Rapport Projet NLP

Exploration des Thématiques Médicales dans PubMed : Scraping, Annotation et Apprentissage

Author :
Nadia BEN YOUSSEF

Email :
nadia.benyoussef@dauphine.tn

2024-2025

Table des matières

Introduction

De nos jours, le domaine de la santé bénéficie de manière significative des avancées de la science des données et du traitement du langage naturel (NLP).

PubMed, une plateforme incontournable pour la recherche biomédicale, offre un large nombre d'articles scientifiques sur une variété de sujets médicaux. L'analyse de ces articles permet non seulement de mieux comprendre les tendances dans la recherche médicale, mais aussi d'extraire des informations cruciales pour la progression des connaissances scientifiques et l'amélioration des soins de santé.

Dans ce projet, l'objectif est de mettre en place un processus de collecte, d'annotation et d'analyse d'articles scientifiques provenant du site **PubMed**.

La première étape consiste à réaliser un scraping des articles scientifiques afin de constituer un corpus textuel riche, qui servira de base pour les phases suivantes du projet. Cette approche repose sur l'extraction de métadonnées pertinentes telles que les titres, les résumés, les auteurs, les mots-clés, et bien d'autres, afin de bâtir un jeu de données exploitable.

Une fois le corpus constitué, il sera nécessaire d'effectuer une phase d'annotation, où chaque article sera classé selon les domaines médicaux spécifiques qu'il couvre, permettant ainsi de structurer les informations.

Enfin, à partir de ce corpus annoté, des techniques d'apprentissage automatique.. seront appliquées pour analyser les articles en fonction de leurs thèmes et de leurs similarités.

Ce projet vise à offrir une approche systématique et reproductible pour la collecte et l'analyse d'articles médicaux, tout en contribuant à l'amélioration des outils d'exploration et de recherche dans le domaine biomédical. Les résultats obtenus permettront non seulement de mieux organiser les informations disponibles sur **PubMed**, mais aussi d'identifier des tendances de recherche et d'assister les chercheurs dans la découverte de nouveaux articles et domaines de recherche.

Voici le lien vers le dépôt GitHub de ce projet :
https://github.com/nadyby/PubMedNLP.

Chapitre 1

Corpus : Implémentation du Scraping

Introduction :

Dans le domaine du traitement automatique du langage naturel (NLP), la construction d'un corpus pertinent et de qualité constitue une étape fondamentale et cruciale.

Pour les applications médicales, disposer d'un ensemble de données fiable est indispensable afin de capturer la richesse et la complexité des informations scientifiques, ainsi que pour assurer la pertinence des résultats des modèles.

PubMed, avec sa vaste collection d'articles médicaux et de recherches spécialisées, représente une ressource incontournable pour la constitution de tels corpus. Ce site, qui regroupe des millions d'articles scientifiques, offre un accès à une mine d'informations couvrant divers domaines de la médecine.

Ce chapitre se concentre sur le processus d'élaboration du corpus à partir de PubMed, en détaillant les outils et méthodes utilisés pour le scraping des articles. Nous aborderons également les défis techniques rencontrés, ainsi que les choix réalisés pour garantir une extraction structurée et exploitable des données scientifiques.

Cette phase est documentée dans le dépôt GitHub du projet, accessible via **ce lien**.

Méthodologie de scraping :

Le scraping de données, également appelé extraction web, est une méthode essentielle pour collecter des informations disponibles en ligne de manière automatisée. Dans le contexte de ce projet, le scraping a permis de constituer un corpus structuré à partir du site de **PubMed**.

Informations extraites :

Dans cette section, nous décrivons les types de données recueillies lors du processus de scraping depuis le site de PubMed. Chaque type d'information joue un rôle crucial pour constituer un corpus structuré, utile pour les analyses ultérieures.

Voici les informations principales extraites :

— **Titre de l'article (Article Title)**
— **Date de publication**
— **La source** : Le nom du journal ou de la revue dans lequel l'article a été publié.
— **Résumé (Abstract)** : facilite la compréhension rapide du contenu d'un article.
— **Mots-clés (Keywords)** : Une liste de termes ou expressions clés associés au contenu de l'article.
— **DOI (Digital Object Identifier)** : Un identifiant unique pour chaque article publié, utilisé pour son référencement et son accès en ligne.
— **PMID (PubMed Identifier)** : Un numéro d'identification unique attribué par PubMed à chaque article référencé.
— **Auteurs (Authors)**

Méthodes et outils utilisés pour le scraping

Dans cette section, nous détaillons les outils et méthodes adoptés pour extraire les données depuis **PubMed**, en expliquant les raisons des choix effectués et les étapes suivies.

