

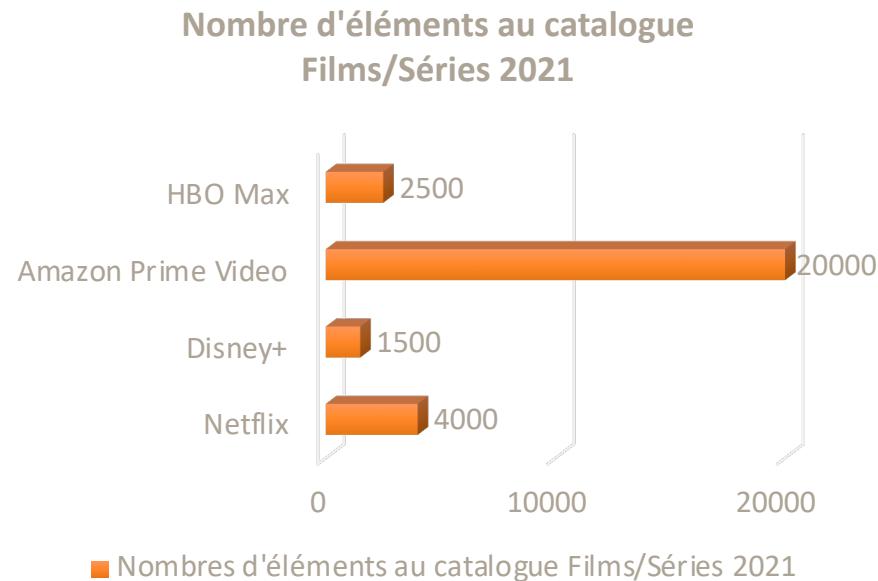
Compte rendu de projet

Algorithme de recommandation de séries

Le 29 juin 2023

Introduction : Le projet et son contexte

Est-il possible de recommander des séries en analysant les sous titres ?



Sommaire

- I) Collecte et préparation des données
- II) Comment représenter les données dans l'espace
- III) Clustering des séries télévisées
- IV) Algorithme de recommandation
- V) Conclusion

Collecte et préparation des données : Analyse et visualisation de la donnée brute, nettoyage

228
00:16:13,530 --> 00:16:14,865
Silver...

229
00:16:16,700 --> 00:16:18,076
or lead.

230
00:16:19,578 --> 00:16:21,121
You decide.

231
00:16:25,500 --> 00:16:27,711
- Let 'em go. Let 'em go.
- All right, then.



