République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique
جامعة عباس لغرور- خنشلة
UNIVERSITE ABBES LAGHROUR
KHENCHELA

Faculté des Science et de la Technologie

Département de Mathématiques et d'Informatique

MEMOIRE
Présenté en vue de l'obtention du Diplôme de
Master en Informatique
Spécialité : Génie Logiciel et Système Distribué

Thème :

Une méthode basée sur les règles d'association pour la mise en évidence de relations entre entités biologiques à partir d'un corpus étiqueté en anglais

Réalisé par :

Saadaoui Amina

Keziz Rachid

Encadré par :

Dr. MOHAMED MAHDI MALIK

Promotion : 2021 / 2022

Remercîment

Nous tenons avant tout à remercier Allah de nous avoir donné la volonté et la détermination pour accomplir ce modeste et humble travail.

Que Dieu soit témoin de nos grands remerciements pour Mr Malik.Mohamed Mahdi qui a sacrifié de son temps précieux pour nous fournir l'aide qu'il nous faut pour mener à bien notre travail. Ainsi, tous les enseignants qui ont contribué par leur collaboration, disponibilité et sympathie à notre formation.

Notre considération pour toute personne ayant contribué de près ou de loin à la réalisation de ce travail.

Nos remerciements les plus sincères et les plus profonds pour nos parents en reconnaissance de leurs sacrifices, aides, soutien et encouragement afin de nous assurer cette formation dans les meilleures conditions.

Enfin que les membres du jury trouvent ici l'expression de notre profonde gratitude pour l'honneur qu'ils nous font en acceptant de juger notre humble travail étant une œuvre humaine qui n'est pas un modèle unique et parfait.

Je dédie cet humble travail à :

Mes très chères parents, Mon Père رحمه الله *et ma Mère qui je souhaite une longue vie et une bonne santé ;*

Que par leurs présences à mes côtés ont rendu chaque moment de ma vie Un merveilleux passage dans le temps :

Mes adorables frères et sœurs

Ainsi je tiens à remercier infiniment tous ceux qui ont contribué de prêt ou de Loin à la réalisation de ce travail :

Ainsi mes amis et toute ma famille.

Amina

C'est avec grand plaisir que je dédie ce modeste travail :

A ma mère.

A mon père

A ma sœur et mon frère

A ma femme et ma fille

Et à mes amis.

Rachid

Résumé

Les masses de données textuelles aujourd'hui disponibles engendrent un problème difficile pour les traiter. Dans ce cadre, des méthodes de Fouille de Texte (Text Mining) sont nécessaires pour extraire les connaissances à partir des textes.

Notre travail consiste à étudier l'une des méthodes d'extraction des connaissances, où on traite le corpus textuel puis on extrait les motifs fréquents pour générer les règles d'association entre les entités biologiques avec leurs supports et leurs confiances à l'aide de l'algorithme APRIORI.

Nous avons finalisé ce mémoire par l'implémentation de cet algorithme et la discussion des résultats.

Les mots clés : Fouille de texte, Text Mining, connaissances, corpus textuel, motifs fréquents, règles d'association, support, confiance, APRIORI.

Abstract

The masses of textual data available today create a difficult problem to process. In this context, Text Mining methods are necessary to extract knowledge from texts.

Our work consists in studying one of the knowledge extraction methods, where we processthe textual corpus then extract the frequent patterns to generate the association rules between some biological entities with their supports and their confidences using the APRIORI algorithm.

We finalized this dissertation by implementing this algorithm and discussing the results.

Keywords: Text Mining, knowledge, textual corpus, frequent patterns, association rules, support, confidence, APRIORI.

ملخص

كتلة وحجم البيانات النصية المتوفرة اليوم تخلق مشاكل و صعوبة في معالجتها. في هذا السياق، تعتبر طرق التنقيب في النص (Text Mining) ضرورية لاستخلاص المعلومات و المعارف من النص.

يندرج عملنا في دراسة إحدى طرق استخلاص المعرفة، حيث نقوم بمعالجة النص (corpus) ثم نقوم باستخراج الكلمات المتكررة (motifs fréquents)لإنشاء قواعد الارتباط بين الوحدات البيولوجية (entités biologiques) مع عاملي الدعم (support) و الثقة (confiance) باستخدام خوارزمية APRIORI.

قمنا في نهاية هذه المذكرة بتطبيق هذه الخوارزمية ومناقشة النتائج.

الكلمات المفتاحية: البيانات النصية، المعرفة، مجموعة النصوص، الكلمات المتكررة، قواعد الارتباط، الدعم، الثقة، خوارزمية APRIORI.

Sommaire

Table of Contents

Sommaire

Table des figures

CHAPITRE I : Fouille de données dans le domaine médical

CHAPITRE II Les règles d'association

CHAPITRE III : Architecture et conception

CHAPITRE IV : Implémentation

Liste des tables

CHAPITRE I : Fouille de données dans le domaine médical

CHAPITRE II Les règles d'association

CHAPITRE III : Architecture et conception

Glossaire

- **Corpus :** Un corpus est un ensemble de documents (textes, images, …) pouvant provenir d'une ou de plusieurs disciplines, regroupés afin d'être soumis à des traitements.

- **Base de donnée textuelle :** Une base de données textuelles, ou base de données en texte intégral, est une compilation de documents ou d'autres informations présentée sous la forme d'une base dans laquelle le texte complet de chaque document référencé peut être visualisé en ligne, imprimé ou téléchargé.

Abréviations :

1. **TM:** **T**ext **M**ining.
2. **DM:** **D**ata **M**ining.
3. **FdT:** **F**ouille **d**e **T**exte.
4. **FdD:** **F**ouille **d**e **D**onnées.
5. **ECD:** **E**xtraction de **C**onnaissance à partir des **D**onnées.
6. **ECT :** **E**xtraction de **C**onnaissance à partir des **T**extes.
7. **NLP :** **N**atural **L**angage **P**rocessing.
8. **RI :** **R**echerche de l'**I**nformation.
9. **EI :** **E**xtraction de l'**I**nformation.
10. **RA :** **R**ègle d'**A**ssociation.

Introduction Générale

Les dernières années, le terme de fouille de textes est apparu dans un bon nombre depublications. En effet, tous les travaux en la matière ont un objectif commun qui consiste à extraire des textes une information plus précise, associée à une sémantique rigoureuse afin d'aider un expert à enrichir son modèle de connaissances ou à effectuer tout autre tâche de raisonnement comme la veille technologique.

Toussaint propose une définition calquée sur celle de l'extraction des connaissances à partir de données : « L'extraction de connaissances à partir de textes est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts. » **[1]**. Ce processus commence par la modélisation des textes afin de les préparer pour la fouille de données, et se termine par l'interprétation des résultats de la fouille et l'enrichissement des connaissances. La fouille des données n'est donc qu'une étape du processus de fouille de texte.

L'extraction des connaissances à partir de textes répond à la problématique de gérer et de traiter une grande masse de textes qui dépasse les capacités humaines.

La problématique générale en FdT est de tirer profit d'éléments d'information extraits afin d'exprimer des connaissances utilisables pour le domaine traité par les textes. Les nouvelles connaissances extraites servent à enrichir les connaissances actuelles d'un domaine contenues, par exemple, dans une base de connaissances.

Notre travail consiste donc à essayer de trouver, à partir d'un corpus médical étiqueté, les liens qui existent entre le cancer de la thyroïde et les facteurs qui peuvent le causer ou ceux qui peuvent aider les scientifiques à trouver un traitement

Ces techniques dans le but d'extraire des informations pertinentes concernant le cancer de la thyroïde. Nous avons choisi ce type de cancer car il est un cancer rare et de bon pronostic. Il touche plus souvent les femmes que les hommes. Son incidence a beaucoup augmenté depuis 1975. **[37]**

Le cancer de la thyroïde touche près de 570 000 personnes chaque année dans le monde selon l'Organisation Mondiale de la Santé (OMS), dont 75 % de femmes. Cette maladie se caractérise par l'apparition de cellules cancéreuses dans la glande thyroïde, formant une tumeur maligne

Les cancers de la thyroïde sont des cancers avec des taux de guérison élevés (Taux de guérison supérieur à 90 %). Le pronostic dépend de plusieurs facteurs, comme l'âge du patient, le type de cancer et le stade de la maladie. Un diagnostic précoce augmente les chances de guérison**. [38]**

Le travail présenté dans ce mémoire est de concevoir et réaliser une application qui permet de faire l'extraction des connaissances médicale à partir d'un texte contenu dans un corpus préalablement indexé et structuré [0]. Ces connaissances vont permettre aux spécialistes de trouver facilement les liens qui existent entre le cancer de la Thyroïde et les facteurs qui peuvent le causer. Ceci nécessite le passage par plusieurs phases : un prétraitement pour sélectionner les mots importants, la mesure les termes pertinents et l'élimination des mots vides puis, enfin, l'application de l'algorithme **APRIORI** afin de générer un ensemble de règles d'association portant sur les termes contenus dans notre corpus.

Notre mémoire est composé de cinq parties principales. Dans l'**Introduction générale** nous présentons le contexte général de l'étude, la problématique, l'objectif qu'on doit atteindre, méthodologie, résultats et le plan de ce mémoire.

Dans le CHAPITRE I, nous introduisons les notions générales liées au Data Mining et au Text Mining en donnant quelques définitions, les tâches principales, les domaines d'application de chacun et surtout les techniques utilisées pour l'extraction des connaissances à partir des textes (règles d'association, etc.)

Dans le CHAPITRE II, l'accent est mis sur l'architecture ainsi que la méthodologie utilisées dans notre application.

Cette dernière est devisée en deux parties principales :

- Application des fonctions de prétraitement et d'indexation de texte (élimination des mots vides, racinisation, lemmatisation et extraction des items.
- Application de l'algorithme APRIORI pour l'extraction des règles d'association.

Dans le CHAPITRE III, on a présenté le langage de programmation utilisé, les outils logiciels, les packages et les bibliothèques et on a exécuté l'application avec un corpus semi-structuré en expliquant les résultats obtenus.

Enfin, dans la Conclusion générale, on a fourni une analyse des résultats en expliquant les problèmes posés ainsi que les perspectives d'ouverture pour ce sujet.

Chapitre I

Fouille de données dans le domaine médical

1. Introduction

La fouille de textes est une discipline qui a pour but le traitement automatique d'une base de données composée exclusivement de texte. Elle offre des perspectives nouvelles pour la statistiqueet répond au défi du traitement des données textuelles.

Nous présentons en premier lieu une définition de la fouille de textes en tant qu'étape particulière d'un processus plus général d'extraction de connaissances à partir des textes afin d'obtenir des informations plus précises pour aider à la décision.

Dans cette partie, nous allons donner un aperçu général sur le **Data Mining** (**DM**) ou **Fouille de données** (**FdD**), le **Text Mining** (**TM**) ou **Fouille de textes** (**FdT**), l'**Extraction de Connaissance à partir des Données** (**ECD**), l'**Extraction de Connaissance à partir des Textes** (**ECT**) et les techniques utilisées (motifs, règle d'association, classification…).

2. Le Data Mining (DM)

" *Le data mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques (souvent grandes), de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données* ". **[2]**

L'analyse de données depuis différentes perspectives est le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns.

2.1 Les techniques de Data Mining :

Il existe plusieurs techniques de DM, parmi ces dernières on cite les trois suivantes :

2.1.1 La catégorisation (Classification supervisée):

Dans l'apprentissage supervisé, on fournit à l'ordinateur des exemples d'entrées qui sont

étiquetés avec les sorties souhaitées. Le but de cette méthode est que l'algorithme puisse 'apprendre' en comparant sa sortie réelle avec les sorties 'enseignées' pour trouver des erreurs et modifier le modèle en conséquence. L'apprentissage supervisé utilise donc des modèles pour prédire les valeurs d'étiquettes sur des données non étiquetées supplémentaires. Parmi les méthodes de classification supervisée, on peut citer : les arbres de décision, les réseaux neurones, la méthode des k plus proche voisins (KNN) ou la classification bayésienne. **[5]**

2.1.2 Le clustering (Classification non supervisée) :

Dans l'apprentissage non supervisé ou le clustering, les données ne sont pas étiquetées à l'avance. L'idée étant que l'algorithme d'apprentissage est censé trouver tout seul des points communs parmi ses données d'entrée. Les données non étiquetées étant plus abondantes que les données étiquetées, les méthodes d'apprentissage automatique qui facilitent l'apprentissage non supervisé sont particulièrement utiles. Parmi les méthodesde classification non supervisée, on peut citer : K-means. **[5]**

2.1.3 Les règles d'association :

La recherche des règles d'association est l'un des sérieux problèmes du ECD. Le principe est de trouver des règles dans les données de type « si *Condition*, alors *Résultats*», notées *Conditions→ Résultats*. Cette technique permet la découverte de règles intelligibles et exploitables dans un ensemble de données volumineux, règles exprimant des associations entre items ou attributs dans une base de données. **[6]**

2.2 Les étapes de Data Mining :

Il est très important de comprendre que le data mining n'est pas seulement le problème de découverte de modèles dans un ensemble de donnée. Ce n'est qu'une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les connaissances à partir des données. En 1996 un groupe d'analystes définit le data mining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (Cross-Industry Standard Process for Data Mining) comme schématisé ci-dessous :

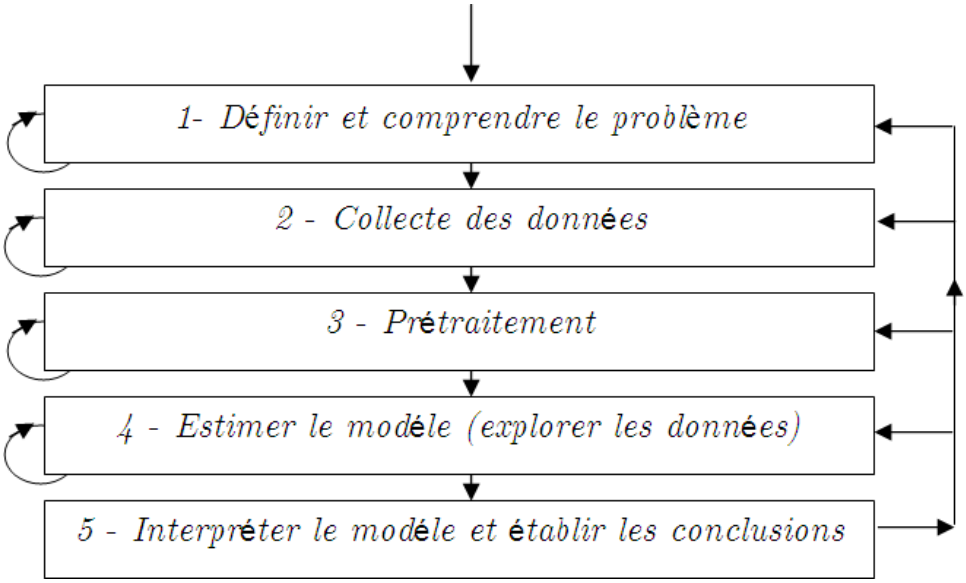

Figure I 1. ***Processus de data mining (CRISP-DM)***

Ce processus, composé de cinq étapes (**Figure I 1. Processus de data mining (CRISP-DM)**), n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données. Par exemple, on peut découvrir à l'étape d'exploration (5) de nouvelles données qui nécessitent d'être ajoutées aux données initiales à l'étape de collection (2). Décrivons maintenant ces étapes :

2.2.1- Définition et compréhension du problème :

Dans la plus part des cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va donner un résultat fiable. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus. Généralement, le data mining est effectué dans un domaine particulier (banques, médecine, biologie, marketing, ...etc) où la connaissance et l'expérience dans ce domaine jouent un rôle très important dans la définition du problème, l'orientation de l'exploration et l'explication des résultats obtenus. Une bonne compréhension du problème comporte une mesure des résultats de l'exploration, et éventuellement une justification de son coût. C'est-à-dire, pouvoir évaluer les résultats obtenus et convaincre l'utilisateur de leur rentabilité.

