Université paris 8

PARIS 8 UNIVERSITE VINCENNES - SAINT-DENIS

LIASD LABORATOIRE INFORMATIQUE AVANCÉE DE SAINT-DENIS Université Paris 8

# FICHE DE LECTURE

**pour le cours « Fouille de données » de Gilles Bernard**

**nom du projet :**

## Clustering sur des paroles de musiques

*faite par*

**Angélique Delevingne et Faycal Bounit**

le 27 février 2024

**Master mention *Informatique* Option *Big data***

MASTER

**Université Paris 8 Vincennes à Saint Denis**

**COMUE Paris Lumières**
**Laboratoire Paragraphe**
**Laboratoire d'Informatique Avancée de Saint Denis**

# Sommaire

# .1 Objectif du projet

L'objectif principal de ce projet est d'effectuer une opération de clustering sur un ensemble de données textuelles, suivi de l'évaluation des clusters obtenus à l'aide de Wordnet. Cette évaluation vise à analyser les résultats afin de tirer des conclusions significatives.

Le jeu de données comprend 227 449 chansons anglaises extraites de l'ensemble Metrolyrics. Ce dernier renferme des informations cruciales telles que le nom de la chanson, l'année, l'artiste, le genre musical, et surtout, les paroles complètes des chansons. Ces données sont plus que suffisantes pour atteindre notre objectif de clustering.

L'objectif ultime est de créer des clusters à partir des paroles des musiques afin de déterminer des groupes distincts de chansons partageant des similarités linguistiques. Cette démarche permettra d'identifier des tendances, des thèmes récurrents, et d'explorer la diversité des expressions artistiques au sein de chaque genre musical. En évaluant ces clusters avec Wordnet, nous cherchons à affiner notre compréhension des relations sémantiques entre les mots dans les chansons, ouvrant ainsi la voie à des insights plus approfondis sur la richesse et la variété du paysage musical.

# .2 Problématique

## .2.1 Description des données

Le jeu de données contient 227 449 entrées, chacune caractérisée par sept attributs, exprimés en latex de la manière suivante :

- `'Unnamed: 0', 'index'` : numéro de chansons,
- `'song'` : titre de la musique,
- `'year'` : année de sortie,
- `'artist'` : chanteur ou groupes,
- `'genre'` : styles musicaux,
- `'lyrics'` : paroles.

Le jeu de données répertorie un total de 11 117 artistes et 11 genres de musique distincts, à savoir Indie, Country, Jazz, Metal, Pop, R&B, Other, Folk, Rock, Electronic, et Hip-Hop.

## .2.2 Analyse exploratoire des données

### .2.2.1 Diversité genre

La diversité de genres suggère une richesse culturelle et artistique dans le jeu de données, offrant des opportunités d'exploration de tendances, de thèmes récurrents, et de compréhension approfondie des différentes expressions musicales présentes.

