IA717 - Visualizing and analyzing AI ethics charters & manifestos with clustering

Alexandre ROCCHI-HENRY
Alexandre MALFOY
Baptiste CERVONI
Damien THAI

2024-2025

Sommaire

Table des figures

Introduction

Avec le développement de l'intelligence artificielle (IA), une prise conscience générale de ses enjeux et de ses problématiques a vu le jour. En effet, de nombreuses institutions et organismes de divers secteurs se sont intéressés à la question de l'éthique et la réglementation dans l'IA.

Durant ce projet, la collection MapAIE de 627 chartes et manifestes autour de l'intelligence artificielle et de l'éthique de l'IA sera étudiée afin d'identifier les principaux thèmes abordés de regrouper les articles selon ceux-ci. Le travail effectué dans ce rapport se basera sur l'article "*Mapping AI ethics*" [1] qui traite de la construction du corpus étudié et fournissant une première analyse des thèmes traités.

Le projet se décompose en quatres grands axes majeurs, le prétraitement des données, la vectorisation des documents, la clusterisation et topic modeling ainsi que la visualisation de ces clusters. Pour chacune des étapes, plusieurs techniques de ont été utilisées afin d'établir une pipeline optimale pour la clusterisation des articles du corpus.

I. Preprocessing

I.1. Récupération des données

Pour cette étape le code issu de l'article [1] à été utilisé et adapté. La récupération des données s'effectue à l'aide de techniques de 'crawling' pour collecter des pages web et des articles PDF. Ce processus consiste à explorer automatiquement une liste de sites portant sur l'IA et l'éthique de l'IA pour trouver des chartes et des manifestes. Ces chartes et manifestes sont alors téléchargés et mis dans le Corpus.

I.2. Nettoyage du texte

Une fois les données collectées il faut alors nettoyer le Corpus. Pour cela, plusieurs étapes sont mises en place.

a) Détection de langue :

Pour une analyse cohérente, il est crucial de déterminer la langue de chaque texte. Pour réaliser cette tâche, la bibliothèque *langid* a été utilisée. Cette bibliothèque permet de récupérer la langue dominante d'un texte avec un indice de confiance. Enfin, chaque document est associé à sa langue dominante, et tous les documents n'étant pas en anglais ont été supprimés.

b) Mise en place de filtre :

Après n'avoir récupéré que les textes en anglais, une étape supplémentaire consiste à ne récupérer que le texte des documents du Corpus. En effet, dans ce Corpus sont présentes des pages webs, et l'utilisation dans ces pages webs de balises HTML ou CSS pose problème pour récupérer le texte utile. Pour ce faire, une fonction vérifie si le texte est principalement constitué de code CSS en recherchant des motifs spécifiques via des expressions régulières. Une autre évalue le ratio de caractères non alphanumériques pour identifier les textes ressemblant à du code plutôt qu'à du langage naturel. L'utilisation de *BeautifulSoup* permet également de détecter les documents où le contenu textuel est significativement inférieur au contenu total, indiquant une prépondérance de balises HTML. De plus, les documents contenant des motifs spécifiques liés à WordPress ont été exclus. Ces filtres combinés assurent que seuls les textes pertinents et exploitables sont conservés pour l'analyse.

c) Nettoyage du Texte :

Pour affiner davantage le corpus, une étape de prétraitement du texte a été effectuée. Le texte a été converti en minuscules et les espaces ou sauts de ligne superflus ont été supprimés pour uniformiser le format. Ensuite, les mots sans importance thématique, connus sous le nom de stopwords—tels que "and", "or" ou "the"—ainsi que la ponctuation, ont été éliminés à l'aide de la bibliothèque *nltk*. Cette étape est cruciale pour se concentrer sur les termes les plus significatifs du corpus et améliorer la pertinence de l'analyse.

Après toutes ces étapes, les données récupérées sont prêtes à être analysées.

II. Embeddings

Dans ce projet, les **embeddings** jouent un rôle crucial en permettant de transformer le texte preprocéssé en données numériques exploitables. Ces données numériques permettent de capturer les relations sémantiques et syntaxiques entre les mots et les documents, ce qui permet d'analyser et de regrouper efficacement les documents.
Différentes méthodes d'**embedding** ont été utilisées :

1. **Singular Value Decomposition & Singular Value Decomposition with Positive Pointwise Mutual Information** : Analyse de la co-occurrence des mots en utilisant la décomposition en valeurs singulières.
2. **GloVe** : Utilisation de vecteurs d'embeddings par mot et calcul de moyenne pour obtenir un vecteur document
3. **TF-IDF** : Méthode pondérant les mots en fonction de leur importance relative dans les documents.
4. **RoBERTa** : Analyse utilisant un modèle pré-entraîné et de l'apprentissage profond
5. Combinaison **SVD** et **N-gram** pour obtenir des expressions de sens plus complexes

II.1. SVD & SVD_PPMI

Singular Value Decomposition

La décomposition en valeurs singulières est utilisée comme méthode de réduction de dimensions appliquée sur des matrices de cooccurrences, où chaque cellule représente la fréquence conjointe de deux mots apparaissant dans un même contexte. Le principe est le suivant :

$M = U\lambda V^T$

U et V sont des matrices orthogonales ($U^T = U^{-1}$ et $V^T = V^{-1}$). Elles contiennent les vecteurs de gauche et de droite de M.
λ est une matrice diagonale, dont les coefficients sont les valeurs propres de M.

Positive Pointwise Mutual Information

La matrice PMI permet de savoir si la cooccurrence entre deux mots est inattendue. Cette mesure est la probabilité jointe des deux mots et le produit de leurs probabilités individuels :

$$PMI\left(M, w_1, w_2\right) = log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

$$PMI\left(M, w_1, w_2\right) = log\frac{M_{w_1,w_2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} M_{i,j}\right)}{\left(\sum_{j=1}^{n} M_{w_1,j}\right)\left(\sum_{i=1}^{n} M_{i,w_2}\right)}$$

La matrice PPMI permet de mettre les valeurs de la matrice PMI négatives à 0, permettant de réduire le bruit de cette matrice, et de ne garder que les cooccurrences importantes.

$$PMI(M, w_1, w_2) = \begin{cases} \mathrm{PMI}(\mathbf{M}, w_1, w_2) & \text{if } \mathrm{PMI}(\mathbf{M}, w_1, w_2) > 0 \\ 0 & \text{otherwise} \end{cases}$$

II.2. GloVe

Global Vectors for Word Representation

GloVe est un algorithme d'apprentissage non supervisé permettant d'obtenir des représentations vectorielles des mots. L'apprentissage est effectué sur les statistiques globales de cooccurrence mot-mot d'un corpus, et les représentations résultantes mettent en évidence des sous-structures linéaires intéressantes de l'espace vectoriel des mots. Un exemple est donné dans la figure suivante :

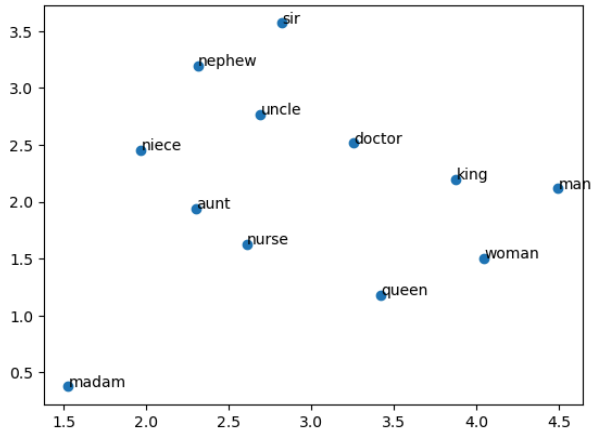

Figure 1 : Exemple de représentation avec GloVe

On peut alors apercevoir que chaque mot forment un couple de mots, par exemple : (“king”;”man”), ou (“queen”;”woman”). Dans ce projet, le modèle *"glove-wiki-gigaword-300"* est celui utilisé pour l’application de GloVe. Il a été entraîné sur 6 milliards de mots et produit des vecteurs de dimension 300 pour chaque mot.

