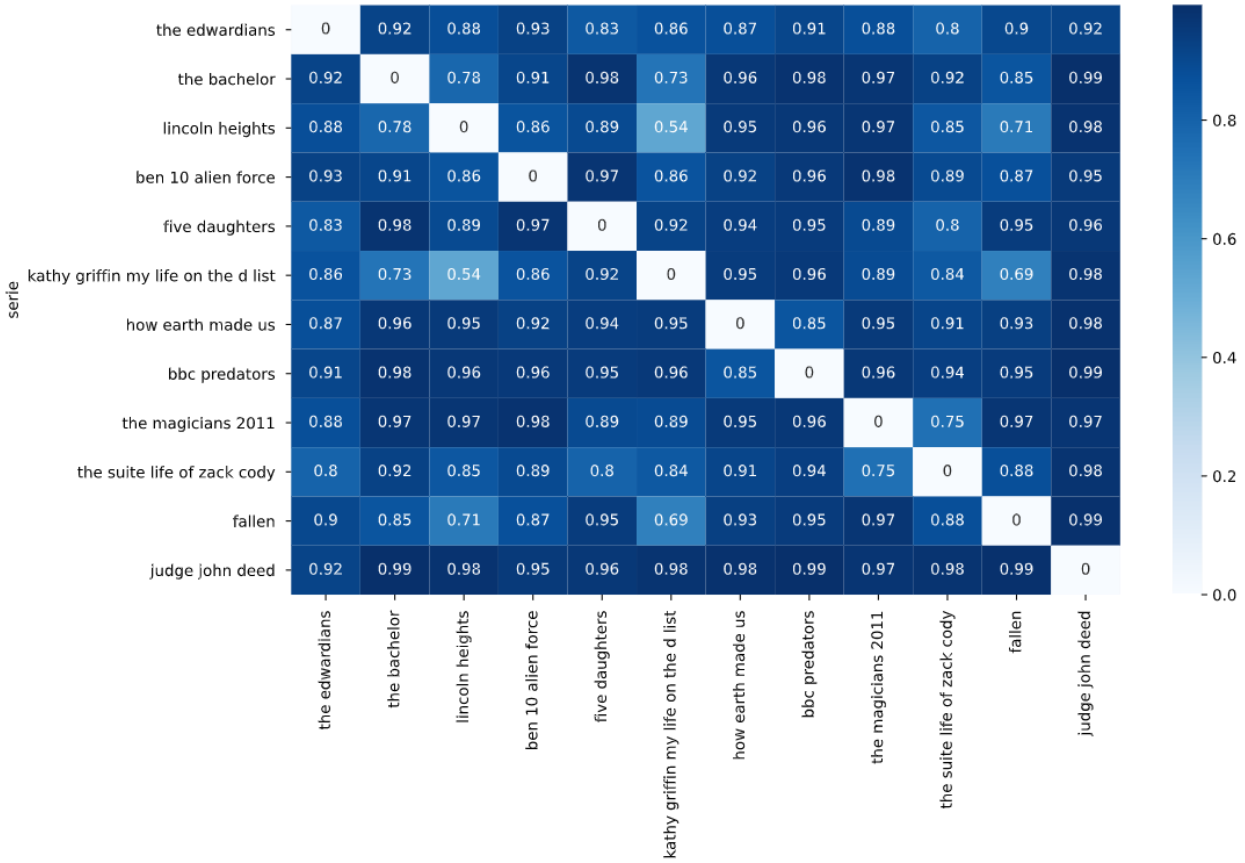
$\log_{10} \frac{4}{1} = 0.602$ $\log_{10} \frac{4}{3} = 0.125$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

	blue	bright	can	see	shining	sky	sun	today
1	**0.301**	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	**0.100**	0.0417	0
4	0	0.0209	**0.100**	**0.100**	**0.100**	0	0.0417	0