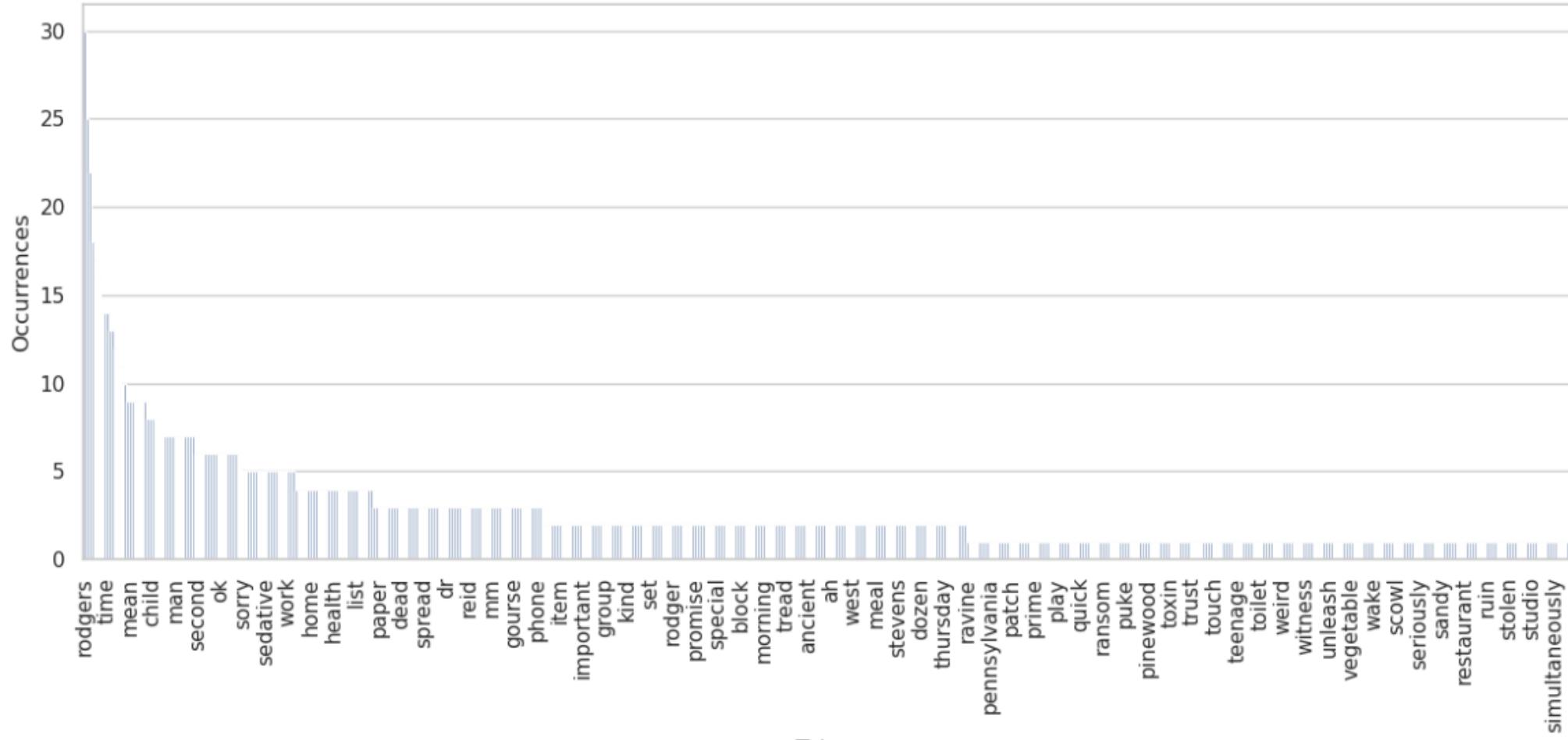
INDEX	VAL
0	silver
1	or
2	lead
...	...
8	let
9	them
10	go
11	all
12	right

Tokenisation et lemmatisation

Stemming vs Lemmatization



Introduction à la **vectorisation** des données



Term Frequency Times Inverse Document Frequency

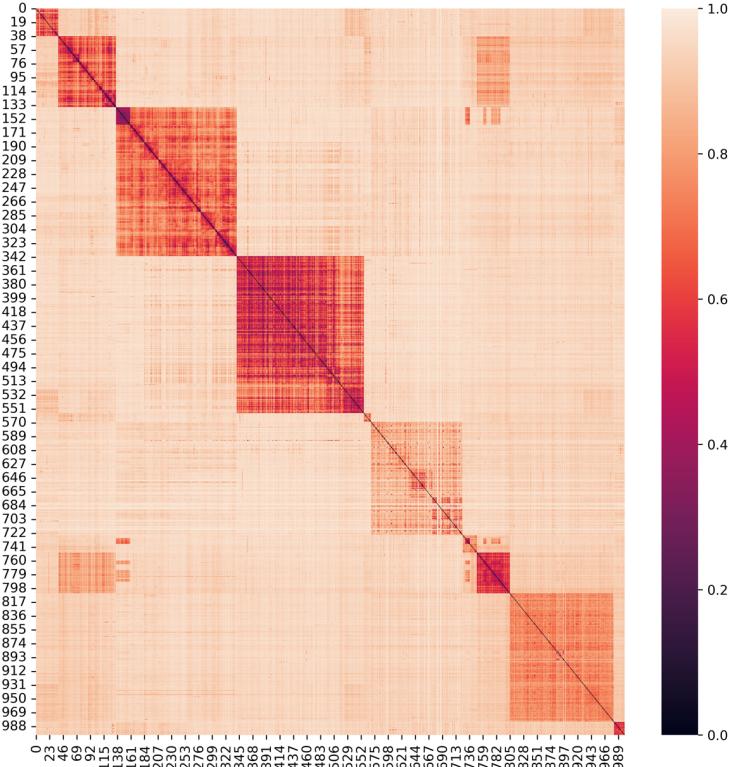
$$TF(i, j) = \frac{\log_2 (1 + Freq(i, j))}{\log_2 (L_j)}$$