Après l’obtention des vecteurs pour chaque mot, ceux-ci sont moyennés sur tout le document afin d’obtenir un vecteur d’embedding par document.

II.3. TF-IDF

La méthode TF-IDF (Term Frequency-Inverse Document Frequency) est une technique courante pour représenter les documents textuels sous forme de vecteurs numériques. Contrairement aux méthodes vu précédemment où chaque mot est vectorisé indépendamment, puis moyenné pour représenter un document, TF-IDF génère directement un vecteur pour chaque document.

La dimension du vecteur correspond à la taille du vocabulaire global. Chaque composante du vecteur représente le score TF-IDF d’un mot spécifique. Ce score est le produit de deux facteurs : la fréquence du mot dans le document (TF) et l’inverse de sa fréquence dans le corpus (IDF). Un score élevé indique que le mot est fréquent dans le document mais rare dans le corpus, ce qui en fait un terme significatif pour ce document. En revanche, les mots très courants dans le corpus, comme les *stop words*, obtiennent des scores faibles.

TF-IDF est simple à implémenter et interprétable. Elle est particulièrement efficace pour des corpus où la pertinence d’un mot repose sur sa fréquence relative. Cependant,

cette méthode ne capture pas les relations sémantiques entre les mots, ce qui peut limiter son utilisation dans certains contextes.

II.4. RoBERTa

RoBERTa (Robustly Optimized BERT Approach) est une méthode d'embedding basée sur un modèle pré-entraîné utilisant des techniques d'apprentissage profond. Ce modèle est un dérivé de BERT.

RoBERTa utilise des transformers pour encoder les relations contextuelles entre les mots dans un document. Contrairement aux méthodes traditionnelles comme TF-IDF ou GloVe, RoBERTa génère des vecteurs en prenant en compte le contexte global et bidirectionnel de chaque mot. Cela signifie qu'un mot sera représenté différemment en fonction des mots qui

Contrairement aux modèles unidirectionnels, RoBERTa encode les relations entre les mots en tenant compte des deux directions, ce qui permet une compréhension plus fine des dépendances linguistiques. Les embeddings alors produits peuvent être utilisés pour diverses tâches NLP, telles que la classification, la génération de texte, ou pour notre cas, le clustering.

Dans ce projet, RoBERTa a été utilisé pour obtenir des représentations vectorielles des documents. Les embeddings ont été extraits à partir du modèle pré-entraîné "RoBERTa-base", qui produit des vecteurs de haute dimension (1024) capturant les relations complexes au sein du texte.

II.5. Combinaison SVD et N-gram

Les n-gram sont courants en NLP pour analyser et représenter des textes afin de représenter des unités de sens plus longues sur plusieurs mots. Un n-gram est une séquence de n-éléments consécutifs permettant de capturer plus facilement des cooccurrences.

Pour pouvoir l'appliquer à la méthode SVD, il faut alors recalculer la matrice de cooccurrence. Chaque ligne représente donc un n-gram et chaque colonne un autre n-gram. Une fois que cette matrice est calculée, la même méthode que pour SVD est appliquée.
L'application des n-gram n'est efficace que pour la méthode SVD, car pour la méthode SVD_PPMI, la matrice PMI peut dévaloriser les n-gram, car elle favorise les cooccurrences inattendues, or les n-gram se basent beaucoup sur les dépendances locales, étant beaucoup moins inattendues.

III. Clustering

III.1. Clustering

Dans le cadre de ce projet, trois méthodes principales de clustering différentes ont été utilisées, à savoir : la méthode des KMeans, la méthode de Hierarchical Clustering ascendant, et la méthode de Gaussian Clustering.

KMeans : les données sont regroupées en k clusters en suivant un processus itératif :

1. Initialisation des centroïdes : k centroïdes (points représentatifs des clusters) sont initialisés aléatoirement.
2. Assignation : Chaque point est assigné au cluster dont le centroïde est le plus proche (en général, selon la distance euclidienne).
3. Mise à jour des centroïdes : De nouveaux centroïdes sont calculés comme la moyenne des points dans chaque cluster.
4. Convergence : L'algorithme s'arrête lorsqu'il n'y a plus de changements significatifs dans les assignations ou les centroïdes.

Cette méthode a pour avantage d'être simple à implémenter et présente une complexité relativement faible d'où sa rapidité sur des données volumineuses. De plus , elle est efficace sur des clusters sphériques, distincts et non chevauchants facilitant l'interprétation.

Clustering hiérarchique ascendant : une hiérarchie de clusters est construite sous forme d'un arbre :

1. Initialisation : Chaque document est initialement traité comme un cluster indépendant.
2. Fusion : À chaque itération, les deux clusters les plus proches (selon une métrique comme la distance euclidienne ou de Manhattan) sont fusionnés.
3. Arbre relationnel : Le processus continue jusqu'à ce qu'il ne reste qu'un seul cluster englobant toutes les données.

Cette méthode de clustering fonctionne bien pour des clusters de tailles, densités et formes variées. De plus, elle n'a pas besoin de définir le nombre de clusters à l'avance et est Robustesse au bruit.

Gaussian Clustering : GMM modélise les données comme un mélange de plusieurs distributions gaussiennes dont Chaque cluster est caractérisé par une moyenne, le centre du cluster, et une matrice de covariance (Σ) définissant la dispersion et l'orientation.

Cette méthode permet d'identifier des formes elliptiques et variées. En effet, GMM est adapté aux clusters de tailles et densités différentes, y compris des formes elliptiques ou irrégulières. De plus, ce clustering permet de traiter les clusters imbriqués ou partiellement chevauchants sont bien gérés car il fournit des probabilités pour chaque point, permettant une interprétation plus nuancée (points pouvant appartenir à plusieurs clusters).

Ces trois méthodes de clusterings sont utilisées en parallèle afin de capturer une plus grande panoplie de clusters de tailles et de formes différentes.

III.2. Scores et Résultats

Explication des scores

● **Silhouette score :** $s_{sil}(i) = \frac{b(i) - a(i)}{max(a(i),b(i))}$

Avec :

- a : distance moyenne de i par rapport aux points du cluster d'appartenance
- b : distance moyenne de i par rapport aux points du cluster voisin

● **Davies score :** $R_{ij} = \frac{s_i + s_j}{d(i,j)}$

Avec :

- s_i et s_j : distance moyennes des points aux centroïdes (taille du cluster)
- d_{ij} : distance entre les centroïdes des clusters

● **Calinski-Harabasz score :** $CH = \frac{trace(B_K)}{trace(W_K)} \times \frac{N-k}{k-1}$

Avec :

- B_k : matrice de dispersion inter clusters
- W_k : matrice de dispersion intra clusters

Métriques	Intervalle de définition	Critère de qualité	Utilisation
Silhouette	$[-1; 1]$	proche de 1	homogénéité intra-clusters et la séparation entre clusters sont essentielles
Davies-Bouldin	$[0; \infty[$	proche de 0	clusters de densité variée, minimiser le chevauchement
Calinski-Harabasz	$[0; \infty[$	très élevé	maximiser la séparation et la densité des clusters

Figure 2 : Tableau récapitulatif des différentes métriques de scores

Plusieurs scores sont utilisés afin d'avoir une validation croisée se basant sur plusieurs critères de qualité des clusters, comme l'homogénéité et la densité des clusters ou bien le non chevauchement et la séparation des clusters. De plus, l'utilisation de plusieurs

scores est préférable car cela permet de minimiser certains biais qu'une méthode de scoring pourrait avoir, en offrant plusieurs points de vue sur la qualité des clusters.