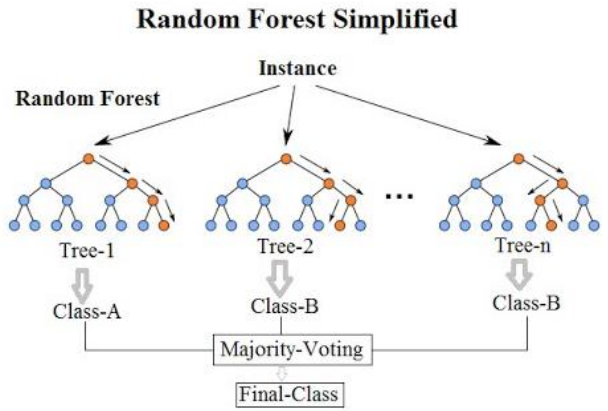
APPROCHE BASÉE LE CONTENU

Mesure de distance



Prédiction des genres

- ❑ Fusion de base données
- ❑ TF-IDF
- ❑ Le classifieur Random Forest :
  - Bootstraping
  - Création
  - Aggregation
  - Optimisation

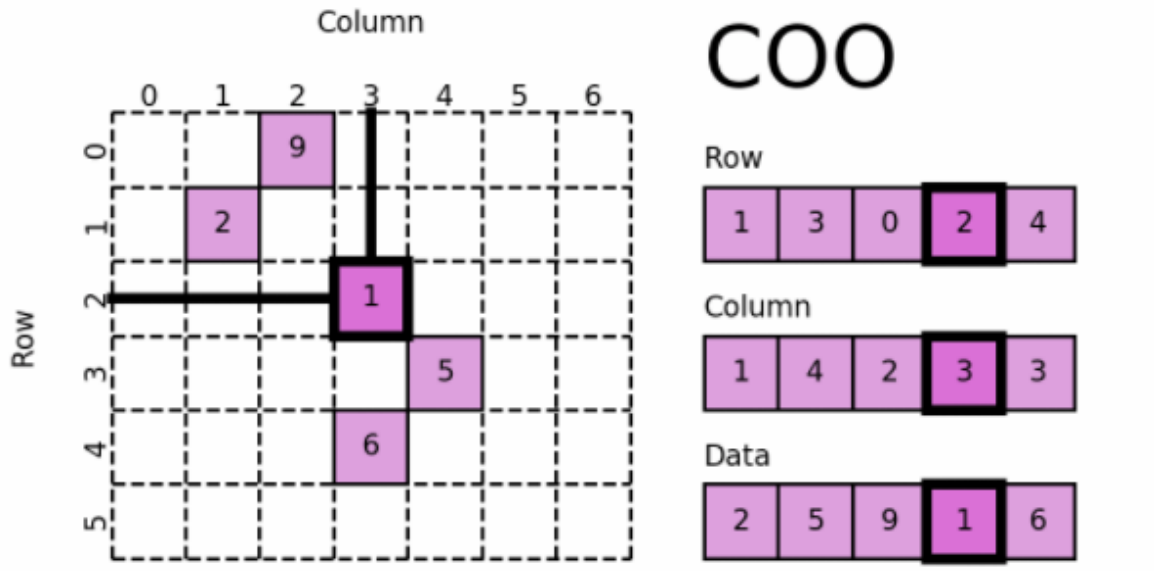


APPROCHE BASÉE SUR LE FILTRAGE COLLABORATIF

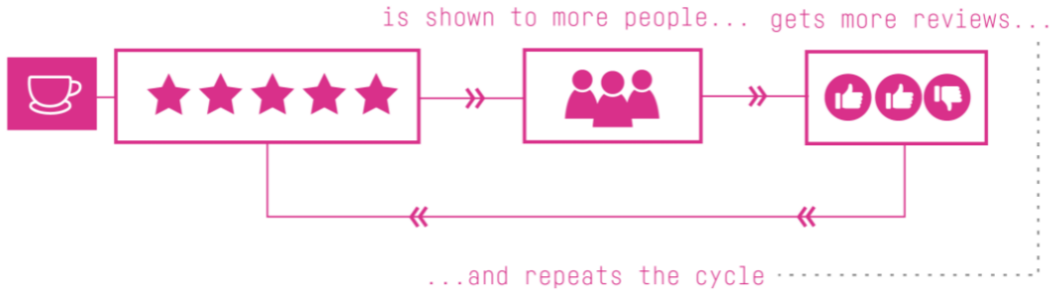
2	4	2	4	5	3
5	4	4	3	5	1
3	4	5	5	2	3
4	1	2	5	3	4
4	3	4	3	5	2
4	5	5	1	2	3

