Visualizing and analyzing AI ethics charters & manifestos with clustering

Alexandre Rocchi-Henry
Alexandre Malfoy
Damien Thai
Baptiste Cervoni

13/11/2024

Introduction

Objectif : fournir une analyse d'un corpus de textes sur l'éthique de l'IA
- réutiliser les outils utilisés dans l'article “Mapping AI ethics” [1] et créer un pipeline pour :
    - Prétraitement
    - Représentation
    - Visualisation/Clustering
- explorer différents modèles pour la partie représentation du pipeline.

Corpus : MapAIE collection de 627 chartes et manifestes autour de l'intelligence artificielle et de l'éthique de l'IA

[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.

Plan de la présentation

I/ Preprocessing

II/ Embeddings

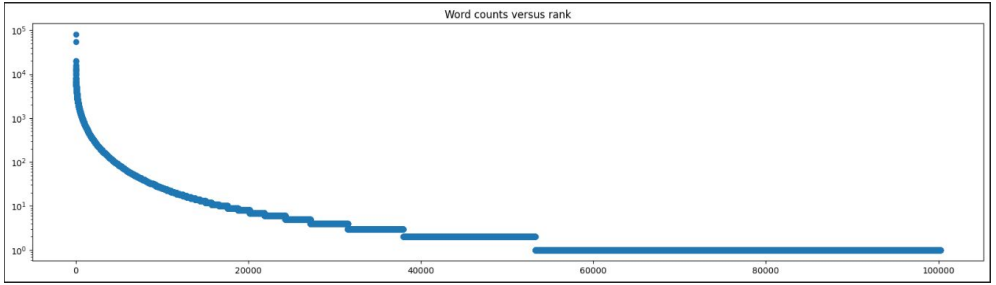
III/ Clustering et Topic modeling

IV/ Résultats et Affichage

V/ Améliorations possibles

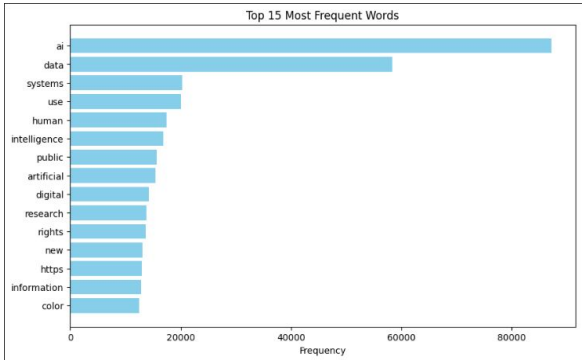
I/ Preprocessing

Récupération des données par ‘Crawling’ : pages html, articles pdf



Nettoyage du texte :

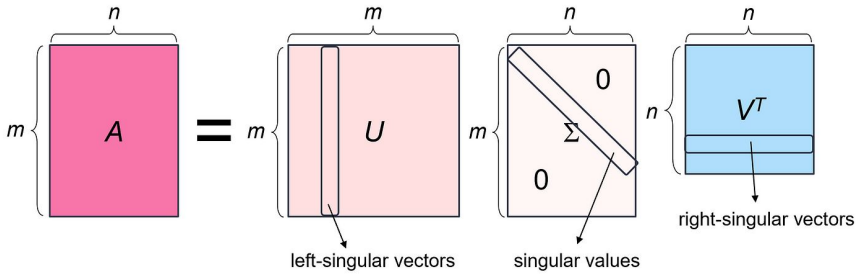
- détection de langue
- filtre des pages webs avec regex sur des marqueurs html, css
- stopwords (nltk)



II/ 1) SVD & SVD_PPMI

Principe :

- Décomposition en valeurs singulières avec une matrice de co-occurence ou une matrice PPMI pour obtenir des relations



Avantages :

- Réduction de dimension efficace
- Réduction de bruit avec la matrice PPMI

Inconvénients :

- Sensibilité au bruit
- Complexité de calcul PPMI
- Sensibilité qualité matrice PPMI

II/ 2) GloVe

Principe :