Deux approches principales ont été explorées pour la phase de scraping :

— Scrapy
— BeautifulSoup (BS4)

Après avoir évalué les deux méthodes, nous avons décidé d'utiliser **BeautifulSoup (BS4)** pour extraire les données finales. Ce choix repose sur les critères suivants :

- La structure relativement simple des pages HTML de PubMed, facilement navigable avec BS4.
- L'objectif du projet nécessitant une extraction ciblée et précise plutôt qu'un scraping massif.
- Une intégration rapide dans un environnement de développement interactif tel que Jupyter Notebook.

Exploration de la structure HTML de PubMed

Comprendre la structure des balises HTML pour identifier les éléments contenant les informations d'intérêt.

Accès à la page web

L'accès initial au site **PubMed** s'est concentré sur la section dédiée aux "Trending Articles" (articles tendances), qui regroupe des publications récentes et pertinentes dans le domaine médical.

Un filtre spécifique a été appliqué pour restreindre les résultats aux **articles publiés** entre début 2024 et aujourd'hui. Cette période a été choisie afin de garantir que le corpus reflète les avancées médicales et scientifiques les plus récentes.

Après l'extraction et la consolidation des données, nous avons stocké les résultats dans un fichier texte structuré pour faciliter les étapes ultérieures d'analyse.

-> Au total, **949 articles** ont été extraits.

Conclusion :

L'extraction d'un corpus structuré à partir de **PubMed** constitue une étape fondamentale dans ce projet. Grâce à l'utilisation de **BeautifulSoup**, nous avons pu collecter efficacement un ensemble d'articles scientifiques contenant des informations clés telles que les titres, les résumés, les auteurs et les identifiants uniques (PMID, DOI).

Cette phase a permis de rassembler un corpus de **949 articles** récents, garantissant une base de données pertinente et actualisée pour les analyses ultérieures. Les défis rencontrés incluent la gestion de la structure dynamique des pages web et la nécessité de filtrer les articles les plus significatifs en fonction de critères bien définis.

L'étape suivante consistera à annoter ce corpus afin d'identifier automatiquement les éléments médicaux pertinents tels que les maladies et les types cellulaires mentionnés dans les articles.

Chapitre 2

Annotation du Corpus

Introduction

L'analyse et l'exploitation des articles scientifiques nécessitent une structuration avancée des informations qu'ils contiennent. Après l'étape de collecte et de structuration des données via le scraping, il est essentiel d'annoter ces articles afin d'identifier automatiquement les entités médicales clés.

L'annotation est une phase déterminante dans le traitement automatique du langage naturel (NLP), car elle permet de structurer les informations contenues dans les articles en identifiant les concepts médicaux pertinents, tels que les maladies (**DISEASE**) et les types cellulaires (**CELL_TYPE**). Cette structuration facilite non seulement l'analyse des tendances scientifiques, mais aussi l'application de techniques de classification et de clustering pour regrouper les articles selon leurs thématiques principales.

Dans ce projet, cette annotation sera réalisée à l'aide de modèles de NLP spécialisés, notamment **BioBERT**, qui sont conçus pour l'extraction d'informations biomédicales. BioBERT permet d'identifier avec précision les entités médicales mentionnées dans les articles, en tirant parti de modèles de langage pré-entraînés sur des corpus biomédicaux.

L'objectif final de cette phase est d'obtenir un corpus annoté exploitable pour les étapes suivantes du projet, notamment l'analyse thématique et la classification automatique des articles scientifiques.

Cette phase est documentée dans le dépôt GitHub du projet, accessible via **ce lien**.

2.1 Méthodologie d'Annotation

2.1.1 Approche de l'annotation automatique

L'annotation automatique des articles scientifiques constitue une étape clé pour structurer et organiser les informations médicales contenues dans le corpus.

Contrairement à l'annotation manuelle, qui est coûteuse en temps et en ressources humaines, l'annotation automatique permet un traitement efficace et reproductible sur un grand volume de documents.

