STAGE :
APPROCHES NLP POUR L'ENRICHISSEMENT DE MÉTA-DONNÉES DANS UN DATA LAKE

Stage co-encadré par Zeenea et LIP6

Tuteurs de stage :
Julien Buret (Zeenea)
Bernd Amann (LIP6)

Référent de stage :
Laure Soulier (LIP6)

AKNOUCHE ANIS

Master 2 DAC **2021/2022**

Sommaire

1. INTRODUCTION

Concepts du Data Catalogue :

- Data process
- Visualization
- Fileds
- Dataset
- Category
- Business-Item

Patrimoine de données :

- Databases
- Datasets
- Tables
- Documents

1.1 OUTILS ZEENEA

zeenea explorer

Faciliter l'exploration des données

zeenea studio

Faciliter la gestion de la documentation

Métadonnées:
-Schéma de BDD
-Noms de tables
-Noms de colonnes
-Documentations

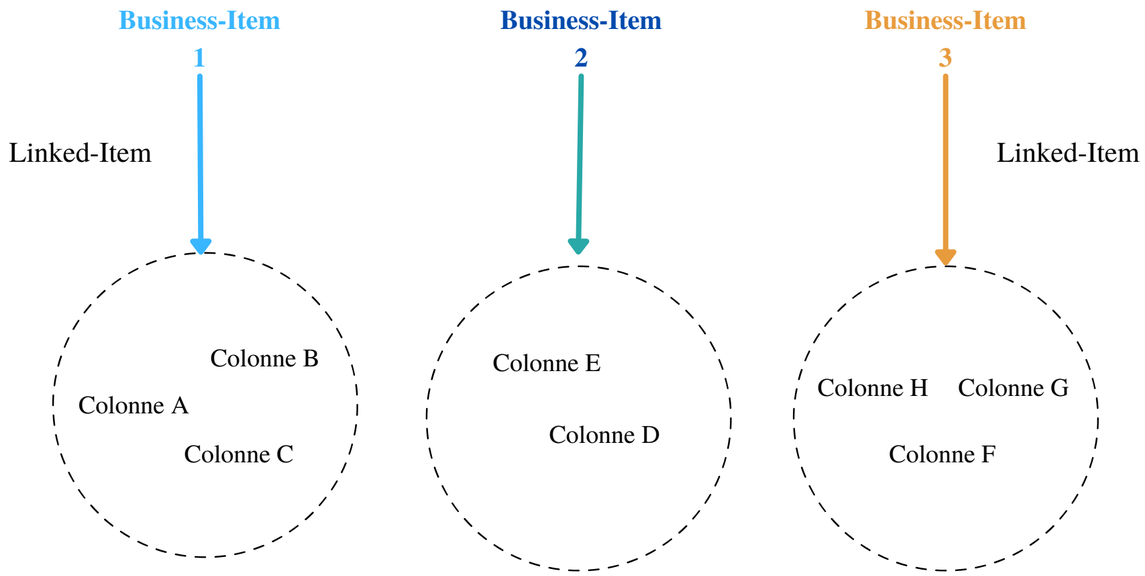
→ Liens entre les concepts du catalogue de données et les objets du patrimoine de données → Enrichir un graphe de connaissances

2. PROBLÉMATIQUE

- Inférer des suggestions de linked-items entre les business-items et les colonnes du patrimoine de données selon la similarité sémantique :

Business-item :
- Nom
- Description textuelle
- Cas d'usage spécifique

Colonne :
-Nom
-Description textuelle (Optionnelle)

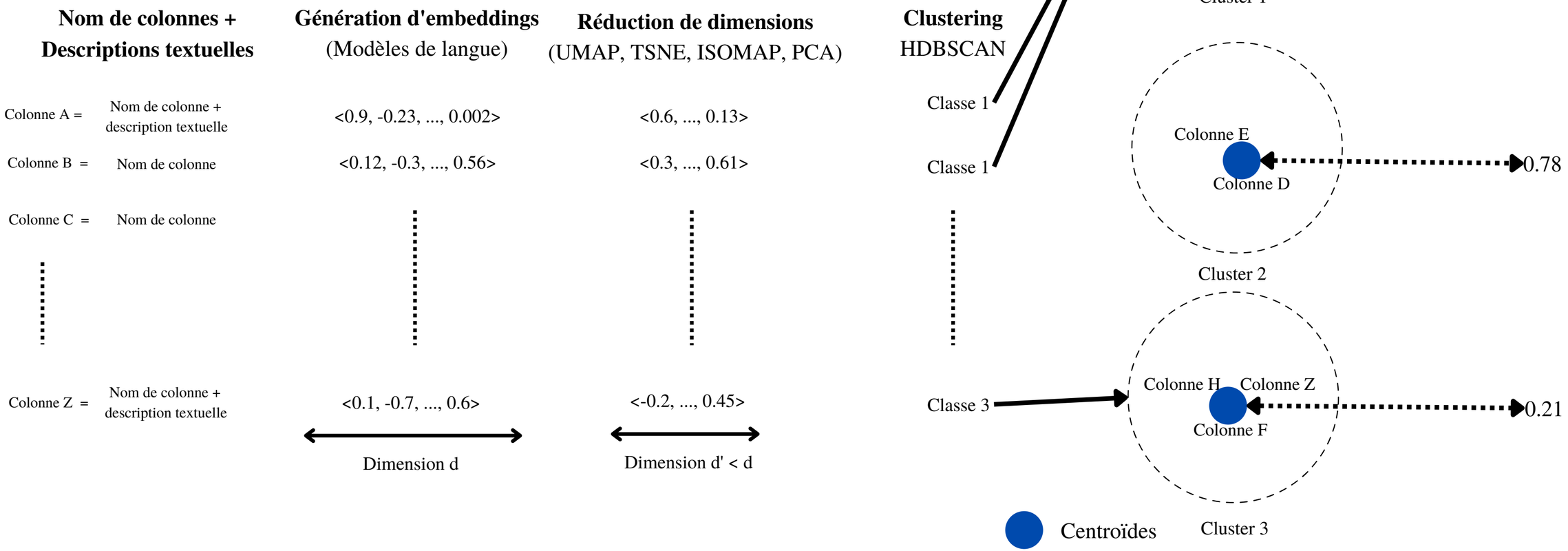


Similarité sémantique
Vecteurs embeddings

3. METHODOLOGIES

3.1 APPROCHE 1 (NOM COLONNE + DESCRIPTION)

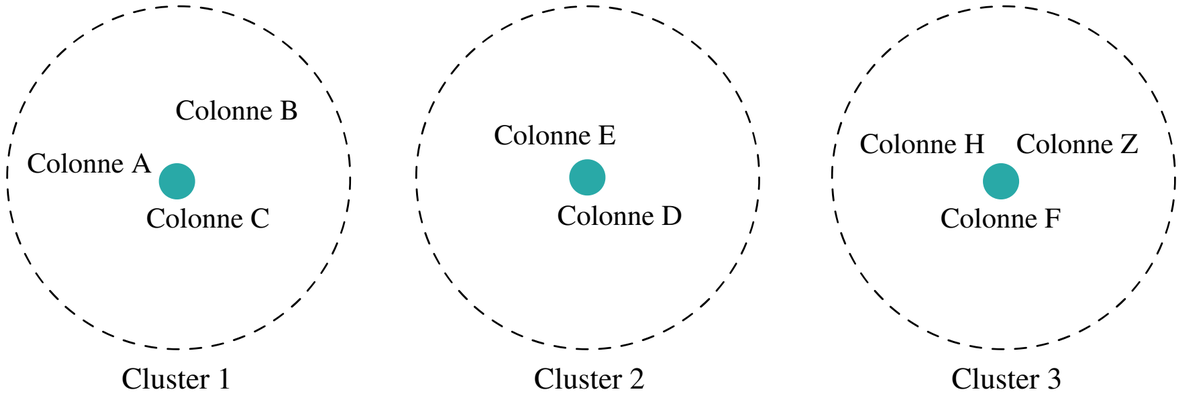
- Clustering des Colonnes avec les noms de colonnes concaténés avec les descriptions textuelles (Optionnelles):



