Conservatoire National des Arts et Métiers

RCP216
Ingénierie de la fouille et de la visualisation de données massives

---

Projet : Classification automatique d'un corpus de résumés d'articles scientifiques en neuroscience

---

Gaëlle Botton-Amiot

Session 1 - Février 2025

TABLE DES MATIÈRES

FIGURES

TABLEAUX

1. Introduction

Présentation du jeu de données

Les données ont été collectées dans le cadre d'un projet d'analyse de données en lien avec l'électroencéphalographie (EEG). Cette technique permet l'enregistrement non-invasif de l'activité cérébrale au moyen d'électrodes placées sur le crâne. Elle est notamment utilisée lors des premières phases d'études cliniques sur sujets volontaires sains, afin de connaitre l'effet sur le cerveau humain de nouveaux médicaments. L'objectif général du projet était d'avoir une vue d'ensemble des effets pharmacologiques de nombreux médicaments et molécules publiés dans la littérature scientifique, comme élément de comparaison avec de nouvelles molécules testées en recherche clinique.

En parallèle d'une collecte et analyse de données soigneusement sélectionnés à plus petite échelle, nous avons engagé une collecte d'information à large échelle issue d'articles scientifiques qui portent sur cette technique

Nous nous intéressons en particulier aux articles scientifiques originaux (excluant les revues ou études de cas) portant sur des sujets humains sains, entre 1990 et aujourd'hui (sélection avec une requête booléenne avec certains mots-clés MeSH, et des filtres spécifiques), et testant une ou plusieurs molécules actives/drogues listées dans la base de données TTD (*Therapeutic Target Database*)[1].

Pour cela, nous avons mis au point un programme qui effectue une recherche croisée avec l'identifiant Pubchem ou le nom de la molécule dans la base TTD, et les articles indexés sur Pubmed au moyen d'un pipeline combinant les EUtils Elink-ESearch-EFetch mis à disposition par NCBI[2].

Les étapes effectuées pour la collecte des articles sont résumées dans la Figure 1. Au total, un corpus regroupant 2101 résumés d'articles a été constitué en novembre 2024 et sera exploré dans le présent projet. Pour chaque article, nous disposons des informations suivantes :

- Titre
- Résumé
- Identifiant unique DOI
- Liste de mots-clés MeSH
- Liste de médicaments/molécules mentionnés
- Année de publication

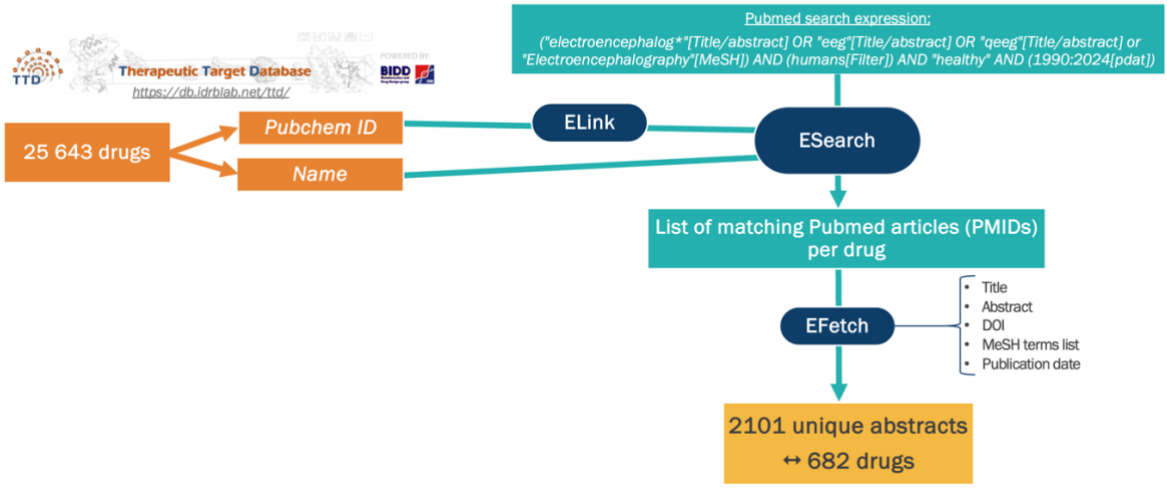


Figure 1: Étapes effectuées pour l'obtention du corpus de résumés d'articles scientifiques sur l'électroencéphalographie avec les APIs de Pubmed (NCBI)[2].

Objectifs

L'objectif de ce projet est d'identifier des thématiques sous-jacentes dans ce corpus de résumés d'articles scientifiques, en mettant en évidence les concepts récurrents et leur évolution au cours du temps.

Le traitement suivra plusieurs étapes : encodage des résumés dans un espace vectoriel, puis regroupement des articles par des méthodes de classification automatique (apprentissage non-supervisé ou clustering). Enfin, une analyse qualitative des regroupements obtenus permettra d'identifier leurs thématiques *a posteriori*, et de comparer les résultats obtenus entre les différentes techniques d'encodage et de classification.

2. Matériel et Méthodes

Système et environnement

Tous les résultats présentés ci-après ont été obtenus avec Python et PySpark[3]. De nombreuses librairies ont été utilisées, elles sont récapitulées dans le tableau suivant.

Usage	***Package***
Visualisation des données	Matplotlib, Seaborn, WordCloud (Python)
Traitement des data frames	Pandas, Numpy (Python)
Classification automatique, Réduction de dimension	Spark MLlib
Traitement de données textuelles	Spark-NLP
Indicateurs de validité de la classification (Rand, Information mutualisée)	Scikit-learn (Python)
Diagramme de Sankey	networkD3 (R)

Table 1: Librairies employées pour le projet.

