Identification des tendances médicales à l'aide de l'analyse de données

Utiliser des techniques d'analyse de données pour identifier les tendances actuelles dans le domaine médical. Cela pourrait impliquer l'analyse de grandes quantités de données provenant de diverses sources, comme les publications de recherche médicale, les données de santé publique, les médias sociaux, etc. Les résultats pourraient aider à anticiper les futures évolutions du domaine médical, à identifier les domaines de recherche prometteurs et à informer les décisions en matière de politique de santé.

Rédigé par :

AL MASSATI Sanae
BEN MUSTAPHA Sarra
ITIR Avave

Enseignant :

OUBENALI Naima

M2 DATA SCIENCE SANTE 2023 – 2024
UFR-3S Faculté d'Ingénierie et de Management de la Santé
42 rue Ambroise Paré, 59129 Loos

I- Introduction

Ce projet vise à exploiter les techniques d'analyse de données pour extraire des informations pertinentes à partir de grandes bases de données médicales, aidant ainsi les professionnels de la santé à rester à jour avec les dernières tendances et découvertes. En se concentrant sur les publications concernant le cancer, l'analyse fournit une vue d'ensemble des tendances actuelles, anticipe les futures évolutions et identifie les domaines de recherche prometteurs.
L'objectif principal de ce projet est d'identifier les tendances actuelles dans les publications médicales sur le cancer. Pour ce faire, nous avons ciblé les objectifs spécifiques suivants :

- Identification des thèmes récurrents
- Analyse temporelle
- Cartographie des collaborations
- Détection des nouvelles approches thérapeutiques

I- Méthodologie

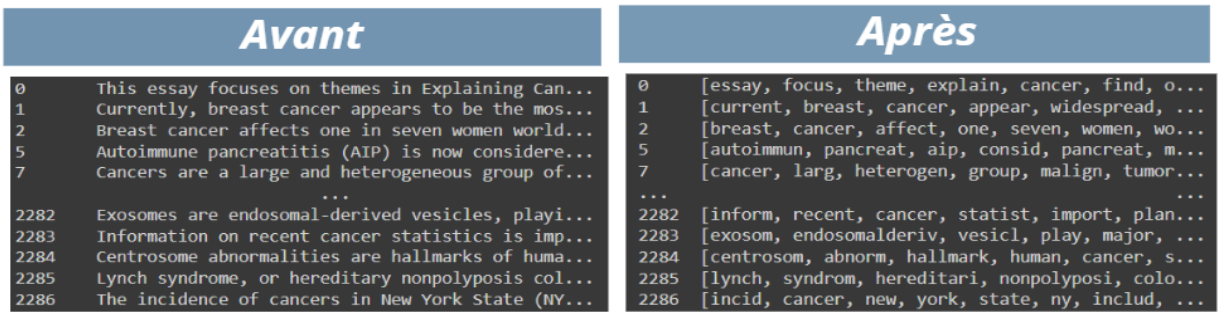
1- Fouille des données et data management

Pour réaliser cette analyse, nous avons utilisé **PubMed**, une base de données de référence dans le domaine biomédical. La méthodologie de collecte des données est ainsi :

- Collecte des données : 150 premières pages de recherche pour le terme "cancer" en ne gardant que les articles en anglais
- Extraction des informations : extraction des titres, des auteurs, des dates de publication et des résumés des articles.
- Filtrage des articles pour ne garder que ceux à partir de 2019 → 2287 articles

2- Traitement de données

Nous avons commencé par la normalisation (conversion en minuscules, suppression de la ponctuation, suppression des emojis, suppression des liens URL et HTML). Ensuite, vient la tokenisation, suivie de la suppression des mots vides (stop words). La troncature (stemming) est également appliquée, bien que la lemmatisation soit une technique plus avancée que la troncature. En utilisant une combinaison de ces techniques, nous avons pu améliorer considérablement la qualité et l'efficacité des processus de traitement du langage naturel. Quand on compare la colonne "abstract" de départ avec celle après l'application de ce processus, la différence est claire (voir figure 1).



	Avant
0	This essay focuses on themes in Explaining Can...
1	Currently, breast cancer appears to be the mos...
2	Breast cancer affects one in seven women world...
5	Autoimmune pancreatitis (AIP) is now considere...
7	Cancers are a large and heterogeneous group of...
...	...
2282	Exosomes are endosomal-derived vesicles, playi...
2283	Information on recent cancer statistics is imp...
2284	Centrosome abnormalities are hallmarks of huma...
2285	Lynch syndrome, or hereditary nonpolyposis col...
2286	The incidence of cancers in New York State (NY...