$$IDF(i) = \log \left(\frac{N_D}{f_i} + 1 \right)$$

$$TF_IDF(i, j) = TF(i, j) \times IDF(i)$$

Étude de la **distance**, différents type de distances ?

MATRICE COSINUS



MATRICE EUCLIDIENNE

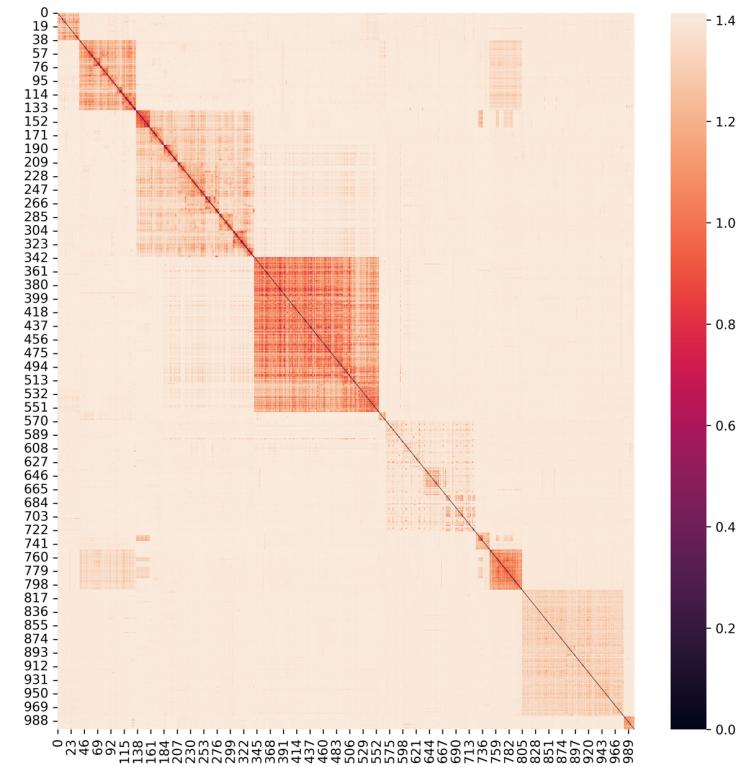
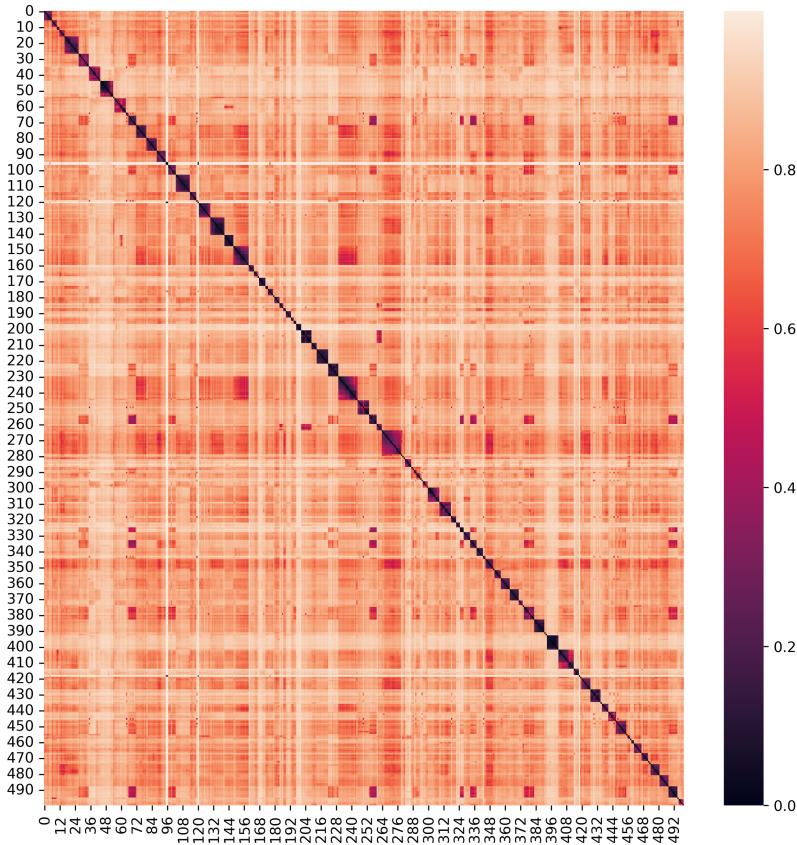
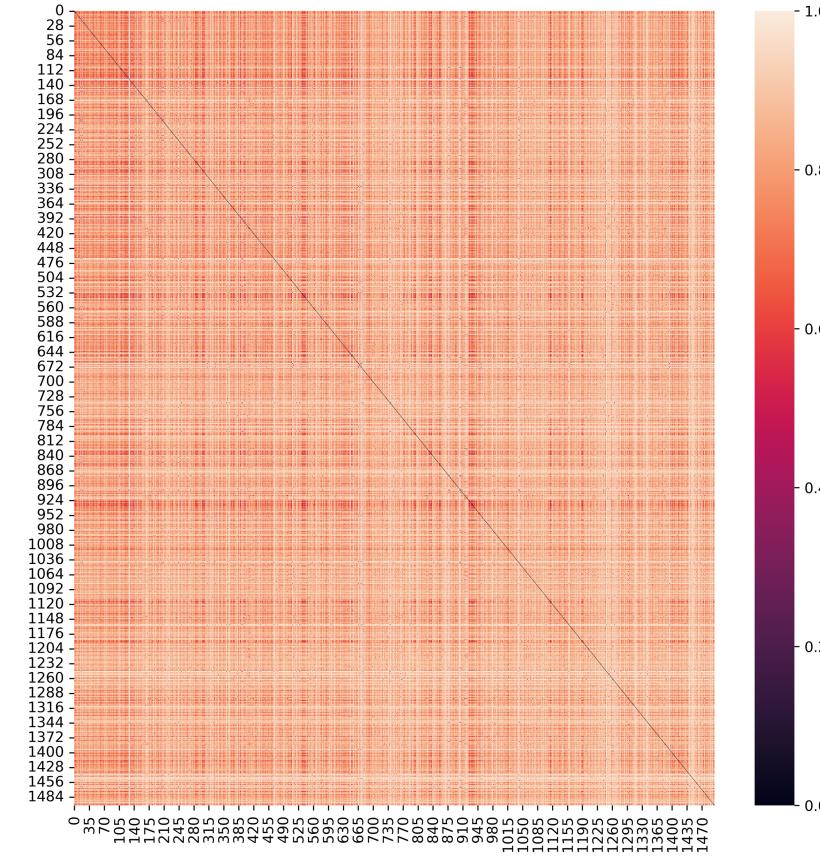