Représentation vectorielle des textes et analyse sémantique latente

Les résumés d'articles sont des textes d'environ 2000 charactères. Ils ont été convertis en des représentations vectorielles au moyen de 3 méthodes différentes qui seront détaillées ci-après:

- Encodage des mots TF-IDF
- Encodage des phrases avec Doc2Vec
- Encodage des phrases avec InstructorEmbeddings

L'ensemble des articles ayant été encodé dans un espace vectoriel de grande dimension, une méthode de réduction de dimension a ensuite été utilisée pour les visualiser. Parmi ces méthodes, l'Analyse en Composantes Principales (ACP) s'est avérée être un choix pertinent pour représenter les abstracts visuellement en deux dimensions. Contrairement à d'autres méthodes de réduction de dimension non linéaires (t-SNE, UMAP, etc…), cette méthode est directement implémentée dans Spark[4].

La visualisation a été réalisée dans des combinaisons des trois premières dimensions afin de mieux comprendre la structure des données. La longueur des abstracts peut avoir un impact important sur la variance expliquée dans les premières dimensions, nous avons donc choisi celle qui présente le moins d'impact. C'est la raison pour laquelle certaines représentations utilisent les dimensions 1 et 3 ou 2 et 3, selon les résultats obtenus (Annexes : Fig1-3).

Classification automatique

D'une part, après la pondération TF-IDF, la classification a été réalisée à l'aide de la Latent Dirichlet Allocation, une méthode spécifique dédiée à l'identification des thèmes dans les documents textuels [5].
D'autre part, après l'encodage des mots ou des phrases (Doc2Vec ou InstructorEmbeddings), les textes, représentés sous forme de vecteurs de grande dimension, ont été classifiés à l'aide de l'algorithme des K-moyennes. Cet algorithme de classification automatique, dans sa version parallélisable (K-means || avec initialisation K-means++), est également implémenté dans Spark [4].

Analyse et validation des classes

Analyse Post-Hoc des classes obtenues

Les termes MeSH de PubMed, collectés en parallèle des articles, ont été utilisés pour analyser les groupes obtenus :

- Le nombre de mentions des termes MeSH a été comptabilisé pour chaque groupe, suivi du calcul de leur proportion relative au sein de celui-ci.
- Un score pondéré a ensuite été établi en multipliant le nombre d'occurrences d'un terme par sa proportion dans le sujet.
- Les termes présents dans plus de 80 % des groupes, considérés comme les plus courants et non spécifiques, ont été éliminés de l'analyse.
- Une liste des 20 termes MeSH les plus représentatifs (score pondéré le plus haut) pour chaque classe a été collectée (Annexes : Tableaux 1-3).

Identification et évaluation des classes

La connaissance métier acquise sur l'électroencéphalographie, ainsi que sur la littérature scientifique associée, a été un atout précieux pour évaluer la qualité des classes obtenues.
Les étapes d'évaluation des classes sont les suivantes :

- Identification de la thématique des classes à partir de la liste des termes MeSH les plus représentatifs, et attribution d'un nom, si possible.
- Comptage des termes MeSH correspondant strictement au thème attribué, parmi la liste des 20 termes les plus représentatifs.
- Échantillonnage aléatoire de 20 articles par classe identifiée, et validation manuelle du nombre d'articles traitant du thème défini parmi l'échantillon.

Les résultats détaillés de cette validation sont visibles dans les dernières colonnes des tableaux en annexes (Annexes : Tableaux 1-3).

Autres partitionnements

Une répartition en 10 groupes a été choisie pour les trois méthodes retenues afin de permettre une comparaison directe de leurs performances. Cependant, des répartitions avec des nombres différents de clusters ont également été réalisées, notamment en 5 clusters (Annexes : Tables 4-6).

4. Encodage TF-IDF des mots et LDA

Encodage TF-IDF

Le procédé d'encodage TF-IDF (Term Frequency-Inverse Document Frequency) permet de pondérer l'importance des termes en fonction de leur fréquence dans un document spécifique et de leur rareté à l'échelle de l'ensemble des documents. Cela permet de mettre en avant les termes discriminants tout en minimisant l'influence des termes communs. Les étapes suivantes de pré-traitement ont d'abord été effectuées :

- **Tokenizer :** Segmentation des textes en tokens individuels.
- **Normalizer :** Nettoyage des textes par l'élimination des caractères spéciaux et des valeurs numériques. Ce nettoyage est important dans le cas d'articles biomédicaux, car des informations telles que les doses administrées ou les durées d'enregistrement (fréquentes dans ce corpus) peuvent altérer l'analyse.
- **Lemmatizer :** Réduction des mots à leur forme canonique (lemme) pour éviter les redondances liées aux différentes flexions ou conjugaisons d'un même terme.
- **StopwordCleaner :** Suppression des mots vides (stopwords) non pertinents pour l'analyse.