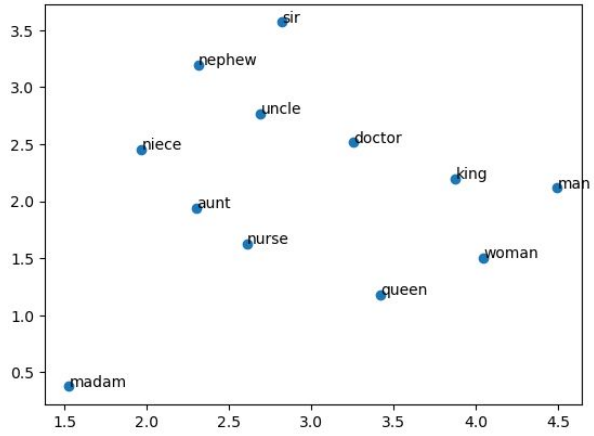
- Capture relations de similarités entre chaque mot

Avantages :

- Efficace pour les similarités
- Rapide à exécuter

Inconvénients :

- Chargement du modèle qui peut être long



II/ 3) TF-IDF

Principe :

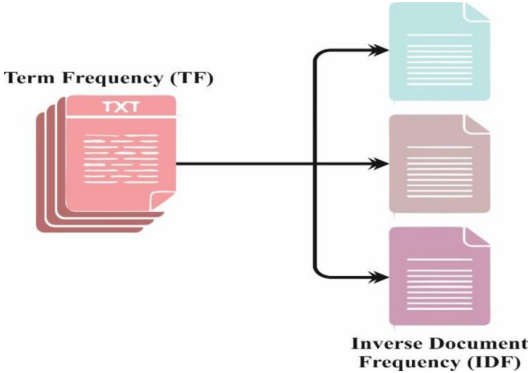
- mesure l'importance d'un mot par sa fréquence dans le texte et sa rareté dans le corpus

Avantages :

- Très simple et rapide
- Bonne interprétabilité

Inconvénients :

- Vecteurs de très hautes dimensions
- Sensibles au bruit et aux mots peu fréquent



II/ 4) Roberta

Principe :

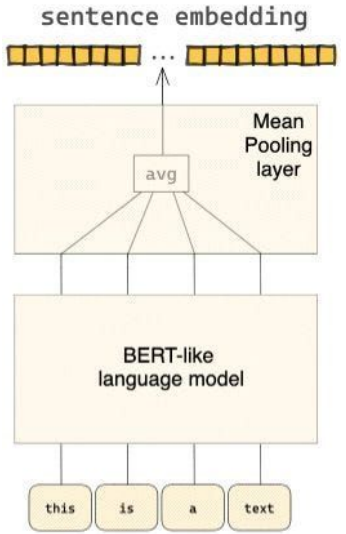
- modèle transformer qui capture le contexte des mots

Avantages :

- Prise en compte du contexte de la phrase

Inconvénients :

- Coût de calcul et d'entraînement très élevé.
- Model complexe, interprétabilité difficile



II/ 4) Comparaison

	TFIDF	Roberta	Glove	SVD	SVD_PPMI
Avantages	Simple Rapide à exécuter	Contexte de la phrase	Modèle déjà entraîné Rapide à exécuter	Efficace pour réduction de dimension	Traite le problème du bruit
Inconvénients	Sensible au bruit	Long à exécuter	Chargement du modèle	Sensible au bruit	Complexité de calcul de la matrice PPMI

III/ 1) Clustering

Utilisation de 3 méthodes principales de clustering (sklearn) :

- **KMeans :** Initialisation des centroïdes => assignation => calcul des points moyens

Avantages : simple et rapide, efficace clusters sphériques, distincts et non chevauchants

- **Hierarchical Clustering ascendant :** initialement, un cluster par document => fusion des clusters les plus proches => arbre relationnel

Avantages : clusters de tailles, formes variées et nombre inconnu, robuste au bruits

- **Gaussian Clustering :** Assimile les clusters à des distributions gaussiennes de moyenne les centroïdes et avec une matrice de covariance

Avantages : clusters elliptiques de tailles variées, irréguliers et des distributions imbriquées

III/ 2) Topic Modelling

Comment donner du sens aux clusters?