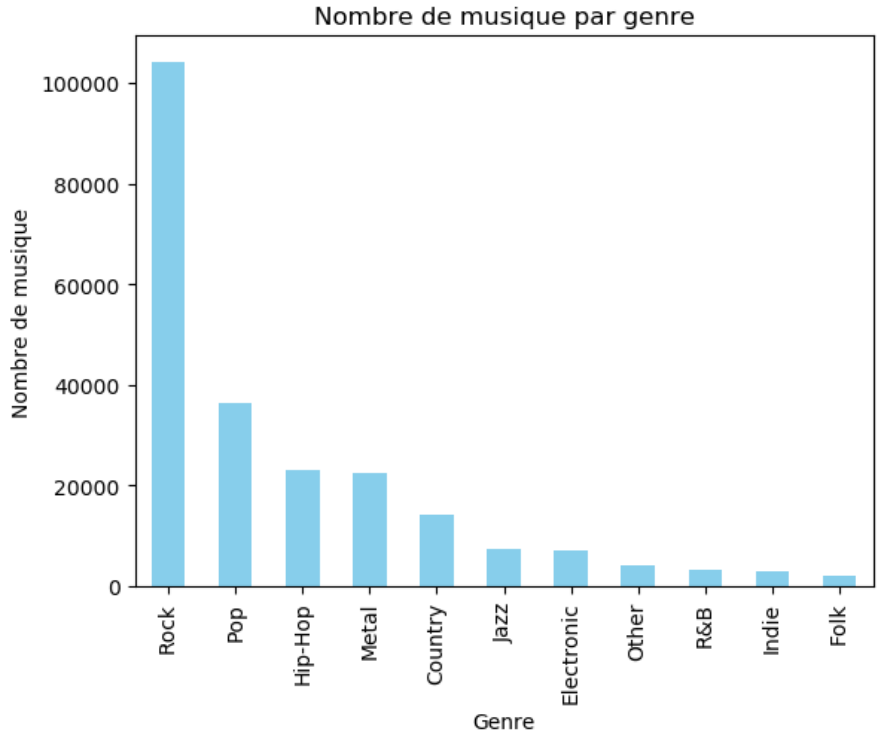


FIG. .21 : Variété de Genres Visualisée à travers un Diagramme en Barres

### .2.2.2 Longueur des musiques

L'analyse des statistiques descriptives pour la longueur des paroles de chansons révèle des informations significatives sur la distribution de cette variable dans l'ensemble de données.

Avec un effectif de 227,449 chansons, la moyenne de la longueur des paroles est de 231.70 unités, indiquant la valeur centrale autour de laquelle la majorité des observations se regroupent. L'écart-type de 157.04 suggère une variabilité notable dans les longueurs des paroles, soulignant la dispersion des données.

Le minimum de 1 unité représente la longueur minimale des paroles, tandis

que le maximum de 6211 unités correspond à la longueur maximale, soulignant la présence d'une grande diversité dans les étendues possibles.

En examinant les quartiles, on observe que 25% des paroles ont une longueur inférieure ou égale à 133 unités (premier quartile), 50% ont une longueur inférieure ou égale à 194 unités (médiane), et 75% ont une longueur inférieure ou égale à 281 unités (troisième quartile). Ces valeurs quartiles offrent une répartition plus détaillée de la longueur des paroles dans l'ensemble de données.

En somme, ces statistiques descriptives fournissent un aperçu complet de la distribution de la longueur des paroles, permettant de saisir la tendance centrale, la variabilité, ainsi que les valeurs minimales et maximales présentes dans le jeu de données.

### .2.2.3 Nuage de Mots

Dans le cadre de notre analyse des paroles de chansons, nous avons opté pour une approche visuelle en générant un nuage de mots. Cette visualisation captivante a été créée à partir des paroles des chansons de notre ensemble de données. L'objectif principal était de mettre en lumière les termes les plus fréquemment utilisés au sein de ces textes musicaux divers.

Le nuage de mots offre une représentation graphique où la taille des mots reflète leur fréquence d'apparition respective. En jetant un coup d'œil à cette visualisation, nous sommes en mesure d'identifier rapidement les mots clés dominants et les thèmes récurrents présents dans l'univers des paroles de chansons que nous explorons. Cette approche visuelle enrichit notre compréhension du langage utilisé dans la musique, nous permettant ainsi de dégager des tendances significatives et d'explorer les nuances linguistiques qui caractérisent ce corpus artistique.

L'analyse des paroles de chansons a révélé la présence de 160 416 mots uniques au sein de l'ensemble de données. Cette diversité lexicale indique une richesse et une variété de vocabulaire au sein des textes musicaux.