2.2.2- Collecte des données :

Dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et des objectifs du data mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles,...). Quelques traitements ne nécessitent qu'une partie des données, on doit alors sélectionner les données adéquates. Généralement les données sont subdivisées en deux parties : une utilisée pour construire un modèle et l'autre pour le tester. On prend par exemple une partie importante (suffisante pour l'analyse) des données (80 %) à partir de laquelle on construit un modèle qui prédit les données futures. Pour valider ce modèle, on le teste sur la partie restante (20 %) dont on connaît le comportement.

2.2.3- Prétraitement :

Les données collectées doivent être "préparées". Avant tout, elles doivent être nettoyées puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements. Des données peuvent être incohérentes c-à-d qui sortent des intervalles permis, on doit les écarter où les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Un exemple de ces transformations est la normalisation des données qui consiste à la projection des données dans un intervalle bien précis [0,1] ou [0,100] par exemple. Un autre exemple est le lissage des données qui considère les échantillons très proches comme étant le même échantillon. Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Une méthode largement utilisée dans ce contexte, est l'analyse en composantes principales (ACP). Une autre méthode de réduction est celle de la

sélection et suppression des attributs dont l'importance dans la caractérisation des données est faible, en mesurant leurs variances. On peut même réduire le nombre de données utilisées par le data mining en écartant les moins importantes. Dans la majorité des cas, le prétraitement doit préparer des informations globales sur les données pour les étapes qui suivent tel que la tendance centrale des données (moyenne, médiane, mode), le maximum et le minimum, le rang, les quartiles, la variance, ... etc. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données (data warehouse).

2.2.4- Estimation du modèle :

Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat

2.2.5- Interprétation du modèle et établissement des conclusions :

Généralement, l'objectif du data mining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des interprétations des modèles obtenus. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que ceux complexes sont plus précis mais difficiles à interpréter.

2.3 Domaines d'utilisation de Data Mining:

Les domaines d'application du DM sont vastes et variés. Cependant un trait commun les lie est le fait qu'ils traitent un volume important de données et extrait des informations qui visent à améliorer la qualité du produit ou du service.

Parmi les domaines ou l'utilisation du DM est devenue monnaie courante:

- Laboratoires pharmaceutiques et médicaux.
- Assurances.

- Banques et grandes administrations.
- Automobiles et grandes industries.
- Les transports à grandes échelles.
- Grande distribution et vente en correspondance. **[3]**

Le data mining s'applique dans différentes spécialités appartenant au domaine médical, notamment pour l'aide à la prise de décision.

Le DM offre un potentiel considérable d'amélioration des systèmes de santé. L'exploitation des données biologiques permet d'extraire des connaissances utiles à partir d'énormes séries de données. Les applications du DM dans la biologie comprennent la découverte de gènes, l'inférence de fonction protéique, le diagnostic de la maladie, le pronostic de la maladie, l'optimisation du traitement de la maladie, la reconstruction de réseaux de protéines et d'interactions géniques, le nettoyage de données et la prédiction d'emplacement sous-cellulaire de protéines, etc.

Il existe également une branche spécialisée de la fouille de données (Data Mining) qui est l'analyse de texte libres ou bien la fouille de données textuelles qui s'appelle « Text Mining »

3. Le Text Mining (TM)

Le Text Mining, également appelé fouille de textes ou extraction de connaissance à partir de textes, est un ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés que sont les textes écrits, comme les fichiers bureautiques de type word, les emails, lesdocuments de présentation de type PowerPoint...etc. Pour extraire du sens de documents non structurés, le TM s'appuie sur des techniques d'analyse linguistique. Le TM est utilisé pour classerdes documents, réaliser des résumés de synthèse automatique ou encore pour assister la veille stratégique ou technologique selon des pistes de recherches prédéfinies. **[3]**

TEXT MINING = LINGUISTIQUES + DATA MINING

« Nous définissons aussi le TM comme étant le DM sur les données textuelles. La fouille de textes est tout ce qui porte sur l'extraction de modèles et d'associations précédemment inconnues à partir de grandes base de données textuelles »

Le TM réfère ainsi à l'ensemble des techniques et méthodes du DM en vue de retrouver dans les textes de documents de grandes base de données textuelles, l'information pertinente, utile et précédemment inconnue

3.1 Les formats possibles de données de Text Mining :

Les données qui font l'objet de taches de fouilles se présentent suivant diffèrent formats. Nous distinguerons trois principaux :

- **Données tabulaires:** les données sont disposées en lignes (une donnée par ligne), les attributs en colonnes, l'ordre des lignes et des colonnes n'a aucune importance, au sens où en changer ne modifiera en rien le résultat des algorithmes de fouille qui y seront appliqués.

- **Textes bruts :** les textes, même numérisés, ne présentent pas du tout les mêmes propriétés que les tableaux de données. Autant les tableaux ont un haut degré d'organisation, autant les textes sont faiblement structurés. La seule structure présente est l'ordre linéaire dans lequel les caractères apparaissent. En revanche, les notions de mots, de phrases, de paragraphe… n'y ont a priori pas de sens, sauf à réaliser un prétraitement qui les identifie.

- **Document semi-structurés :** ce format est intermédiaire entre les précédents : il est plus structuré qu'un texte brut, mais moins qu'un tableau : c'est celui des documents XML, les éléments propres au langage utilisé (principalement les balises ouvrantes et fermantes) sont considérés comme des « caractères » indivisibles supplémentaires qui s'ajoutent aux autres. Le prétraitement consistant à identifier les balises est trivial. Les balises, en effet, respectent une syntaxe qui décrit une structure.

3.2 Les taches de Text Mining :

Le Text Mining ne se substitue pas à la recherche d'information ou le traitement du langage naturel. Les techniques qui permettent d'organiser un corpus de documents textuels selon leur contenu ont un spectre d'utilisation très large. Le TM cherche des réponses aux questions difficiles ou impossibles à résoudre avec les seuls moteurs de recherche.

3.3 Le processus de Text Mining :

Les étapes nécessaires pour effectuer le processus de text mining (**Figure I 2**) sont :

3.3.1 L'acquisition ou sélection: Source de données telle que : corpus textuels, bibliothèques électroniques, Web…

3.3.2 Le prétraitement du corpus :

- **Nettoyage:** Variable selon la source des données, cette phase consiste à réaliser des tâches telles que la L'élimination des caractères non alphabétiques comme l'URL, l'emoji, les caractères spéciaux et dans le cas des données semi-structuré (HTML ou XML) nous éliminons les tags et les balises.
  Cette étape peut aussi inclure la suppression des chiffres, ponctuation, symboles et *stopWords*, passage en minuscule. **[31]**
  - ✓ **Suppression des mots outils (vide) :** Les mots qui apparaissent très souvent dans tous les textes sont appelé les mots vides (pronoms, prépositions, déterminants, etc.). Ils constituent la majorité des mots d'un texte mais sont faiblement informatifs. L'élimination de ces mots est nécessaire, elle est effectuée par l'intermédiaire d'une liste prédéfinie pour chaque langue étudié. Par exemple, pour l'anglais : the, that, after, one, are, above, few…….ect. sont considérés comme étant des mots vides.
  - ✓ **Suppression des mots rare :** Généralement, les mots rares sont les mots qui n'apparaissent qu'une ou deux fois dans l'ensemble des textes. Ces mots doivent être éliminés puisque d'après la loi de zipf ils sont inutiles dans la phase de modélisation.
- **Normalisation des données:**
  - ✓ **Tokenization** : la tokenisation est le processus de séparation d'un flux de texte en mots, phrases, symboles et d'autres éléments significatifs appelés jetons ou tokens

    Exemple: ***« Vous trouverez en pièce jointe le document en question » ; « Vous », « trouverez », « en pièce jointe », « le document », « en question ».***
  - ✓ **Stemming:** un même mot peut se retrouver sous différentes formes en fonction du genre (masculin féminin), du nombre (singulier, pluriel), la personne (moi, toi, eux…) etc. cette technique est basée

sur le regroupement des mots ayant la même racine (stem). Le stemming désigne généralement le processus heuristique brut qui consiste à découperla fin des mots afin de ne conserver que la racine du mot.

Exemple: ***trouverez → trouv***

Cette représentation réduit de 30% la taille moyenne d'un texte

- ✓ **Lemmatisation :** cela consiste à réaliser la même tâche mais en utilisant un vocabulaire et une analyse fine de la construction des mots. La lemmatisation permet donc de supprimer uniquement les terminaisons inflexibles et donc à isoler la forme canonique dumot, connue sous le nom de lemme.

Exemple: ***trouverez → trouver***

3.3.3 L'indexation : permet de créer une représentation des documents dans le système, sonobjectif est de trouver les concepts les plus importants du document (ou de la requête),qui forme le descripteur de document.

3.3.4 Le Data Mining : La fouille de données est l'étape centrale du processus d'extraction de connaissance. Elle consiste à découvrir de nouveaux modèles au sein de grandes quantités de données.

3.3.5 L'extraction des connaissances : Application de l'un des algorithmes de la fouille de textes. **[5]**

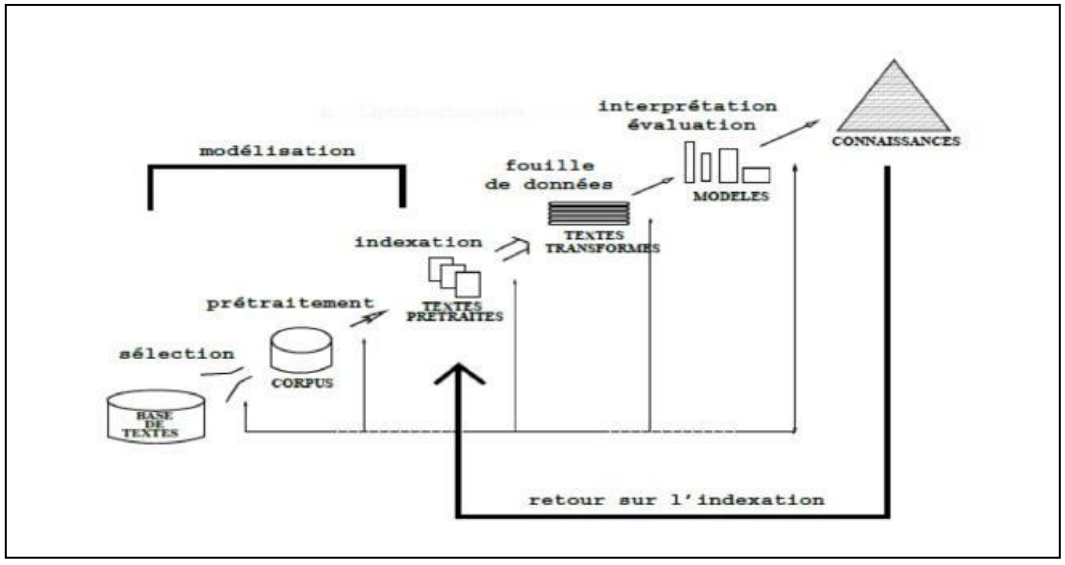

Figure I 2. ***Le processus de Text Mining [17]***

3.4 Les techniques de Text Mining :

Les techniques de Text Mining (TM) sont: Le traitement de langage naturel (NLP), larecherche d'information (RI) et l'extraction d'information (EI).

3.4.1 Le traitement du langage naturel (NLP) :

Un traitement de langage naturel est :

- Une suite d'actions ou calculs à faire par la machine. Le Traitement Automatique des Langues a pour objectif de traiter des données linguistiques (textes) exprimées dans une langue dite "naturelle".
- La Conception de programmes capables de traiter automatiquement des données linguistiques de type : textes écrits ; dialogues écrits ou oraux ; unités linguistiques (mots, phrases, énoncés, ...).

Les tâches impliquées dans cette technique peuvent inclure le nettoyage et la normalisation des données tels que : La tokenization, élimination des mots vides et filtrage de textes, Lemmatisation, La racinisation (ou troncature). **[7]**

3.4.2 La recherche d'information (RI) :

La recherche d'information « RI » s'intéresse aux documents dans leur globalité et aux thèmes qu'ils abordent, pour comparer les documents et détecter des typologies. Elle cherche à détecter tous les thèmes présents.

3.4.3 L'extraction d'information (EI) :

L'extraction d'information « EI » est la recherche automatisée d'informations sur un sujet précis dans le corps d'un texte ou un corpus documentaire.

Les outils d'EI permettent de récupérer des informations dans des documents textuels, des bases de données, des sites Web ou des sources diverses. Les informations sont extraites de textes non structurés, semi-structurés ou structurés, et lisibles par ordinateur. Toutefois, cette technique est surtout employée dans le traitement automatique du langage naturel où elle sert à extraire du texte structuré d'un texte qui ne l'est pas. **[36]**

3.5 Comparaison entre 'RI' et 'EI' :

La table ci-dessous (**Table I 1**), montre les principales différences qui existent entre la Recherche d'Informations et l'Extraction d'Informations.