-Regrouper les textes par types d'institutions

-> pré définir une liste d'entités pour chaque type

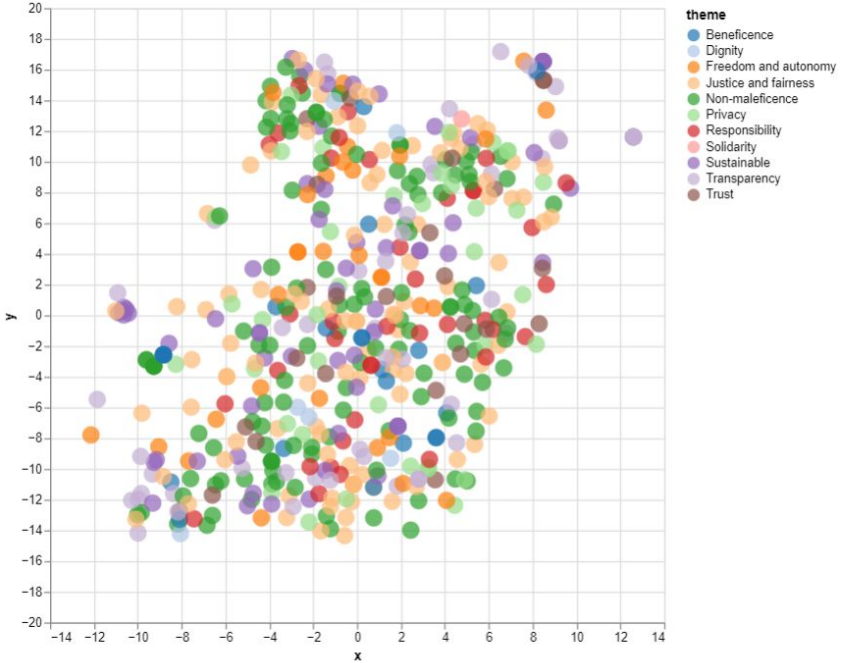
-Regrouper les textes par thème

-> un texte aborde un thème si des mots clés sont utilisés plusieurs thèmes par texte

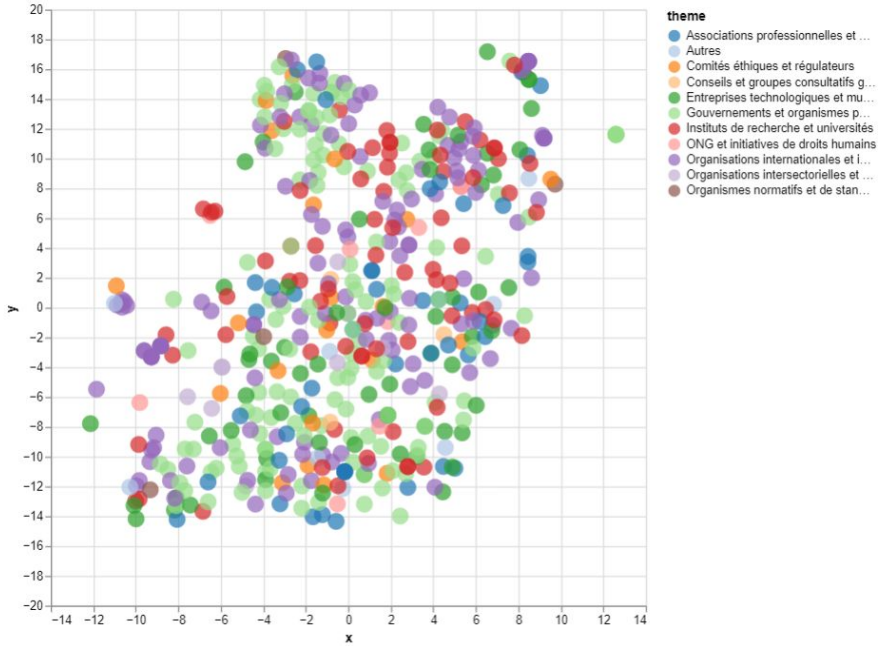
-> on peut choisir le thème principal qui a le plus de mots clés un thème est omniprésent