	Après
0	[essay, focus, theme, explain, cancer, find, o...
1	[current, breast, cancer, appear, widespread, ...
2	[breast, cancer, affect, one, seven, women, wo...
5	[autoimmun, pancreat, aip, consid, pancreat, m...
7	[cancer, larg, heterogen, group, malign, tumor...
...	...
2282	[inform, recent, cancer, statist, import, plan...
2283	[exosom, endosomalderiv, vesicl, play, major, ...
2284	[centrosom, abnorm, hallmark, human, cancer, s...
2285	[lynch, syndrom, hereditari, nonpolyposi, colo...
2286	[incid, cancer, new, york, state, ny, includ, ...

Figure 1 : Exemple des résumés avant et après le traitement

3- Exploration des données

Pour cette partie nous avons généré les graphiques ci-dessous (voir figure 2) permettant une vue d'ensemble de la distribution temporelle des résumés (abstract des articles) et des termes les plus courants dans ces résumés. Cela permet de comprendre les volumes de publication au fil des ans et les sujets les plus abordés dans votre corpus de données, en particulier autour de la thématique du cancer.

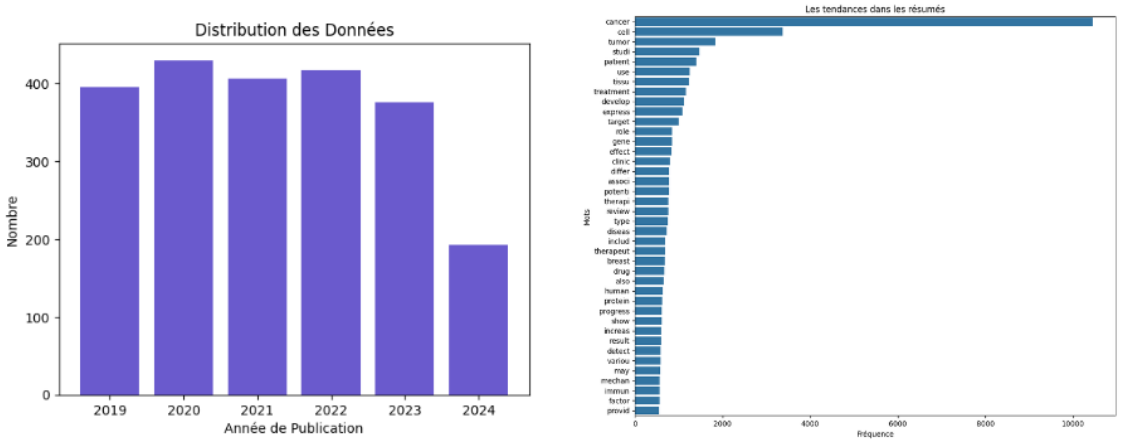


Figure 2 : Distributions des données

4- Modélisation

a- TF-IDF : term frequency-inverse document frequency

Nous avons utilisé La méthode TF-IDF pour évaluer l'importance des termes dans les plus tendances pour chaque année (de 2019 à 2024).

Ci-dessous, l'exemple de 2024 (voir figures 3). Ces 2 graphiques nous montrent l'évolution des termes importants dans les résumés scientifiques au fil des ans, en mettant en évidence les tendances et les priorités de recherche.

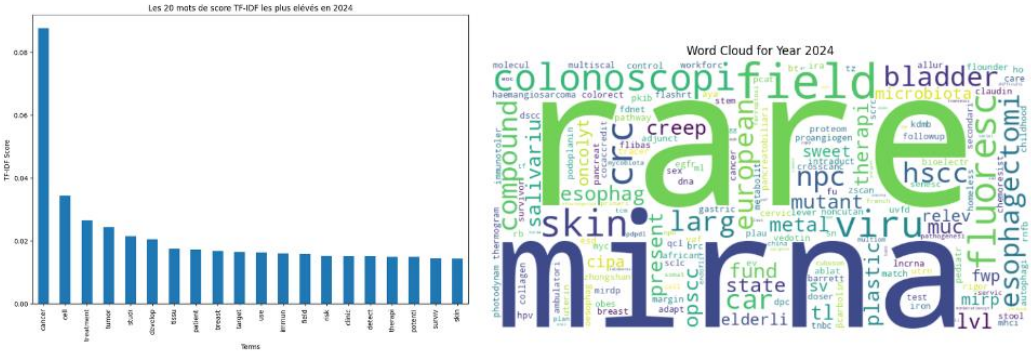


Figure 3 : Exemple de l'année 2024

b- Utilisation de LDA et Validation du Modèle

La méthode LDA (Latent Dirichlet Allocation) est une technique de modélisation thématique qui permet d'identifier les sujets principaux d'un ensemble de documents en analysant les co-occurrences de mots. Pour notre projet, nous avons utilisé le package `gensim` en Python pour appliquer LDA sur les résumés des articles.