Illustration des résultats obtenu avec exemples



Matrice similarité des saisons



Matrice similarité des séries

K-Moyens

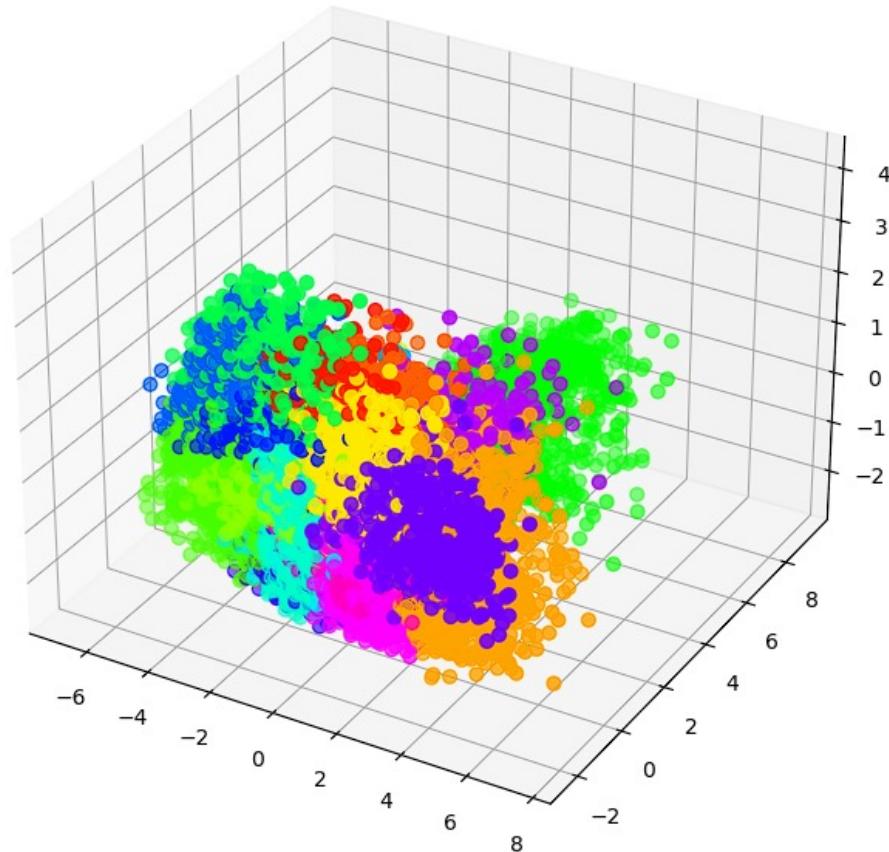
Soit $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, on cherche à partitionner les n points en k ensembles $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ ($k \leq n$)

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

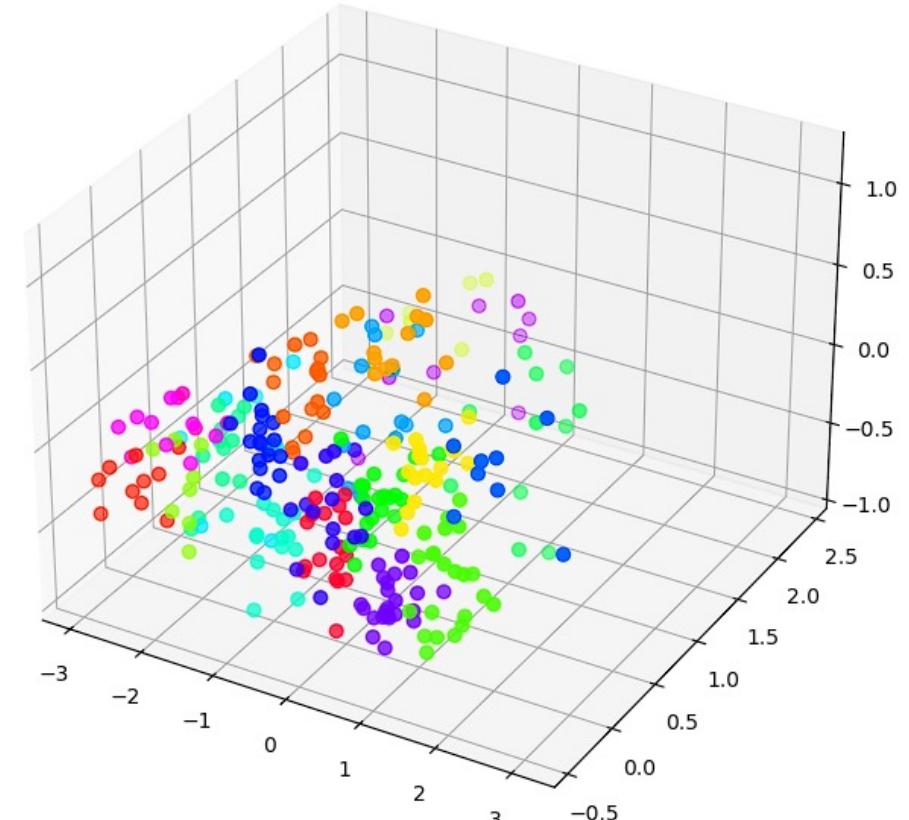
$\boldsymbol{\mu}_i$: barycentre des points dans S_i .