3.2 APPROCHE 2

1- Clustering des colonnes avec descriptions textuelles :

Colonne = Nom de colonne + Description textuelle



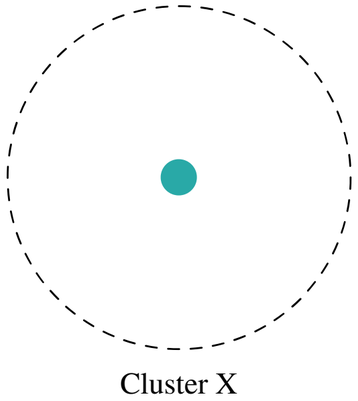
2- Clustering incrémental des colonnes sans descriptions textuelles :

Column name
(sans description textuelle)

<0.28, 0.65, 0.78, 0.001, ...>

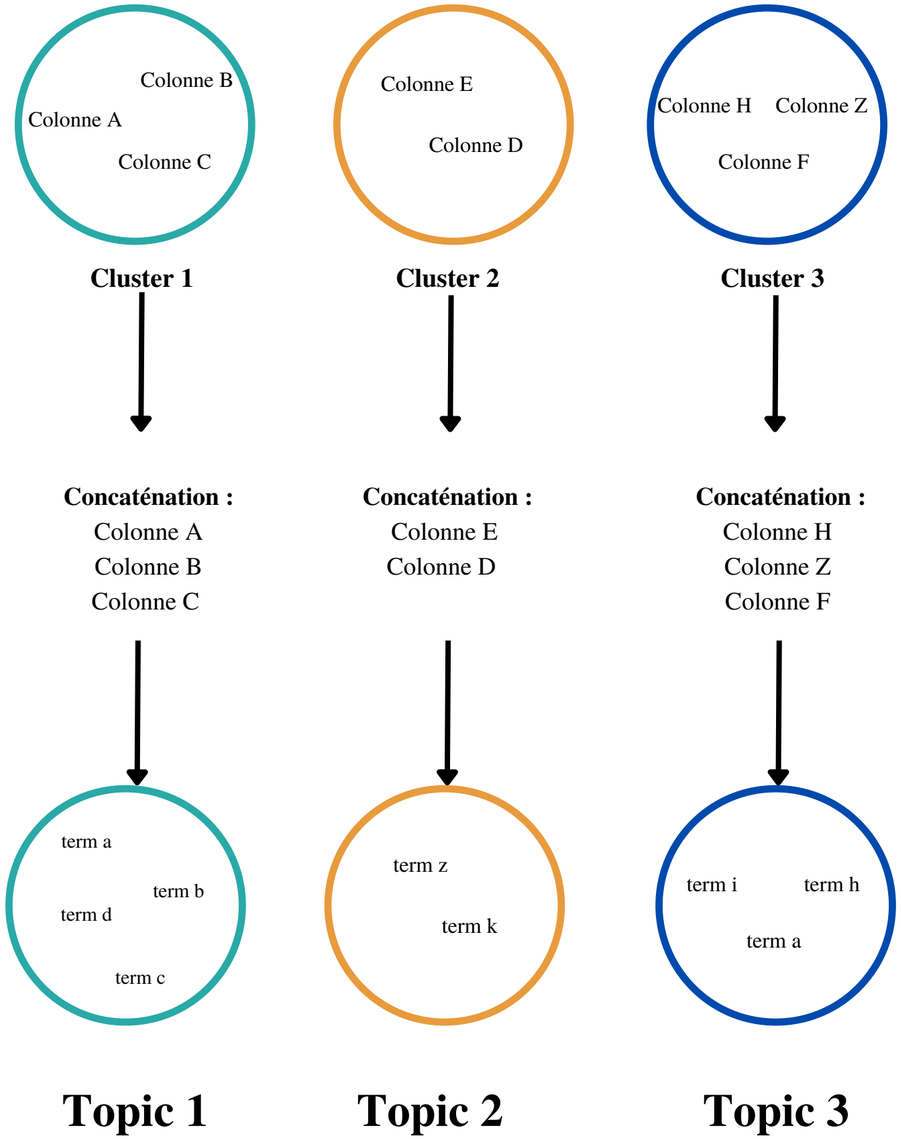
K-Nearest Neighbors

Nearest Centroid



Centroïde des noms de colonnes avec descriptions textuelles

3.3 GÉNÉRATION DE REPRÉSENTATIONS TEXTUELLES DES CLUSTERS



Cluster-based TF-IDF :

$$W_{t,c} = tf_{t,c} * log(1 + \frac{A}{tf_t})$$

$tf_{t,c}$: Fréquence de t dans le cluster c .

A : Moyenne du nombre de mots dans tous les clusters .

tf_t : Fréquence de t tout le corpus de données .

(Implémentée dans BertTopic)