L'objectif principal est d'identifier deux types d'entités médicales essentielles :
— **DISEASE** : Maladie principale étudiée dans l'article.
— **CELL_TYPE** : Types cellulaires mentionnés dans l'étude.

Pour cela, nous avons opté pour une approche basée sur des modèles de **traitement du langage naturel (NLP)** entraînés sur des corpus biomédicaux, permettant d'extraire ces informations de manière fiable.

2.1.2 Présentation de BioBERT

BioBERT (*Bidirectional Encoder Representations from Transformers for Biomedical Text Mining*) est un modèle de NLP pré-entraîné sur des corpus médicaux, notamment **PubMed** et **PMC**. Il a été conçu pour améliorer la compréhension du langage biomédical en s'appuyant sur l'architecture de **BERT** (*Bidirectional Encoder Representations from Transformers*).

Les principaux avantages de BioBERT sont :
— Une meilleure compréhension des textes biomédicaux grâce à un entraînement sur des millions d'articles médicaux.
— Une capacité à reconnaître des entités complexes telles que les noms de maladies, de protéines ou de cellules.
— Une compatibilité avec les tâches d'extraction d'information et d'annotation sémantique.

BioBERT surpasse les modèles NLP classiques dans les tâches biomédicales, notamment grâce à sa capacité à exploiter un contexte bidirectionnel, ce qui le rend particulièrement adapté à l'annotation des articles scientifiques de notre corpus.

2.1.3 Processus d'annotation

L'annotation du corpus suit plusieurs étapes essentielles afin de garantir une extraction précise et fiable des entités médicales.

L'ensemble de ces étapes permet de générer un corpus annoté structuré, facilitant l'analyse ultérieure des articles scientifiques.