Clustering des séries télévisés

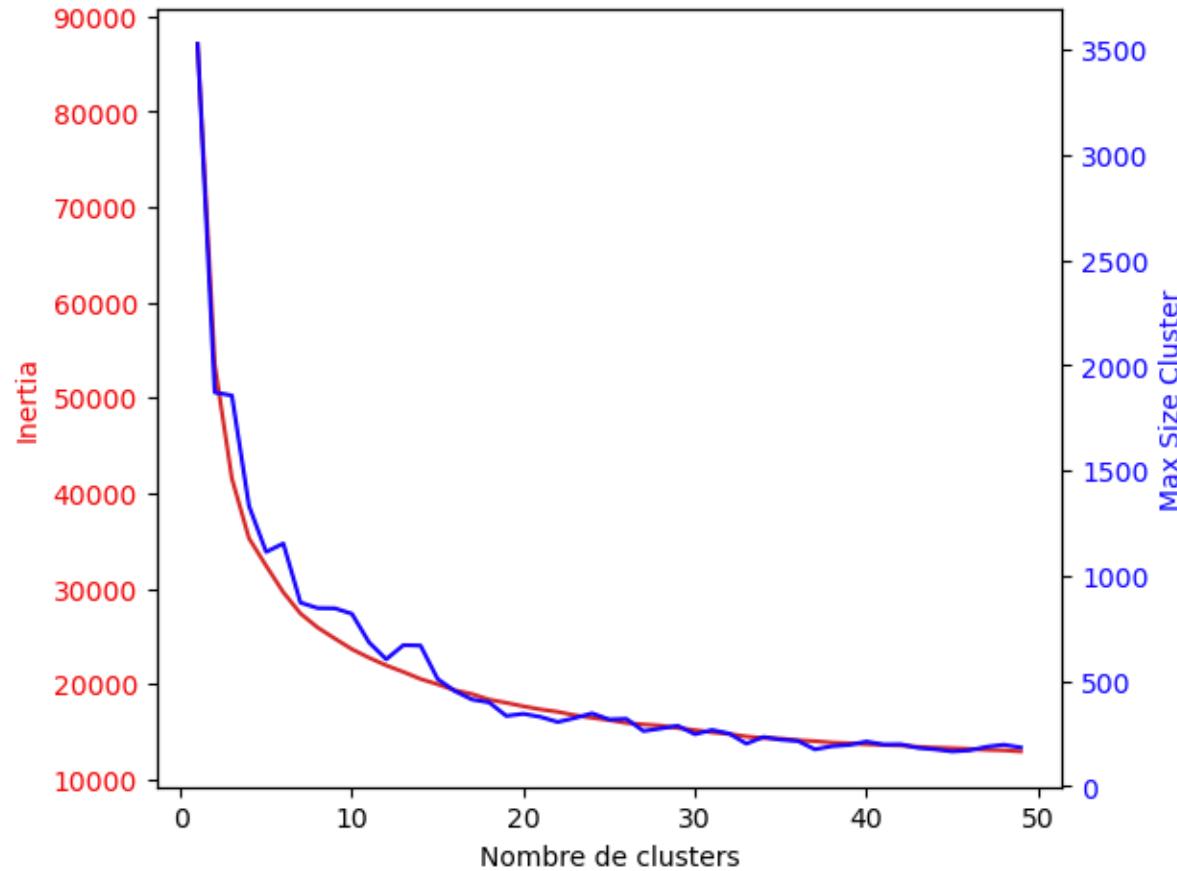
Représentation dans l'espace des épisodes clustérés