	RI	EI
01	Récupère des informations précieuses à partir de texte non-structuré.	Extraire les informations des bases de données structurées
02	Tâche de recherche des documents textuels qui sont pertinents pour le besoin d'information d'utilisateur.	L'objectif est d'extraire les fonctionnalités pré-spécifié des documents ou d'affichage information.
03	Récupération des documents	Récupération des fonctionnalités
04	La sortie du RI est un sous-ensemble de documents qui sont pertinent pour la requête de l'utilisateur.	Plus difficile car cela nécessite des connaissances plus détaillées sur un document. Cela nécessite souvent d'établir des relations entre les caractéristiques.
05	**Outils:** Intelligent Miner Text Analyst	**Outils:** Text Finder Clear Forest Text

Table I 1. ***Comparaison entre 'RI' et EI'***

4. Extraction des connaissances (EC) :

4.1 A partir des données (ECD):

L'extraction de connaissances à partir de bases de données est un processus non trivial quiconstruit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données.

4.2 A partir des textes (ECT):

L'extraction de connaissances à partir de textes est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts**.**

L'ECT est à l'intersection de plusieurs domaines de recherche. En cherchant à présenter un schéma un peu plus détaillé de fouille de textes, il apparaît clairement que la représentation linéaire ne suffit pas et qu'il s'agit d'un processus itératif, incrémental dans lequel

un même outil ou un même algorithme (comme la classification) peut être utilisée sur des mots comme sur des structures plus complexes d'objets. En réalité, les premières boucles du processus peuvent s'effectuer avec très peu de connaissances et sur des structures très simples avant de s'enrichir progressivement. Nous proposons donc un schéma général qui suit (*Cf.* **Figure I 3**) dans lequel un certain nombre d'étapes peuvent être court-circuitées pour permettre l'accès à des outils ou méthodes relevant des étapes suivantes.

L'apprentissage intervient presque à chaque étape et parfois sur des choses aussi simples que la constitution d'une liste de mots. On observe également que les connaissances sont utilisées dans le processus de fouille, ce qui est, pour nous, la garantie qu'au fur et à mesure que les connaissances s'enrichissent, le processus ne stagne pas en proposant des classifications toujours identiques mais au contraire qu'il permette l'identification de connaissances nouvelles.

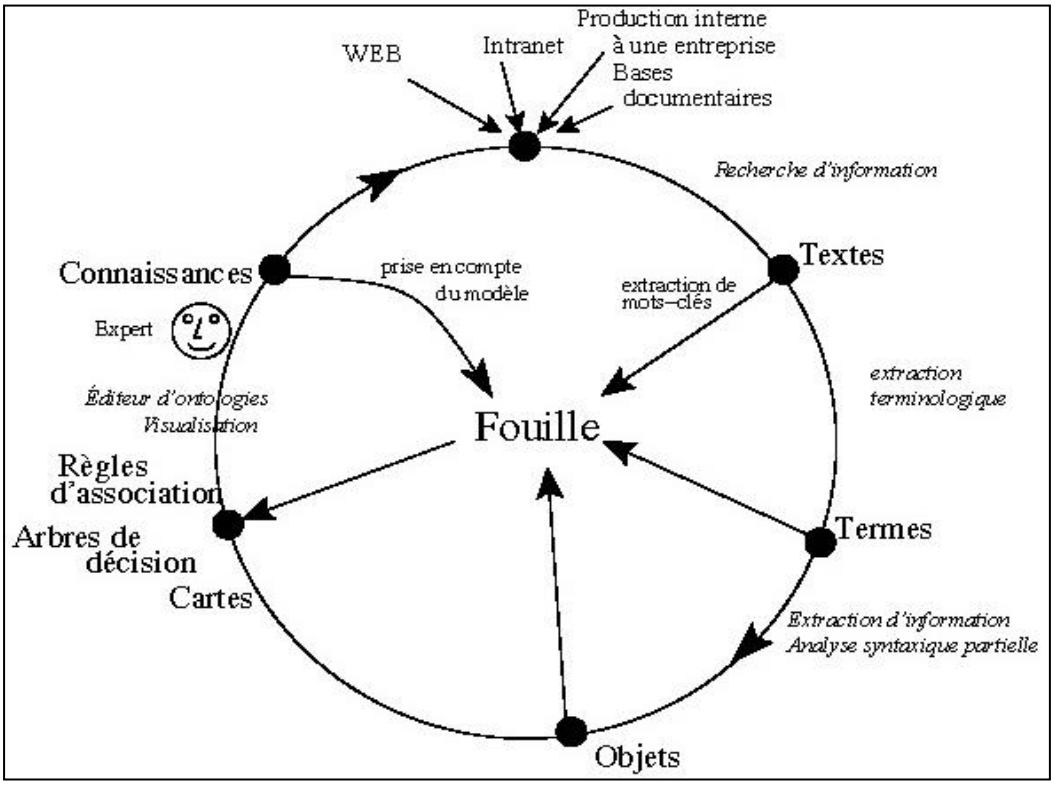

Figure I 3. ***Extraction des connaissances à partir des textes [17]***

Chapitre II

Les règles d’association

1. Etat de l'art sur l'utilisation des règles d'association dans le domaine médical

Nous survolons dans cette partie quelques travaux qui portent sur l'utilisation des règles d'association dans le domaine médical en général. Par exemple, en diabétologie, **[17]** ont appliqué l'algorithme APRIORI sur une base de données contenant les dossiers des patients diabétiques afin d'extraire des règles d'association des paramètres réels stockés. Les résultats indiquent que la méthodologie suivie peut être utile à la procédure de diagnostic, en particulier lorsqu'il s'agit de grandes quantités de données.

L'approche proposée dans **[18]** a pour but d'explorer les connaissances et les règles sur la compatibilité des médicaments à partir des prescriptions pour le traitement de l'arythmie dans la base de données de la médecine traditionnelle chinoise. Le résultat expérimental montre que la compatibilité de la prescription obtenue correspond généralement à la loi fondamentale de la médecine traditionnelle chinoise pour l'arythmie. L'expérience a également permis de découvrir certaines compatibilités spéciales non signalées, qui pourraient servir de base à l'élaboration de nouvelles prescriptions en matière d'arythmie.

Le travail de **[19]** vise à trouver des associations entre le diagnostic et les traitements. La ressemblance entre la facture médicale et la facture d'achat est la motivation de l'utilisation de l'algorithme APRIORI dans ce travail de recherche.

[20] l'algorithme Fp-growth pour identifier des modèles intéressants dans les données d'audiologie médicale. Ce travail de recherche a proposé un modèle de découverte des connaissances en cinq étapes qui est ensuite mis en œuvre à l'aide de la technique Fp-growth afin de découvrir des informations précieuses à partir d'ensembles de données audiométriques.

[21] APRIORI pour générer les règles pour les patients cardiaques, cette recherche a permis de découvrir les facteurs qui causent les problèmes cardiaques chez les hommes et les femmes. Après avoir analysé les règles, les auteurs concluent que les femmes ont moins de risques d'avoir une maladie coronarienne que les hommes.

La recherche de **[22]** vise à extraire des règles d'Association en se basant sur une nouvelle idée pour trouver les cooccurrences de maladies transmises par un patient qui utilise le dépôt de soins de santé. Les auteurs ont développé un prototype qui extrait des données à partir de base de données des soins de santé des patients, transforme les données OLTP dans un Data Warehouse en générant des règles d'association avec l'algorithme APRIORI. Leur prototype prédit les corrélations parmi les maladies primaires (la maladie pour laquelle le patient visite le médecin) et les maladies secondaires (qui sont autres maladies associées transmises par le même patient ayant la maladie primaire).

L'étude de **[23]** vise à analyser les règles d'association de l'injection Fufang Kushen en combinaison avec d'autres médicaments modernes dans le traitement du cancer du poumon sur la base des dossiers médicaux dans des situations cliniques extraites du système d'information hospitalière dans l'Institut des sciences médicales de l'Académie chinoise des sciences médicales chinoises.

Le cancer du sein est le deuxième plus fréquent néoplasme humain qui représente un quart de tous les cancers chez les femmes. Dans la plupart des pays, il est considéré comme la principale cause de décès chez les femmes. Différents travaux ont été effectués dans ce domaine. **[24]** tentent d'améliorer le processus de classification grâce à une classification pondérée efficace basée sur l'algorithme des règles d'association, appelé WCBA. Ils présentent également une nouvelle technique d'élagage et de prédiction basée sur des mesures statistiques pour générer des règles d'association plus précises afin d'améliorer le niveau de précision des classificateurs. **[24]** utilisent le WCBA pour classer les cas de cancer du sein avec l'aide des experts du King Hussein Cancer Center (KHCC) situé à Amman, en Jordanie. Les auteurs visent à réduire la crainte d'une récidive de la maladie et prendre les mesures nécessaires pour prévenir la progression de la maladie et pour prédire le cancer du sein chez les patientes. L'algorithme peut être généralisé pour travailler sur différents domaines avec l'aide d'experts en la matière.

[25] a développé une méthode basée sur les règles d'association dans le domaine du cancer du sein. Des liens ont été découverts entre la récidive tumorale et certains facteurs en faisant l'exploration de données sur les patientes atteintes de cancer du sein. Les résultats correspondant aux connaissances de base sur le diagnostic des maladies du sein peuvent être utilisés comme références importantes dans les maladies du sein.

[26] tentent à explorer les associations entre les effets indésirables et la pharmacothérapie chez les patients atteints d'un cancer du poumon. Un algorithme basé sur les règles d'association a été mis au point et utilisé pour étudier les associations entre les médicaments et les événements indésirables. Les trois effets indésirables les plus fréquents étaient l'hypocalcémie, une élévation de la créatine phosphokinase et l'hypertriglycéridémie. De plus, en utilisant l'algorithme APRIORI modifié, 380 règles d'association ont été trouvées entre les événements indésirables et la chimiothérapie.

Tous ces travaux nous permettent de constater que les règles d'associations sont très utilisées dans l'aide à la décision médicale. La fouille de données par règles d'association permet d'aider à construire une base de connaissances de prévention des maladies qui va aider les fournisseurs de soins de santé dans le suivi. Nous remarquons aussi que l'algorithme Apriori est très utilisé parmi les différents algorithmes d'extraction de règles association. Pour cette raison notre choix s'est porté sur l'intégration des règles d'association dans notre système interactif d'aide à la décision médicale en utilisant l'algorithme Apriori.

2. Les règles d'associations :

Les règles d'association constituent un des modèles les plus puissants en fouille de données. Elles ont été utilisées avec succès dans de nombreux domaines, tel que l'aide à la décision médicale. Elles sont en mesure de détecter les tendances et les relations cachées et peuvent faire l'exploration des corrélations à partir des données.

2.1 Définition :

Les règles d'association (RA) sont des instructions if-then qui aident à montrer la probabilité de relations entre des éléments de données au sein d'ensembles de données volumineux dans divers types de bases de données. L'exploration de règles d'association a plusieurs applications et est largement utilisée pour aider à découvrir les corrélations des ventes dans les données transactionnelles ou dans les ensembles de données médicales.

Une RA comporte deux parties : un antécédent (si) et un conséquent (alors). Un antécédent est un élément trouvé dans les données. Un conséquent est un élément trouvé en combinaison avec l'antécédent. Les RAs sont créés en recherchant dans les données des modèles fréquents de type if-then et en utilisant les critères de prise en charge et de confiance pour identifier les relations les plus importantes.

2.2 Représentation d'un tableau de données binaire

Les données transactionnelles à explorer peuvent être représentées sous la forme d'un tableau binaire de dimension n * m ; soit un ensemble T={t1,t2,..,tn} de n transactions avec un ensemble I, de m attributs booléens, appelés items (I={i_1,i_2,i_3,… ,i_m })· Chaque transaction est associée à un identifiant unique (noté Tid).

Le tableau (a) (gauche de la **Figure II 1**) représente ce type de données, pour un ensemble de 6 transactions {1, 2, 3, 4, 5, 6} décrit par 4 items {A, B, C, D}. Chaque transaction est représentée par l'ensemble des attributs observés. Dans le tableau (b) (droite de la **Figure II 1**), ces transactions sont présentées sous une forme binaire (1 pour la présence d'un attribut, 0 sinon).

T_{id}	Transactions
1	{A, B, C}
2	{A, B}
3	{A, B, C}
4	{A, C}
5	{B, C}
6	{D}

T_{id}	A	B	C	D
1	1	1	1	0
2	1	1	0	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

Le codage disjonctif complet (**Figure II 2**) s'agit de détecter les cooccurrences des modalités (attribut = valeur). Certaines associations sont impossibles par construction (ex. on ne peut pas être « petit » et « grand » en même temps). Et dès que l'on peut se ramener à des données 0/1, il est possible de construire des règles d'association.

Figure II 1. ***Exemple de base de données(a), Représentation binaire (b)***

Obs	Taille	Corpulence
1	Petit	Mince
2	Grand	Enveloppé
3	Grand	Mince

Obs	=Petit	=Grand	=Mince	=Enveloppé
1	1	0	1	0
2	0	1	0	1
3	0	1	1	0

Figure II 2. ***Tableau individus * variables (c), Tableau en binaire (d)***

2.3 Principe de base

- **Item :** Soit $(I = \{i_1, i_2, i_3, \ldots \ldots \ldots, i_m\})$ un ensemble de m items i où chaque item est une variable binaire de la base de données.

- **Itemset :** les items peuvent être regroupés de multiples manières pour former des itemsets**,** donc un itemset est un ensemble de n Items noté X: {A, B, C, D….}. L'ensemble de tous les Itemsets possiblement formés par les éléments d'Items est $2^{(m)}$.

- **Règle d'association :** est une implication de la forme X ⇒ Y, où X ⊆ I, Y ⊆ I, et X ∩ Y =∅.

Une règle X ⇒ Y indique que les transactions possédant le motif X ont tendance à posséder le motif Y. Cependant, il n'existe aucune relation de causalité entre X et Y : la présence de X ne cause pas la présence de Y.

- **Prémisse ou antécédent, conclusion ou conséquent :** La partie gauche de la règle est appelée la prémisse ou l'antécédent et la partie droite est la conclusion ou le conséquent. Pour une règle X ⇒ Y, X est donc la prémisse ou l'antécédent et Y est donc la conclusion ou le conséquent.