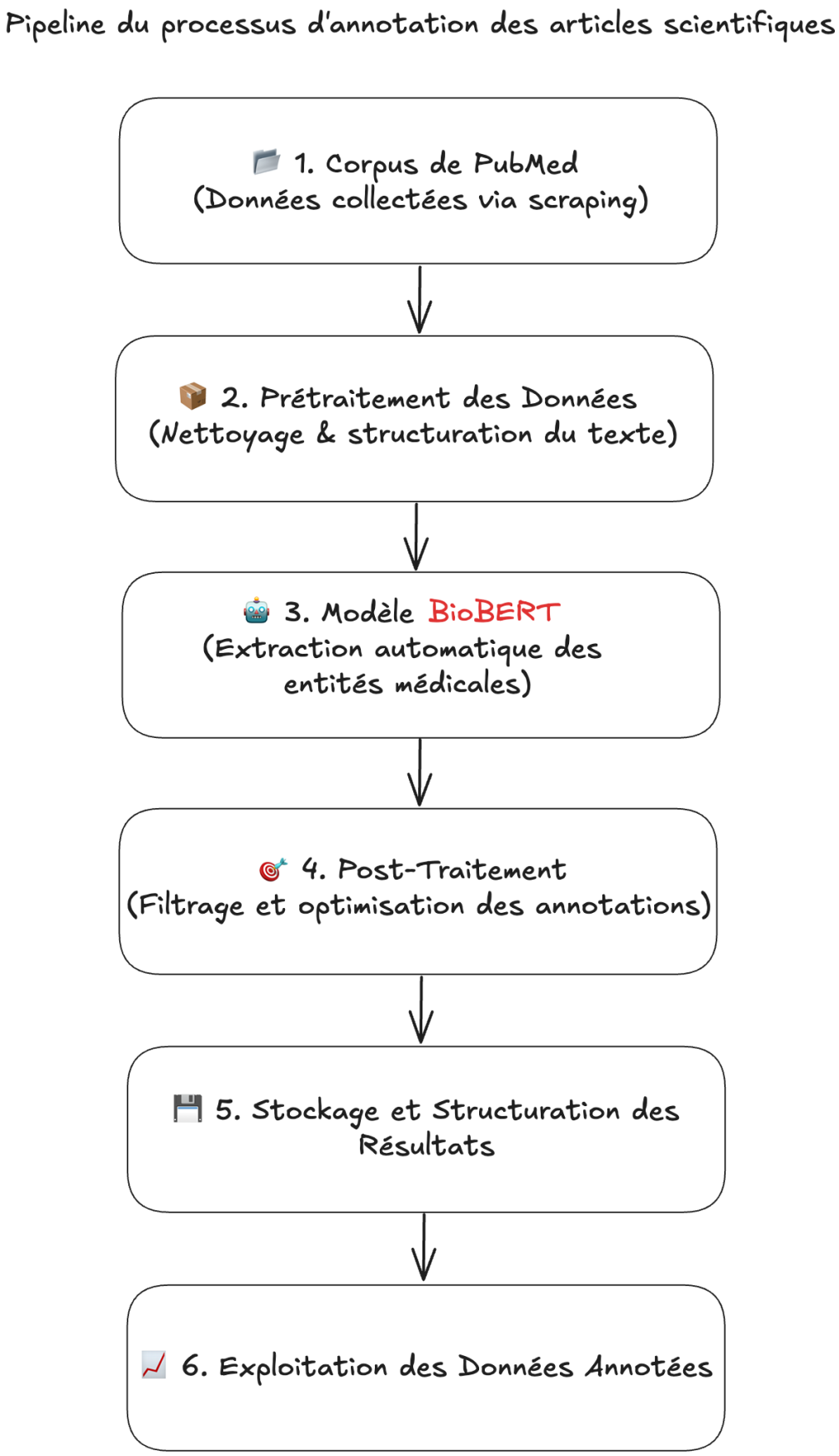

FIGURE 2.1 – Pipeline général du processus d'annotation

1. **Prétraitement des articles** : Cette étape consiste à nettoyer et normaliser les textes extraits de PubMed (suppression des caractères spéciaux, normalisation des majuscules/minuscules, etc.).
2. **Tokenization et segmentation du texte** : Chaque article est segmenté en unités linguistiques traitables par BioBERT.
3. **Application du modèle BioBERT** : BioBERT est utilisé pour identifier les entités **DISEASE** et **CELL_TYPE** à partir des titres et des résumés des articles.
4. **Post-traitement des annotations**
5. **Stockage et structuration des résultats** : Les entités annotées sont enregistrées dans un format exploitable (JSON) pour les étapes suivantes du projet.

2.2 Implémentation Technique

2.2.1 Préparation des Données

Après la phase de scraping, les articles scientifiques extraits sont stockés sous un format JSON. Chaque article contient les métadonnées mentionnées au niveau du chapitre 1. Ces données sont ensuite chargées et prétraitées avant l'annotation.

2.2.2 Application de BioBERT

L'annotation des entités médicales repose sur le modèle **BioBERT**, spécialisé dans l'analyse biomédicale. Son implémentation suit les étapes suivantes :

1. **Chargement du modèle** : Initialisation de BioBERT et de son tokenizer via la bibliothèque `transformers`.
2. **Tokenization et préparation du texte** : Segmentation des articles en tokens exploitables par le modèle.
3. **Extraction des entités** : Application de BioBERT pour détecter les entités médicales selon deux catégories :
   — **DISEASE** : Identification des maladies principales étudiées.
   — **CELL_TYPE** : Détection des types cellulaires mentionnés.
4. **Filtrage des résultats** : Suppression des annotations à faible score de confiance.

L'exécution du modèle est réalisée via un pipeline de type `question-answering`, qui applique une série de requêtes spécifiques aux articles.

2.2.3 Post-traitement des annotations

Une fois les entités extraites, plusieurs étapes de post-traitement sont appliquées :

— **Nettoyage et normalisation** : Suppression des caractères spéciaux et des redondances.
— **Filtrage des faux positifs** : Vérification de la pertinence des entités détectées.
— **Consolidation des entités** : Regroupement des termes similaires (*ex : "lung cancer" et "cancer du poumon"*).

2.2.4 Stockage et Structuration des Résultats

Les annotations finales sont enregistrées sous un format JSON structuré dans un fichier intitulé `"annotated_articles.json"` facilitant leur exploitation ultérieure. Un exemple de sortie est présenté ci-dessous :

Listing 2.1 – Exemple d'annotation JSON

```
{
  "39743589": {
  "pmid": "39743589",
  "title" : "Aspartate signalling drives lung
    metastasis via alternative translation.",
 "annotations": {
    "DISEASE": ["breast cancer"],
    "CELL_TYPE": ["immune cells"]
  }
  }
}
```

2.3 Conclusion

Ce chapitre a détaillé le processus d'annotation automatique des articles scientifiques extraits de **PubMed**. En exploitant le modèle **BioBERT**, nous avons pu identifier deux types d'entités médicales essentielles : les **maladies (DISEASE)** et les **types cellulaires (CELL_TYPE)**.