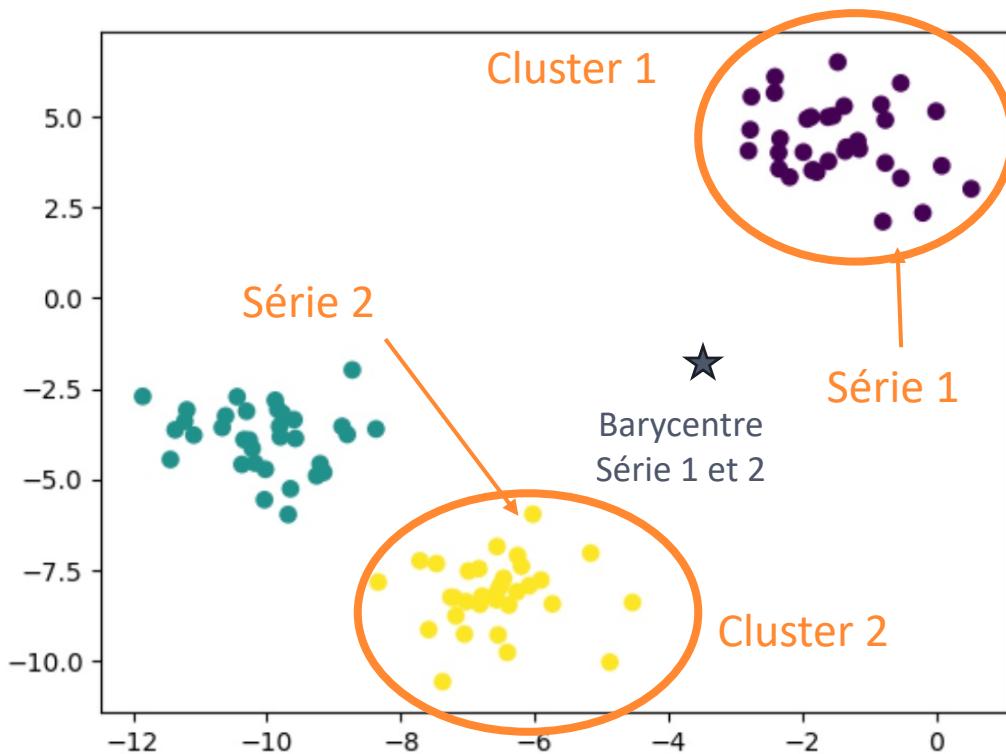
Représentation dans l'espace des séries clustérés



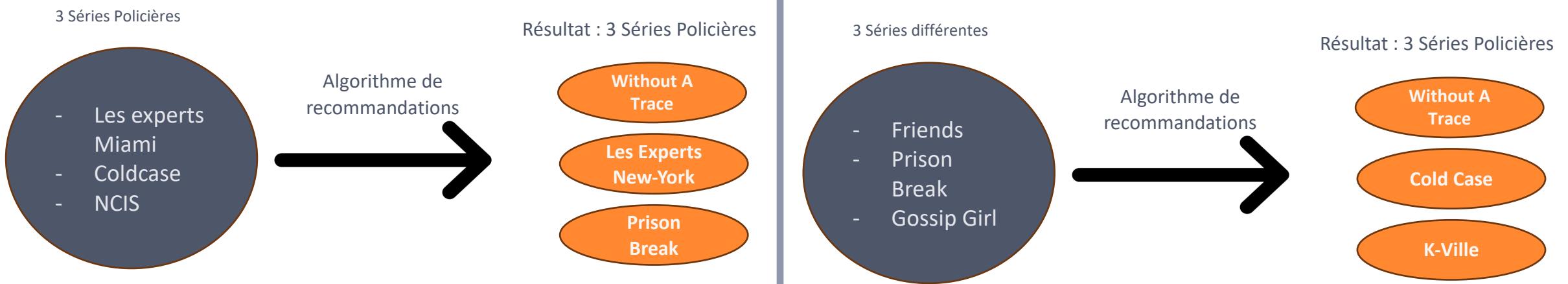
Résultats quantitatives



Algorithme de recommandations



Résultats qualitatifs



Conclusion:

Évaluation et amélioration de l'algorithme

- Amélioration :
 - NLP
 - Poids tokens
- Conclusion :
 - Recommandation par analyse de sous-titre piste envisageable
 - Manque de données nécessitant de scrapper (Avis utilisateur)