Tableau des scores

Une pipeline générale a été mise en place pour tester chaque combinaison de méthode d'embedding et de clusterisation afin de trouver la solution optimale. En effet, cette pipeline prend en entrée une fonction d'embedding, de clustering et de réduction de dimension avec affichage. Pour chacune de ces combinaisons, les trois scores explicités ci-dessus sont calculés. De plus, cette pipeline prend en compte différentes méthodes d'affichage. Le tableau ci-dessous correspond aux scores obtenus.

Embedding Method	Clustering Method	silhoutte score	davies score	calinski score
roberta_embeddings	Kmeans_clustering	0.061486	2.653828	24.272067
roberta_embeddings	gaussian_clustering	0.052942	3.040456	27.060936
roberta_embeddings	hierarchical_clustering	0.060462	2.377261	24.241228
sentence_transformer_embeddings	Kmeans_clustering	0.054105	3.310796	13.229069
sentence_transformer_embeddings	gaussian_clustering	0.054217	3.295601	13.241650
sentence_transformer_embeddings	hierarchical_clustering	0.047386	3.352618	12.389337
SVD_embeddings	Kmeans_clustering	0.377596	0.682031	832.075427
SVD_embeddings	gaussian_clustering	-0.041704	3.530118	113.659759
SVD_embeddings	hierarchical_clustering	0.375453	0.753683	831.809830
SVD_embeddings_PPMI	Kmeans_clustering	0.096428	1.861922	134.229278
SVD_embeddings_PPMI	gaussian_clustering	0.028220	2.581765	101.721296
SVD_embeddings_PPMI	hierarchical_clustering	0.084271	1.815039	127.610776
glove_embeddings	Kmeans_clustering	0.069872	2.110408	32.057324
glove_embeddings	gaussian_clustering	0.031471	2.599862	26.490819
glove_embeddings	hierarchical_clustering	0.068912	1.936615	30.354342
tfidf	Kmeans_clustering	0.028223	3.847906	6.393479
tfidf	gaussian_clustering	0.027997	3.834598	6.383273
tfidf	hierarchical_clustering	0.012780	4.291518	6.967200
SVD_embeddings_ngram	Kmeans_clustering	0.533492	0.439364	2552.508780
SVD_embeddings_ngram	gaussian_clustering	-0.194399	6.942059	75.214374
SVD_embeddings_ngram	hierarchical_clustering	0.535574	0.425816	2657.826176
SVD_embeddings_PPMI_ngram	Kmeans_clustering	0.100276	2.025401	41.529202
SVD_embeddings_PPMI_ngram	gaussian_clustering	-0.003021	3.863803	26.199328
SVD_embeddings_PPMI_ngram	hierarchical_clustering	0.111967	1.602433	47.509360

Figure 3 : tableau des scores pour les combinaisons d'embedding et de clustering

D'après ces résultats, les meilleurs scores sont obtenus pour les combinaisons entre embeddings glove et SVD avec les clusterings Kmeans et hierarchical clustering. Cela peut s'expliquer par le fait que SVD soit favorisé lorsque l'on traite de petit corpus et que Glove est un modèle complexe pré-entraîné.

Après avoir appliqué les améliorations mentionnées lors de la soutenance, les résultats n-grams sont les meilleurs avec SVD, cela s'explique par le fait que SVD améliore la sémantique en consolidant les cooccurrences fréquentes entre n-grams voisins dans un espace dense, ce qui renforce la qualité des représentations.

III.3. Affichages

En sortie des méthodes de clusterings, nous obtenons des vecteurs dans de grandes dimensions. Afin de pouvoir visualiser en deux dimensions nos clusters, une réduction de dimension est nécessaire. Pour cela nous avons utilisé deux méthodes principales : la PCA et la t-SNE. Ces techniques permettent de réduire les dimensions élevées des embeddings tout en conservant des informations pertinentes pour la structuration des données.

a) PCA - Principal Component Analysis

La PCA repose sur une approche linéaire pour réduire les dimensions. Elle identifie les directions principales, ou composantes principales, dans lesquelles la variance des données est maximale. Chaque composante principale est une combinaison linéaire des dimensions initiales. Cette méthode est particulièrement efficace pour simplifier des données qui présentent une structure fortement linéaire. Dans ce projet, la PCA a été utilisée pour explorer la variabilité globale dans les données et réduire les dimensions des embeddings de haute dimension (par exemple, ceux générés par RoBERTa). Grâce à son efficacité, la PCA s'est montrée utile pour des tâches de prétraitement avant le clustering ou comme base pour une visualisation rapide des clusters. Cependant, elle peut perdre en précision pour des structures plus complexes où les relations entre données ne sont pas linéaires.

b) t-SNE

La t-SNE, en revanche, se distingue par sa capacité à capturer des relations complexes et non linéaires entre les points de données. Contrairement à la PCA, t-SNE optimise une fonction de coût qui préserve les distances locales dans l'espace réduit, favorisant ainsi la séparation visuelle des clusters. Elle est donc particulièrement adaptée pour visualiser des clusters dans des ensembles de données à haute dimension. Cependant, cette méthode est sensible aux paramètres comme la perplexité et le taux d'apprentissage, ce qui peut influencer les résultats obtenus.

III.4. Optimisation des clusters

Après avoir choisi les meilleures combinaisons pour les méthodes d'embedding et de clusterings, il serait intéressant de chercher à déterminer le nombre optimal de clusters. Le notebook detailed_analysis.ipynb détaille la méthode et les résultats obtenus en traçant des graphes des scores en fonction du nombre de clusters mis en paramètre. Les combinaisons testées sont celles déterminées dans la partie précédente pour un nombre de clusters situé entre 5 et 15. D'après les graphes obtenus, les meilleurs combinaisons embeddings, clustering, taille de clusters sont:

- SVD Kmeans 11 clusters (silhouette : 0.39, Davies : 0.68, Calsinski : 853)
- SVD Hierarchical 12 clusters (silhouette : 0.33, Davies : 0.79, Calsinski : 782)

Voici un exemple des graphes tracé pour SVD Kmeans 11 clusters :