-> on choisit le thème principal qui a le plus grand TFIDF

III/ 2) Topic Modelling



Avec les thèmes



Avec les catégories d'institutions

IV/ Résultats et Affichage

explication des métriques :

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

a distance moyenne intra clusters
b distance moyenne cluster voisin

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

si et sj : distances moyennes des points aux centroïdes (= taille du cluster). d(i,j) est la distance entre les centroïdes des clusters

Calinski-Harabasz score : $$\mathrm{CH} = \frac{\mathrm{trace}(B_k)}{\mathrm{trace}(W_k)} \times \frac{N - k}{k - 1}$$

Bk matrice de dispersion inter clusters
Wk intra clusters => variance inter/intra

IV/ Résultats et Affichage

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

IV/ Résultats et Affichage

Tableau des scores :

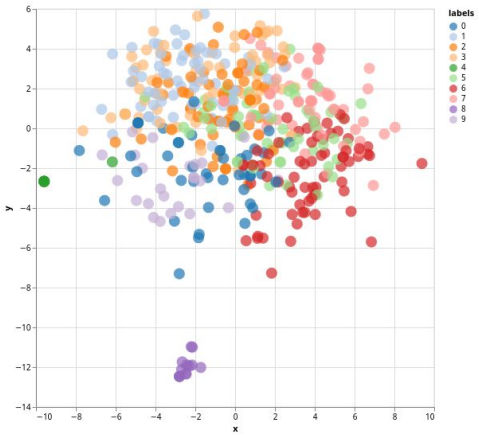
Embedding Method	Clustering Method	silhoutte score	davies score	calinski score
tfidf	Kmeans_clustering	0.019838	4.077358	7.525365
tfidf	gaussian_clustering	0.019838	4.077358	7.525365
tfidf	hierarchical_clustering	0.028784	3.803905	8.374446
glove_embeddings	Kmeans_clustering	0.106650	1.850116	241.352233
glove_embeddings	gaussian_clustering	0.079459	2.473966	224.174702
glove_embeddings	hierarchical_clustering	0.090080	1.505558	258.526031
SVD_embeddings	Kmeans_clustering	0.312837	0.728765	447.107176
SVD_embeddings	gaussian_clustering	-0.048261	3.787660	42.585890
SVD_embeddings	hierarchical_clustering	0.326010	0.709568	428.390897
SVD_embeddings_PPMI	Kmeans_clustering	0.146530	1.419538	166.732513
SVD_embeddings_PPMI	gaussian_clustering	0.009214	3.766121	108.226332
SVD_embeddings_PPMI	hierarchical_clustering	0.132991	1.290737	160.427599
roberta_embeddings	Kmeans_clustering	0.071225	2.631717	31.673343
roberta_embeddings	gaussian_clustering	0.106496	2.261490	37.670936
roberta_embeddings	hierarchical_clustering	0.090034	1.996821	32.604742
sentence_transformer_embeddings	Kmeans_clustering	0.048589	3.136269	12.926073
sentence_transformer_embeddings	gaussian_clustering	0.047638	3.167610	12.932860
sentence_transformer_embeddings	hierarchical_clustering	0.067191	3.075668	13.975881

IV/ Résultats et Affichage

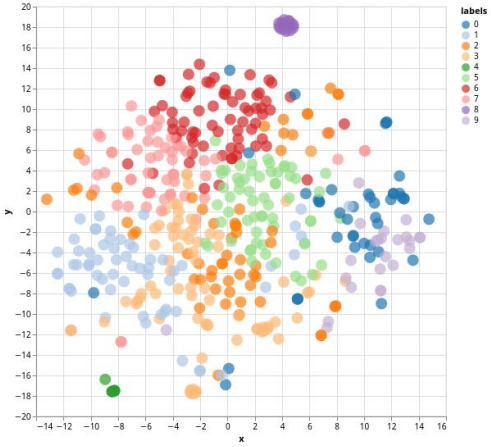
- Nécessité de Réduire la dimension en sortie des modèles pour une visualisation 2D.