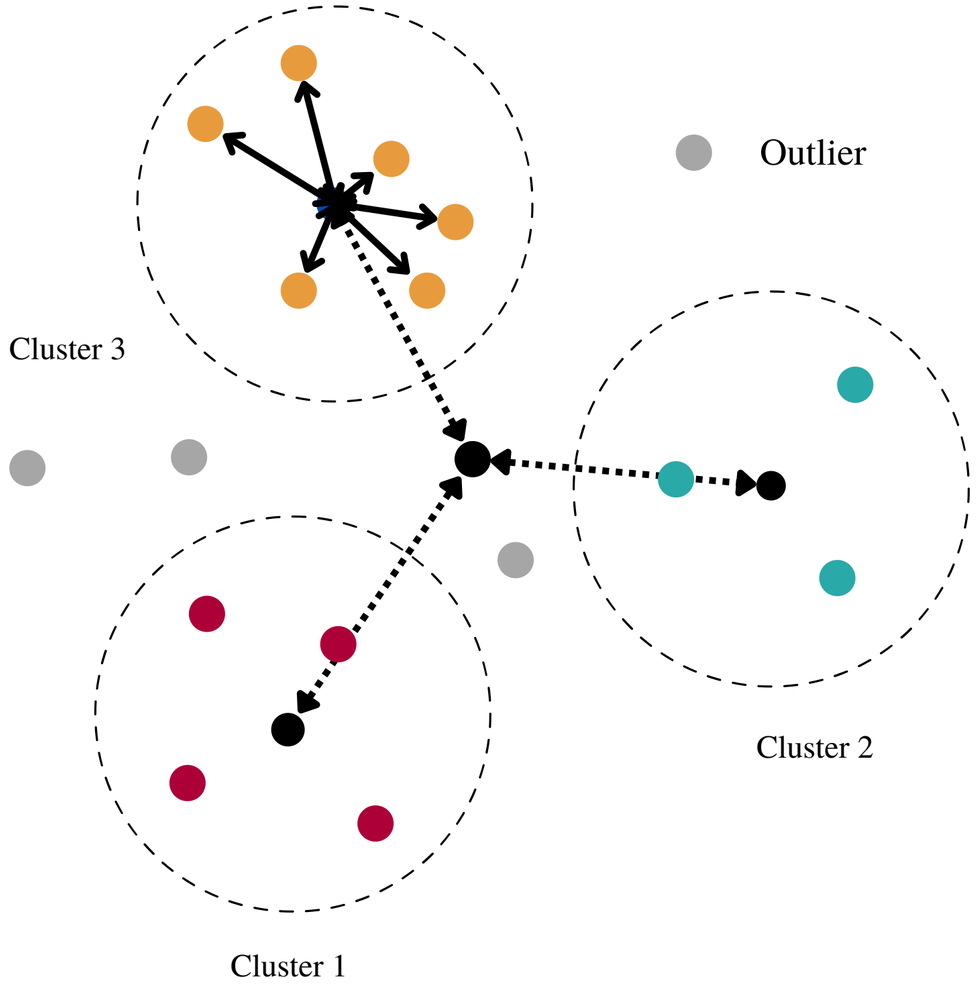
3.4 MÉTRIQUES DE MESURE DE LA QUALITÉ DES CLUSTERS :

Forme des clusters :

- Homogénéité des clusters : score d'inertie intra-cluster ⟷
- Séparabilité des clusters : score d'inertie inter-cluster ⟷
- Rapport d'inertie : Inertie intra-cluster / Inertie inter-cluster

Taux d'outliers :

Représente le taux de données classées comme étant des données aberrantes ou du bruit.



4. EXPÉRIMENTATIONS

4.1 DATASET

Données utilisées :
Métadonnées issues de BigQuery

	table_catalog	table_schema	table_name	column_name	field_path	data_type	description
0	bigquery-public-data	chicago_taxi_trips	taxi_trips	unique_key	unique_key	STRING	Unique identifier for the trip.
1	bigquery-public-data	chicago_taxi_trips	taxi_trips	taxi_id	taxi_id	STRING	A unique identifier for the taxi.
2	bigquery-public-data	chicago_taxi_trips	taxi_trips	trip_start_timestamp	trip_start_timestamp	TIMESTAMP	When the trip started, rounded to the nearest ...
3	bigquery-public-data	chicago_taxi_trips	taxi_trips	trip_end_timestamp	trip_end_timestamp	TIMESTAMP	When the trip ended, rounded to the nearest 15...
4	bigquery-public-data	chicago_taxi_trips	taxi_trips	trip_seconds	trip_seconds	INT64	Time of the trip in seconds.

Table 1 : Métadonnées issue de BigQuery. (Schéma de la table Chicago_taxi_trips))

Statistiques sur les données issues de BigQuery

257 datasets
258 813 colonnes

→

Après suppression des doublons
et des champs vides
et nettoyage du texte

- Suppression des stops words
- Suppression des caractères spéciaux
- Mise en minuscule
- Lemmatization

→

5488 colonnes **avec descriptions textuelles**

12524 colonnes **sans descriptions textuelles**

4.2 APPROCHE 1 (NOM + DESCRIPTION)

4.2.1 CHOIX DU MODÈLE DE LANGUE

Modèles de génération d'embeddings	Rapport d'inertie	
	Sans réentraînement MLM	Avec réentraînement MLM
all-mpnet-base-v2	1.78	**1.9**
all-distillroberta-v1	**1.83**	1.63
all-MiniLM-L12-v2	1.73	1.68
paraphrase-multilingual-MiniLM-L12-v2	1.73	1.34

Table 2 : Rapport d'inertie des partitionnements générés avec HDBSCAN après une réduction de dimensions avec UMAP (5 dimensions) et en utilisant différents modèles de langue (Avec et sans réentraînement MLM sur les descriptions textuelles des colonnes).

4.2.2 RÉSULTATS DES DIFFÉRENTS MODÈLES DE L'APPROCHE 1

Taux de colonnes **avec** descriptions textuelles : **30%**
Taux de colonnes **sans** descriptions textuelles : **70%**

Modèles	EMBEDDINGS	Modèle de réduction de dimensions	Dimensions	Clustering	Inertie Intra-cluster	Inertie Inter-cluster	Taux d'outliers	Nombre de clusters
KMEANS (ELBOW)	Doc2Vec (PV-DM)	/	384	KMEANS	0.68	0.52	/	24
LDA	BagOfWords	/	/	LDA	0.42	0.16	/	200
CONFIG 1	all-mpnet + MLM	**UMAP**	6	HDBSCAN	0.71	0.38	**30%**	250
CONFIG 2	all-mpnet + MLM	PCA	700	HDBSCAN	0.84	0.32	69%	67
CONFIG 3	all-mpnet + MLM	ISOMAP	300	HDBSCAN	0.82	0.37	79%	96
CONFIG 4	all-mpnet + MLM	/	768	HDBSCAN	0.84	0.33	71%	65

Table 3 : Résultats de l'approche 1 (Mesures de qualité des partitionnements générés).

4.3 APPROCHE 2

4.3.1 RÉSULTATS DE L'APPROCHE 2

1- Clustering des descriptions textuelles : Taux de colonnes **avec** descriptions textuelles : **100%**

Modèles	Génération d'embeddings	Modèle de réduction de dimensions	Dimensions	Clustering	Inertie Intra-cluster	Inertie Inter-cluster	Taux d'outliers	Nombre de clusters
CONFIG 1	all-mpnet + MLM	**UMAP**	4	HDBSCAN	0.71	0.40	**23%**	98
CONFIG 2	all-mpnet + MLM	PCA	600	HDBSCAN	0.83	0.37	78%	31
CONFIG 3	all-mpnet + MLM	ISOMAP	400	HDBSCAN	0.82	0.41	81%	27
CONFIG 4	all-mpnet + MLM	/	768	HDBSCAN	0.83	0.37	78%	30

Table 4 : Résultats de l'approche 2 (Mesures de qualités des partitionnements générés sur les descriptions des colonnes).