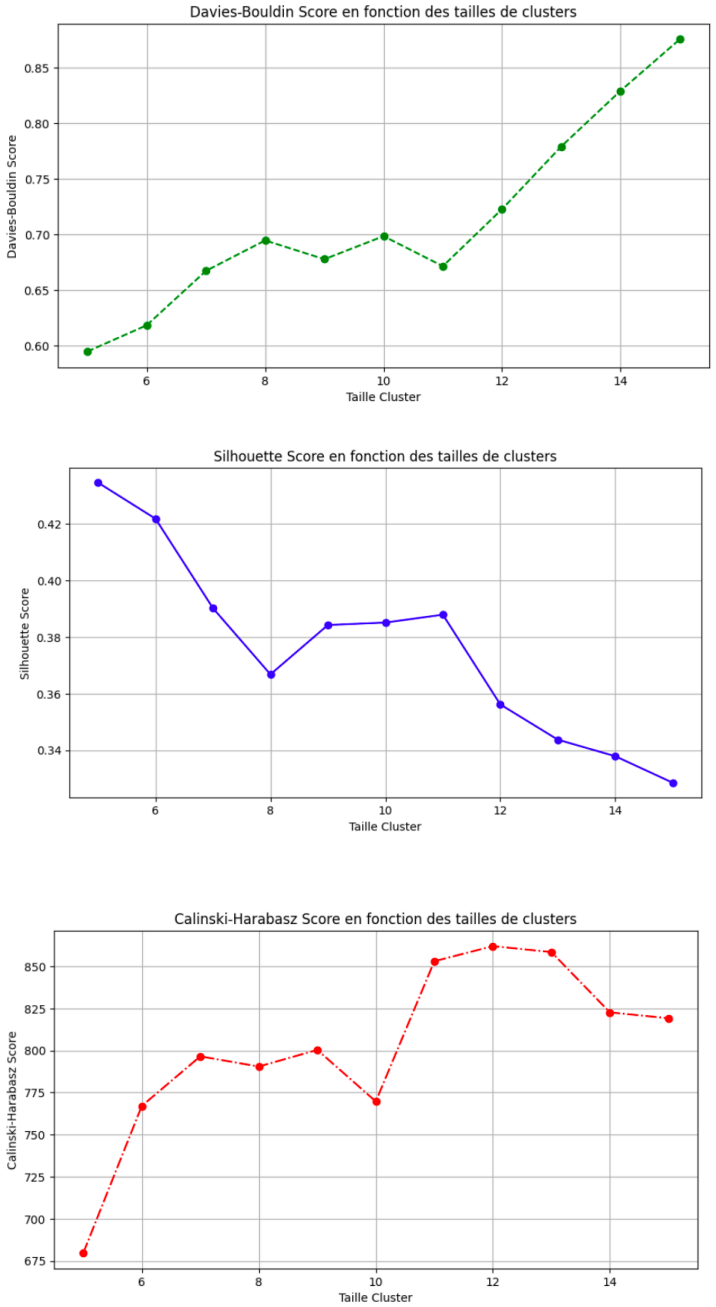

Figure 4 : Optimisation nombre de clusters pour la combinaison SVD Kmeans

IV. Topic Modeling

IV.1. Utilisation des metadatas

Dans un second temps, il est intéressant d'essayer d'interpréter les clusters en trouvant un sujet commun aux documents d'un même cluster.

Pour ce faire, il a d'abord été décidé d'utiliser les métadonnées disponibles, en particulier le nom des organismes émetteurs. La première étape a été de regrouper ces organismes en différents types d'institutions. Étant donné le nombre important d'organismes, un LLM a été utilisé pour créer 10 catégories. Ensuite, en utilisant les méthodes d'embedding vues précédemment et en réduisant la dimension, il a été possible de tracer sur un plan les documents avec un label différent pour chaque catégorie d'organismes (Figure 5).

Le résultat espéré était idéalement de retrouver les clusters obtenus grâce aux méthodes évoquées plus tôt ou bien de trouver des clusters différents mais intéressants à analyser. Malheureusement aucun cluster ne s'est dégagé grâce à cette méthode. Ce phénomène peut tout de même être analysé de la manière suivante: si toutes les catégories d'institutions sont mélangées, cela peut montrer que les différentes institutions n'abordent pas forcément des sujets spécifiques mais qu'elles abordent plutôt toutes des sujets variés sur l'IA. Cette analyse est cohérente avec la littérature.

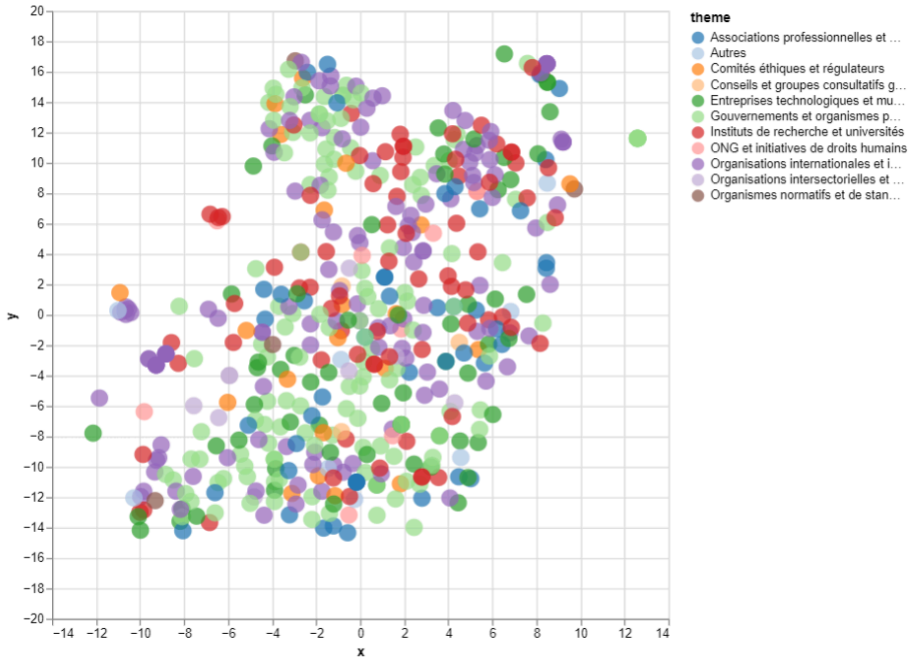

Figure 5 : Visualisation des documents labellisés par type d'organisme (embedding RoBERTa, réduction de dimension t-SNE)

La deuxième méthode mise en place se base davantage sur la littérature [1]. En effet, 11 thèmes récurrents ont été définis et des mots clés leur ont été associés. Pour regrouper les textes par thèmes, différentes méthodes ont été mises au point.

La première était d'attribuer un thème à un texte si un des mots clés apparaît au moins une fois dans le texte. Cependant avec cette méthode chaque texte pouvait avoir plusieurs thèmes, ce qui rendait difficile la visualisation.

Ainsi pour ne garder que le thème principal de chaque texte, il a d'abord été décidé de garder celui dont les mots clés apparaissent le plus de fois dans le texte. Mais cette méthode ne donnait pas de résultats satisfaisants car certains mots clés sont beaucoup plus fréquents que d'autres dans le corpus. A cause de cela, un seul thème était attribué à plus de la moitié des textes.

Pour remédier à ce problème, une sorte de normalisation a été appliquée au nombre d'occurrences des mots clés en gardant le thème dont les mots clés avaient la plus grande TF-IDF. Cette méthode a permis de mieux répartir les thèmes sur les différents textes, et des méthodes d'embedding et de visualisation ont été utilisées comme précédemment, en utilisant cette fois le thème principal comme label (Figure 6).

Encore une fois, ces labels n'ont pas permis d'obtenir des clusters intéressants.