Après avoir appliqué LDA, nous avons validé notre modèle en calculant le score de cohérence, qui s'élève à 0.4, indiquant que notre modèle est raisonnablement valide. Cela signifie que les sujets identifiés sont sémantiquement cohérents.

index	0	1	2	3
Topic 1	0.049,cancer	0.017,role	0.013,express	0.013,protein
Topic 2	0.037,cancer	0.031,cell	0.015,tumor	0.012,therapi
Topic 3	0.029,oral	0.027,gastric	0.016,thyroid	0.015,lesion
Topic 4	0.046,cancer	0.022,cell	0.012,target	0.009,develop
Topic 5	0.074,cancer	0.02,age	0.015,year	0.011,incid
Topic 6	0.046,cancer	0.017,tissu	0.016,gene	0.013,cell
Topic 7	0.041,cancer	0.028,express	0.018,gene	0.017,tumor
Topic 8	0.051,cancer	0.026,infect	0.022,skin	0.015,viru
Topic 9	0.05,cancer	0.015,detect	0.013,use	0.013,tissu
Topic 10	0.057,cancer	0.022,patient	0.015,treatment	0.012,diseas

Figure 4 : Exemple des topics

Pour la visualisation nous avons généré Le graphe de la distance intertopics qui montre les relations entre les sujets identifiés par LDA. Par exemple, les sujets 3, 4, 6, et 8 sont proches, suggérant des similitudes thématiques, tandis que les sujets 1 et 9 sont plus distincts.

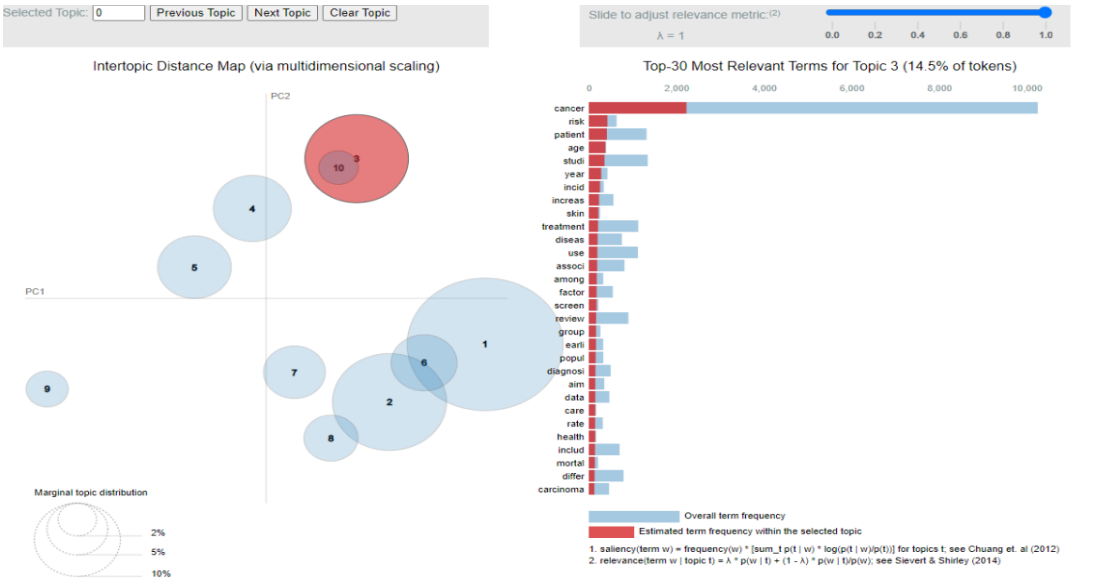


Figure 5 : Topics générés en fonction des occurrences des mots

Perspectives

Les perspectives de ce projet incluent l'amélioration des modèles d'analyse textuelle, comme l'affinage des modèles LDA et l'intégration de techniques d'apprentissage automatique avancées. Il est également bénéfique d'élargir l'analyse à d'autres domaines médicaux et de diversifier les sources de données pour maximiser l'impact et la pertinence des résultats.

Conclusion

L'analyse des publications sur le cancer a révélé des tendances et thèmes significatifs grâce à des techniques avancées de traitement du langage naturel et des méthodes d'analyse comme TF-IDF et LDA. Ces résultats offrent une compréhension approfondie des tendances émergentes, pouvant orienter les futures recherches, informer les décideurs et favoriser les collaborations entre chercheurs.