Ensuite, le calcul de la matrice de fréquence des termes (TF) est calculé avec le **CountVectorizer**. Comme visible dans le code suivant, le vocabulaire est limité à 1000 termes afin de réduire la complexité de l'analyse tout en maintenant les informations clés. De même, les termes trop rares (présents dans moins de 20 documents) ou trop fréquents (présents dans plus de 40 % des documents) sont éliminés, car ils apportent peu d' information discriminante.

```
from pyspark.ml.feature import CountVectorizer, IDF
from pyspark.sql.functions import udf, col, explode
# Calcul de la fréquence des termes (TF)
cv = CountVectorizer(inputCol="output", outputCol="rawFeatures", vocabSize=1000, minDF=20, maxDF=0.4)
cv_model = cv.fit(result_lemma)
# Calcul des valeurs IDF
idf = IDF(inputCol="rawFeatures", outputCol="features")
# Pipeline
pipeline_tf_idf = Pipeline(stages=[cv,idf])
# Application du pipeline
model_tfidf = pipeline_tf_idf.fit(result_lemma)
result_tfidf = model_tfidf.transform(result_lemma)
```

Latent Dirichlet Allocation

La Latent Dirichlet Allocation (LDA) est une méthode probabiliste de classification automatique des textes, souvent utilisée pour l'extraction de thèmes latents dans des ensembles de documents textuels [5].

La LDA a été utilisée avec les représentations vectorielles issues du modèle TF-IDF pour identifier les thèmes présents dans les articles. Les distributions des termes dans chaque thème ont permis de sélectionner les mots les plus discriminants pour décrire chaque groupe. Les mots ayant les plus fortes probabilités dans chaque groupe sont interprétés comme caractéristiques principales du thème, ils sont présentés dans le tableau 2.

Classe	**10 termes les plus représentatifs**
0	['comparison', 'serotonergic', 'serotonin', 'genotype', 'gene', 'paradigm', 'dependence', 'genetic', 'acth', 'accord']
1	['pd', 'network', 'erp', 'connectivity', 'motor', 'process', 'couple', 'coherence', 'enhance', 'stimulus']
2	['sleep', 'night', 'power', 'rem', 'frequency', 'band', 'theta', 'delta', 'group', 'density']
3	['patient', 'epilepsy', 'seizure', 'child', 'temporal', 'adhd', 'image', 'lobe', 'control', 'mri']
4	['mg', 'dose', 'placebo', 'administration', 'hour', 'plasma', 'day', 'concentration', 'sleep', 'drug']
5	['rest', 'cognitive', 'method', 'woman', 'gh', 'performance', 'severity', 'intake', 'test', 'sustain']
6	['pain', 'stimulation', 'response', 'interval', 'process', 'task', 'potential', 'evoke', 'stimulus', 'anxiety']
7	['schizophrenia', 'gate', 'quantitative', 'nicotine', 'auditory', 'amplitude', 'score', 'deficit', 'gamma', 'gaba']
8	['administer', 'sedation', 'infusion', 'ad', 'concentration', 'heart', 'bis', 'consciousness', 'anesthesia', 'min']
9	['learn', 'memory', 'oscillatory', 'model', 'receptor', 'opioid', 'scopolamine', 'pharmacodynamic', 'human', 'relation']

Table 2: Termes les plus représentatifs des 10 classes obtenues avec la LDA.

Si la LDA repose sur l'hypothèse que chaque document est une combinaison de plusieurs thèmes, et donne une distribution probabiliste des thèmes pour chaque document, nous nous sommes contentés ici d'attribuer la classe dominante (i.e. avec la probabilité la plus élevée) à chaque article.

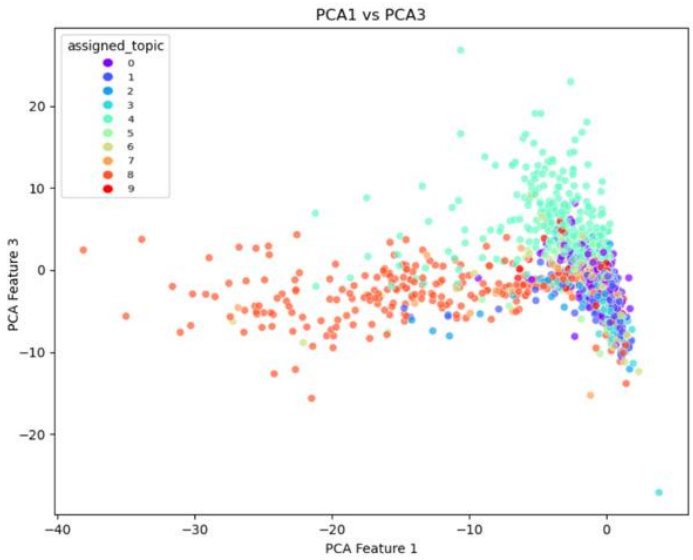
Visualisation et description des classes obtenues

Enfin, nous avons attribué un nom et analysé la qualité (Table 3, Annexe : Table 1) et la répartition des classes obtenues (Figure 2).

On constate que 9 sur 10 classes ont pu avoir un thème identifié à partir des mots-clés et termes les plus courants.

Classe	**Total nb. articles**	**Nom attribué**	**Score validation MeSH**	**Score validation Articles**
0	142	"Monoamines and genetic variation"	80%	60%
1	237	"Parkinson and motor control"	35%	40%
2	275	"Sleep"	75%	75%
3	265	"Other neural imaging methods: PET, MRI in diagnosis"	50%	25%
4	260	"Drugs effect on EEG - healthy volunteers"	65%	90%
5	113	"Stimulants and attention"	40%	65%
6	357	"Sensory processing and pain"	55%	45%
7	153	"Schizophrenia"	30%	50%
8	177	"Anesthesia and sedation"	95%	95%
9	115	-	-	-

Table 3: Classes obtenues avec TF-IDF et LDA et le nom attribué. Les scores de validation MeSH et articles est calculé en comptant respectivement le nombre du top 20 des termes MeSH et de 20 articles sélectionnés aléatoirement dans le groupe qui sont en accord avec le nom de classe attribué. Voir Table 1 en annexe pour le détail et la liste complète des termes MeSH.