- 2 techniques employées :
  - PCA
  - t-SNE

Roberta, Kmeans, PCA

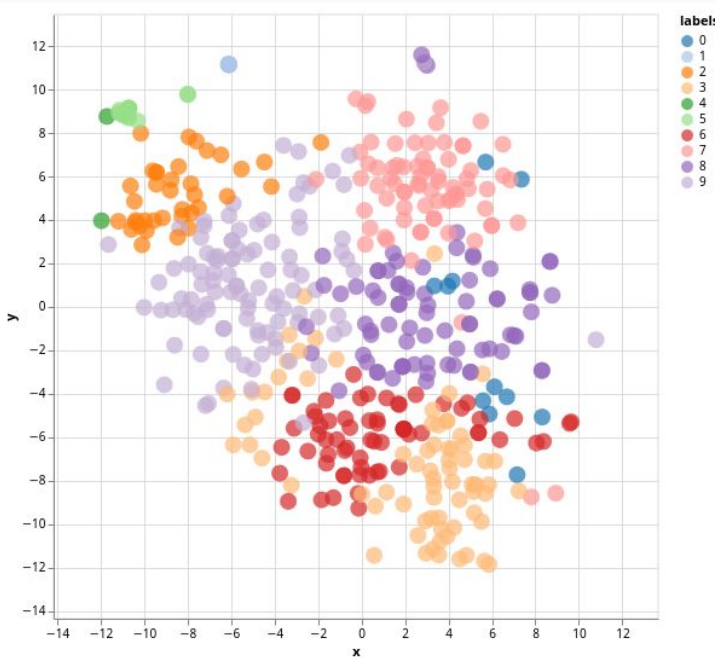


Roberta, Kmeans, t-SNE

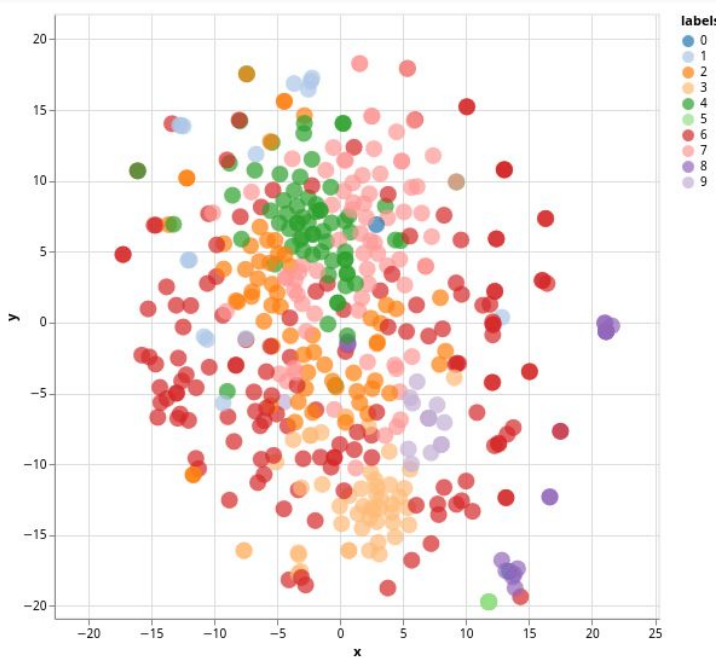


IV/ Résultats et Affichage

Glove K-Means



TF-IDF K-Means



V/ Améliorations possibles

- Tester réduction de dimension avant ou après le clustering
- Ngram (2,3)
- Topic modeling automatisé (LDA)
- Topic modeling par clusters
- Tester différentes tailles de clusters pour trouver le nombre de clusters optimal
- Cluster vs topics