- **Support d'un Itemset** : représente le nombre total des transactions d'une base de données comportant cet Itemset divisé par le nombre total des observations de cette base de données,Par exemple, soit une base de données D et soit X un Itemset de n éléments. Dans une base de données transactionnelle D, le support de l'itemset X est le nombre de transactions dans D incluant X, divisé par le nombre total des transactions de D. **Support(X)** $= \frac{\mathbf{card(X)}}{\mathbf{card(D)}}$

- **Itemset Fréquent :** Un itemset est dit fréquent si son support est supérieur à un *seuil* donné. Cette notion permet de filtrer les itemsets, pour ne garder que les plus intéressants dans le processus de fouille de données.

- **Mesures d'intérêt**: Il existe deux mesures importantes, le support et la confiance, la robustesse d'une règle d'association est déterminée grâce à ces deux métriques. Une règle d'association qui a un support faible va être observée rarement. La confiance mesure la pertinence de l'inférence dans une règle, par exemple plus grande est la mesure de confiance de la règle X ⟹ Y, plus pertinente sera cette dernière.

- **Le support d'une règle d'association**

Le support d'une règle d'association s'exprime par le nombre de transactions qui contiennent les éléments de X et les éléments de Y divisé par le nombre total des transactions de la base des transactions. Dans une base de données D, le support d'une règle d'association X ⇒ Y est le nombre de transactions qui contiennent X et Y divisé par le nombre total des transactions.

$$\textbf{Support}(X \Rightarrow Y) = \frac{\textbf{card}(XUY)}{\textbf{card}(D)} \; . \; \textbf{Support}(X \Rightarrow Y) \in [0,1]$$

Prenons la règle d'association « céréales ⇒ lait », littéralement, Si céréales Alors lait. Le support représente le nombre de transactions dans lesquelles on trouve les Items céréales et Lait, divisé par le nombre total des transactions.

- **La confiance d'une règle d'association**

La confiance d'une règle d'association s'exprime par le nombre de transactions qui contiennent la relation d'union entre la transaction X et la transaction Y divisé par le nombre des transactions qui contiennent la transaction X; elle est définit par

$$\textbf{Confiance}(X \Rightarrow Y) = \frac{\textbf{card}(XUY)}{\textbf{support}(X)} \cdot \textbf{conf}(X => Y) \in [0,1]$$

Et elle représente la probabilité qu'une transaction supportant X supporte également Y ou bien la probabilité que la partie droite de la règle soit vérifiée, si la partie gauche de la règle est vérifiée.

3. Extraction des règles d'association :

Rappelons qu'une règle d'association est une relation d'implication X ⇒Y entre deux ensembles d'Items X et Y tel que X∩Y =Ø et X≠Ø. X est appelé corps de la règle et Y est la tête, ou ainsi X est appelé condition ou prémisse et Y résultat ou conclusion. Cette règle indique que les transactions qui contiennent les articles de l'ensemble X ont tendance à contenir les articles de l'ensemble Y. Ces règles sont de la forme :

Si {Item 1, Item 2, ..., Item i} Alors {Item j,….,Item t}.

Par exemple : Si {symptôme 4 = "1"} Alors {symptôme 5 ="1"}. Cette règle est interprétée de la manière suivante :"Si la séquence possède symptôme 4 Alors elle possède symptôme 5".

Le processus d'extraction des règles d'association est constitué de quatre étapes allant de la sélection et la préparation des données jusqu'à l'interprétation des résultats, en passant par la phase de recherche des connaissances (extraction des ensembles fréquents d'attributs et génération des règles d'association).

Ces étapes sont représentées dans la figure suivante (**Figure II 3**) :

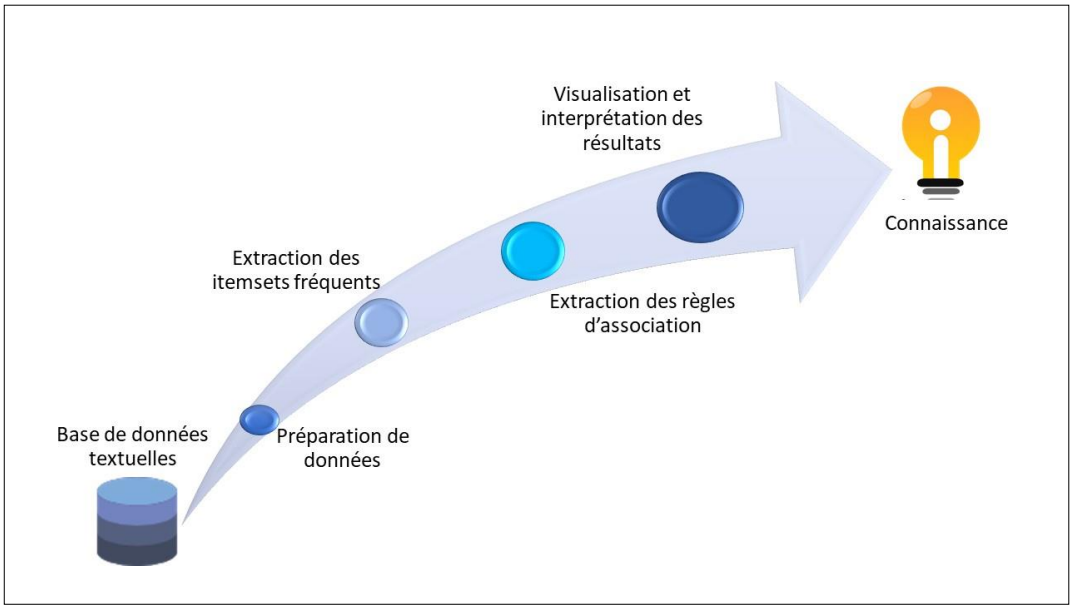

Figure II 3. ***Les étapes d'extraction des règles d'association***

3.1. Préparation de données

Cette étape est très importante avant de démarrer un processus d'exploration de données, car la qualité des résultats dépend de la qualité des entrées. Il est souvent nécessaire d'appliquer un processus de nettoyage car les informations sont souvent bruitées et incomplètes. Cette étape consiste à sélectionner les données (attributs et objets) de la base de données utiles à l'extraction des règles d'association et les transformer par la suite en un contexte d'extraction, c'est-à-dire une transformation en triplets constitués : d'un ensemble d'objets, d'un ensemble d'Itemsets ainsi que d'une relation binaire entre les deux. Cette transformation est nécessaire afin qu'il soit possible d'appliquer les algorithmes d'extraction de règles d'association sur divers types de données.

3.1.1. Recherche d'itemsets fréquents

L'extraction des Itemsets fréquents permet de séparer, depuis une base de données, des ensembles d'attributs, qui satisfont un *seuil* de fréquence minimale *min.sup* spécifié par l'utilisateur. La recherche des itemsets fréquents est un problème non trivial car le nombre

d'itemsets fréquents potentiels est exponentiel en fonction du nombre d'items de la base de données.

Cette étape est très coûteuse en temps d'exécution et en espace. Pour un ensemble de m items par exemple, le nombre d'Itemsets fréquents qui peut être générés est de 2m.

3.1.2. Génération des règles d'association

Sur la base des Itemsets fréquents extraits à l'étape précédente, il est possible d'extraire des règles d'association, de forme générale $X \Rightarrow Y$ qui associent un sous ensemble d'Itemsets X avec un second sous-ensemble d'Itemsets Y. Afin de limiter l'extraction aux règles d'association les plus informatives, seules celles qui possèdent une confiance supérieure ou égale au seuil minimal défini par l'utilisateur sont générées. En général, la génération des règles d'association est réalisée de manière directe, et le coût de cette étape en temps d'exécution est donc faible par rapport au coût de l'extraction des itemsets fréquents.

3.1.3. Visualisation et interprétation des règles d'association

C'est la phase finale du processus d'extraction de règles d'association qui consiste en la visualisation par l'utilisateur des règles d'association découvertes et leur interprétation afin d'en déduire des connaissances utiles pour comprendre une situation donnée. La forme de présentation de règles peut être textuelle, graphique ou bien une combinaison de ces deux formes. L'expert du domaine peut juger la pertinence et l'utilité des règles, mais vu le nombre important des règles générées par les algorithmes, il est parfois difficile aux experts du domaine de les exploiter dans leur intégralité, car cela engendre un travail cognitif très important. Donc leur premier souhait est de réduire cet ensemble pour diminuer le temps d'expertise correspondant. Dans le domaine médical par exemple, les experts n'ont pas forcément beaucoup de temps à consacrer à l'analyse des résultats.

Dans le chapitre suivant, nous allons décrire cette problématique, et présenter les propositions faites dans ce domaine.

4. Algorithmes d'extraction des règles d'association :

Dans cette section, nous présentons les algorithmes d'extraction de règles d'association les plus utilisés.

4.1. L'algorithme APRIORI

Proposé par [14] ; APRIORI représente l'algorithme pionnier pour la recherche des itemsets fréquents, c'est une approche révolutionnaire dans l'apprentissage et l'exploration des règles d'association. Il tire son nom de son heuristique qui utilise l'information connue a priori sur la fréquence des items. Cette heuristique définit que si un ensemble d'items est fréquent, alors tous ses sous-ensembles sont aussi fréquents. La réciproque est que si un ensemble {i1, i2, etc.} est peu fréquent, alors ses super-ensembles sont aussi peu fréquents. Cette stratégie s'appelle *l'élagage basé sur le support*.

L'algorithme APRIORI est constitué de deux étapes : une étape de jointure et une autre d'élagage.

L'étape de jointure : pour trouver les Itemsets fréquents dans la base de données transactionnelle, l'algorithme APRIORI effectue plusieurs balayages de la base de données. Le premier balayage sert à identifier les candidats C_k, un ensemble d'Itemsets, et à compter le nombre de fois qu'apparaît chaque item, c'est-à-dire leur support respectif. Tous les items dont le support est plus grand qu'une valeur prédéterminée appelée *min.sup*, sont conservés afin de former L_k, l'ensemble des k Itemsets fréquents. Cet ensemble sert d'amorce pour générer l'ensemble de candidats C_{k+1}. L'ensemble C_{k+1}, qui regroupe les (k+1)-Itemsets, est généré en liant L_k avec lui-même. Pour que deux k-Itemsets puissent être liés, ils doivent posséder k-1 items en commun. Par conséquent la liaison de deux 1-itemsets ne requiert aucun élément en commun, alors que la liaison de deux 3-itemsets requiert 2 éléments en commun. Les deux 1-itemsets {1} et {2} peuvent être liés ensemble pour générer le 2-itemset {1,2}. Le 3-itemset {1,2,3} peut être lié avec {2,3,4} pour générer {1,2,3,4}, mais ne peut pas être lié avec {3,4,5} puisque seul l'items {3} est commun aux Itemsets {1,2,3} et {3,4,5}. Il est fort possible qu'un Itemset généré ne respecte pas le seuil de support minimal. Si c'est le cas, cet Itemset est éliminé lors de l'étape d'élagage.

L'étape d'élagage : Une fois l'ensemble des candidats C_{k+1} généré, le support de tous les (k+1)-Itemsets est calculé. Tous les (k+1)-Itemsets appartiennent à C_{k+1} dont le support ne dépasse pas le min.sup sont retirés de la liste des candidats. Comme la liste des candidats C_{k+1} est réalisée à partir de la liste antérieure des candidats C_k, tout candidat retiré à l'étape k n'est plus considéré dans l'étape k+1.

La complexité d'Apriori est $O(n2^m)$, où n est le nombre de transactions, et m le nombre d'attributs. Elle montre que la durée d'exécution de l'algorithme croît linéairement avec le nombre de transactions, et exponentiellement avec le nombre d'attributs.

4.2. L'algorithme Close

L'algorithme Close **[27]** repose sur l'extraction de générateurs d'ensemble de mots fermés fréquents et le nombre d'ensembles de mots fermés fréquents est généralement bien inférieur au nombre d'ensembles de mots fréquents.

Cet algorithme passe par les étapes suivantes :

- Initialisation de l'ensemble des générateurs avec l'ensemble des singletons formés par les mots du corpus ;
- Calcul de la fermeture des générateurs de niveau k et de leur support ;
- Ajout des fermetures des générateurs à l'ensemble des ensembles de mots fermés fréquents ;
- Génération des générateurs de niveau k + 1 ;
- A la fin, Les générateurs de niveau k + 1 sont obtenus de la même manière que dans l'algorithme APRIORI, mais ceux appartenant à la fermeture d'un générateur de niveau k sont supprimés.
- La fermeture d'un ensemble de mots A est un ensemble de mots B tel que B apparaît dans les mêmes textes que A. Pour la calculer on utilise deux fonctions :
  - ❖ f : associe à un ensemble de mots les textes où il apparaît ;
  - ❖ g : associe à un ensemble de textes les mots qu'ils ont en commun ;
- Soit A un ensemble de mots : fermeture (A) = g o f(A)

4.3. Algorithme Fp-Growth [28]

L'algorithme Fp-growth permet la découverte des itemsets fréquents sans génération des itemsets candidats. Le processus se déroule en deux étapes, une étape de construction des arbres FP-tree et une étape d'extraction des itemsets fréquents directement de ces arbres. La méthodeconsiste d'abord à compresser la base de données en une structure compacte appelée FP tree (Frequent Pattern tree), puis à diviser cette base ainsi compressée en sous projections de la base de données appelées bases conditionnelles. Chacune de ces projections est associée à un item fréquent. L'extraction des itemsets fréquents se fera sur chacune des projections séparément **[29]**

La construction de l'arbre FP-tree s'effectue suivant les étapes ci-dessous **[28] et [9]** :

1. Calculer le support minimal.
2. Calculer chacune des occurrences d'un item constituant la base de transactions.
3. Établir un critère de priorité pour ces items.
4. Faire le tri des items en fonction de leur priorité.
5. Établir le nœud racine.
6. À partir de chaque nœud père insérer les enfants en partant du nœud racine
7. Valider la structure de l'arbre FP-Growth.

Nous avons dressé une comparaison entre les trois algorithmes :
APRIORI, CLOSE, FP- Growth qui sera présentée dans la table suivant (**Table II 1**) :

Algorithme	Avantages	Inconvénients
APRIORI	• Découverte rapide de règles d'association pertinentes entre objets. • Facile à implémenter	• Génère un grand nombre de règles d'association. • Recherche de règles impose un temps considérable.
CLOSE	• Meilleure temps de réponse • Exhaustive	• Beaucoup de ressources de calcul
FP-Growth	• L'algorithme est considéré comme étant complet • La structure contient uniquement les objets fréquents classés par ordre de fréquence décroissante	• Construction de l'arbre peut être très longue. • Beaucoup de ressources de calcul

*Table II 1. **Comparaison entre les algorithmes d'extraction de règles d'association***

L'algorithme APRIORI permet la découverte rapide d'association néanmoins il produit un nombre important de règle d'association et les algorithmes CLOSE et FP-Growth nécessitent beaucoup de ressources de calcul.