L'approche adoptée repose sur l'application de questions ciblées aux articles, permettant ainsi d'extraire des informations pertinentes de manière efficace.

Grâce à cette méthodologie, nous avons obtenu un corpus structuré et annoté, qui servira de base aux prochaines analyses. Le chapitre suivant se concentrera sur l'exploitation de ces annotations pour regrouper et classifier les articles selon leurs thématiques médicales principales.

Chapitre 3

Phase d'Apprentissage

Introduction

Après avoir construit un corpus de 949 articles scientifiques à partir de **PubMed** (Chapitre 1) et annoté automatiquement les entités médicales **DISEASE** et **CELL_TYPE** avec BioBERT (Chapitre 2), la phase d'apprentissage vise à exploiter ces annotations pour identifier des thématiques médicales sous-jacentes dans le corpus.

Cette étape s'inscrit dans une démarche exploratoire en traitement automatique du langage naturel (NLP), cherchant à regrouper les articles selon leurs sujets principaux sans supervision préalable.

Deux approches complémentaires ont été adoptées : le **clustering** avec l'algorithme K-means pour former des groupes distincts, et le **topic modeling** avec Latent Dirichlet Allocation (LDA) pour extraire des thèmes probabilistes.

Ce chapitre détaille la méthodologie, les résultats obtenus, et leur analyse, en se concentrant principalement sur les annotations **DISEASE** pour capturer les maladies étudiées dans les articles.

Cette phase est documentée dans le dépôt GitHub du projet, accessible via **ce lien**.

3.1 Préparation des Données pour l'Apprentissage

Avant d'appliquer les techniques d'apprentissage, une préparation minutieuse des données a été nécessaire pour transformer le corpus brut en un format exploitable. Les deux fichiers JSON issus des phases précédentes `articles.json` contenant les métadonnées (titre, résumé, auteurs, etc.) et `annotated_articles.json` avec les annotations BioBERT ont été fusionnés en un seul tableau structuré.

Cette fusion a permis d'associer chaque article (identifié par son PMID) à ses annotations correspondantes, créant un DataFrame avec les colonnes `PMID`, `Title`, `Abstract`, `DISEASE`, et `CELL_TYPE`.

Un filtrage a ensuite été appliqué pour exclure les articles annotés avec `"unspecified disease"` dans la colonne `DISEASE`, car ces cas manquent de spécificité et risquent de brouiller les résultats.

Sur les 949 articles initiaux, cette étape a réduit le corpus à 586 articles pertinents. Enfin, pour se concentrer sur les thématiques liées aux maladies objectif principal de cette phase , la colonne `DISEASE` a été transformée en une chaîne textuelle unique par article (ex. "Hepatocellular carcinoma metabolic-dysfunction-associated steatohepatitis") via une concaténation des termes annotés.

3.2 Clustering avec K-means

3.2.1 Méthodologie

Le clustering avec l'algorithme K-means a été choisi pour regrouper les 586 articles annotés en clusters thématiques basés sur la feature `Disease`. Cette méthode non supervisée vise à partitionner les données en k groupes en minimisant la distance intra-cluster, ici calculée à partir d'une représentation vectorielle des annotations.

Pour transformer les chaînes textuelles de `Disease` en données numériques, une vectorisation par **TF-IDF** (*Term Frequency-Inverse Document Frequency*) a été appliquée. Les paramètres retenus incluent `ngram_range=(1, 3)` pour capturer des expressions composées (ex. "hepatocellular carcinoma") et `min_df=3` pour exclure les termes rares, aboutissant à une matrice de 586 articles et 213 termes uniques.

L'algorithme K-means a été exécuté avec un nombre initial de clusters $k = 5$, ajusté après optimisation. Pour déterminer le k optimal, une évaluation quantitative a été réalisée en testant k de 3 à 10, en calculant l'**inertie** (somme des distances au carré aux centroides) et le **score de silhouette** (mesure de cohérence des clusters).