2- Clustering incrémental des colonnes sans descriptions textuelles :

Taux de colonnes **avec** descriptions textuelles : **30%**

Taux de colonnes **sans** descriptions textuelles : **70%**

Modèles	Génération d'embeddings	Dimensions	Méthode de clustering incrémental	Inertie Intra-cluster	Inertie Inter-cluster	Taux d'outliers	Nombre de clusters
CONFIG 1	all-mpnet + MLM	768	KNN	0.56	0.48	44%	98
CONFIG 1	all-mpnet + MLM	768	**Nearest Centroid**	0.54	0.51	**11%**	98

Table 5 : Clustering incrémental des colonnes sans descriptions textuelles
(Mesures de qualités des partitionnements générés)

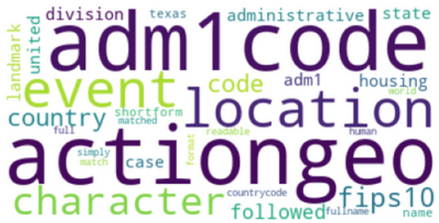
5. CAS D'USAGE

Approche 2 avec Nearest Centroid (NC)

Le business-item : Le business-item fait référence à une position géographique ainsi qu'à des informations financières et institutionnelles.

Les 3 topics les plus proches :

Topic 118 (0.47)



Topic 162 (0.33)

Topic 164 (0.32)

Figure 3 : Les wordClouds des 3 topics les plus similaires au business-item (Avec le score de similarité cosinus).

Les 10 colonnes les plus similaires avec le business-item dans chaque topic (Avec le score de similarité cosinus):

1. institution_class: 0.47
2. bank_charter_class: 0.47
3. insured_commercial_bank: 0.43
4. iba: 0.42
5. chartering_agency: 0.40
6. insured_savings_institute: 0.40
7. qbp_region: 0.40
8. federal_charter: 0.39
9. top_holder: 0.38
10. docket: 0.38

1. Actor1Geo_ADM1Code: 0.34
2. ActionGeo_ADM1Code: 0.32
3. Actor2Geo_Type: 0.32
4. Actor1Geo_Type: 0.31
5. Actor2Geo_ADM1Code: 0.30
6. ActionGeo_FullName: 0.24
7. ActionGeo_CountryCode: 0.24
8. Actor2Geo_CountryCode: 0.23
9. Actor1Geo_CountryCode: 0.21
10. Actor2Geo_FullName: 0.20

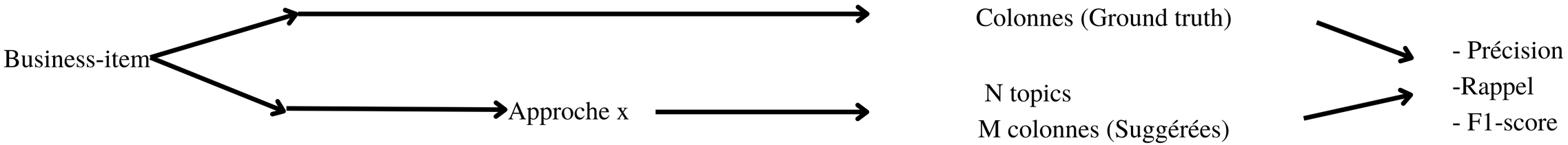
1. facility_sub_region_2: 0.35
2. sub_region_3: 0.34
3. state_province_inc: 0.33
4. country_subdivision_secondary: 0.32
5. facility_country_region_code: 0.30
6. facility_sub_region_1: 0.30
7. country_subdivision_primary: 0.29
8. country_iso_code_2: 0.29
9. province_abbreviation: 0.29
10. countries_and_territories: 0.27

6. ÉVALUATION

Simuler un utilisateur :

- 1000 colonnes avec descriptions textuelles (Business-items)
- Chaque business-item contient 25 liens vers des colonnes qui sont similaires sémantiquement.

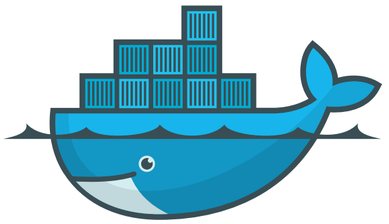
- Pour un nombre de topic N allant de 2 à 10
- Pour un nombre de colonnes M allant de 2 à 15



Approches	Précision	Rappel	F1-score
Approche 1	0.19	0.72	0.298
Approche 2 (KNN)	0.27	0.73	0.392
Approche 2 (NC)	**0.28**	**0.79**	**0.410**

Table 6 : Moyenne des scores de précision, rappel et f1-score selon le nombre de topics et le nombre de colonnes pour chaque approche.

7. DÉPLOIEMENT :



Docker

Environnement de développement : Elastic Compute Cloud (EC2) d'AWS

Bucket S3 du client
Base de données
(JSON)

Container 1 :
-Chargement des données, nettoyage et structuration
-Entraînement du modèle de langue (MLM)

Stockage Local
- Données nettoyées et structurées
- Modèle entraîné

Container 2 :
-Trouver les meilleurs hyperparamètres pour faire une réduction de dimension et faire du clustering
-Génération des représentations de topics
-Inférence des linked-items entre les colonnes du catalogue de données et les business-items

Bucket S3 Stockage
- Restitution des résultats
(Mesures de qualité des clusters et précision, rappel et f1-score entre les linked-items suggérés et les linked-items existants.

Prochaine étape : Mise en production

8. CONCLUSION

- La première approche donne de bons résultats dans le cas où on mélange les colonnes ayant des descriptions textuelles et les colonnes sans descriptions textuelles et les résultats sont meilleurs en utilisant uniquement des colonnes avec descriptions textuelles.

- La deuxième approche permet de couvrir plus de données, en revanche elle est très sensible au biais des descriptions textuelles lors de la génération des clusters.

9. PERSPECTIVES

- Inférer des suggestions de linked-items entre les autres concepts du catalogue de données et les différents objets (tables, datasets, etc) du patrimoine de données

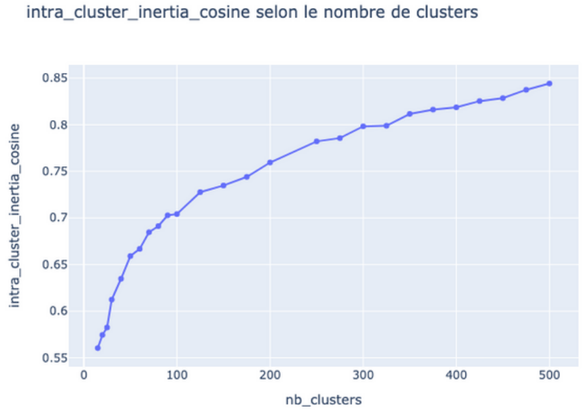
- Générer et enrichir les descriptions textuelles des différents concepts du catalogue de données

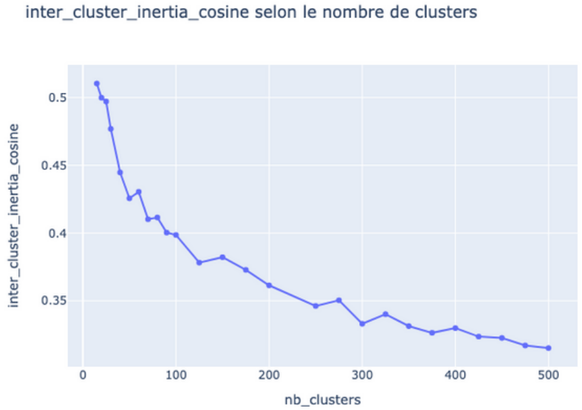
MERCI POUR VOTRE ATTENTION !