Rappelons que notre domaine d'application est purement médical, nous avons donc jugé bon de jeter un coup d'œil sur les travaux dans le domaine médical qui utilisent les règles d'associations.

5. Conclusion

Nous venons de présenter les fondements de la recherche de règles d'association en citant également quelques domaines médicaux dans lesquels les règles d'association sont appliquées tel que la cancérologie, le diabète, etc... Nous avons cité le processus de l'extraction des règles, nous avons conclu que l'étape de recherche des itemsets fréquents constitue une phase importante dans ce processus et nous avons fait un tour d'horizon sur les algorithmes les plus utilisées dans la recherche des règles d'associations. L'algorithme APRIORI a présenté une solution pour beaucoup de cas d'étude.

La réalisation de ce chapitre a mis en exergue une suite logique à nos travaux sur l'extraction de règles d'association à partir des textes, à savoir la recherche des motifs fréquents, et la génération de règles d'association au moyen d'une mesure de qualité plus pertinente, par rapport à la dite mesure **confiance** d'Agrawal.

Chapitre III

Architecture et conception

1. Introduction

L'extraction des règles d'association est une méthode qui a vu le jour avec la recherche en bases de données (Documents textuels) pour retrouver les relations entre les termes d'une collection de documents.

Par exemple on sera capable de dire que 70% des clients qui achètent du lait achètent en même temps des œufs (lait → œuf : 0.70), une telle constatation est très intéressante puisqu'elle aide le gestionnaire d'un supermarché à ranger ses rayons de telle sorte que le lait et les œufs soient à proximité.

Plusieurs travaux basés sur la recherche des règles d'association ont été appliqués dans des applications réelles comme :

- La planification commerciale.
- Les réseaux de télécommunication.
- Les applications biomédicales.
- Les données de web.
- Les applications de sécurité.
- La gestion des connaissances.
- La recherche d'information.

Dans notre application on a utilisé l'algorithme **'APRIORI'** pour extraire ces règles d'associations en passant par plusieurs étapes **(Figure III 1)** :

- Sélection et prétraitement des données textuelles.
- Conversion du texte en matrice d'éléments
- Génération des règles d'association (En appliquant l'algorithme APRIORI).

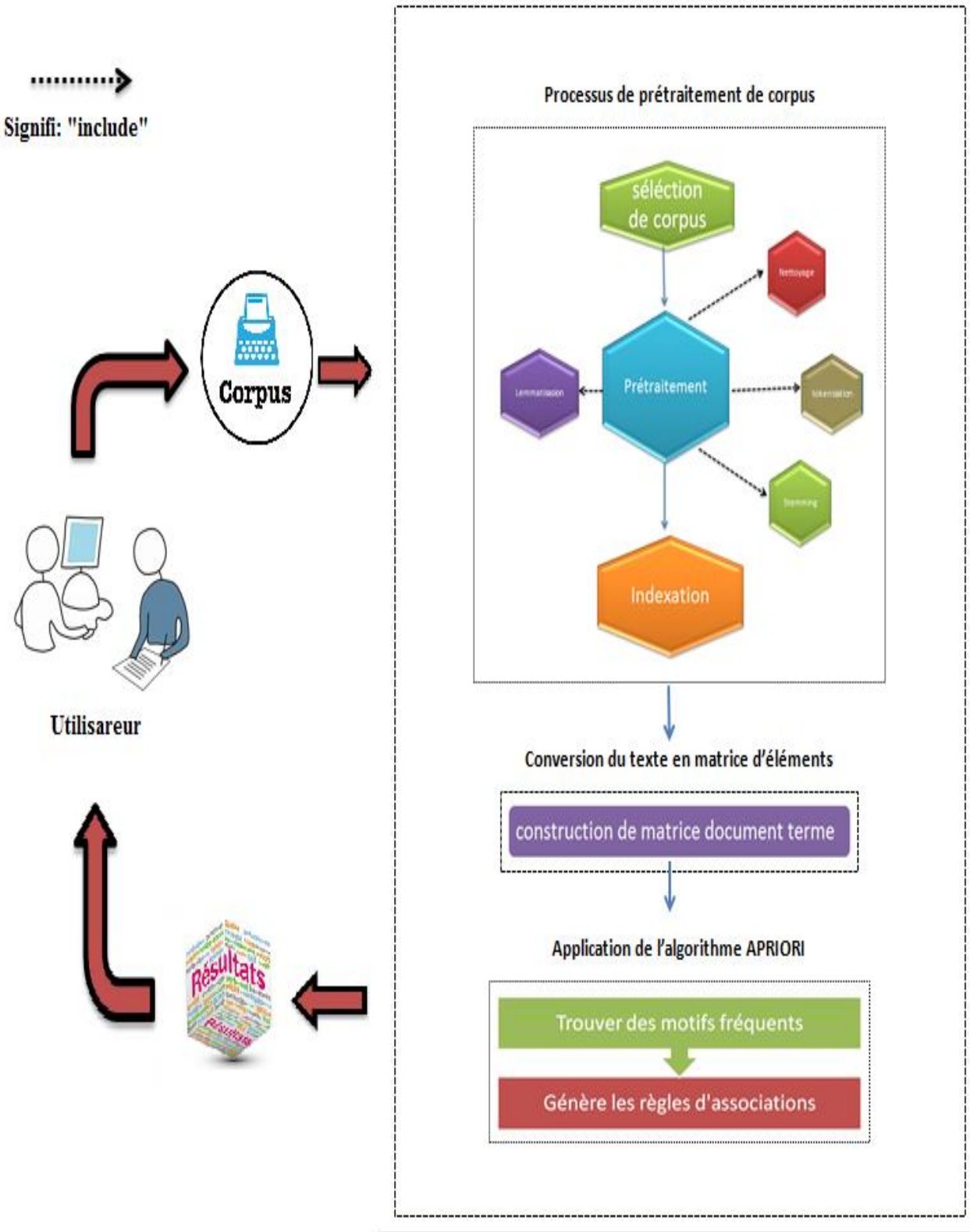

Figure III 1. ***Schéma du processus de l'application avec APRIORI***

2. Description du corpus indexé

Notre corpus portant sur les gênes et les protéines impliquées dans le cancer de la thyroïde et différents autres cancers associés. Le corpus est constitué de 6253 notices bibliographiques extraites de la base de donnée Medline — qui contient plus 10 millions de résumés et approximativement 40000 nouveaux résumés chaque mois — couvrant la période allant de 1965 jusqu'en septembre 2001. Ce corpus a été indexé dans le cadre d'un projet consistant à mettre en évidence les relations entre différentes entités biologiques [0]. Cette indexation a été réalisée avec deux importantes ressources lexicales dans les domaines de la médecine, la biologie et la génomique : UMLS (Unified Medical Language System) et LocusLink. [16]

Le résultat de l'indexation a été mis dans des balises SGML dans le but d'associer à chaque terme des catégories sémantiques bien spécifique. Les catégories jugées pertinentes sont au nombre de neuf :

Org	"Organisms",
Anat	"Anatomy",
Biol	"Biology, Physiology and sociology",
Bioc	"Biochemistry and Molecular Biology",
Mal	"Diseases",
Tech	"Analytical, Diagnostic and Therapeutic Techniques and Equipment",
Chem	"Chemicals and Drugs",
	"Psychology and sociology"
Misc	"Miscellaneous".

Nous donnons, ci-dessous, un exemple d'indexation :

tumorigenesis.

Dans cet exemple la séquence tumorigenesis est la séquence textuelle identifiée rencontrée dans le texte. Elle est mise à l'intérieur d'une balise SGML**. [16]**

La balise de début est:

– , elle contient deux attributs :

1– Le premier attribut "mc" représente l'appellation standard d'UMLS ou LocusLink pouvant être un synonyme de la séquence textuelle identifiée dans le texte (nommé aussi terme préférentiel). La valeur de cet attribut est donc "carcinogenesis".

2 – Le second attribut "class" représente la catégorie sémantique à laquelle appartient la séquence textuelle identifiée dans le texte. Elle a pour valeur "mal" (maladie).

– représente la fin de la balise SGML.

Comme nous avons pu l'observer en examinant les résultats de l'indexation, la plupart des termes ne sont pas standardisés et possèdent plusieurs appellations (synonymes). Si on n'arrive pas à faire le lien entre ces termes et leurs synonymes, la mise en évidence des relations liant les différentes catégories serait plus compliquée. Ainsi le terme "carcinogenesis" est le nom d'une maladie et a comme synonyme le terme "tumorigenesis". De plus, on notera que les traitements linguistiques de l'indexation permettent de repérer les séquences textuelles référant à des termes. Ainsi, le terme "cell transplant" correspondant à l'attribut "mc" peut référer par le jeu de transformations linguistiques à une séquence textuelle du type : "transplantation of monodispersed rat thyroid cells". Dans **[30]** il est montré que l'échec dans la reconnaissance des noms et des synonymes est le problème le plus important, et il est responsable de la non détection des interactions à hauteur de 44%. Pour cette raison, le balisage SGML effectué est une solution intéressante pour remédier à ce problème. Il règle en grande partie le problème de la synonymie et permet de rattacher chaque séquence textuelle identifiée (synonyme) à son terme préférentiel et à sa catégorie sémantique.

3. Description des étapes faisant partie du processus de génération des règles d'association

Comme le montre assez clairement la Figure III 1. **Schéma du processus de l'application avec APRIORI** pour pouvoir appliquer l'algorithme APRIORI sur le texte et extraire l'ensemble des règles d'association, nous sommes obligés de préparer notre texte en le faisant passer par un ensemble de prétraitements. Ces derniers sont :

- Sélection des données utiles à traiter,
- Le nettoyage de ces données : suppression des tags et des balises, ainsi que la suppression des mots vides.
- Normalisation des données : tokenization, stemming, lemmatization, etc.
- Conversion du texte en matrice d'éléments.

- Génération des règles d'association en appliquant l'algorithme APRIORI

Concernant ce dernier point, nous pouvons montrer en premier lieu, les étapes suivies par l'algorithme Apriori à travers cet organigramme (**Error! Reference source not found.**), puis à travers un exemple général son fonctionnement étape par étape.

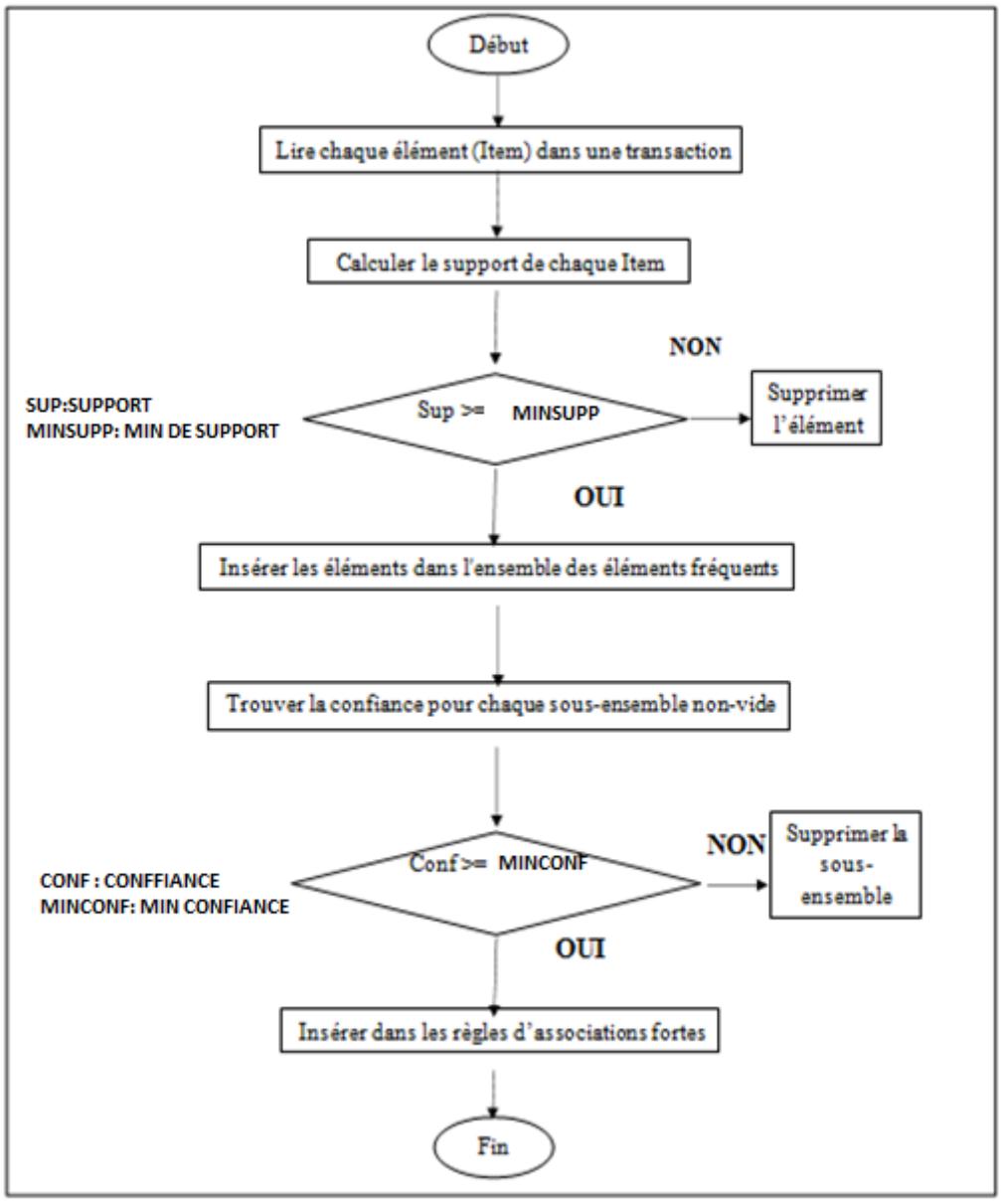

Figure III 2 ***Les étapes d'extraction des règles d'associations en utilisant l'algorithme Apriori***

La première étape consiste à lire le Fichier ligne par ligne pour récupérer son contenu. (Table III. 7)

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Table III. 7. ***Exemple de Data Set***

Exemple : On considère que le minimum support = 2 % et le minimum confiance = 60 %.