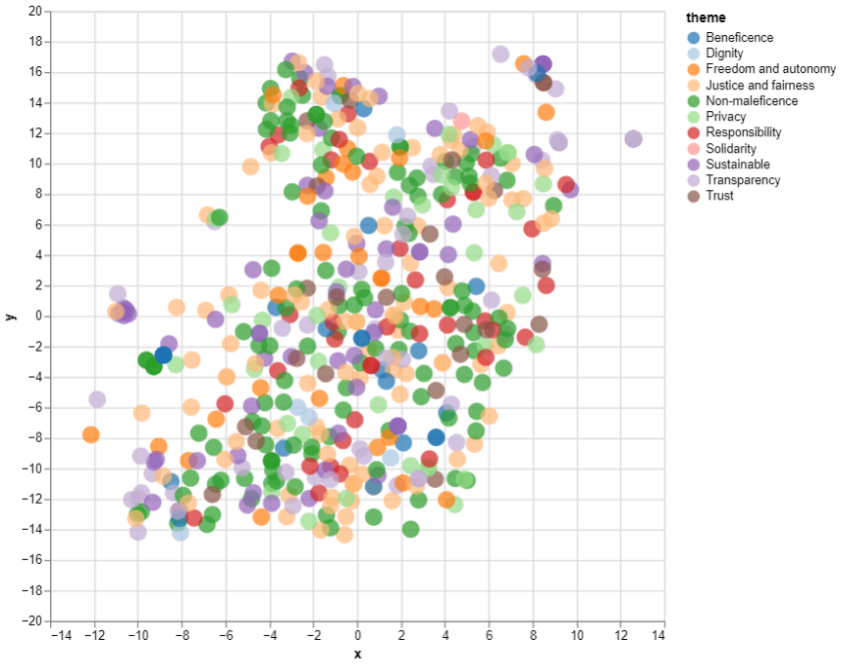

Figure 6 : Visualisation des documents labellisés par thème principal (embedding RoBERTa, réduction de dimension t-SNE)

IV.2. Latent Dirichlet Allocation

Pour essayer de trouver des thèmes plus pertinents, une autre méthode qui n'utilise pas les métadonnées a été choisie: la LDA (cf LDA.ipynb). Elle permet de créer des thèmes contenant des mots clés et de regrouper les documents par thèmes.

Tout d'abord, pour choisir le nombre de thèmes à créer, plusieurs valeurs ont été testées et la perplexité a été calculée à chaque fois (Figure 7) pour mesurer à quel point le modèle prédit les données avec incertitude. Plus la perplexité est basse, plus les thèmes prédits sont censés être pertinents. Cependant la valeur de la métrique ne fait que diminuer lorsque le nombre de topics augmente mais si le nombre de topics est trop important, ils risquent d'être moins intéressants. Ainsi la valeur 12 a été choisie pour le nombre de thèmes

car la perplexité y atteint un minimum local et cela reste cohérent avec les 11 thèmes mentionnés dans la littérature.

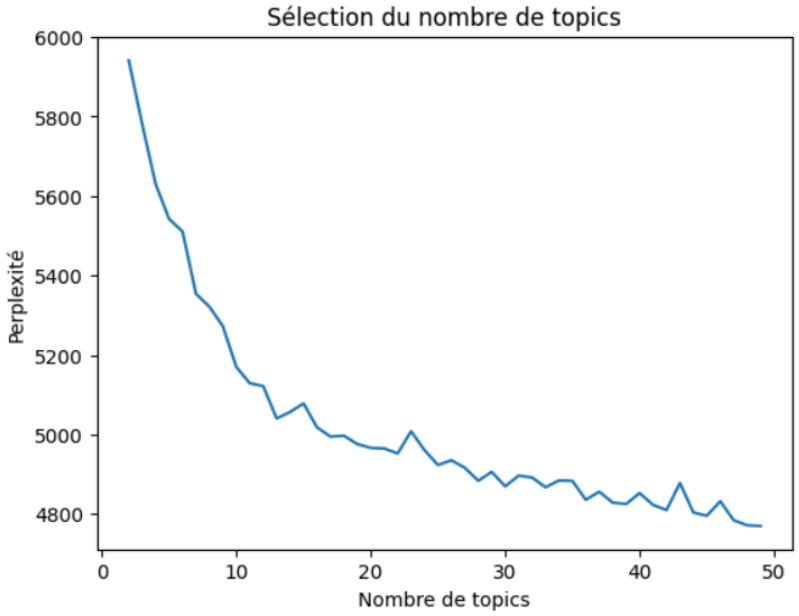

Figure 7 : Perplexité du modèle LDA en fonction du nombre de topics

Les mots les plus importants de chaque thème ont été affichés (Figure 8). A partir de ces mots, il ressort que les thèmes obtenus semblent assez cohérents même si certains mots qui sont assez génériques dans le corpus (comme “artificial” et “intelligence”) apparaissent dans plusieurs thèmes ainsi que certains mots (comme “also”) qui n’apportent pas d’information et pourraient être considérés comme des stop words.

```
Topic  1  :  ['health', 'care', 'use', 'technologies', 'technology', 'healthcare', 'may', 'patient', 'medical',
Topic  2  :  ['intelligence', 'index', 'report', 'artificial', 'number', 'figure', 'chapter', 'papers', 'learni
Topic  3  :  ['eu', 'public', 'systems', 'european', 'artificial', 'use', 'intelligence', 'digital', 'commissio
Topic  4  :  ['tion', 'ts', 'al', 'wha', 'ed', 'es', 'ethics', 'ation', 'pr', 'also', 'resear', 'tions', 'ch',
Topic  5  :  ['learning', 'machine', 'systems', 'ethics', 'research', 'human', 'ethical', 'technology', 'intell
Topic  6  :  ['eu', 'cdn', 'effect', 'de', 'segoeui', 'regulation', 'brussels', 'net', 'facto', 'office', 'act',
Topic  7  :  ['digital', 'research', 'government', 'development', 'new', 'national', 'strategy', 'public', 'inn
Topic  8  :  ['decision', 'use', 'systems', 'system', 'may', 'algorithmic', 'algorithms', 'model', 'making', 'pu
Topic  9  :  ['oecd', 'et', 'skills', 'information', 'de', 'training', 'al', 'les', 'news', 'using', 'related',
Topic  10  :  ['system', 'systems', 'model', 'intelligence', 'based', 'artificial', 'human', 'level', 'use', 'al
Topic  11  :  ['intelligence', 'artificial', 'media', 'european', 'human', 'technology', 'also', 'use', 'online
Topic  12  :  ['rights', 'human', 'law', 'protection', 'use', 'systems', 'system', 'may', 'also', 'information',
```

Figure 8 : Listes des mots les plus importants des thèmes obtenus avec la LDA

En se basant sur les 15 premiers mots de chaque thème, un titre leur a été donné à la main pour pouvoir les visualiser plus simplement. Enfin, en utilisant la LDA comme embedding et en réduisant la dimension, les données ont pu être visualisées (Figure 9). En les labellisant avec les thèmes obtenus, des clusters assez nets ont pu être obtenus.