Annexe

APPROCHE 1

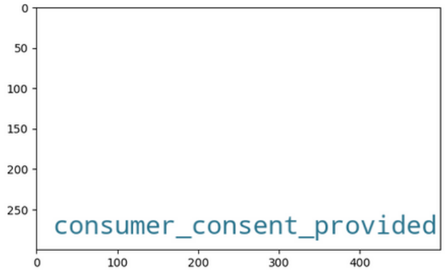
CLUSTERING AVEC LE K-MEANS



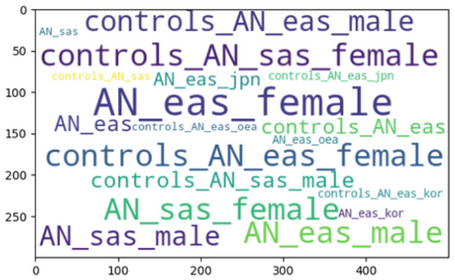


Avec un nombre de clusters de 500 :
- score d'inertie intra-cluster 0.85
- score d'inertie inter-cluster 0.31

Figure 1 : Évolution des scores d'inertie intra-cluster et inter-cluster selon un nombre de clusters allant de 2 à 500.



Topic 0 (1 élément)



Topic 45 (18 éléments))

Figure 2 : WordCloud de deux topics.

CLUSTERING AVEC LE K-MEANS

En utilisant la méthode de l'ELBOW :

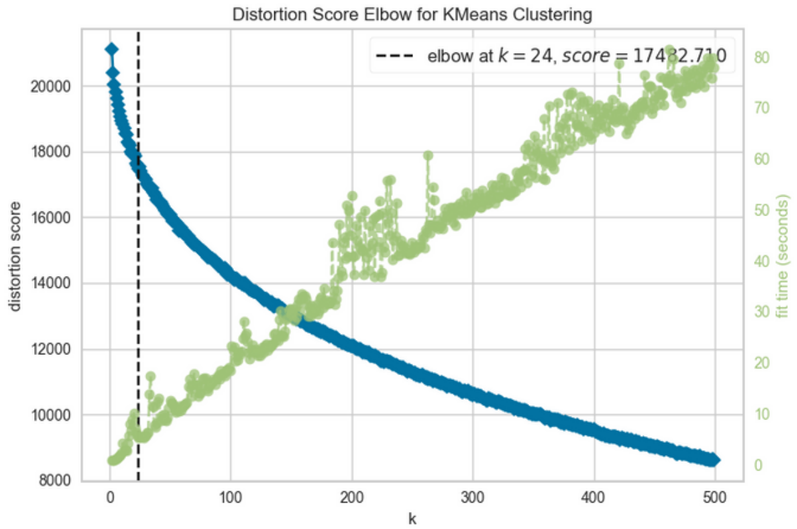
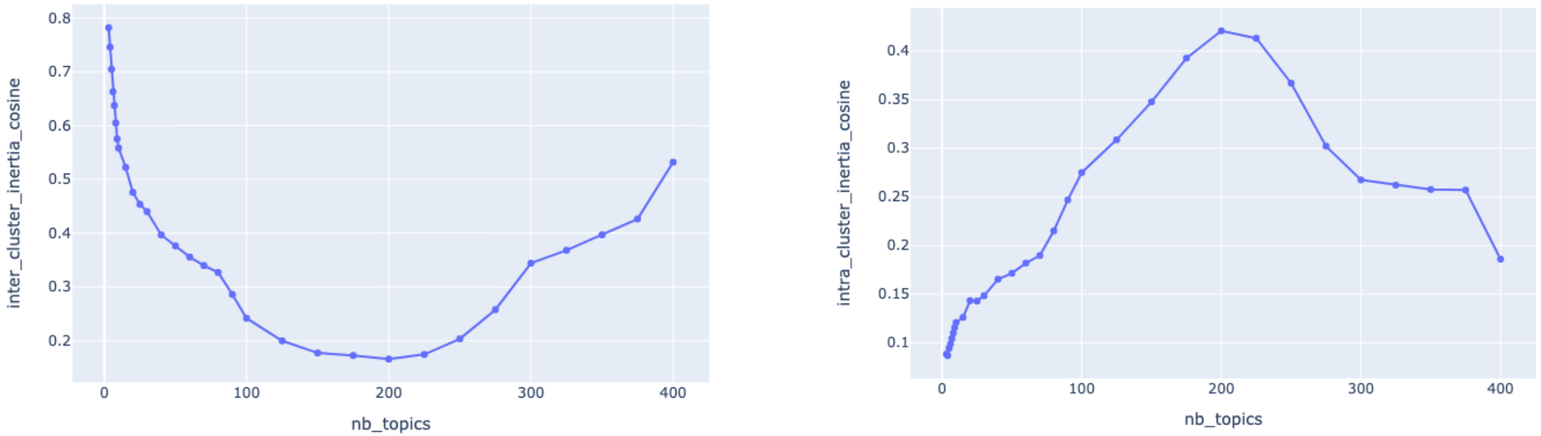


Figure 3 : Évolution du score de distorsion et du temps d'exécution du KMeans selon un nombre de clusters allant de 2 à 500 clusters.

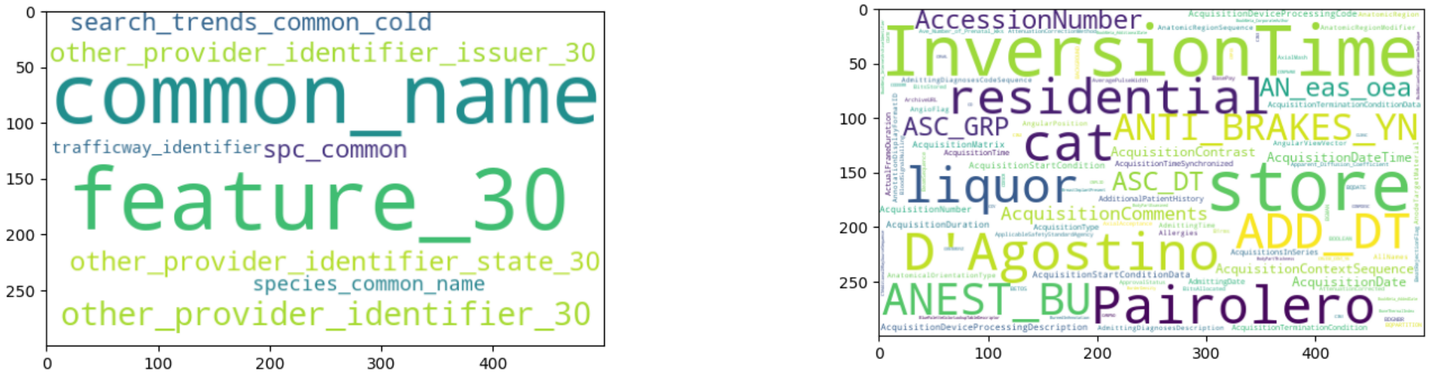
Inertie Intra -cluster : 0.68
Inertie inter-cluster : 0.52
Nombre de clusters : 24

LATENT DIRECHLET ALLOCATION (LDA)



Nombre de clusters : 200
Score intra-cluster : 0.42
Score inter-cluster : 0.16

Figure 3 : Évolution des scores d'inertie intra-cluster et inter-cluster selon un nombre de clusters allant de 2 à 500 clusters.