Les clusters obtenus ont été analysés via leurs termes dominants (issus des centroides) et visualisés en deux dimensions avec **t-SNE** (*t-distributed Stochastic Neighbor Embedding*).

```
Termes dominants par cluster (nettoyés) :
Cluster 0: disorders, injury, syndrome, osteoarthritis, disorder, fibrosis, acute, ferroptosis, cancers, failure
Cluster 1: disease, cardiovascular disease, cardiovascular, kidney disease, kidney, liver disease, pulmonary disease, pulmon
ary, liver, alzheimer
Cluster 2: cancer, breast cancer, breast, lung, lung cancer, cell, gastric cancer, gastric, prostate, prostate cancer
Cluster 3: carcinoma, hepatocellular, hepatocellular carcinoma, cell carcinoma, cell, squamous, squamous cell, squamous cell
carcinoma, renal, esophageal
Cluster 4: diabetes, apoptosis, atherosclerosis, arthritis, cardiovascular, cell, fibroblasts, ferroptosis ferroptosis
```

FIGURE 3.1 – Termes dominants par cluster (k=5)

3.2.2 Résultats et Analyse

L'application de K-means avec $k = 5$ a produit cinq clusters distincts, chacun reflétant une thématique médicale identifiable. Les termes dominants et la liste des `Disease` par cluster sont résumés ci-dessous :

- **Cluster 0 (268 articles)** : Regroupe des troubles variés, incluant des maladies neurologiques et des conditions diverses.
- **Cluster 1 (86 articles)** : Thématique des maladies systémiques affectant les organes.
- **Cluster 2 (161 articles)** : Regroupe les cancers localisés par organe.
- **Cluster 3 (35 articles)** : Focalisé sur les carcinomes.
- **Cluster 4 (36 articles)** : Spécifique au diabète et ses variantes.

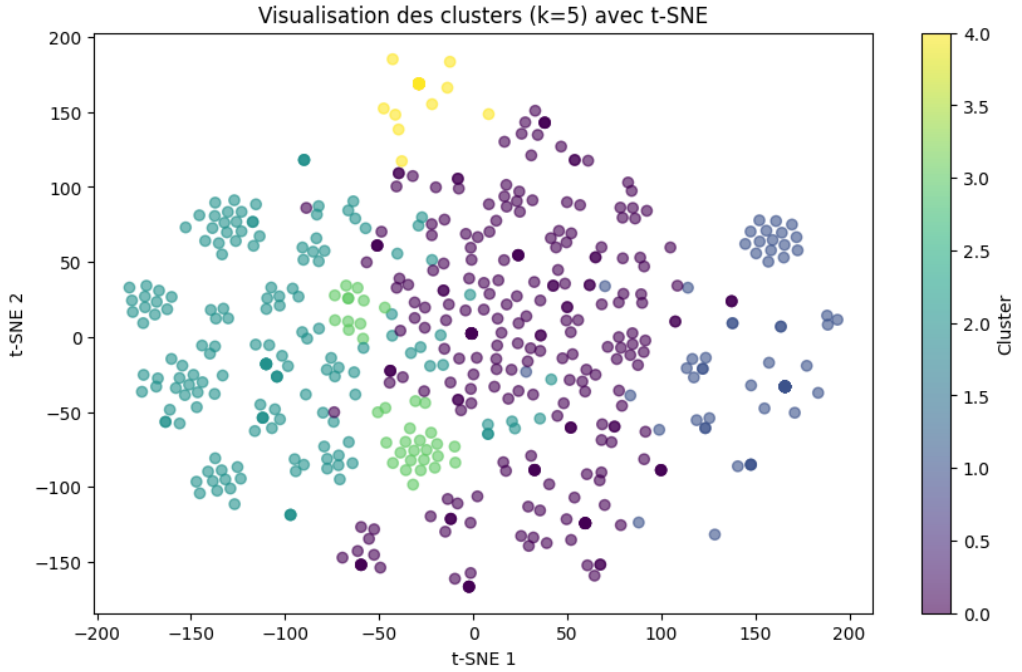

FIGURE 3.2 – Visualisation des clusters (k=5) avec t-SNE

Le score de silhouette moyen pour $k = 5$ est de 0.159, indiquant une cohérence modérée, typique des données textuelles éparses. Une analyse comparative avec $k = 6$ (score de silhouette 0.162) a révélé une fragmentation excessive, notamment un Cluster 0 mélangeant "hepatocellular carcinoma" et "multiple sclerosis", justifiant le choix de $k = 5$ pour sa simplicité et sa lisibilité.

```
Top 5 Disease les plus fréquents du Cluster 2 (161 articles):
  - gastric cancer: 5
  - breast cancer: 4
  - Gastric cancer: 4
  - lung cancer: 4
  - colorectal cancer: 4
```