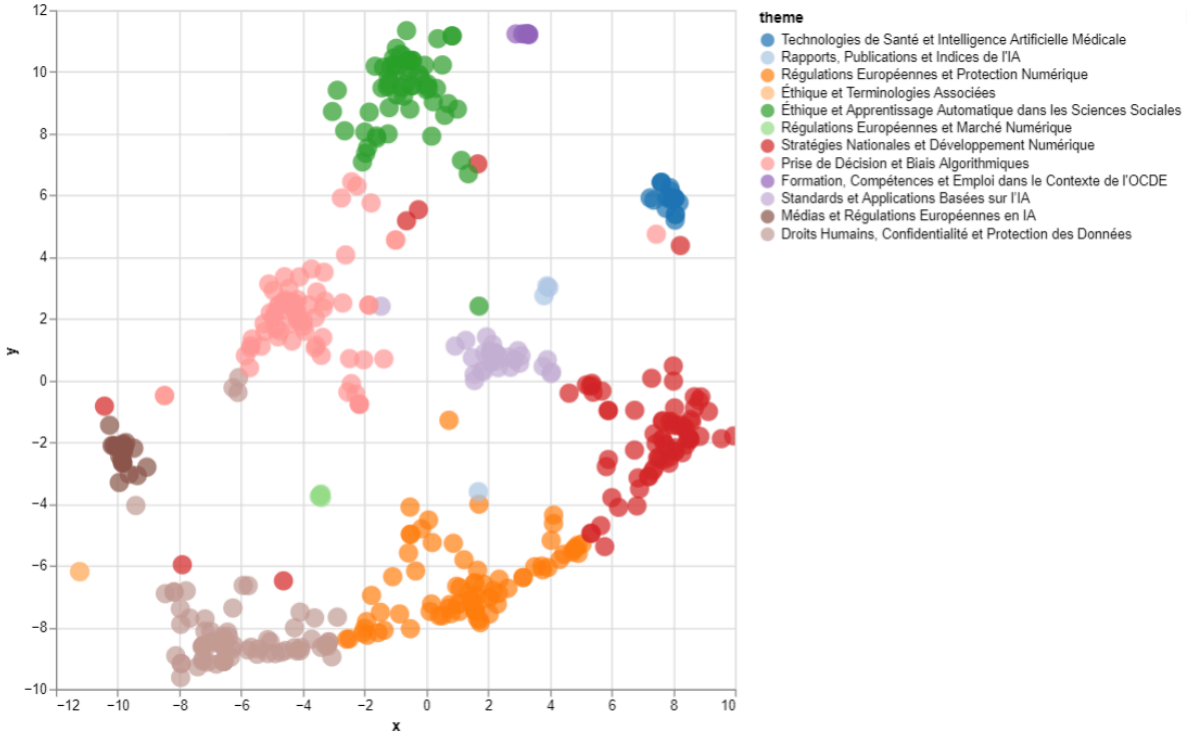

Figure 9 : Visualisation des documents labellisés avec les thèmes trouvés grâce à la LDA (réduction de dimension t-SNE)

Néanmoins, même avec des nombres de thèmes différents, la LDA permet de trouver des clusters assez nets. De plus, les titres donnés à chaque thème sont assez subjectifs. Enfin, pour cette méthode, les calculs de scores de clustering grâce aux métriques silhouette, Davies-Bouldin et Calinski-Harabasz donnent des valeurs correctes (respectivement 0.43, 0.79 et 125.85) sans être les meilleures obtenues. C'est pourquoi il est difficile d'affirmer que ces thèmes sont ceux qui représentent le mieux le corpus.

IV.3. Génération des thèmes par clusters avec un LLM

Une seconde approche de déterminer les thèmes des clusters en partant directement de la répartition obtenue dans la partie III à l'aide d'un modèle génératif LLM. Pour ce faire, les clusters et les titres des articles sont fournis au LLM accompagnés d'une prompt (instruction) qui donne des précisions quant à la tâche à accomplir. La solution cloud Azure OpenAI a été utilisée au travers le modèle GPT-4o-mini. L'implantation de cette méthode se trouve dans le notebook *GPTopics.ipynb*. Les clusters choisis pour être traité par cette méthode sont la meilleure combinaison déterminée dans III.4., soit embedding SVD, clustering Kmean et 11 clusters.

Une première idée est de fournir les titres de chaque clusters individuellement mais les titres des clusters retournés sont très identiques. En effet, les thématiques de l'éthique, la régularisation et la gouvernance de l'IA sont très présentes dans le corpus. Pour remédier à ce problème, l'ensemble des clusters et titres sont fournis tous ensemble au LLM et une tâche de différenciation est précisée dans le prompt fourni au LLM. L'objectif de cela est que le LLM distingue les spécificités de chaque cluster. Ci-dessous sont présentés les résultats obtenus avec deux prompts.

Premier prompt utilisé :
"You are an analyst tasked with finding a precise and differentiating theme from a provided list of groups of article names. Please, for each group, I need you to find the theme of the articles inside the group. It is very important that the themes return show the specificity and uniqueness of each group and differentiate it from the other groups. Try your maximum to not repeat words between themes as they all should be unique and different."


```
[
    {"labels": 8, "theme": "Comprehensive policy frameworks and ethical considerations for AI development and govern
ance"},
    {"labels": 6, "theme": "Addressing human rights implications and ethical standards in AI and algorithmic decisio
n-making"},
    {"labels": 11, "theme": "Regulatory strategies and international cooperation for AI governance in Europe"},
    {"labels": 1, "theme": "Evaluation and implementation of ethical frameworks in AI technologies and practices"},
    {"labels": 2, "theme": "Promoting responsible AI practices and citizen-centric applications in digital environme
nts"},
    {"labels": 5, "theme": "Strategic initiatives for AI development, governance, and ethical implications across se
ctors"},
    {"labels": 10, "theme": "Building national AI strategies with a focus on trust, ethics, and regional cooperatio
n"},
    {"labels": 7, "theme": "Safeguarding ethical standards and responsibilities in advanced technology application
s"},
    {"labels": 4, "theme": "Exploring inclusive AI practices and ethical frameworks across various jurisdictions"},
    {"labels": 9, "theme": "Ensuring ethical data governance and protection principles in the context of AI"},
    {"labels": 0, "theme": "Integrating ethics into big data practices and fostering responsible data usage"},
    {"labels": 3, "theme": "Establishing guidelines for ethical considerations in AI design and deployment"}
]
```


Figure 10 : Génération de thèmes associés au clusters avec prompt 1

Le premier prompt utilisé, demandant au LLM de générer des thèmes uniques pour chaque clusters, retourne des sujets très similaires avec de légères variations indiquant une certaine homogénéité dans les thématiques principales traitées dans le corpus. En effet, les sujets de la gouvernance de l'IA, de l'éthique sont présents dans la majorité des clusters, avec quelques variations comme un focus sur l'Europe pour le cluster 11, les applications centrés sur les citoyens pour le cluster 2 ou bien les algorithmes de décision label 6.

Deuxième prompt utilisé :

“You are an analyst tasked with finding a precise and differentiating theme from providing a list of groups of article names.Please, for each group, I need you to find what makes it different from group. It is very important that you show the specificity and uniqueness of each group and differentiate it from the other groups. Try your maximum to not repeat words between group unique traits as they all should be unique and different.”


```
[
    {"labels": 8, "theme": "Focus on public policy, governance, ethical frameworks, and human rights implications of
AI across various sectors."},
    {"labels": 6, "theme": "Emphasis on bias, transparency, accountability, and ethical frameworks specifically rela
ted to algorithmic decision-making and public health."},
    {"labels": 11, "theme": "Concentration on regulatory proposals, international cooperation, and fostering trust i
n AI technologies with a European perspective."},
    {"labels": 1, "theme": "Engagement with societal implications, accountability, and responsibility of technologie
s, particularly in legislative and justice contexts."},
    {"labels": 2, "theme": "Prioritization of citizen engagement, national strategies for AI deployment, and establi
shing frameworks for trustworthy AI practices."},
    {"labels": 5, "theme": "Developing comprehensive governance strategies, ethical frameworks, and international co
llaboration to address AI's societal impacts."},
    {"labels": 10, "theme": "Creation of ethical guidelines and principles focused specifically on trustworthy AI in
itiatives in regional contexts."},
    {"labels": 7, "theme": "Exploration of ethical implications, professional conduct, and safety standards related
to the impacts of AI across industries."},
    {"labels": 4, "theme": "Global and inclusive approaches to AI ethics, promoting rights and best practices to har
ness AI's benefits responsibly."},
    {"labels": 9, "theme": "In-depth exploration of data governance, quality control, and ethical principles in the
intersection of big data and AI."},
    {"labels": 0, "theme": "Concern for data ethics, privacy alignment, and legislative implications of AI and big d
ata on fundamental rights."},
    {"labels": 3, "theme": "Focused on establishing specific ethical guidelines by a single organization, underlinin
g organizational responsibility in AI ethics."}
]
```