DIFFICULTÉS

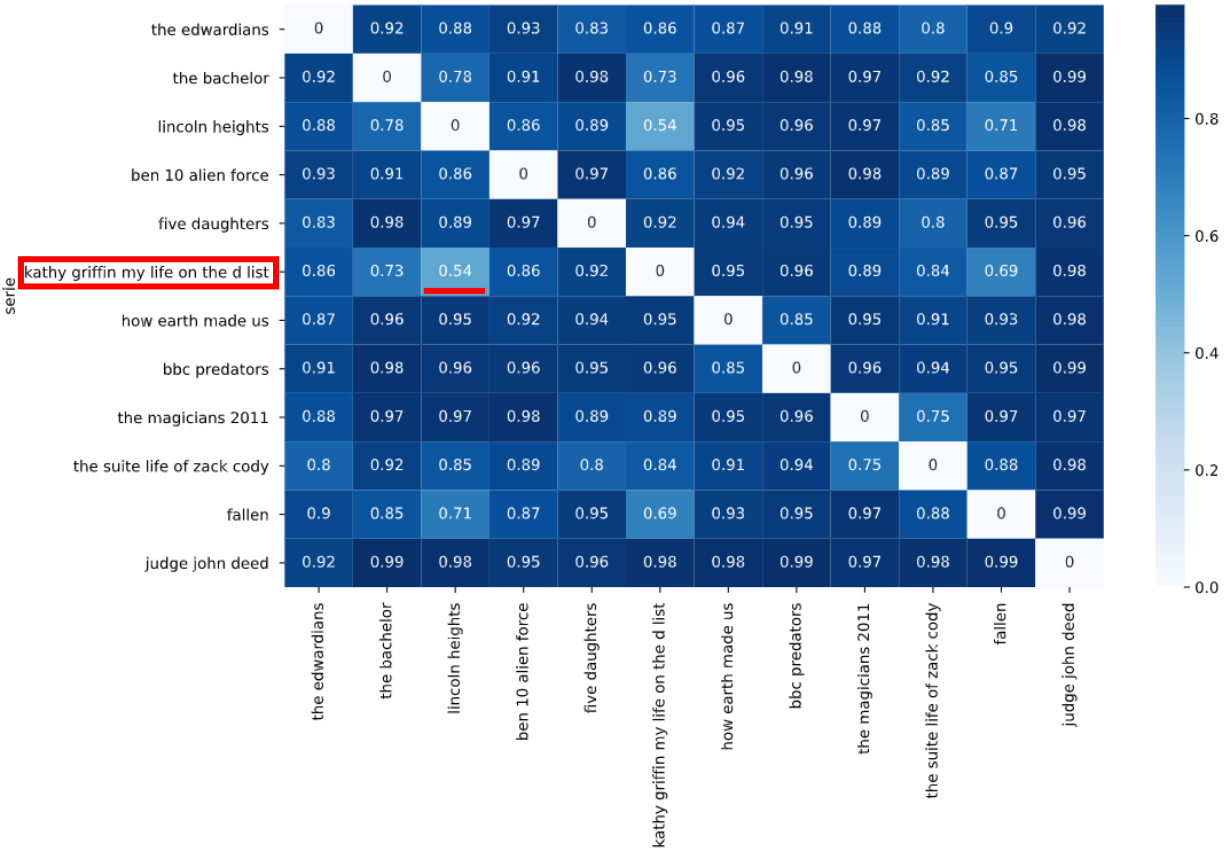
Sparsité



Démarrage à froid

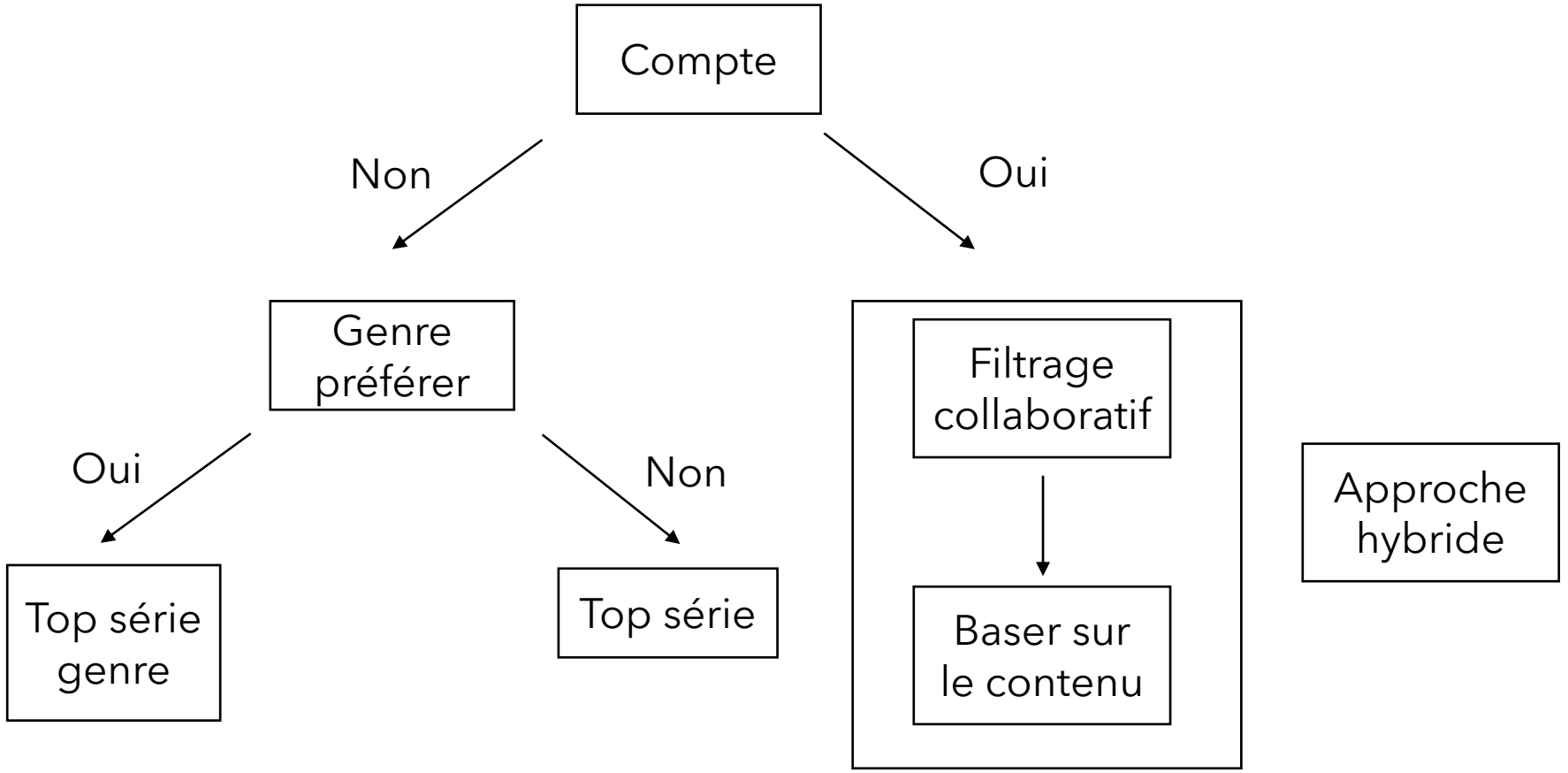


APPROCHE HYBRIDE



	2	4	2	4	5	3
	5	4	4	3	5	1
	3	4	5	5	2	3
		1		5		4
			4			2
	4	5		1		

CONCLUSION


Compte
Non
Oui
Genre préférer
Oui
Non
Top série genre
Top série
Filtrage collaboratif
Baser sur le contenu
Approche hybride


MERCI POUR VOTRE
ATTENTION