FIGURE 3.3 – Top 5 des 'Disease' présents dans le cluster 2

3.2.3 Discussion

Les résultats du clustering K-means avec $k = 5$ démontrent une capacité à identifier des thématiques médicales pertinentes dans le corpus.

Les Clusters 1, 2, 3 et 4 offrent des regroupements cohérents ; maladies systémiques, cancers localisés, carcinomes, et diabète, alignés avec les objectifs du projet.

Cependant, le Cluster 0, bien qu'incluant des maladies notables comme "multiple sclerosis" et "glioblastoma", reste un groupe résiduel hétérogène, suggérant une diversité non capturée par les autres clusters.

La présence de termes comme "in" et "or" dans le Cluster 4 (avant nettoyage) indique un léger bruit dans la vectorisation, corrigé par un post-traitement pour l'affichage.

3.3 Topic Modeling avec LDA

3.3.1 Méthodologie

Pour compléter l'analyse par clustering, une approche de **topic modeling** a été appliquée en utilisant l'algorithme **Latent Dirichlet Allocation (LDA)**.

Contrairement à K-means, qui assigne chaque article à un seul cluster, LDA permet une répartition probabiliste des articles sur plusieurs thématiques, offrant une vue plus nuancée des sujets présents dans le corpus.

La matrice `disease_tfidf` (586 articles, 213 termes), déjà utilisée pour le clustering, a servi d'entrée pour LDA, conservant ainsi une cohérence dans la représentation des données `Disease`.

Le nombre de topics a été fixé à 5, en alignement avec le choix de $k = 5$ pour K-means. Les paramètres de LDA incluent `n_components=5` pour définir les 5 topics et `random_state=42` pour assurer la reproductibilité. Après ajustement, les mots clés les plus probables par topic ont été extraits à partir des composantes

du modèle, et chaque article s'est vu attribuer un topic dominant basé sur sa probabilité maximale.

3.3.2 Résultats et Analyse

L'application de LDA a généré cinq topics distincts, chacun défini par une liste de mots clés reflétant des thématiques médicales sous-jacentes dans le corpus. Les résultats sont présentés ci-dessous :

- **Topic 0** : Thème varié mêlant cancers (colorectal) et maladies systémiques (cardiovasculaire).
- **Topic 1** : Focalisé sur les carcinomes (ex. "hepatocellular carcinoma") et certains cancers (ex. "breast cancer").
- **Topic 2** : Dominé par les cancers pulmonaires (ex. "lung cancer").
- **Topic 3** : Centré sur le diabète et des conditions associées (ex. "atherosclerosis").
- **Topic 4** : Mélange de cancers digestifs (ex. "gastric cancer") et maladies systémiques.