Figure 11 : Génération de thèmes associé au clusters avec prompt 2

Dans le second prompt un accent est porté sur les différences entre les clusters pour orienter le LLM vers la génération de thèmes uniques. En sortie du LLM, on retrouve les thématiques spécifiques trouvées dans le premier prompt avec en plus, des thèmes plus distinct permettant de mieux différencier les sujets et de dégager des sous thèmes plus développés. Les sujets simplifiés suivants peuvent être extrait de cette analyse :

- “policy and governance taking account ethics and human rights”
- “bias and transparency for algorithms in the health sector”
- “regulation and trust of AI in the context of international cooperation in Europe”
- “societal implication, accountability and responsibility of technologies in legislative and justice”
- “citizen engagement and national strategies for AI development”
- “comprehensive guidelines and principles for trustworthy AI”
- “ethical implication, professional conduct, and safety standards of AI across industries”
- “Global and inclusive approach to AI ethics”
- “data governance, quality control and ethics in big data and AI”
- “data ethics, privacy alignment and legislative implications of AI”
- “ethical guidelines and organizational responsibility”

Concernant cette méthode de topic modeling, une piste d'amélioration possible de cette méthode, la caractérisation des thèmes pourrait être améliorée en utilisant aussi l'ensemble du contenu de l'article mais cela poserait un problème de consommation de tokens.

IV.4. BERTopics

La dernière méthode mise en place consiste à utiliser BERTopic, une technique de modélisation permettant de créer des clusters thématiques sur un corpus de documents. Cette méthode s'appuie sur BERT pour transformer les documents en vecteurs, puis regroupe les vecteurs similaires en clusters grâce à une réduction de dimension réalisée avec UMAP et un algorithme de clustering HDBSCAN. Enfin, une analyse TF-IDF sur les documents permet à BERTopic de générer des thèmes en identifiant les mots les plus représentés dans chaque cluster.

La première étape a consisté à préparer les documents pour la clusterisation. Pour cela, après avoir récupéré les 457 documents prétraités, les stop words tels que les balises HTTP ou CSS ont été retirés par mesure de sécurité.

Afin d'obtenir des mots-clés pertinents pour les clusters en sortie de l'algorithme, un traitement supplémentaire a été effectué pour éliminer divers stop words jugés non pertinents.

Les 20 mots les plus fréquents dans le corpus, ainsi que les dates (années), les mois et les jours de la semaine, ont également été supprimés des documents.

Il est possible de choisir l'utilisation ou non des n-gram dans BERTopic. Les deux cas sont étudiés ci-dessous :

a) BERTopic sans n-gram

BERTopic est exécuté avec un paramétrage d'HDBSCAN pour garantir un minimum de 20 documents par cluster. Cela permet de limiter les clusters peu représentatifs et de réduire le nombre de documents regroupés dans le cluster "-1", qui contient tous les documents non classifiables. Les graphiques suivants sont ainsi obtenus :

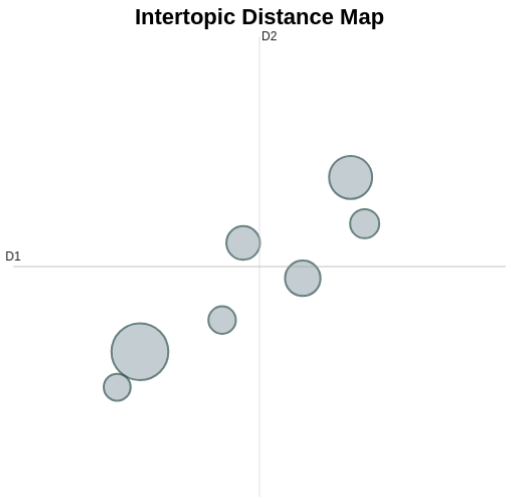

Figure 12 : Clusters obtenus par BERTopic sans n-gram

Les 7 topics obtenus sont alors représenté par les mots suivants :

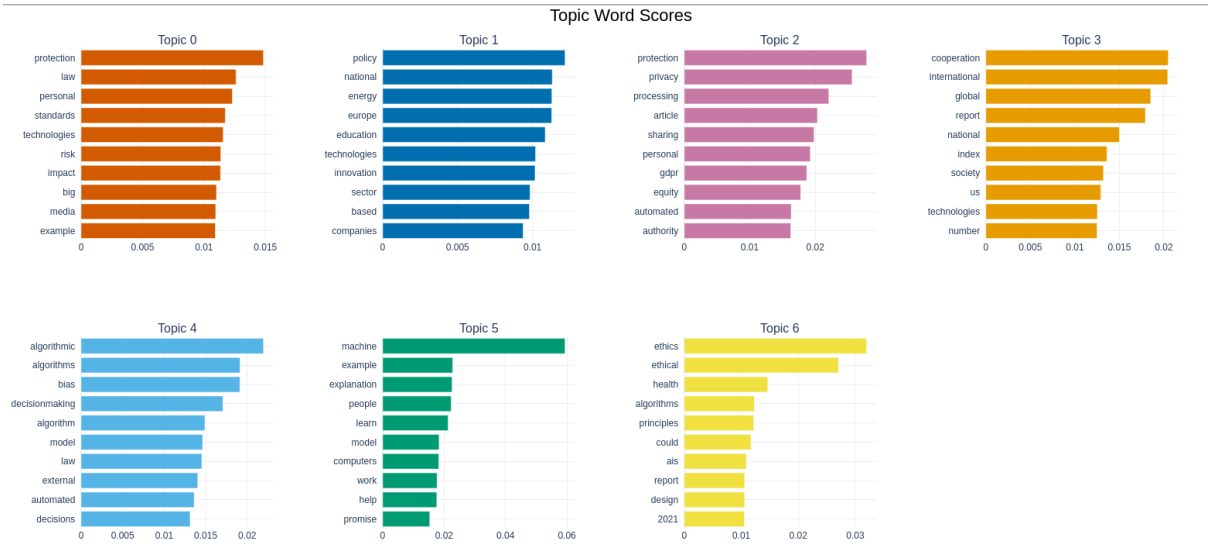

Figure 13 : Thèmes des clusters

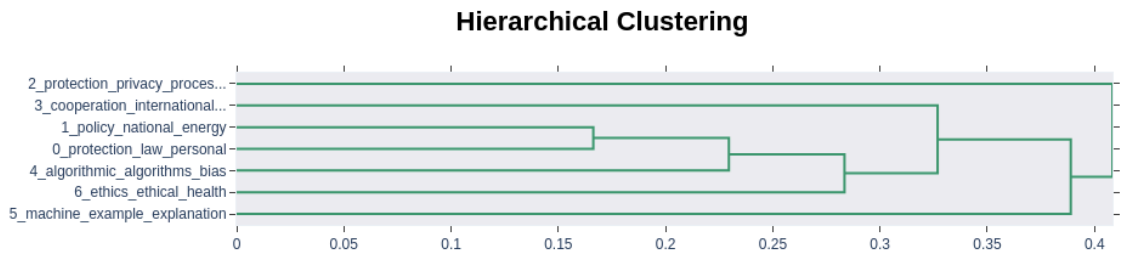

Figure 14 : Lien entre les clusters

Sept clusters distincts ont été identifiés, chacun caractérisé par des mots-clés spécifiques. Leurs relations sont mises en évidence par une vue hiérarchique, notamment le lien apparent entre les clusters 0 et 1, ce qui se reflète également dans les mots-clés communs tels que "technologies" ou les mots proches comme "law" et "policy".