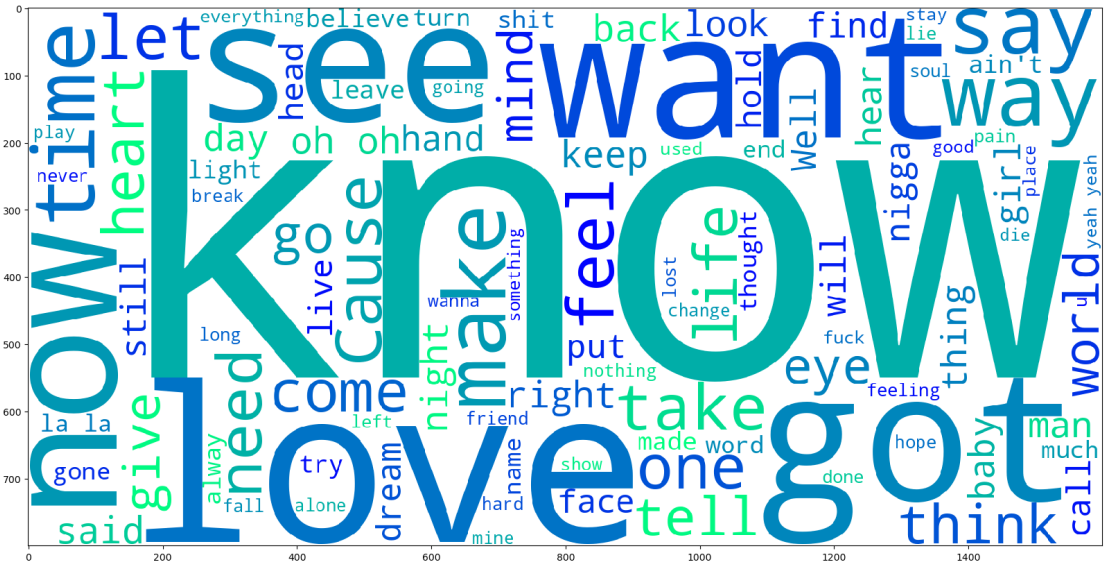


FIG. .22 : Nuages de mots

# .3 État de l'art

### .3.0.1 Algorithme clustering : K-Means

Étapes de l'algorithme :

1. **Initialisation des centres de cluster :** Sélectionnez aléatoirement $k$ points de données comme centres initiaux des clusters, appelés centroids.

2. **Attribution des points aux clusters :** Chaque point de données est attribué au cluster dont le centre (centroid) est le plus proche, souvent en utilisant la distance euclidienne.

3. **Mise à jour des centres des clusters :** Recalcul des centroids comme la moyenne des points attribués à chaque cluster.

4. **Répétition des étapes 2 et 3 :** Les étapes d'attribution et de mise à jour sont répétées jusqu'à ce qu'une convergence soit atteinte, c'est-à-dire lorsque les centres de clusters ne changent plus significativement.

L'intégration de l'algorithme K-Means dans le clustering des paroles de chansons offre une méthode efficace pour découvrir des motifs linguistiques et regrouper les œuvres musicales similaires. En initialisant des centroids et en itérant sur l'assignation et la mise à jour des clusters, K-Means segmente automatiquement les

paroles en groupes significatifs. Cette approche permet une exploration rapide des thèmes récurrents et des variations stylistiques au sein des genres musicaux.

La thèse intitulée "Groupement des thèmes des paroles de chansons à l'aide du clustering K-Means" de Dionisia Bhisetya Rarasati a suscité une inspiration pour notre propre projet, motivant l'utilisation de l'algorithme K-Means sur notre jeu de données spécifique.

### .3.0.2 Evaluation des clusters avec Wordnet

L'évaluation des clusters avec WordNet constitue une étape cruciale de notre projet, visant à affiner la qualité des regroupements obtenus à partir des paroles de chansons. WordNet, une ressource lexicale de référence, jouant un rôle essentiel dans cette évaluation en offrant une structure sémantique qui permet de mesurer la similarité entre les termes.

WordNet organise les mots en synsets, des ensembles de mots synonymes représentant une idée ou un concept particulier. Cette hiérarchie sémantique facilite l'évaluation des regroupements en mesurant la proximité conceptuelle entre les termes inclus dans un même cluster. L'utilisation de WordNet permet d'appréhender la richesse des relations sémantiques entre les mots, dépassant ainsi une simple comparaison basée sur la similarité lexicale.

Pour évaluer les clusters, nous utilisons deux mesures spécifiques dans WordNet : la mesure de similarité WordNet d'Upper Bound (Wu-Palmer) et la mesure de similarité Lin (Path Length).

La mesure Wu-Palmer quantifie la similarité en évaluant la profondeur des concepts dans la hiérarchie de WordNet. Plus les concepts sont proches dans la hiérarchie, plus la similarité est élevée. Cette mesure offre une perspective plus fine en tenant compte de la position relative des termes dans la taxonomie sémantique.