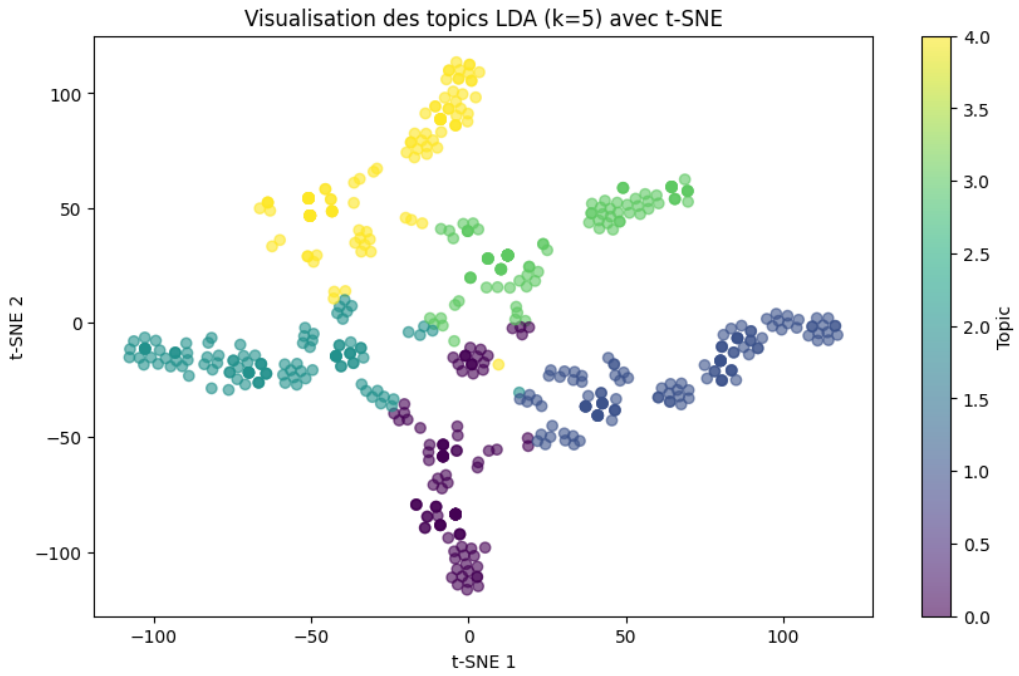

FIGURE 3.4 – Visualisation des Topics LDA (k=5) avec t-SNE

Une matrice de correspondance entre clusters (K-means) et topics (LDA) a été calculée pour évaluer leur alignement (Tableau 3.1). Les Clusters 3 (carcinomas) et 4 (diabète) montrent une forte concentration dans les Topics 1 (20/35 articles) et 3 (33/36 articles), respectivement, tandis que le Cluster 2 (cancers localisés) se disperse entre les Topics 0 (40 articles), 1 (46 articles), et 2 (47 articles). Le Cluster 0 (troubles variés) reste diffus (50 à 64 articles par topic), confirmant son hétérogénéité.

Topic	0	1	2	3	4
Cluster 0	50	54	64	63	37
Cluster 1	16	7	2	8	53
Cluster 2	40	46	47	7	21
Cluster 3	1	20	0	0	14
Cluster 4	2	1	0	33	0

TABLE 3.1 – Matrice de correspondance entre clusters (K-means) et topics (LDA)

3.3.3 Discussion

LDA enrichit l'analyse par clustering en offrant une perspective probabiliste sur les thématiques médicales. Le Topic 3 ("diabetes") et le Topic 1 ("carcinoma") reproduisent fidèlement les Clusters 4 et 3, confirmant leur spécificité.

Le Topic 2 ("lung cancer") affine le Cluster 2 en isolant les cancers pulmonaires, une distinction perdue dans K-means.

Cependant, les Topics 0 et 4, plus hétérogènes, mélangent des thématiques (ex. "cardiovascular" et "colorectal" dans Topic 0), reflétant une moindre séparation par rapport aux clusters.

La dispersion du Cluster 2 entre plusieurs topics (0, 1, 2) indique que les cancers localisés partagent des termes communs avec d'autres thématiques (ex. "cancer" dans Topic 1), tandis que l'hétérogénéité du Cluster 0 persiste en LDA.

Cette complémentarité entre K-means (groupes fixes) et LDA (répartition flexible) met en évidence des nuances dans le corpus, comme la prédominance des cancers et du diabète, tout en soulignant les limites d'une analyse basée uniquement sur `Disease` pour capturer des relations complexes.

Conclusion

Ce projet a permis de construire et d'exploiter un corpus de 949 articles scientifiques extraits de **PubMed** pour analyser les thématiques médicales actuelles. La phase de scraping (Chapitre 1) a utilisé **BeautifulSoup** pour collecter des métadonnées riches, suivie par l'annotation automatique avec **BioBERT** (Chapitre 2) pour identifier les entités `DISEASE` et `CELL_TYPE`, structurant ainsi les données en un corpus annoté. La phase d'apprentissage (Chapitre 3) a exploité ces annotations via **K-means** (k=5) et **LDA**, révélant cinq thématiques principales : troubles variés, maladies systémiques, cancers localisés, carcinomes, et diabète, avec une prédominance des cancers et du diabète dans les recherches récentes (2024-2025). Ces résultats démontrent l'efficacité des approches non supervisées pour explorer un corpus médical.

Voici le lien vers le dépôt GitHub de ce projet :
https://github.com/nadyby/PubMedNLP.