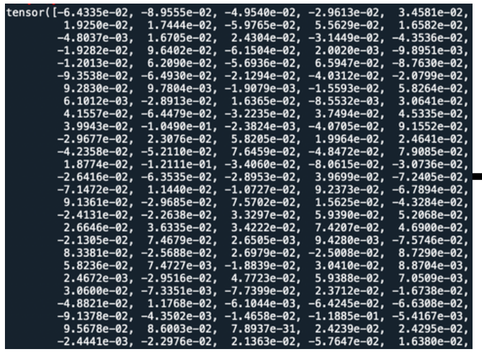


Topic 1 Topic 2

Figure 4 : WordCloud de deux topics.

BERTOPIC

Schéma du modèle BerTopic :



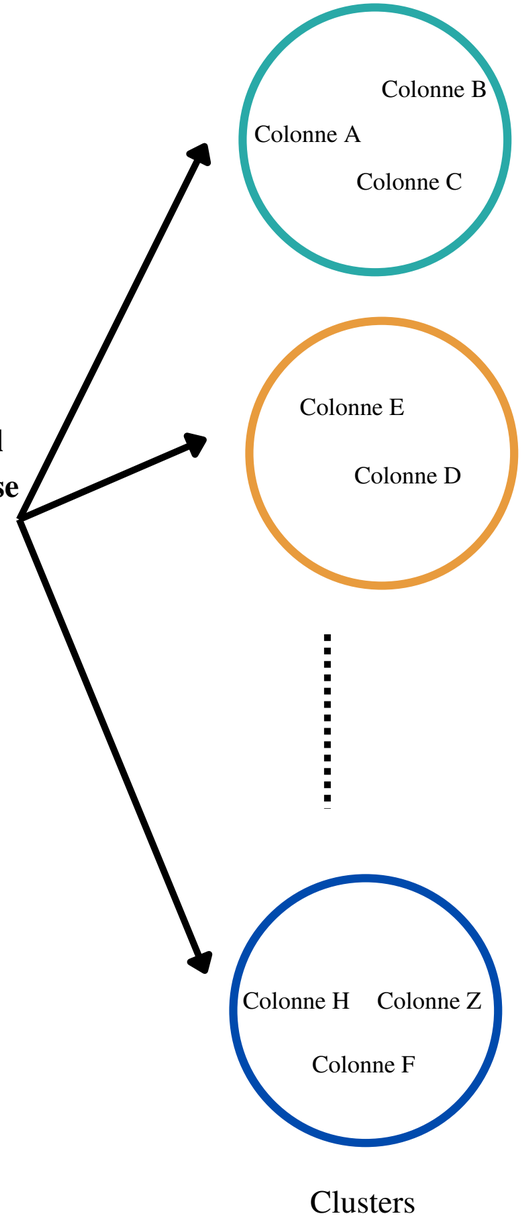
Embeddings des noms de colonnes concaténés avec leurs descriptions textuelles si elles sont disponible.

Uniform Manifold Approximation and Projection for Dimension Reduction

UMAP
(5 DIMENSIONS)

Hierarchical Density-Based Spatial Clustering of Applications with Noise

HDBSCAN



Clusters

BERTOPIC:

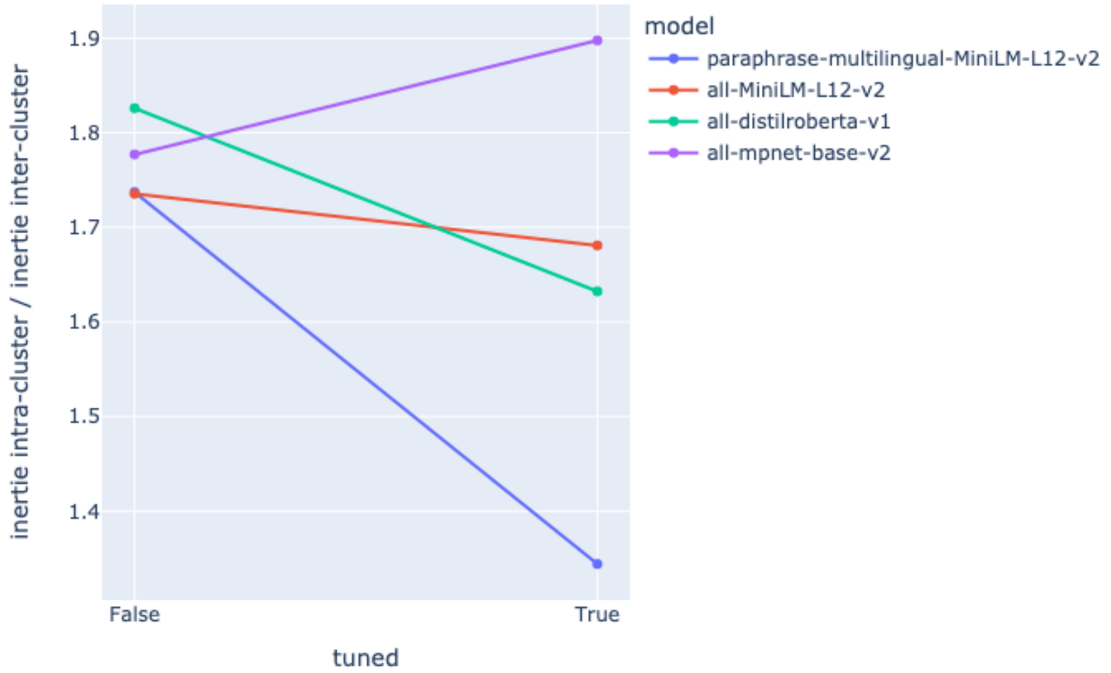


Figure 5 : Rapport d'inertie des clusters générés par un partitionnement avec HDBSCAN en utilisant différents modèles de langue (Avec et sans domaine adaptation) pour la génération des vecteurs embeddings.