D'autre part, la mesure de similarité Lin compare la quantité d'information partagée entre deux concepts. Elle considère non seulement la profondeur des concepts, mais également la quantité d'information dans les sous-arbres de ces concepts. Ainsi, elle fournit une évaluation plus nuancée de la similarité entre les termes.

La mesure de similarité LCH (Leacock-Chodorow) est également une mesure de similarité sémantique qui calcule la similarité en fonction du chemin le plus court entre les noeuds. La valeur du chemin est ensuite normalisée en utilisant le double de la profondeur maximale du graphe, nous n'utilisons pas cette mesure car l'implémentation de wordnet plante durant son éxecution.

En somme, l'évaluation des clusters avec WordNet et l'utilisation des mesures Wup (Wu-Palmer) et Lin offrent une approche approfondie pour évaluer la co-

hérence sémantique au sein des regroupements. Ces mesures enrichissent notre analyse en considérant la signification sous-jacente des termes, permettant une évaluation plus contextuelle et précise de la qualité des clusters formés à partir des paroles de chansons.

# .4 Système réalisé

## .4.1 Description du Programme

### .4.1.1 Librairies

Les librairies utilié par le programme sont les suivantes :

- matplotlib
- pandas
- numpy
- nltk
- re
- string
- sklearn
- pywsd

pywsd est utilisé pour désambiguation dans la partie validation, ainsi que nltk pour la partie processing et validation.

### .4.1.2 Programme livré

Le code livré **Clustering_projet.ipynb** est au format notebook réalisant tout les étapes du système qui sont décrite dans la suite de cette section, également un fichier indépendant python **Scalculate.py** de la partie validation par wordnet est fournie sous forme d'un outil appelable dans son programme individuel.