Sur la représentation sémantique latente (Fig.2), il est intéressant de constater que les résumés d'articles traitant d'« Anesthésie et sédation » (groupe 7) sont dispersés sur la gauche, indiquant que dans cette représentation, ces résumés sont plus éloignés des autres.

Figure 2: Analyse sémantique latente. Projection sur les composantes 1 et 3 de l'ACP effectuée sur les résumés encodés en vecteurs avec la méthode TF-IDF. Les couleurs correspondent aux clusters déterminés par LDA.

5. Encodage avec Doc2Vec et K-means

Représentation vectorielle des mots ou des phrases

Une autre approche consiste en la représentation vectorielle des mots (approche "bag-of-words") ou des phrases dans les résumés. Plusieurs approches ont été comparées en termes de performance sur la classification automatique (K-means), mesurée par le coefficient de silhouette. Ce dernier quantifie la cohérence interne des clusters, une valeur plus élevée indiquant des clusters bien séparés et homogènes. Les résultats obtenus pour plusieurs méthodes sont présentés dans le Tableau 4, et détaillés en Annexe (Tables 7 – 9).

Le modèle **Doc2Vec**, une variante de Word2Vec, propose une vectorisation adaptée à des unités plus larges, comme des phrases, des paragraphes ou des documents entiers. Contrairement à Word2Vec, qui génère des vecteurs pour des mots individuels basés sur leur contexte local, Doc2Vec incorpore des informations contextuelles sur l'ensemble du corpus, ce qui le rend particulièrement utile pour des tâches impliquant des relations globales dans les données.

Selon une étude, Word2Vec permet d'obtenir de meilleurs résultats sur des textes dans le domaine biomédical [6]. Cependant, dans ce projet, le clustering réalisé après un encodage par Doc2Vec a donné des résultats plus satisfaisants, selon le coefficient de silhouette (Table 4).

Méthode d'encodage	Taille du vecteur obtenu	Score de silhouette pour 10 clusters avec K-means
Doc2Vec	100	0.22
Word2Vec	100	0.13
Small Bert	128	0.07
Glove 100d	100	0.04

Table 4: Comparaison du coefficient de silhouette obtenu après partition en 10 clusters avec l'algorithme des K-means pour chacune des méthodes d'encodage de mots testée.

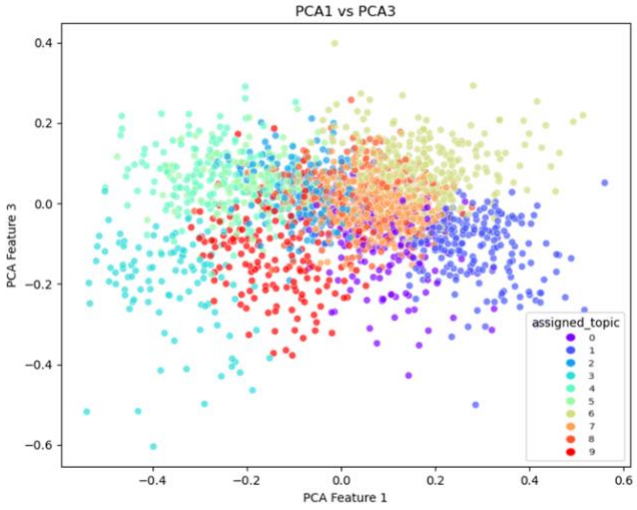
Pour l'application des méthodes Doc2Vec comme Word2Vec, un pré-traitement similaire à celui effectué avant la transformation TF-IDF a été effectué : Tokenizer, Normalizer, Lemmatizer, StopwordCleaner. Enfin, chaque document est représenté par l'encodage des phrases qu'il contient, obtenu en prenant la moyenne des embeddings des mots ou des phrases au sein de chaque abstract.

Classification automatique avec l'algorithme des K-means

Ensuite, une fois la représentation des résumés en vecteurs de dimension 100 obtenus, l'algorithme des K-means est utilisé pour effectuer une classification en 10 groupes. Le nom attribué à ces groupes et les scores de validation obtenus sont résumés dans le Tableau 5 et détaillé en annexe (Table 2).

Classe	Total nb. articles	Nom attribué	Score validation MeSH	Score validation Articles
0	175	-	-	-
1	209	-	-	-
2	195	"Anesthesia and sedation"	65%	70%
3	103	"Sleep 1"	70%	100%
4	160	"Drugs effect on EEG - healthy volunteers"	55%	95%
5	159	-	-	
6	296	"Other neural imaging methods: PET, MRI, etc... in diagnosis"	50%	25%
7	435	"Evoked potentials"	65%	55%
8	208	"Brain waves – spectral analysis"	40%	75%
9	154	"Sleep 2"	65%	95%

Table 5 : Classes obtenues avec l'encodage Doc2Vec et une partition en 10 classes avec les K-Means, et le nom attribué aux classes. Les scores de validation MeSH et articles est calculé en comptant respectivement le nombre du top 20 des termes MeSH et de 20 articles sélectionnés aléatoirement dans le groupe qui sont en accord avec le nom de classe attribué. Voir Table 2 en annexe pour le détail et la liste complète des termes MeSH.