Etape 1 : k = 1

- Générer des ensembles d'éléments fréquents de longueur 1.
- Répétez jusqu'à ce qu'aucun nouvel ensemble d'éléments fréquents ne soit identifié.
  - Générer des ensembles d'éléments candidats de longueur (k + 1) à partir de la longueur k fréquente.
  - Elaguer les ensembles d'éléments candidats contenant des sous-ensembles de longueur k qui sont peu fréquents.
  - Comptez le support de chaque candidat en scannant le DB.
  - Éliminez les candidats non-fréquents, ne laissant que ceux qui sont fréquents.

Résultat de l'exemple :

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Table III. 8. ***La première liste candidate C1***

On compare le support des éléments de l'ensemble des candidats avec le support minimum (ici min-support = 2). Si support-count d'éléments de l'ensemble candidat est inférieur à min-support, supprimez ces éléments). Cela nous donne le itemset L1.

Etape 2 : k = 2

- Génération de l'ensemble candidat C2 à l'aide de L1 (c'est ce qu'on appelle l'étape de jointure). La condition de jonction de L(k-1) et L(k-1) est qu'il doit avoir des éléments (k-2) en commun.
- Vérification que tous les sous-ensembles (sub-sets) d'un ensemble d'éléments (Itemsets) sont fréquents ou non, et s'ils ne le sont pas, supprimez cet ensemble d'éléments. (Exemple de sous-ensemble de {I1, I2} sont {I1}, {I2} ils sont fréquents. Vérifiez pour chaque ensemble d'éléments).
- Recherche du support de ces itemsets en effectuant une recherche dans l'ensemble de données.

Résultat de l'exemple :

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

Table III. 3. ***La deuxième liste candidate C2***

Maintenant, on doit comparer le support candidat (C2) avec le nombre minimum de support (ici min-support = 2). Si support-count de l'itemset candidat est inférieur à min-support, donc ces itemsets seront supprimés). Cela nous donne l'ensemble d'éléments L2.

Résultat de l'exemple :

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

Table III. 4. ***La liste des itemsets L2***

Etape 3 : k = 3

- Génération de l'ensemble candidat C3 à l'aide de L2 (étape de jointure). La condition de jonction de L(k-1) et L(k-1) est qu'il doit avoir des éléments (k-2) en commun. Ici pour L2, le premier élément doit correspondre.

Donc l'ensemble d'éléments généré en rejoignant L2 est {I1, I2, I3} {I1, I2, I5} {I1, I3, i5} {I2, I3, I4} {I2, I4, I5} {I2, I3, I5}

- Vérification si tous les sous-ensembles de ces ensembles d'éléments sont fréquents ou non, et s'ils ne le sont pas, supprimez cet ensemble d'éléments. (Ici, les sous-ensembles de {I1, I2, I3} sont {I1, I2}, {I2, I3}, {I1, I3} qui sont fréquents. Pour {I2, I3, I4}, le sous-ensemble {I3, I4} n'est pas fréquent, alors supprimez-le. De même, vérifiez pour chaque ensemble d'éléments).
- Détermination du support de ces éléments restants en effectuant une recherche dans l'ensemble de données.

Résultat de l'exemple :

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Table III. 5. ***La troisième liste candidate C3***

Ensuite, il compare le support candidat (C3) avec le nombre minimum de support (ici min-support = 2). Si support-count de l'élément de l'ensemble candidat est inférieur à min-support, supprimez ces éléments) cela nous donne l'ensemble d'éléments L3.

Résultat de l'exemple :

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Table III. 6 ***La liste des itemsets L3***

Etape 4 : k = 4

- Génération de l'ensemble candidat C4 à l'aide de L3 (étape de jointure). La condition de jonction de L(k-1) et L(k-1) (ou k = 4) est qu'ils doivent avoir des éléments (k-2) en commun. Ici pour L3, les 2 premiers éléments (items) doivent correspondre.
- Il vérifie que tous les sous-ensembles de ces ensembles d'éléments sont fréquents ou non (ici l'ensemble d'éléments formé en joignant L3 est {I1, I2, I3, I5} donc son sous-ensemble contient {I1, I3, I5}, ce qui n'est pas fréquent). Donc pas d'itemset dans C4.
- Nous nous arrêtons ici car aucun itemsets fréquents sont trouvés.

Maintenant c'est l'étape pour extraire les règles d'associations.

La génération de toutes les règles d'associations possibles. Considérons le **min-sup** et le **min-conf** fournis par l'utilisateur.

Nous avons découvert tous les itemsets fréquents. Maintenant, la génération d'une règle d'association forte entre en scène. Pour cela, l'algorithme va calculer la confiance de chaque règle.

Confiance

Une confiance de 60% signifie que 60% des clients qui ont acheté du lait et du pain ont également acheté du beurre.

Résultat de l'exemple :

Si en prenant un exemple de n'importe quel itemset fréquent, nous montrerons la génération de règle.

Itemset {I1, I2, I3} // à partir de L3

Donc les règles générées :

[I1^I2]=>[I3] // confiance = sup(I1^I2^I3)/sup(I1^I2) = 2/4*100=50%
[I1^I3]=>[I2] // confiance = sup(I1^I2^I3)/sup(I1^I3) = 2/4*100=50%
[I2^I3]=>[I1] // confiance = sup(I1^I2^I3)/sup(I2^I3) = 2/4*100=50%
[I1]=>[I2^I3] // confiance = sup(I1^I2^I3)/sup(I1) = 2/6*100=33%
[I2]=>[I1^I3] // confiance = sup(I1^I2^I3)/sup(I2) = 2/7*100=28%
[I3]=>[I1^I2] // confiance = sup(I1^I2^I3)/sup(I3) = 2/6*100=33%

Remarque : Si la confiance minimale est = 50%, les 3 premières règles peuvent être considérées comme des règles d'association fortes.

Conclusion :

Plusieurs algorithmes existent pour l'extraction des règles d'association, Le plus connu est l'algorithme Apriori

Dans ce travail nous avons l'implémentation l'algorithme A-priori et applique sur corpus indexé pour l'extraction des règles d'association par suivies plusieurs étapes.

Dans ce chapitre nous avons calcules le support et confiance pour itemset pour génères les règles plus fort.

Chapitre IV

Implémentation

1. Implémentation

L'implémentation est la phase la plus importante après celle de l'architecture et la conception. Le choix des outils de développement influence énormément sur le coût en temps de programmation, ainsi que sur la flexibilité du produit à réaliser.

2. Outils et langages utilisés

2.1 Langage de programmation

Le langage de programmation Python a été créé en 1989 par Guido van Rossum, aux Pays-Bas. Le nom *Python* vient d'un hommage à la série télévisée *Monty Python's Flying Circus* dont G. van Rossum est fan. La première version publique de ce langage a été publiée en 1991.

La dernière version de Python est la version 3. Plus précisément, la version 3.7 a été publiée en juin 2018. La version 2 de Python est désormais obsolète et cessera d'être maintenue après le 1er janvier 2020. Dans la mesure du possible évitez de l'utiliser.

La *Python Software Foundation* est l'association qui organise le développement de Python et anime la communauté de développeurs et d'utilisateurs.

Ce langage de programmation présente de nombreuses caractéristiques intéressantes :

— Il est multiplateforme. C'est-à-dire qu'il fonctionne sur de nombreux systèmes d'exploitation : Windows, Mac OS X, Linux, Android, iOS, depuis les mini-ordinateurs Raspberry Pi jusqu'aux supercalculateurs.

— Il est gratuit. On peut l'installer sur autant d'ordinateurs qu'on veut (même sur le téléphone!).

— C'est un langage de haut niveau. Il demande relativement peu de connaissance sur le fonctionnement d'un ordinateur pour être utilisé.

— C'est un langage interprété. Un script Python n'a pas besoin d'être compilé pour être exécuté, contrairement à des langages comme le C ou le C++.

— Il est orienté objet. C'est-à-dire qu'il est possible de concevoir en Python des entités qui miment celles du monde réel (une cellule, une protéine, un atome, etc.) avec un certain nombre de règles de fonctionnement et d'interactions.

— Il est relativement *simple* à prendre en main.

— Enfin, il est très utilisé en bio-informatique et plus généralement en analyse de données. **[13]**

2.2 Outil de développement :

Jupyter est une application web utilisée pour programmer dans plus de 40 langages de programmation, dont Python, Julia, Ruby, R, ou encore Scala2. C'est un projet communautaire dont l'objectif est de développer des logiciels libres, des formats ouverts et des services pour l'informatique interactive. Jupyter est une évolution du projet IPython. Jupyter permet de réaliser des calepins ou notebooks, c'est-à-dire des programmes contenant à la fois du texte en markdown et du code. Ces calepins sont utilisés en science des données pour explorer et analyser des données.

Jupyter Notebook (anciennement IPython Notebooks) est un environnement de programmation interactif basé sur le Web permettant de créer des documents Jupyter Notebook. Le terme "notebook" peut faire référence à de nombreuses entités différentes, adaptées au contexte, telles que l'application web Jupyter, le serveur web Jupyter Python ou le format de document Jupyter.

Un document Jupyter Notebook est un document JSON. Il suit un schéma contenant une liste ordonnée de cellules d'entrée/sortie. Celles-ci peuvent contenir du code, du texte (à l'aide de Markdown), des formules mathématiques, des graphiques et des médias interactifs. Ce document se termine généralement par l'extension ".ipynb".

2.3 Les packages utilisés :

On a utilisé différentes bibliothèques de Python dans notre application tels que :

‘NLTK’, ‘WORDCLOUD’, ‘MLXTEND’…

2.3.1 Natural Language Toolkit (NLTK)

C’est une bibliothèque logicielle en Python permettantun traitement automatique des langues, développée par Steven Bird et Edward Loper du département d'informatique de l'université de Pennsylvanie. En plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'interface de programmation (API). **[32]**

2.3.2 WORDCLOUD:

Le nuage de mots-clés, ou nuage de tags (en anglais tag cloud, word cloud ou keyword cloud2) est une représentation visuelle des mots-clés (tags) les plus utilisés sur un texte. Généralement, les mots s'affichent dans des tailles et graisses de caractères d'autant plus visibles qu'ils sont utilisés ou populaires.**[33]**

Le package Wordcloud nous aide à connaître la fréquence d'un mot dans un contenu textuel à l'aide de la visualisation.

Pour implémenter cela, nous devons d'abord installer certains packages, tels que pandas, matplotlib et Wordcloud**.[34]**

2.3.3 MLXTEND:

Machine Learning extensions (extensions d'apprentissage automatique) est une bibliothèque Python d'outils utiles pour les tâches quotidiennes de science des données. Il se compose de nombreux outils utiles pour les tâches de science des données et d'apprentissage automatique, par exemple : Sélection de fonctionnalité, Extraction de caractéristiques, Visualisation, Assemblage, Et beaucoup plus**.[35]**

Cette bibliothèque nous a été très utile, elle nous a permis de réduire considérablement le temps de chargement des données. Sans elle, le chargement non complet a pris plus de 2 heures. ()

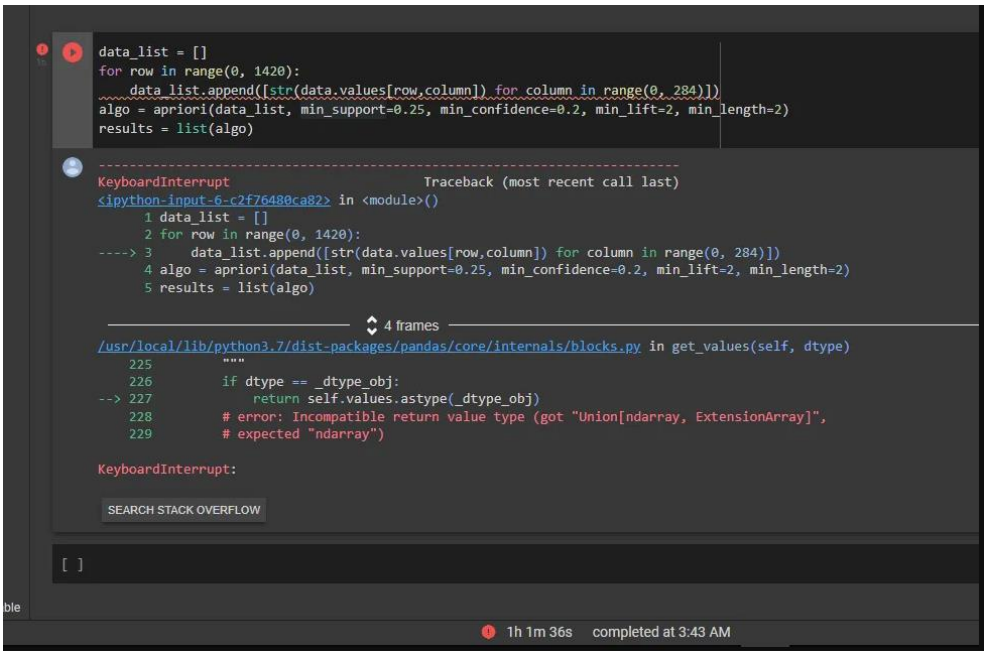

Figure IV 1. Lenteur du chargement des données sans la bibliothèque MlExtend

3. La génération des règles d'association

3.1 Sélection et prétraitement des données textuelles:

3.1.1 Sélection des données textuelles:

Cette étape permet de préparer les données afin de leur appliquer les algorithmes d'extraction des règles d'association.

La seule chose qu'on a en entrée est un fichier texte (**.TXT**), pour le manipuler, nous avons décidé d'adopter l'approche habituelle, et le convertir en un fichier (**.CSV**), nous l'avons fait manuellement. **(**Figure IV **)**

```
Entrée [5]: data.head()
```

Out[5]:

	id	text
0	0	Life threatening giant mediastinal goiter : a surgical challenge . Mediastinal goiter is a well known benign disease , usually resectable through a cervical approach with minimal morbidity and mortality.
1	0	Only occasionally a median sternotomy or a lateral thoracotomy may be required.
2	0	The present case is worthy of presentation because of the exceptional dimension of the disease and the surgical challenge that it presented.
3	0	In a 72 year old woman a large intrathoracic goiter of the right thorax caused a severe dyspnoea due to an important contralateral mediastinal shift with compression of the lung , superior vena cava system and trachea.
4	0	At surgical exploration , through a cervico sternotomic approach , the mediastinal structures dislocation and the strong adherences between the anomalous neovascularized capsula of the mass and the surrounding structures , complicated the surgical dissection.