RÉDUCTION DE DIMENSIONS AVEC (UMAP, PCA, ISOMAP) SUIVIE D'UN D'UN CLUSTERING AVEC HDBSCAN

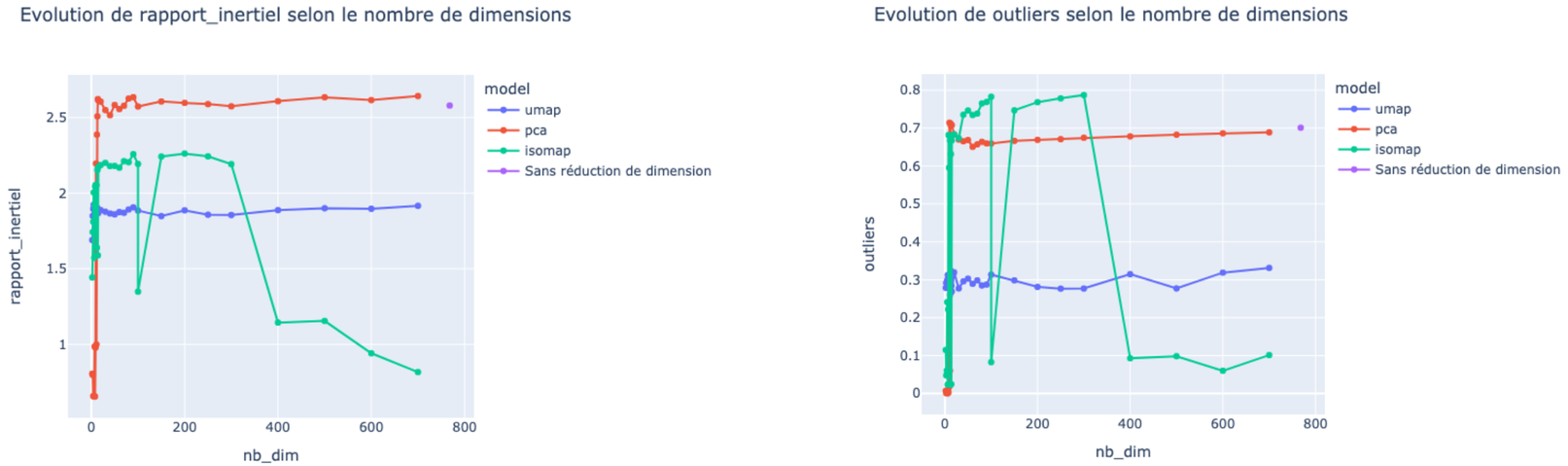


Figure 8 : Rapports inertiels et taux d'outliers des partitionnements générés par HDBSCAN avec différentes réductions de dimensions des vecteurs embeddings.

RÉSULTATS DE L'APPROCHE 2 (CLUSTERING INCRÉMENTAL DES COLONNES SANS DESCRIPTIONS TEXTUELLES)

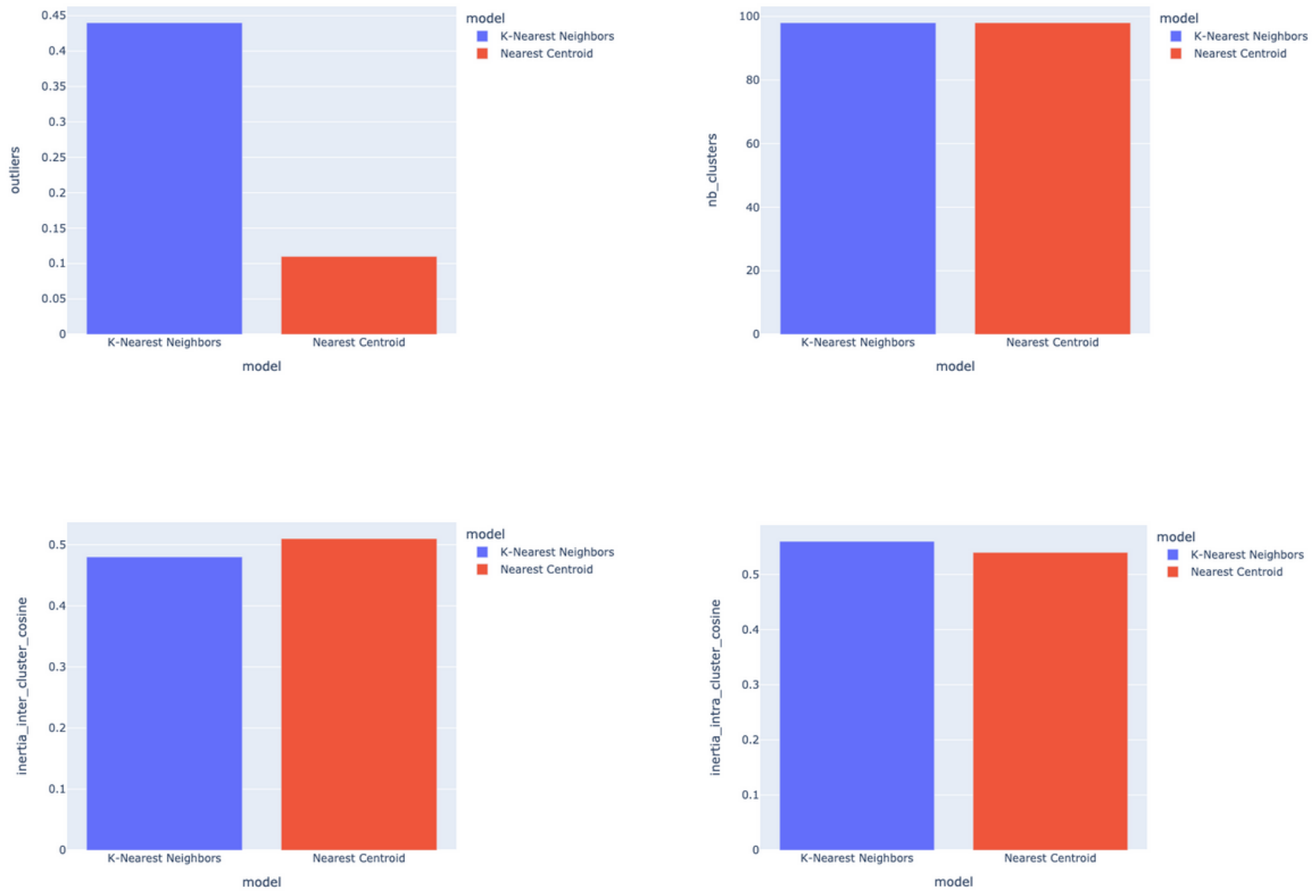


Figure 9 : Clustering incrémental des colonnes sans descriptions textuelles.