Sur la représentation sémantique latente correspondante (Fig.3), il est intéressant de constater que les résumés d'articles traitant du sommeil (groupes 3 et 9) sont localisés l'un à côté de l'autre sur la gauche, indiquant que dans cette représentation, ces résumés sont bien proches, comme le suggère leur thématique commune.

Figure 3 : Analyse sémantique latente. Projection sur les composantes 1 et 3 de l'ACP effectuée sur les résumés encodés en vecteurs avec la méthode Doc2Vec. Les couleurs correspondent aux clusters déterminés par l'algorithme des K-moyennes.

6. Encodage des phrases avec *InstructorEmbeddings* et K-means

Représentation vectorielle des résumés avec InstructorEmbeddings

Les I**nstructorEmbeddings** permettent de générer des vecteurs représentant les résumés d'articles tout en intégrant des instructions explicites pour orienter l'encodage, comme la description du type de textes ou l'usage de l'encodage [7]. Cette méthode se distingue par sa capacité à produire des encodages adaptés aux spécificités du corpus et offrant ainsi une meilleure discrimination pour la tâche à effectuer ensuite. L'instruction employée, adaptée de la documentation, est la suivante :

```
embeddings = InstructorEmbeddings.pretrained() \
    .setInputCols(["document"]) \
    .setInstruction("Represent the biomedical abstracts for clustering: ") \
    .setOutputCol("instructor_embeddings")
```

Ensuite, comme l'InstructorEmbeddings produit un encodage (vecteur) pour chaque phrase, la moyenne des vecteurs obtenus pour chaque résumé a été calculée avec une fonction UDF dédiée :

```
# UDF pour effectuer la moyenne des vecteurs
def average_vectors(vectors):
    avg_vector = np.mean(vectors, axis=0).tolist()
    return avg_vector


average_vectors_udf = udf(average_vectors, ArrayType(FloatType()))
# Application de la fonction et conversion en type vecteur pour l'ACP
result_modif = result_embed.withColumn("mean_output", average_vectors_udf(result_embed["output"]))
result_modif = result_modif.withColumn("features", array_to_vector("mean_output"))
```

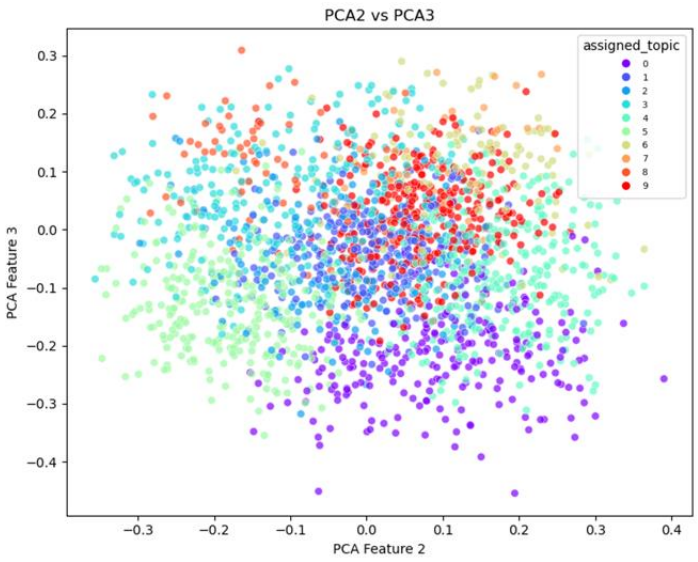
Classification automatique avec l'algorithme des K-means

L'encodage avec l'InstructorEmbeddings produit des vecteurs à 768 dimensions, une réduction dimensionnelle est donc nécessaire avant d'appliquer l'algorithme des K-means. On retient donc les 100 premières valeurs singulières (SVD) sur lesquelles sont employées les K-means pour une classification en 10 groupes :

Classe	Total nb. articles	Nom attribué	Score validation MeSH	Score validation Articles
0	237	"Drugs effect on EEG - healthy volunteers"	40%	50%
1	275	-	-	-
2	202	"Pain"	55%	25%
3	225	"Sleep"	100%	100%
4	313	-	-	-
5	300	"Pharmakodynamic (PD) studies"	45%	90%
6	124	"Schizophrenia"	55%	90%
7	37	-	-	-

8	67	“Hormones studies”	85%	95%
9	314	-	-	-

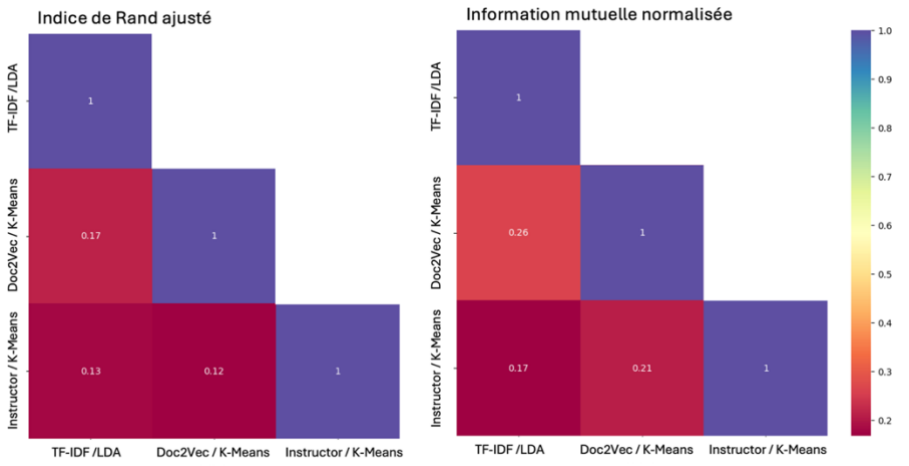
Table 6 : Classes obtenues avec l’encodage InstructorEmbeddings et une partition en 10 classes avec les K-Means, et le nom attribué aux classes. Les scores de validation MeSH et articles est calculé en comptant respectivement le nombre du top 20 des termes MeSH et de 20 articles sélectionnés aléatoirement dans le groupe qui sont en accord avec le nom de classe attribué. Voir Table 3 en annexe pour le détail et la liste complète des termes MeSH.