Figure IV 2. ***Lecture de fichier bio.csv***

6.1.2. Nettoyage de données textuelles:

Comme le montre La figure précédente, notre fichier a besoin de beaucoup de nettoyage (cas des données semi-structuré)

La première étape est l'élimination des tags et des balises (gestion du contexte), cette étape inclut aussi la suppression des mots vides **(Annexe A)**.

En se basant sur ces informations et dans le but de réduire la taille du fichier, nous avons décidé de supprimer toutes les parties qui se trouvent à l'intérieur des balises, c'est-à-dire du début de la balise jusqu'à la fin de cette dernière (voir Figure IV)

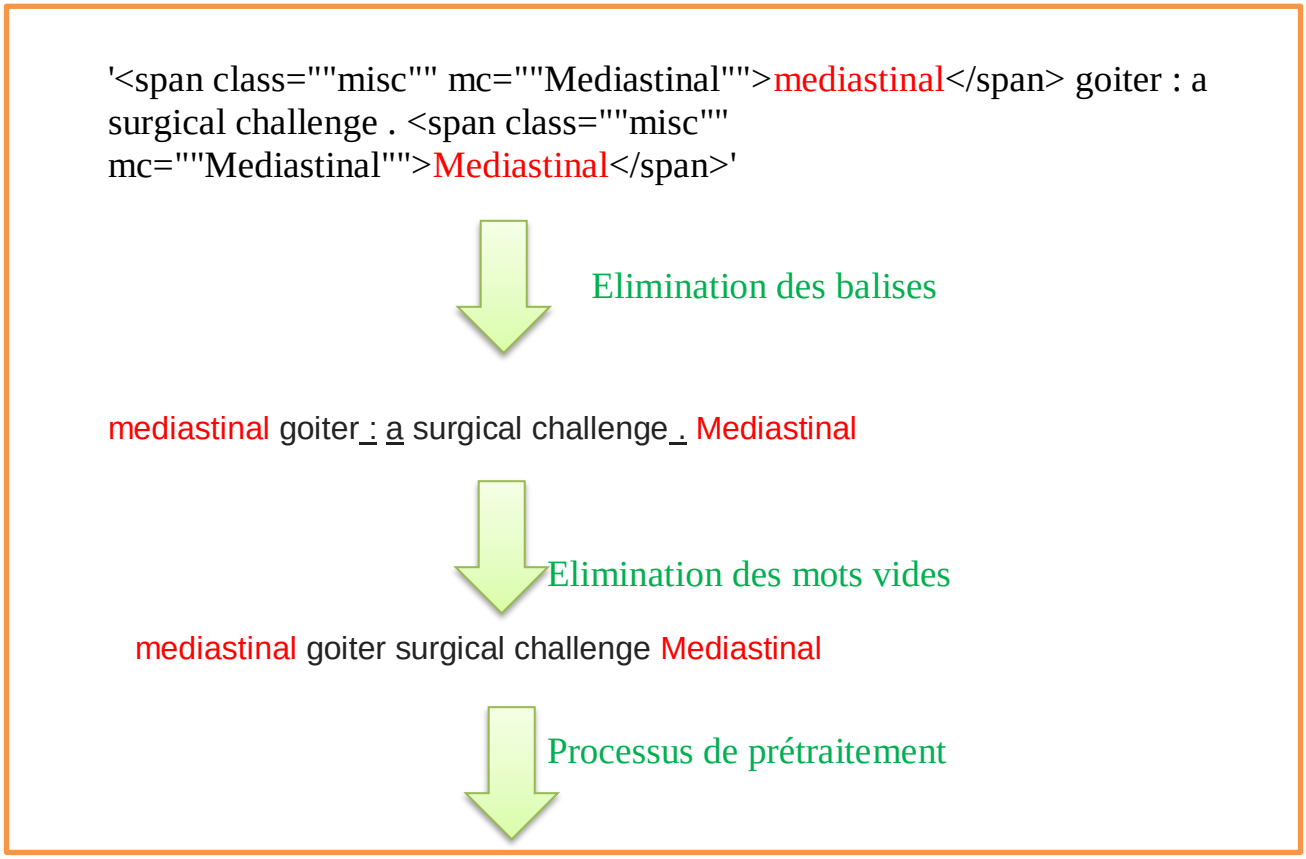

Figure IV 3. Elimination des tags et des balises

Ensuite, on passe à la reconnaissance des termes, des ponctuations, des fins de paragraphe et des phrases ainsi que l'unification de l'écriture des lettres en minuscule afin de faciliter la mise en correspondance lors de la phase de programmation. Pour cela, nous avons utilisé plusieurs fonctions pour le nettoyage et la normalisation du corpus. **(Annexe B)**

Une fois l'étape précédente terminée, on efface les données bruyantes et on extrait le contenu jugé pertinent se trouvant entre les balises **(**Figure IV **)**.

Out[25]:

	id	text	classes	mcs
0	0	Life threatening giant mediastinal goiter a surgical challenge Mediastinal goiter is a well known benign disease usually resectable through a cervical approach with minimal morbidity and mortality	misc misc misc	Mediastinal Mediastinal Cervical approach
1	0	Only occasionally a median sternotomy or a lateral thoracotomy may be required	tech misc tech	Sternotomy Lateral Thoracotomy
2	0	The present case is worthy of presentation because of the exceptional dimension of the disease and the surgical challenge that it presented		
3	0	In a 72 year old woman a large intrathoracic goiter of the right thorax caused a severe dyspnoea due to an important contralateral mediastinal shift with compression of the lung superior vena cava system and trachea	org mal anat misc mal misc misc tech anat anat anat	Elderly woman Intrathoracic Goiter Thorax Severe Dyspnea Contralateral Mediastinal Compression procedure Lung Superior Vena Cava Trachea
4	0	At surgical exploration through a cervico sternotomic approach the mediastinal structures dislocation and the strong adherences between the anomalous neovascularized capsula of the mass and the surrounding structures complicated the surgical dissection	misc misc tech	Mediastinal Circumferential Dissection

*Figure IV 4.**Résultat d'extraction du contenu des balises du corpus***

6.1.3. Normalisation de données :

Ci-dessous **(**Figure IV 5. **Nuage des mots les plus fréquents dans le corpus),** les mots les plus fréquents dans notre corpus :

"thyroide", "carcinoma", "tumer"

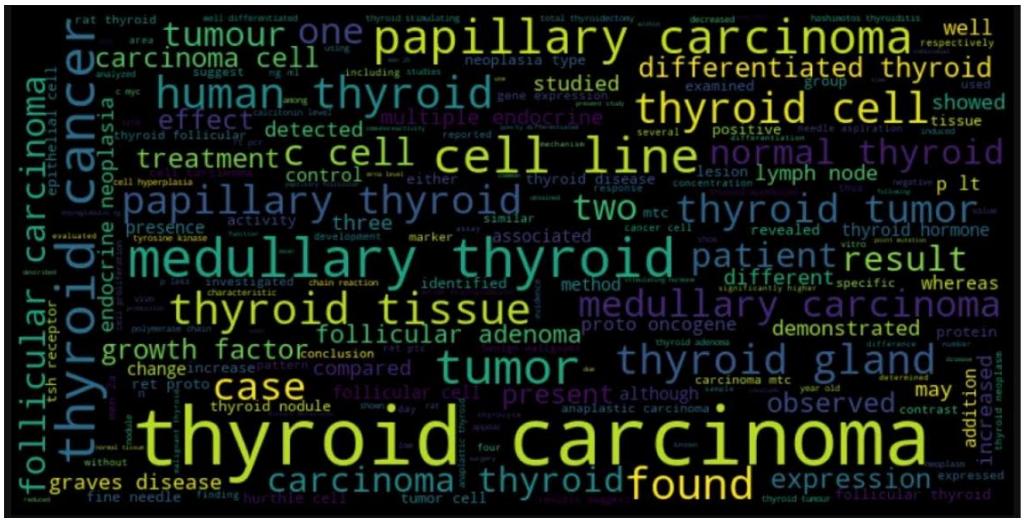

*Figure IV 5. **Nuage des mots les plus fréquents dans le corpus***

La normalisation passe par la tokenization. Ce dernier processus permet de découpé un flux de caractères en phrases, symboles, mots. Nous nous intéressons plus particulièrement aux mots *'tokens'*. **(Annexe C)**

Cette fonction permet de découpée le texte en plusieurs pièces appelés *token*, et voilà ci-dessous le résultat de notre corpus **(**Figure IV **).**

Out[55]:

	new
0	[life, threatening, giant, mediastinal, goiter, surgical, challenge, mediastinal, goiter, well, known, benign, disease, usually, resectable, cervical, approach, minimal, morbidity, mortalitymisc, misc, miscmediastinal, mediastinal, cervical, approach, occasionally, median, sternotomy, lateral, thoracotomy, may, requiredtech, misc, techsternotomy, lateral, thoracotomy, present, case, worthy, presentation, exceptional, dimension, disease, surgical, challenge, presented, 72, year, old, woman, large, intrathoracic, goiter, right, thorax, caused, severe, dyspnoea, due, important, contralateral, mediastinal, shift, compression, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anatelderly, woman, intrathoracic, goiter, thorax, severe, dyspnea, contralateral, mediastinal, compression, procedure, lung, superior, vena, cava, trachea, surgical, exploration, cervico, sternotomic, ...]
1	[c, cell, thyroid, epithelial, tumours, altered, follicular, development, transgenic, mice, expressing, long, isoform, men, 2a, ret, biol, org, protdevelopment, aspects, transgenic, mice, ret, gain, function, mutations, gene, encoding, receptor, tyrosine, kinase, ret, identified, aetiological, factor, multiple, endocrine, neoplasia, type, 2a, men2a, mal, prot, malmutation, ret, multiple, endocrine, neoplasia, type, 2a, men2a, dominantly, inherited, cancer, predisposition, syndrome, characterized, medullary, thyroid, carcinoma, tumour, calcitonin, producing, thyroid, c, cellsbiol, mal, protpredisposition, medullary, thyroid, carcinoma, calcitonin, three, isoforms, ret, ret9, ret43, ret51, although, vitro, evidence, suggests, vary, cellular, transformation, activities, little, known, function, tumorigenesis, vivo, prot, misc, mal, miscret, vitro, carcinogenesis, vivo, address, used, ...]
2	[multivariate, analysis, clinicopathologic, parameters, insular, subtype, differentiated, thyroid, carcinoma, hypothesis, insular, carcinoma, represents, aggressive, subtype, differentiated, thyroid, cancer, multivariate, analysis, controlling, various, clinicopathologic, parametersmalthyroid, cancer, design, retrospective, analysis, setting, tertiary, referral, center, university, hospital, misc, misctertiary, university, hospital, patients, one, hundred, twenty, seven, consecutive, patients, histological, diagnosis, follicular, variant, papillary, thyroid, carcinoma, follicular, thyroid, carcinoma, main, outcome, measure, logistic, regression, model, used, examine, relationship, various, clinicopathologic, parameters, insular, subtypetech, mal, mal, misc, mischistological, diagnosis, papillary, thyroid, carcinoma, follicular, thyroid, carcinoma, primary, treatment, outcome, results, insular, subtype, involved, 14, 127, tumors, unlike, extrathyroidal, extension, nodal, metastasis, primary, tumor, diameter, gpt, 40, ...]
3	[thyroid, hormones, thyroid, antibodies, infertile, malesprot, bioc, malthyrotropin, thyroid, antibody, infertility, objective, investigate, incidence, thyroid, dysfunction, thyroid, antibodies, correlation, semen, hormonal, parameters, infertile, menmisc, mal, bioc, bioc, misc, malincidence, thyroid, dysfunction, thyroid, antibody, semen, hormonal, infertility, design, prospective, study, techprospective, study, setting, university, based, andrology, laboratory, misc, miscandrology, laboratory, patient, three, hundred, five, infertile, men, idiopathic, infertilitymalinfertility, intervention, medical, history, clinical, examination, semen, analysis, measurement, free, thyroxin, ft4, free, triiodothyronine, ft3, basal, thyroid, stimulating, hormone, btsh, lh, fsh, free, testosterone, ft, prl, e2, sex, hormone, binding, globulin, shbg, dheas, thyroid, antibodies, thyreoglobulin, antibody, tga, thyroid, peroxidase, antibody, tpo, ab, thyroid, ...]
4	[papillary, thyroid, carcinoma, thyroglossal, duct, cyst, comparative, cytohistologic, immunochemical, study, 2, new, cases, review, literaturemal, mal, techpapillary, thyroid, carcinoma, thyroglossal, cyst, comparative, study, report, cytohistologic, immunohistochemical, study, 2, cases, papillary, thyroid, carcinoma, occurring, thyroglossal, duct, cyst, mal, malpapillary, thyroid, carcinoma, thyroglossal, cyst, patients, 21, year, old, woman, 48, year, old, man, org, orgelderly, woman, elderly, man, needle, aspiration, cytology, smears, consistent, papillary, thyroid, carcinoma, misc, malcytology, papillary, thyroid, carcinoma, sistrunk, procedure, done, papillary, carcinoma, found, within, thyroglossal, duct, cyst, mal, malpapillary, thyroid, carcinoma, thyroglossal, cyst, 1, case, tumor, spread, outside, cyst, anatcyst, follow, uneventful, patients, 2, 9, years, respectively, results, ...]

Figure IV 6. ***Résultat de la Tokenization du corpus***

La lemmatisation ainsi que le stemming sont deux techniques (ou méthodes) qui ont pour but de réduire la dimension du vecteur document.

Cette réduction permet de réduire le nombre de mots considérés. Pour cela, il est possible de rassembler les mots faisant partie de la même famille ou possédant la même racine. C'est le but de la *lemmatisation* et du *stemming*.

En effet, la lemmatisation permet d'extraire les formes canoniques des mots. Par contre, le stemming utilise la racine des mots plus la dérivation (préfixe, suffixe).

La lemmatisation et le stemming peuvent induire une perte d'information. En effet, des degrés de précision disparaissent par rapport au texte initial tel que le pluriel ou le temps de conjugaison.