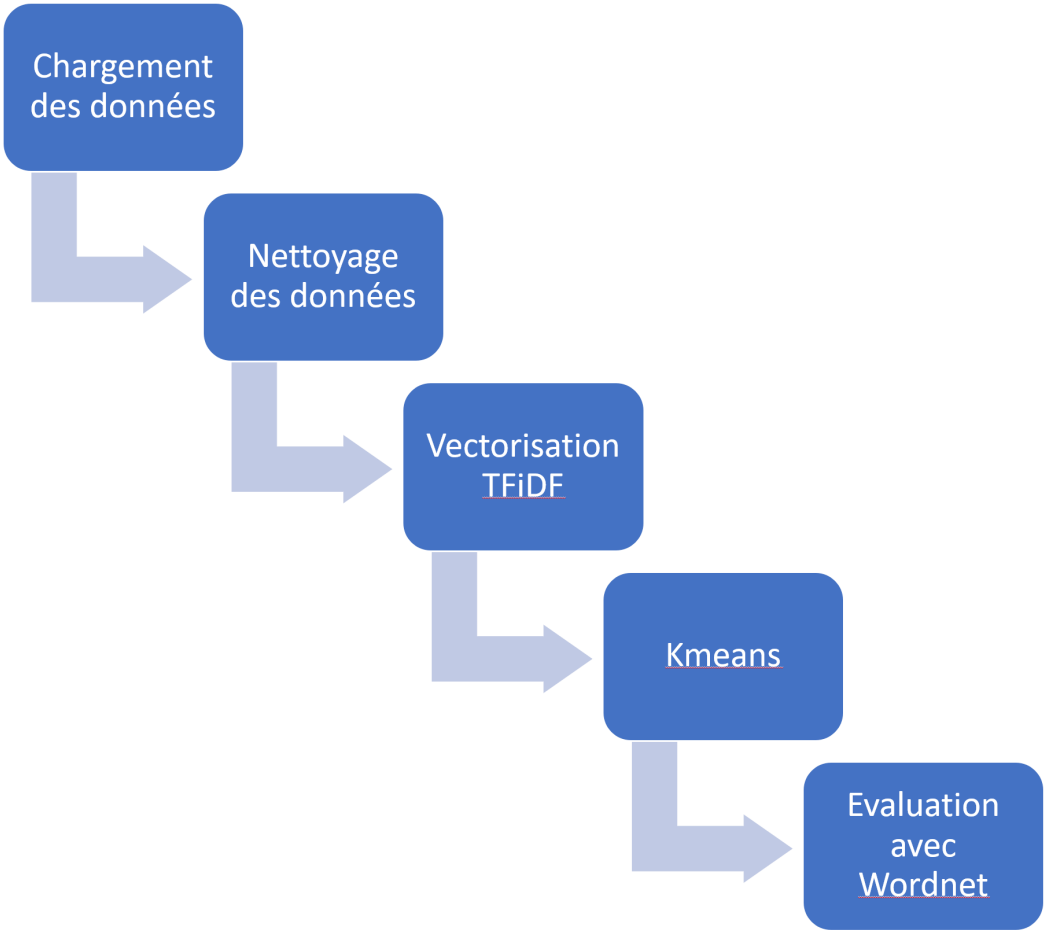


FIG. .43 : Etapes du système réalisé

## .4.2 Etapes

## .4.3 Nettoyage des données

Le prétraitement des données est important, car il permet d'atteindre les points suivants : qualité des données, modèles plus précis, réduction des erreurs, temps de traitement réduit, consistance et simplification.

La fonction cleantext est une fonction de pré-traitement du texte qui effectue plusieurs opérations pour préparer les données textuelles avant de les utiliser dans un modèle d'apprentissage automatique. Voici une explication des différentes étapes de pré-traitement dans cette fonction :

1. Supprimer la ponctuation

2. Mettre en minuscules
3. Supprimer les caractères numériques
4. Supprimer les espaces multiples
5. Supprimer les balises HTML
6. Filtrer les caractères non ASCII
7. Tokenization
8. Vérifier si le mot existe dans WordNet : Seuls les mots qui ont des synsets dans WordNet (un dictionnaire lexical) sont conservés. Cela permet d'éliminer des mots inexistants ou mal orthographiés.
9. Racinisation (Stemming)
10. Lemmatisation
11. Supprimer les mots vides (stopwords)

Enfin, cette fonction est appliquée à la colonne 'lyrics', et les résultats sont stockés dans une nouvelle colonne 'textclean'. Cette séquence de pré-traitements vise à rendre le texte plus adapté à l'analyse et à la classification, en éliminant le bruit et en conservant les informations pertinentes. L'objectif global est d'obtenir une représentation textuelle uniforme et dépourvue de bruit. Le processus de nettoyage de texte a entraîné une réduction significative de la diversité lexicale, ramenant le nombre de mots à 151 250 par rapport à l'original qui en comptait 160 416.

## .4.4 Vectorisation TF-IDF

La vectorisation des données avec TF-IDF (Term Frequency-Inverse Document Frequency) est une méthode essentielle de transformation des documents textuels en représentations numériques pour les algorithmes d'apprentissage automatique. TF-IDF attribue des poids aux termes en fonction de leur fréquence dans le document et de leur importance dans l'ensemble du corpus. Ce processus crée une matrice où chaque ligne représente un document et chaque colonne représente un terme avec son score TF-IDF, offrant ainsi une représentation numérique précieuse pour l'analyse et le traitement ultérieur.

### .4.5 Kmeans

Nous avons opté pour l'algorithme K-Means pour le clustering des paroles de chansons, une décision motivée par la simplicité et l'efficacité de cette méthode dans la segmentation de données textuelles. K-Means offre une approche intuitive pour regrouper les paroles en clusters distincts en fonction de leurs similarités lexicales.

Quant au choix du nombre de clusters, après une analyse approfondie et des expérimentations préliminaires, nous avons délibérément fixé ce nombre à 7. Cette décision découle d'une volonté de capturer une diversité significative tout en maintenant une interprétation et une gestion pratiques des résultats. Le nombre 7 a été jugé approprié pour subdiviser les paroles de chansons en catégories distinctes, offrant ainsi une granularité suffisante sans compromettre la clarté des regroupements. Nous croyons que cette configuration de 7 clusters permettra de mettre en évidence des thèmes et des motifs variés présents dans notre ensemble de données, facilitant ainsi une analyse approfondie et des insights significatifs.