Ici, seulement 6 thèmes peuvent être clairement identifiés, mais avec une remarquable qualité pour 4 d’entre eux, comme en témoignent les scores de validation élevés pour les articles testés (Tableau 6, détails dans la Table 3 en annexe).

Les dimensions 2 et 3 ont été utilisées pour représenter le nuage des résumés classés (Fig.4). On peut noter le recouvrement partiel des articles traitant du sommeil (groupe 3), et du système hormonal (groupe 8), thématiques qui sont en pratique souvent associées.

Figure 4 : Analyse sémantique latente. Projection sur les composantes 2 et 3 de l'ACP effectuée sur les résumés encodés en vecteurs avec la méthode InstructorEmbeddings. Les couleurs correspondent aux clusters déterminés par l’algorithme des K-moyennes.



7. Discussion

Comparaison des résultats obtenus avec les 3 méthodes

On s’interesse aux similitudes dans la composition des groupes entre les méthodes. **L’indice de Rand** et l’**information mutuelle normalisée (NMI)** évaluent respectivement le degré de similitude et la quantité d’information partagée entre les partitions obtenues (Fig.5). Dans les deux cas, une valeur proche de 1 indique une forte similarité dans la composition des groupes. Les valeurs étant ici plutôt autour de zéro, on s’aperçoit donc que la composition exacte de tous les groupes varie fortement entre les 3 méthodes.

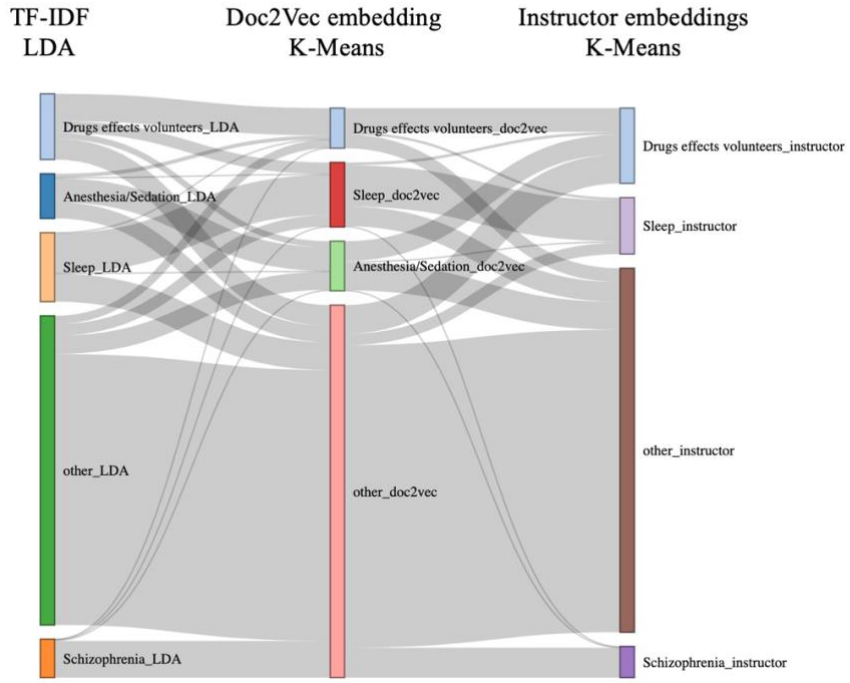
Figure 5 : Comparaison de la composition des groupes formés avec les 3 méthodes retenues : Indice de Rand ajusté (gauche) et information mutuelle normalisée (droite).

Certains groupes générés lors du clustering présentent pourtant des thèmes relativement similaires, ce qui témoigne de la robustesse globale des résultats obtenus par les algorithmes de clustering (voir Tables 1 à 3 pour une description détaillée des groupes et des thèmes). En particulier, les classes regroupant notamment les articles sur le sommeil, ou des études pharmacologiques sur sujets sains, sont visibles avec toutes les méthodes, ce qui est un indicateur de la pertinence de ces thématiques dans le corpus.

On peut ainsi identifier 4 thématiques qui ressortant avec une certaine confiance dans 2 des 3 méthodes testées. Elles sont listées dans le tableau 7, et les articles communs entre ces 4 groupes identifiés avec les 3 méthodes sont représentés dans un diagramme de Sankey (Fig.6).

	TF-IDF / LDA		Doc2Vec / K-means		Instructor / K-means	
Thématiques identifées	Cluster	Nb. Articles	Cluster	Nb. Articles	Cluster	Nb. Articles
"Sleep"	2	275	3 + 9	257	3	225
"Drugs effect on EEG - healthy volunteers"	4	260	4	160	5	300
"Schizophrenia"	7	153	-	-	6	124
"Anesthesia and sedation"	8	177	2	195	-	-

Table 7 : Nombre d'articles dans les groupes similaires obtenus avec les 3 méthodes d'encodage et de partitionnement retenues.