Les thèmes des clusters peuvent être analysés comme suit :

- **Cluster 0** : Centré sur la protection des droits des individus dans le contexte de l'IA, notamment l'autonomie et l'impact des technologies sur les vies humaines.
- **Cluster 1** : Axé sur la politique publique et la réglementation de l'IA.
- **Cluster 2** : Focalisé sur la confidentialité et le traitement des données.
- **Cluster 3** : Concerné par la coopération internationale et l'impact des technologies de l'IA.
- **Cluster 4** : Porté sur les biais algorithmiques, les discriminations qu'ils peuvent induire et leurs impacts négatifs.
- **Cluster 5** : Lié à l'exploitation et à l'innovation technologique.
- **Cluster 6** : Concerne l'éthique dans la conception et l'utilisation des algorithmes.

b) Utilisation de n-gram

En employant des n-gram, des relations plus pointues entre les documents peuvent être analysées, permettant de dégager des thèmes formés par des unités de sens.

Les clusters suivants sont alors obtenus :

Figure 15 : Clusters n-gram (2,3)

On peut alors regrouper les clusters obtenus précédemment et ceux la dans des thèmes globaux.

Le cluster 0 axé sur la “protection” et le “personnel” dans la première méthode s’aligne avec le cluster 0 et le cluster 4 de la seconde, qui examinent la sauvegarde des droits individuels face aux technologies de l’IA, avec une attention particulière à la gouvernance et aux répercussions sociétales.

Le cluster 2, centré sur la “protection” et la “privacy”, trouve un écho dans le cluster 5 de la seconde méthode, qui met en lumière les efforts mondiaux pour garantir une gestion éthique des données, en s’appuyant sur des cadres collaboratifs comme Creative Commons.

Le cluster 3, tourné vers la “coopération” et les problématiques globales, correspond au cluster 2 de la seconde analyse, qui explore les dynamiques de coopération internationale et les discussions sur l’impact global des technologies de l’IA, notamment sur des enjeux comme le climat.

Le cluster 4, focalisé sur “algorithmes” et “bias”, se reflète dans le cluster 4 de la seconde méthode, où sont analysés les biais dans les décisions algorithmiques, ainsi que les outils législatifs comme l’Equality Act pour prévenir les discriminations.

Le cluster 5, portant sur les “machines” et leur influence, converge avec le cluster 6 et le cluster 1 de la seconde méthode, qui traitent de l’apprentissage automatique, de son pouvoir transformateur, et des évolutions liées à la révolution industrielle.

Le cluster 6 sur l’“ethics” et le “design” obtenu sans n-gram peut être relié au cluster 0 de la seconde méthode, qui mettent en avant l’éthique dans la conception et l’utilisation des technologies de l’IA.

Le cluster 1 est un nouveau thème parlant de l’IA dans l’industrie ainsi que le secteur de la santé.

En conclusion de cette partie, certains sous thèmes ressortent de toutes les méthodes de topic modeling explorées. Ceux peuvent alors considérer comme des thématiques valides du corpus :

- Droits humains (gpt-4o-mini, LDA)
- Régulations Européennes (gpt-4o-mini, LDA, BERTopics)
- Secteur de la santé (gpt-4o-mini, LDA, BERTopics)
- Prise de décision et biais algorithmiques (gpt-4o-mini, LDA, BERTopics)
- Confidentialité (gpt-4o-mini, LDA, BERTopics)
- Industries (gpt-4o-mini, BERTopics)
- Justice (gpt-4o-mini, LDA, BERTopics)

V. Améliorations possibles

Avec la pipeline de clustering et le topic modeling utilisée, des thèmes et des clusters divers ont été identifiés. Cependant, il reste très important d'avoir des critères et des méthodes de validation pour déterminer la qualité des informations obtenues. Pour les clusters, des métriques sont utilisées pour déterminer la qualité des clusters. Cependant, ceux-ci ne traitent les clusters qu'en tant que nuages des points (embeddings) mesurant leur bonne séparation et leur densité. C'est pourquoi pour vérifier la qualité de l'embedding et du clustering, une validation supplémentaire de la qualité sémantique des clusters reste à traiter. En d'autres termes, il faudrait vérifier si les clusters sont bien formés d'articles traitant du même sujet tout en étant distinct des articles d'autres clusters. De même, une étape de validation des topics retournés par les méthodes de topic modeling reste à concevoir pour vérifier la fiabilité des résultats.

De plus, il serait intéressant d'étudier l'impact de la réduction de dimension si elle est effectué avant ou après l'étape de clustering. En première analyse, la réduction effectuée avant entraînerait la perte de l'information contenu dans les embeddings au profit d'un clusterings plus visuel car effectué sur des données en deux dimensions avec un affichage aussi en deux dimensions. Un travail d'expérimentation pourra être effectué confrontant l'impact de cette perte sur le clustering et le gain visuel. Un autre argument justifiant l'intérêt de la réduction à priori est que celle-ci pourrait même améliorer la qualité sémantique des clusters en éliminant le bruit contenu dans les embeddings. Cet argument reste à vérifier dans notre cas.

D'autre part, pour améliorer la visualisation des clusters, il est possible de réduire et d'afficher les vecteurs d'embedding à trois dimensions pour mieux observer les clusters.

Conclusion

Ce projet a permis de développer une pipeline adaptable pour l'analyse de corpus textuels autour de l'éthique de l'intelligence artificielle. En partant d'un processus de prétraitement des données, les données ont été collectées et nettoyées un corpus conséquent. Diverses techniques d'embedding ont ensuite été mises en œuvre, allant des approches classiques (TF-IDF, GloVe, SVD) aux modèles avancés comme RoBERTa. Les étapes de clustering ont permis de regrouper les documents selon des thématiques principales, et les méthodes de topic modeling, comme LDA et BERTopic ou GPT, ont affiné l'identification des thèmes spécifiques.

Pour mesurer la qualité de la pipeline, une analyse approfondie accompagnée de scores et de visualisations en réduction de dimensions a été effectuée. Les résultats ont mis en lumière le fait qu'une majorité des articles traitent de thématiques récurrentes comme la gouvernance et l'éthique de l'IA. De plus, des sous thèmes différenciant les clusters d'articles peuvent être identifiés comme les biais algorithmiques, la protection des données, ou bien la régularisation européenne.

Pour aller plus loin, des pistes d'amélioration incluent l'intégration d'une validation sémantique des clusters et des thèmes, ainsi qu'une exploration plus poussée de la réduction de dimension avant clustering. Enfin, une ouverture intéressante consisterait à appliquer cette méthodologie à des corpus liés à d'autres domaines de l'IA ou à des questions interdisciplinaires, afin de mieux comprendre l'évolution des préoccupations éthiques à travers différents contextes.

Répartition des tâches

Baptiste : étude de l'article "Mapping AI ethics", topic modeling avec metadata, LDA, pipeline_startpoint
Alexandre M : SVD, SVD_PPMI, ngram sur SVD&SVD_PPMI, Glove, aide preprocessing
Damien : aide preprocessing, mis en place de la pipeline générale, méthodes de clustering et de scoring, optimisation des clusters, Thèmes par cluster avec LLM
Alexandre R : preprocessing, TF_IDF, RoBERTA, BERTopics, réduction de dimension, PCA, TSNE, CA

Références

[1] : GORNET Mélanie, DELARUE Simon, BORITCHEV Maria, VIARD Tiphaine, *Mapping AI ethics: a meso-scale analysis of its charters and manifestos*, The 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024. p. 127-140.