### .4.6 Evaluation avec Wordnet

Calculer la similarité sémantique entre des documents constitue un problème ancien dans le domaine du traitement du langage naturel, car la similarité par WordNet est une valeur calculée entre des mots individuels et non entre des groupes de mots. Le champ de l'analyse sémantique joue un rôle crucial dans la recherche liée à l'analyse de texte. La similarité sémantique varie en fonction du domaine d'application.

Nous avons décidé de reprendre les travaux de l'article de Pawar et Mago, 2018. Il présente une méthodologie qui aborde cette problématique en incorporant la similarité sémantique entre phrases et les statistiques du corpus. Pour calculer la similarité sémantique entre les mots et les phrases, la méthode proposée adopte une approche basée sur l'utilisation d'une base de données lexicale. Cette méthodologie peut être appliquée dans divers domaines. Elle a été testée à la fois sur des normes de référence et sur un ensemble de données moyennes de similarité humaine.

Leur méthode consiste tout d'abord à décider de retirer l'ambiguïté des mots à partir de leur catégorie grammaticale en ne gardant que les verbes et noms qui sont les mots fournissant le plus d'information sur le contenu de la phrase. Les synsets associés les plus probables sont récupérés pour chaque mot, utilisés par la suite en calculant la similarité entre les deux phrases en itérant sur elles, récupérant les similarités pertinentes à l'aide des différentes méthodes de calcul du plus court chemin fournies par WordNet dans un graphe de similarité.

Nous prenons cela en compte en créant des paires de phrases, récupérant toutes

leurs évaluations de similarité afin de calculer la similarité globale du cluster à partir de ses éléments.

Un problème majeur rencontré est le temps pris par notre méthode afin d'obtenir les résultats des clusters, en raison du grand nombre d'éléments dans les clusters et de la taille importante des chansons qui devaient être traitées.

Pour résoudre ce problème, nous avons décidé de mettre en place une méthode basée sur l'algorithme TF-idf afin d'identifier les mots les plus importants de chaque élément du cluster. En traitant uniquement les éléments les plus importants, nous cherchons à raccourcir au minimum les membres qui constituent la partie majeure de chaque chanson, afin qu'ils puissent être comparés avec les autres composants du cluster pour déterminer si la similarité est présente.

# .5 Résultats

## .5.1 Analyse des Clusters de Paroles de Chansons

**Cluster 1 - Exploration de la vie :** Les mots tels que "like," "come," "life," et "world" suggèrent des thèmes liés à l'exploration de la vie et du monde. Il pourrait s'agir de chansons abordant des sujets variés de la vie quotidienne.


Fig. .54 : Exemple : Nuages de mots obtenues pour le cluster 1

**Cluster 2 - Sentiments et relations :** Les termes tels que ”love,” ”like,” ”say,” et ”youre” pointent vers des chansons centrées sur les sentiments et les relations. Le cluster pourrait englober des chansons évoquant des aspects émotionnels et relationnels.

**Cluster 3 - Émotions et expériences :** Les mots ”said,” ”love,” et ”day” suggèrent une exploration d'émotions et d'expériences diverses. Ce cluster pourrait inclure des chansons exprimant des sentiments et des expériences vécues.

**Cluster 4 - Rythmes entraînants :** Les termes ”got,” ”baby,” ”yeah,” et ”girl” indiquent des thèmes liés à des rythmes entraînants et à des sujets joyeux. Il pourrait s'agir de chansons énergiques et festives.

**Cluster 5 - Langage explicite :** Le cluster est caractérisé par des termes explicites tels que ”nigga,” ”shit,” ”bitch,” et ”fuck,” suggérant des chansons avec un langage plus direct et cru.


Fig. .55 : Exemple : Nuages de mots obtenues pour le cluster 5

**Cluster 6 - Thèmes romantiques :** Les mots ”love,” ”baby,” ”heart,” et ”youre” pointent vers des thèmes romantiques et affectifs. Ce cluster pourrait regrouper des chansons centrées sur l'amour.


Fig. .56 : Exemple : Nuages de mots obtenues pour le cluster 6

**Cluster 7 - Réflexion personnelle :** Le cluster semble caractérisé par des mots reflétant la réflexion personnelle tels que ”know,” ”time,” ”youre,” et ”say.” Il pourrait regrouper des chansons explorant des introspections personnelles.