On constate qu'une majorité des articles identifiés pour la thématique du sommeil sont communs entre les 3 méthodes, sans être total. Au contraire, il semble que seulement la moitié des articles groupés sous les autres thématiques soit partagé entre les 3 méthodes, illustrant la divergence mesurée avec l'indice de Rand et le score d'information mutuelle normalisée.

Figure 6 : Diagramme de Sankey représentant les articles communs entre les catégories similaires entre méthodes (voir Table 7).

Des études plus approfondies des classes obtenues pourraient certainement permettre de déterminer la méthode la plus adaptée, ainsi que le nombre de classes idéal. En effet, il a été noté que des thématiques distinctes sont parfois regroupées au sein d'un même groupe. L'étude de la distribution probabiliste des thèmes obtenue avec la LDA pour chaque article mériterait par exemple d'être approfondie.

Évolution des thématiques au cours du temps

Dans la Figure 7, on peut observer une tendance générale d'évolution similaire des thématiques obtenues avec les différentes méthodes, ce qui suggère que les classes formées partagent une composition similaire. Cela indique une certaine stabilité dans l'identification des tendances majeures à travers les différentes techniques d'encodage.

D'un point de vue scientifique, plusieurs tendances intéressantes ressortent :

- Pic de publications sur l'anesthésie/sédation entre 2000 et 2010 (Fig.7, violet)
- Baisse des articles sur les essais sur sujets sains et sur le sommeil (Fig.7, vert et bleu, respectivement). Cela pourrait être attribué à l'évolution et diversification des méthodologies expérimentales sur ces sujets, avec des techniques plus avancées, et donc un vocabulaire plus technique et spécifique.
- Augmentation du nombre d'articles concernant la schizophrénie (Fig.7, rouge). Cela reflète sans doute l'intérêt croissant et les avancées dans les recherches sur le traitement et le diagnostic des troubles psychiatriques.

Ces observations montrent que, malgré les différences dans les méthodes d'encodage, les tendances globales restent cohérentes, ce qui suggère une pertinence continue des thématiques sur la durée.

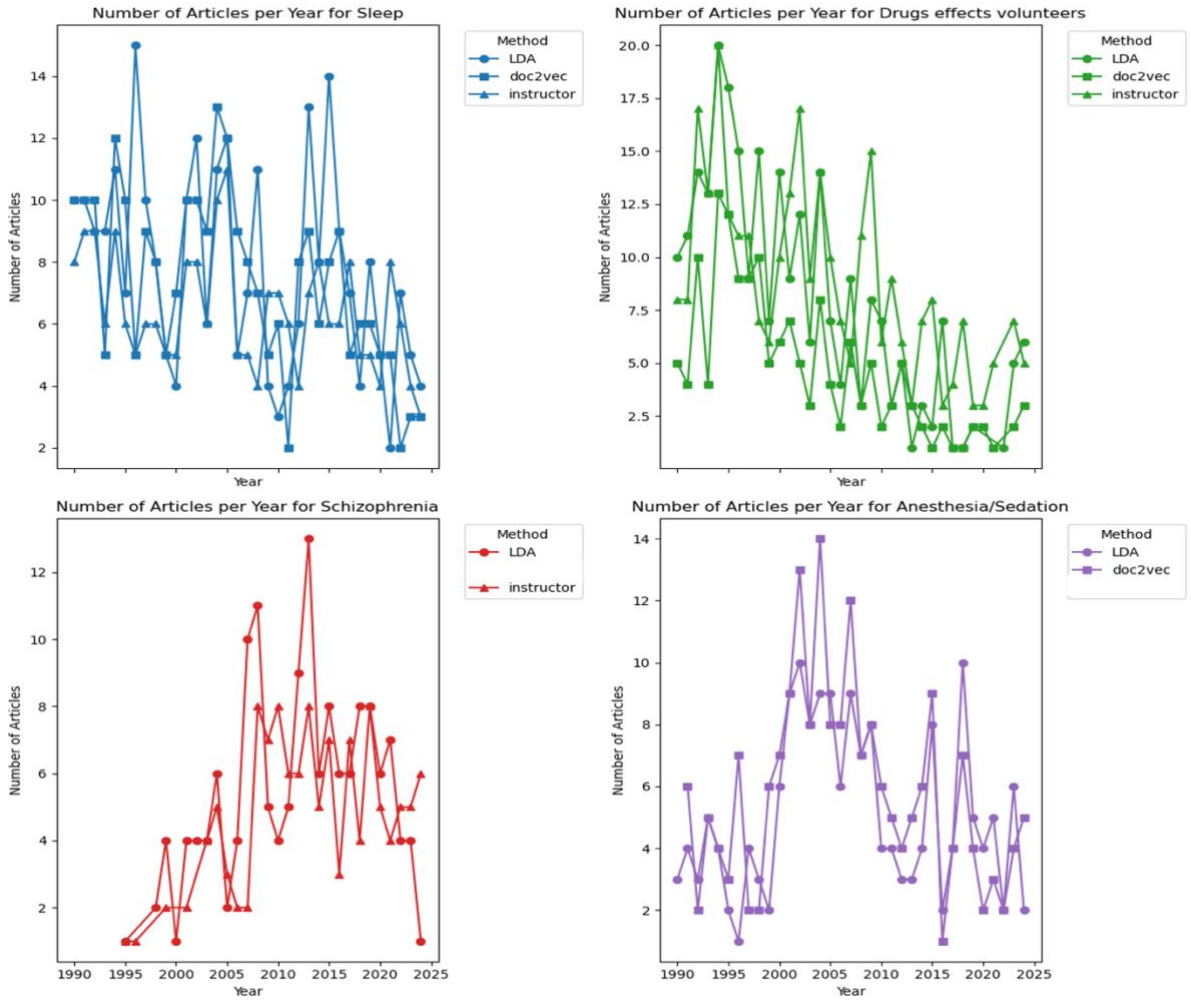


Figure 7 : Évolution du nombre d'article par thématique en fonction de leur année de publication.