RÉDUCTION DE DIMENSIONS (UMAP ET TSNE) DES EMBEDDINGS SUIVIE D'UN CLUSTERING AVEC HDBSCAN

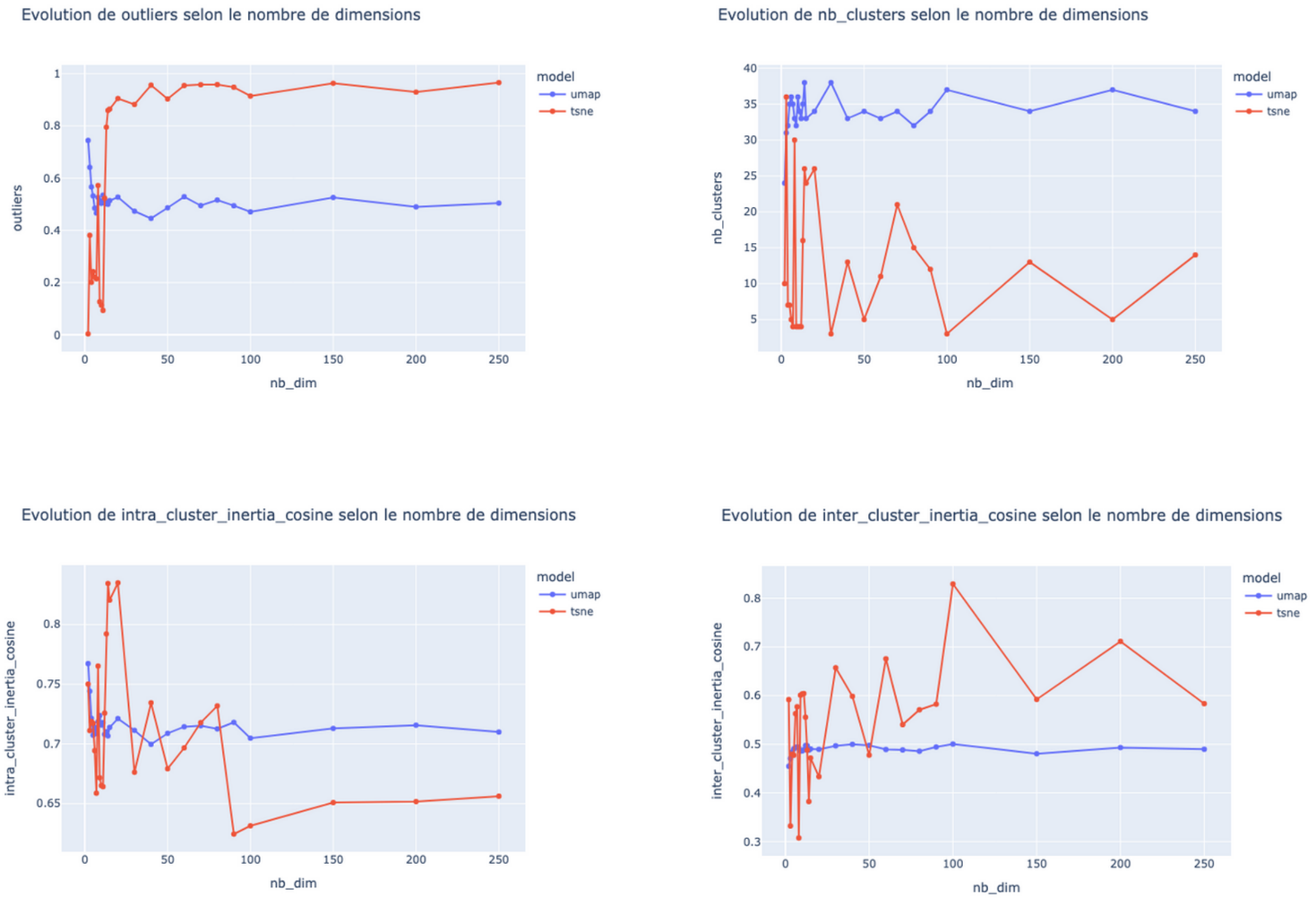


Figure 1 : Réduction de dimensions (UMAP , T-SNE) et mesure de la qualité des partitionnements générés avec HDBSCAN. (Dataset de taille 2K)

ÉVOLUTION DU TEMPS D'EXÉCUTION DE DIFFÉRENTS MODÈLES DE RÉDUCTION DE DIMENSIONS

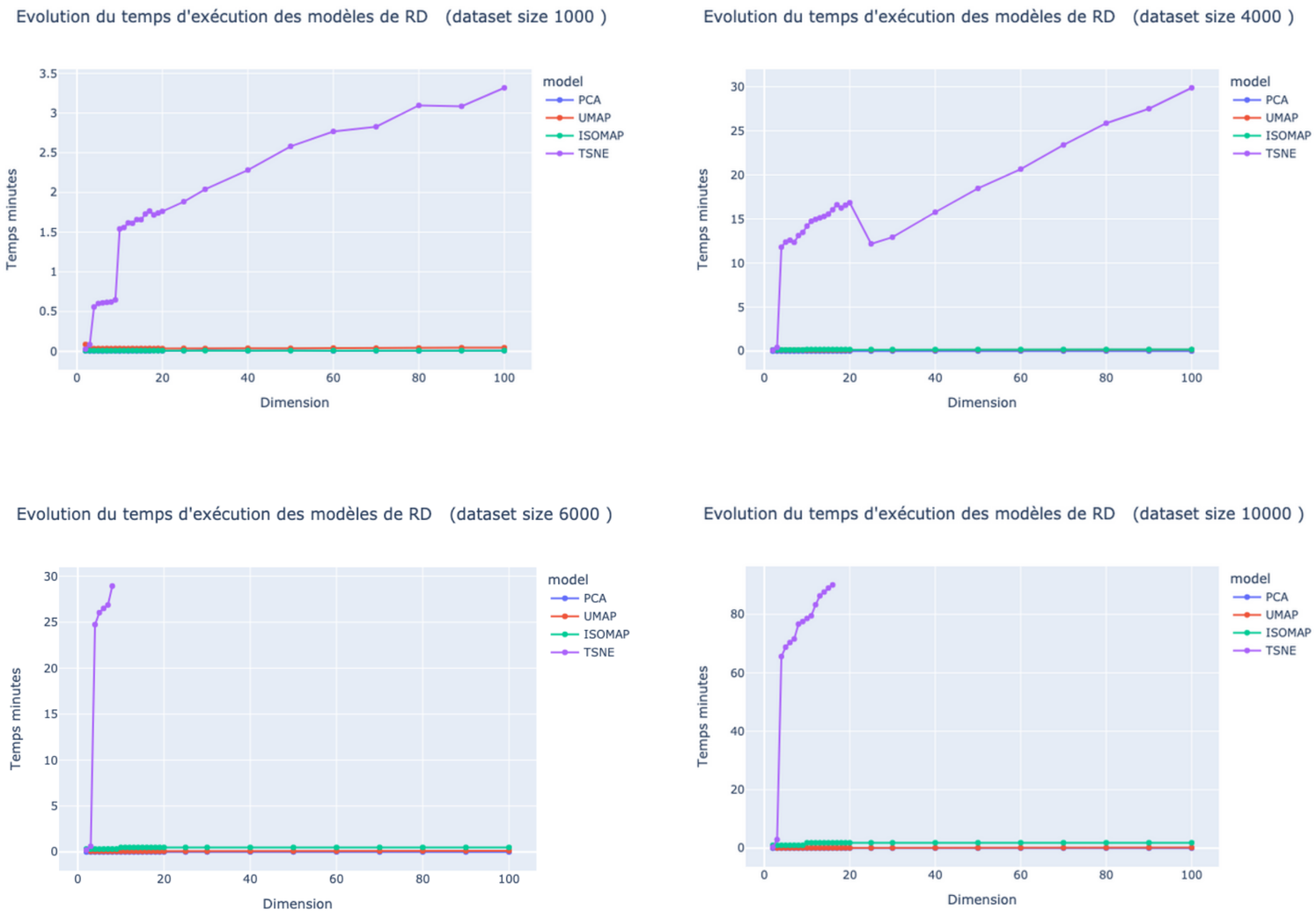


Figure 2 : Évolution du temps d'exécution des modèles de réduction de dimensions (UMAP, T-SNE, PCA, ISOMAP)