## .5.2 Validation des clusters avec Wordnet

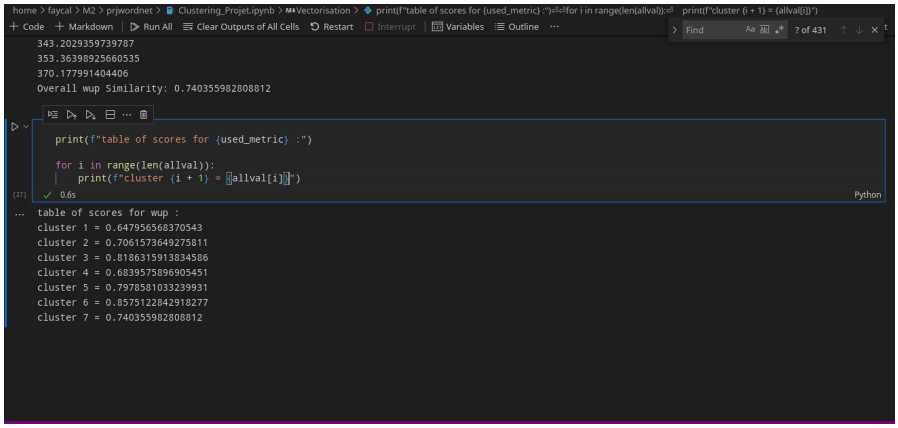


Fig. .57 : Capture de l’éxecution du programme

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|---|---|---|---|---|---|---|---|
| **Wup Similarity** | 0.65 | 0.70 | 0.81 | 0.68 | 0.79 | 0.85 | 0.74 |
| **Path Similarity** | 0.12 | 0.17 | 0.19 | 0.15 | 0.211 | 0.218 | 0.158 |

Tab. 1 : Similarités entre les clusters selon différentes mesures

L’analyse des résultats se concentre sur l’évaluation de la similarité entre les clusters (C1 à C7) à l’aide de deux mesures distinctes : la similarité Wup et la similarité de chemin (path similarity). Voici une interprétation de ces résultats :

1. **Similarité Wup :**
   - Les valeurs de similarité Wup varient entre 0.65 et 0.85, indiquant la proximité sémantique entre les paires de clusters. Une similarité plus élevée suggère une proximité sémantique plus forte.
   - Les clusters C3 et C6 montrent la plus forte similarité Wup avec une valeur de 0.81, indiquant une forte proximité sémantique entre ces deux clusters.
   - Les clusters C1, C4, et C5 présentent également des valeurs de similarité

Wup élevées, indiquant des similitudes sémantiques significatives entre ces clusters.

2. **Similarité de Chemin :**

   - Les valeurs de similarité de chemin varient entre 0.12 et 0.218, mesurant la proximité sémantique en termes de distance dans la hiérarchie sémantique de WordNet.
   - Les clusters C5 et C6 présentent la plus forte similarité de chemin avec une valeur de 0.218, indiquant une proximité sémantique plus prononcée selon cette mesure.
   - Les clusters C3 et C6, bien que similaires selon la mesure Wup, montrent une similarité de chemin légèrement plus faible (0.19), suggérant une nuance dans la nature de leur proximité sémantique.

En résumé, ces résultats suggèrent des tendances intéressantes dans la similarité sémantique entre les clusters. La combinaison de ces deux mesures offre une perspective complète sur les relations sémantiques entre les regroupements, permettant une évaluation approfondie de la qualité du modèle.

# .6 Conclusion

Le projet a été mené à bien, avec pour objectif d'explorer la diversité des paroles de chansons à travers une approche de clustering basée sur le traitement du langage naturel. L'utilisation de l'algorithme K-Means a permis de regrouper les chansons en clusters distincts, révélant des tendances sémantiques et des thèmes communs au sein de notre vaste ensemble de données.

L'évaluation des clusters à l'aide de mesures de similarité, telles que Wup et la similarité de chemin, a fourni des insights précieux sur la proximité sémantique entre les regroupements. Ces résultats permettent de mieux comprendre la structure sous-jacente des paroles de chansons et ouvrent la voie à des analyses plus approfondies.