Etudes en lien avec le sommeil

Enfin, comme la classe qui regroupe les études en lien avec le sommeil semble la plus robuste, nous avons essayé de représenter visuellement les mots les plus courants dans cette classe avec la méthode « nuage de mots » (WordCloud). Le résultat, visible dans la figure 8, montre que les mots les plus représentatifs de ce groupe sont extrêmement similaires entre les 3 méthodes. L'étude plus avancée de la composition de cette classe permettrait sans aucun doute de dégager des tendances principales sur l'évolution des recherches dans ce domaine.

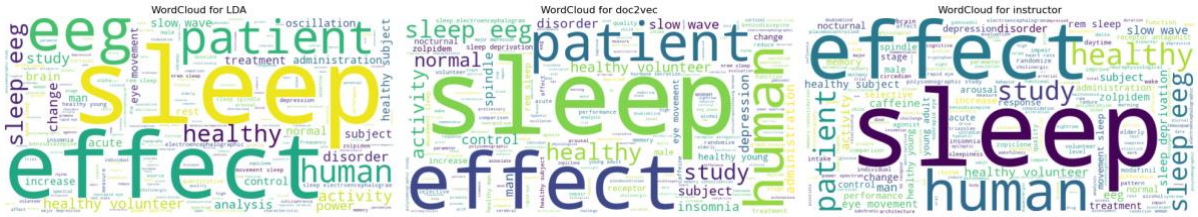


Figure 8 : Nuages de mots pour le groupe « sommeil » tels que déterminé par chacune des 3 méthodes retenues : TF-IDF+LDA (gauche), Doc2Vec+K-means (centre) et InstructorEmbeddings+K-means (droite).

8. Conclusion

Ce projet a permis d'identifier différentes thématiques scientifiques dans le domaine de l'électroencéphalographie (EEG), à partir des résumés d'articles et à l'aide de techniques de d'encodage du texte et de classification automatique (apprentissage non supervisé).
Affinées, ces catégories thématiques associées aux articles pourraient à l'avenir être utilisées pour entraîner un modèle d'apprentissage supervisé, permettant ainsi de classifier automatiquement de nouveaux articles publiés en fonction de leurs thématiques, à partir de leur résumé mais aussi de la liste des termes MeSH, et autres variables collectées concomitamment. Cela faciliterait par exemple l'indexation et l'organisation d'une base de données sur l'EEG en continu.

Un intérêt non anticipé de ce projet réside dans la possibilité d'identifier et grouper les études pharmacologiques sur volontaires sains ("Drugs effect on EEG - healthy volunteers", Table 7). Dans le cadre d'un projet plus général sur l'EEG, ces articles étaient la cible principale des recherches bibliographiques et devaient être sélectionnés manuellement avant d'être étudié plus en détail et inclus dans une base de données. L'une des méthodes employées ici pourrait donc permettre d'automatiser cette étape de tri et offrir une pré-sélection d'articles à inclure dans la base.

Enfin, un autre aspect intéressant concerne l'identification des molécules/médicaments les plus couramment mentionnés au sein de chaque catégorie, et leur évolution au cours du temps. Cela pourrait être effectué à l'aide de la liste collectée parallèlement aux résumés d'articles, mais nécessiterait un travail important de sélection et exclusion des noms de familles de molécules et/ou de composés pharmacologiques non pertinents (voir le nuage de mots pour le groupe « sommeil » pour exemple dans la figure 4 en annexe).

Malgré certaines limitations, ce projet ouvre des pistes prometteuses pour l'optimisation de l'analyse et de l'indexation d'articles scientifiques dans le domaine biomédical, notamment en facilitant le tri automatique des publications, et permettant l'identification et le suivi de tendances dans la littérature scientifique.

9. Références

[1] Y. Zhou *et al.*, “TTD: Therapeutic Target Database describing target druggability information,” *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1465–D1477, Jan. 2024, doi: 10.1093/nar/gkad751.

[2] E. Sayers, “The E-utilities In-Depth: Parameters, Syntax and More,” in *Entrez Programming Utilities Help [Internet]*, National Center for Biotechnology Information (US), 2022. Accessed: Jan. 06, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25499/

[3] “PySpark Overview — PySpark 3.5.4 documentation.” Accessed: Jan. 16, 2025. [Online]. Available: https://spark.apache.org/docs/latest/api/python/index.html

[4] “KMeans — PySpark 3.5.4 documentation.” Accessed: Jan. 09, 2025. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.KMeans.html

[5] D. M. Blei, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, no. 3, pp. 993–1022, 2003.

[6] Q. Chen and M. Sokolova, “Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts,” *Sn Comput. Sci.*, vol. 2, no. 5, p. 414, 2021, doi: 10.1007/s42979-021-00807-1.

[7] H. Su *et al.*, “One Embedder, Any Task: Instruction-Finetuned Text Embeddings,” May 30, 2023, *arXiv*: arXiv:2212.09741. doi: 10.48550/arXiv.2212.09741.