C'est pourquoi nous avons essayé les deux méthodes, nous verrons la différence entre eux afin pour pouvoir faire le bon choix afin de réduire l'espace de dimension en une dimension moindre tout en perdant le moins d'information possible. **(Annexe D)**

Ci-dessous le résultat d'exécution de ces fonctions sur notre corpus (Figure IV) :

```
Entrée [59]: data.head(2)
```

Out[59]:

	new	stemmed	lemmed
0	[life, threatening, giant, mediastinal, goiter, surgical, challenge, mediastinal, goiter, well, known, benign, disease, usually, resectable, cervical, approach, minimal, morbidity, mortalitymisc, misc, miscmediastinal, mediastinal, cervical, approach, occasionally, median, sternotomy, lateral, thoracotomy, may, requiredtech, misc, techsternotomy, lateral, thoracotomy, present, case, worthy, presentation, exceptional, dimension, disease, surgical, challenge, presented, 72, year, old, woman, large, intrathoracic, goiter, right, thorax, caused, severe, dyspnoea, due, important, contralateral, mediastinal, shift, compression, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anatelderly, woman, intrathoracic, goiter, thorax, severe, dyspnea, contralateral, mediastinal, compression, procedure, lung, superior, vena, cava, trachea, surgical, exploration, cervico, sternotomic, ...]	[life, threaten, giant, mediastin, goiter, surgic, challeng, mediastin, goiter, well, known, benign, diseas, usual, resect, cervic, approach, minim, morbid, mortalitymisc, misc, miscmediastin, mediastin, cervic, approach, occasion, median, sternotomi, later, thoracotomi, may, requiredtech, misc, techsternotomi, later, thoracotomi, present, case, worthi, present, except, dimens, diseas, surgic, challeng, present, 72, year, old, woman, larg, intrathorac, goiter, right, thorax, caus, sever, dyspnoea, due, import, contralater, mediastin, shift, compress, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anateld, woman, intrathorac, goiter, thorax, sever, dyspnea, contralater, mediastin, compress, procedur, lung, superior, vena, cava, trachea, surgic, explor, cervico, sternotom, ...]	[life, threatening, giant, mediastinal, goiter, surgical, challenge, mediastinal, goiter, well, known, benign, disease, usually, resectable, cervical, approach, minimal, morbidity, mortalitymisc, misc, miscmediastinal, mediastinal, cervical, approach, occasionally, median, sternotomy, lateral, thoracotomy, may, requiredtech, misc, techsternotomy, lateral, thoracotomy, present, case, worthy, presentation, exceptional, dimension, disease, surgical, challenge, presented, 72, year, old, woman, large, intrathoracic, goiter, right, thorax, caused, severe, dyspnoea, due, important, contralateral, mediastinal, shift, compression, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anatelderly, woman, intrathoracic, goiter, thorax, severe, dyspnea, contralateral, mediastinal, compression, procedure, lung, superior, vena, cava, trachea, surgical, exploration, cervico, sternotomic, ...]
1	[c, cell, thyroid, epithelial, tumours, altered, follicular, development, transgenic, mice, expressing, long, isoform, men, 2a, ret, biol, org, protdevelopment, aspects, transgenic, mice, ret, gain, function, mutations, gene, encoding, receptor, tyrosine, kinase, ret, identified, aetiological, factor, multiple, endocrine, neoplasia, type, 2a, men2a, mal, prot, malmutation, ret, multiple, endocrine, neoplasia, type, 2a, men2a, dominantly, inherited, cancer, predisposition, syndrome, characterized, medullary, thyroid, carcinoma, tumour, calcitonin, producing, thyroid, c, cellsbiol, mal, protpredisposition, medullary, thyroid, carcinoma, calcitonin, three, isoforms, ret, ret9, ret43, ret51, although, vitro, evidence, suggests, vary, cellular, transformation, activities, little, known, function, tumorigenesis, vivo, prot, misc, mal, miscret, vitro, carcinogenesis, vivo, address, used, ...]	[c, cell, thyroid, epitheli, tumour, alter, follicular, develop, transgen, mice, express, long, isoform, men, 2a, ret, biol, org, protdevelop, aspect, transgen, mice, ret, gain, function, mutat, gene, encod, receptor, tyrosin, kinas, ret, identifi, aetiolog, factor, multipl, endocrin, neoplasia, type, 2a, men2a, mal, prot, malmut, ret, multipl, endocrin, neoplasia, type, 2a, men2a, domin, inherit, cancer, predisposit, syndrom, character, medullari, thyroid, carcinoma, tumour, calcitonin, produc, thyroid, c, cellsbiol, mal, protpredisposit, medullari, thyroid, carcinoma, calcitonin, three, isoform, ret, ret9, ret43, ret51, although, vitro, evid, suggest, vari, cellular, transform, activ, littl, known, function, tumorigenesi, vivo, prot, misc, mal, miscret, vitro, carcinogenesi, vivo, address, use, ...]	[c, cell, thyroid, epithelial, tumour, altered, follicular, development, transgenic, mouse, expressing, long, isoform, men, 2a, ret, biol, org, protdevelopment, aspect, transgenic, mouse, ret, gain, function, mutation, gene, encoding, receptor, tyrosine, kinase, ret, identified, aetiological, factor, multiple, endocrine, neoplasia, type, 2a, men2a, mal, prot, malmutation, ret, multiple, endocrine, neoplasia, type, 2a, men2a, dominantly, inherited, cancer, predisposition, syndrome, characterized, medullary, thyroid, carcinoma, tumour, calcitonin, producing, thyroid, c, cellsbiol, mal, protpredisposition, medullary, thyroid, carcinoma, calcitonin, three, isoforms, ret, ret9, ret43, ret51, although, vitro, evidence, suggests, vary, cellular, transformation, activity, little, known, function, tumorigenesis, vivo, prot, misc, mal, miscret, vitro, carcinogenesis, vivo, address, used, ...]

Figure IV 7. ***Résultat du Stemming et Lemmatisation du corpus***

Nous avons choisi d'utiliser la lemmatization car le *stemming* nous fait perdre le sens réel des mots dans certains cas (Figure IV **).**

Etant donné que le domaine d'étude est médical, on ne peut pas risquer de perdre le sens des mots, parce que cela affectera négativement l'exactitude des informations ainsi que le résultat attendu.

	new	stemmed	lemmed
0	[life, threatening, giant, mediastinal, goiter, surgical, challenge, mediastinal, goiter, well, known, benign, disease, usually, resectable, cervical, approach, minimal, morbidity, mortalitymisc, misc, miscmediastinal, mediastinal, cervical, approach, occasionally, median, sternotomy, lateral, thoracotomy, may, requiredtech, misc, techsternotomy, lateral, thoracotomy, present, case, worthy, presentation, exceptional, dimension, disease, surgical, challenge, presented, 72, year, old, woman, large, intrathoracic, goiter, right, thorax, caused, severe, dyspnoea, due, important, contralateral, mediastinal, shift, compression, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anatelderly, woman, intrathoracic, goiter, thorax, severe, dyspnea, contralateral, mediastinal, compression, procedure, lung, superior, vena, cava, trachea, surgical, exploration, cervico, sternotomic, ...]	[life, threaten, giant, mediastin, goiter, surgic, challeng, mediastin, goiter, well, known, benign, diseas, usual, resect, cervic, approach, minim, morbid, mortalitymisc, misc, miscmediastin, mediastin, cervic, approach, occasion, median, sternotomi, later, thoracotomi, may, requiredtech, misc, techsternotomi, later, thoracotomi, present, case, worthi, present, except, dimens, diseas, surgic, challeng, present, 72, year, old, woman, larg, intrathorac, goiter, right, thorax, caus, sever, dyspnoea, due, import, contralater, mediastin, shift, compress, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anateld, woman, intrathorac, goiter, thorax, sever, dyspnea, contralater, mediastin, compress, procedur, lung, superior, vena, cava, trachea, surgic, explor, cervico, sternotom, ...]	[life, threatening, giant, mediastinal, goiter, surgical, challenge, mediastinal, goiter, well, known, benign, disease, usually, resectable, cervical, approach, minimal, morbidity, mortalitymisc, misc, miscmediastinal, mediastinal, cervical, approach, occasionally, median, sternotomy, lateral, thoracotomy, may, requiredtech, misc, techsternotomy, lateral, thoracotomy, present, case, worthy, presentation, exceptional, dimension, disease, surgical, challenge, presented, 72, year, old, woman, large, intrathoracic, goiter, right, thorax, caused, severe, dyspnoea, due, important, contralateral, mediastinal, shift, compression, lung, superior, vena, cava, system, trachea, org, mal, anat, misc, mal, misc, misc, tech, anat, anat, anatelderly, woman, intrathoracic, goiter, thorax, severe, dyspnea, contralateral, mediastinal, compression, procedure, lung, superior, vena, cava, trachea, surgical, exploration, cervico, sternotomic, ...]

Figure IV 8. ***Comparaison entre Stemming et Lemmatisation***

Une fois toutes ces méthodes de réduction de dimension appliquées, l'étape de la conversion du texte en matrice d'éléments peut commencer.

6.2. Conversion du texte en matrice d'éléments

La base de données (texte) sera transformée en matrice, dont la dimension est optimisée. Le code utilisé pour effectuer cette tâche est montré dans l'**Annexe E**

Out[84]:

	0	1	2	3	4	5	6	7	8	9	...	262	263	264	265
0	life	threatening	giant	mediastinal	goiter	surgical	challenge	well	known	benign	...	None	None	None	None
1	c	cell	thyroid	epithelial	tumour	altered	follicular	development	transgenic	mouse	...	None	None	None	None
2	multivariate	analysis	clinicopathologic	parameter	insular	subtype	differentiated	thyroid	carcinoma	hypothesis	...	None	None	None	None
3	thyroid	hormone	antibody	infertile	malesprot	bioc	malthyrotropin	infertility	objective	investigate	...	None	None	None	None
4	papillary	thyroid	carcinoma	thyroglossal	duct	cyst	comparative	cytohistologic	immunochemical	study	...	None	None	None	None

Figure IV 9. ***Résultat de la conversion du texte en matrice d'éléments***

Nous pouvons voir que nous avons deux problèmes à résoudre (Figure IV) :

1. Le premier est la présence des valeurs nulles, parce que la longueur de chaque document (transaction) est différente de l'autre.

   Réglons ce problème en remplissant les valeurs nulles avec des chaînes vides (caractère blanc), cela évitera les erreurs lors de l'appel de l'algorithme.

2. Le deuxième est le singulier et le plurielle, par exemple : [Cell , Cells] ces deux mots veulent dire la même chose. Mais l'algorithme sera différent entre eux. Nous devrons singulariser les mots en plurielle et supprimer les doublons par ligne. Cela optimisera également l'ensemble de données.

Le résultat est assez significatif, nous sommes partis d'un corpus contenant ***6254*** documents, et après l'application de toutes ces méthodes de réduction de dimension, on obtient finalement une matrice de ***5136*** lignes et ***271*** colonnes.

Différents algorithmes d'extraction des règles d'association existent. Notre choix s'est porté sur l'algorithme apriori.

6.3. Application de l'algorithme APRIORI pour la génération des règles d'association :

Comme déjà mentionné dans les sections précédentes, les règles d'association nous permettent d'identifier l'ensemble d'éléments ou attributs qui se produisent ensemble dans une table ou dans un texte. La génération de cet ensemble de règles passe par plusieurs étapes.

Après application de l'algorithme sur nos données textuelles

Par exemple, nous pouvons imprimer tous les éléments d'une longueur de 3 et le support minimum est supérieur ou égale à 0,03

```
Entrée [58]: # printing the frequntly items
             #distinct_l = set(l)
             (frequent_itemsets[ (frequent_itemsets['length'] == 3) &
                                 (frequent_itemsets['support'] >= 0.03) ]).drop_duplicates()
```

Out[58]:

	support	itemsets	length
376	0.128115	(adenoma, carcinoma, thyroid)	3
377	0.106893	(adenoma, cell, thyroid)	3
378	0.114097	(adenoma, follicular, thyroid)	3
379	0.106114	(cell, carcinoma, also)	3
380	0.159852	(carcinoma, also, thyroid)	3
...	...	...	...
558	0.148949	(study, tumor, thyroid)	3
559	0.102998	(two, study, thyroid)	3
560	0.133178	(tissue, tumor, thyroid)	3
561	0.111176	(two, tumor, thyroid)	3
562	0.102804	(type, tumor, thyroid)	3

187 rows × 3 columns

Création des règles d'association

Nous savons que les règles d'association sont simplement les instructions if-else. Le composant IF d'une règle d'association est appelé antécédent. La composante ALORS est connue sous le nom de conséquente. L'antécédent et le conséquent sont disjoints ; ils n'ont aucun élément en commun. Alors, créons des antécédents et des conséquents

Entrée [59]:
```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=0.1)
rules["antecedents_length"] = rules["antecedents"].apply(lambda x: len(x))
rules["consequents_length"] = rules["consequents"].apply(lambda x: len(x))
rules.sort_values("lift",ascending=False)
```

Out[59]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedents_length	consequents_length
1673	(medullary, thyroid)	(carcinoma, calcitonin)	0.280179	0.129478	0.124026	0.442669	3.418865	0.087749	1.561946	2	2
1672	(carcinoma, calcitonin)	(medullary, thyroid)	0.129478	0.280179	0.124026	0.957895	3.418865	0.087749	17.095746	2	2
1668	(carcinoma, calcitonin, thyroid)	(medullary)	0.128894	0.282321	0.124026	0.962236	3.408305	0.087637	19.004143	3	1
1677	(medullary)	(carcinoma, calcitonin, thyroid)	0.282321	0.128894	0.124026	0.439310	3.408305	0.087637	1.553633	1	3
591	(carcinoma, calcitonin)	(medullary)	0.129478	0.282321	0.124416	0.960902	3.403582	0.087861	18.356024	2	1
...	...	...	...	...	...	...	...	...	...	...	...
1845	(cell)	(patient, carcinoma, thyroid)	0.572625	0.279206	0.124611	0.217613	0.779401	-0.035269	0.921276	1	3
714	(cell)	(patient, carcinoma)	0.572625	0.282516	0.125974	0.219993	0.778694	-0.035802	0.919844	1	2
711	(patient, carcinoma)	(cell)	0.282516	0.572625	0.125974	0.445899	0.778694	-0.035802	0.771296	2	1
189	(cell)	(patient)	0.572625	0.424065	0.188084	0.328460	0.774550	-0.054746	0.857632	1	1
188	(patient)	(cell)	0.424065	0.572625	0.188084	0.443526	0.774550	-0.054746	0.768006	1	1

2100 rows × 11 columns

La sortie affiche les valeurs de divers composants de support de l'algorithme Apriori.

Pour obtenir plus d'informations à partir des données, trions-les données par la valeur de « Lift » :

Entrée [60]:
```
# Sort values based on confidence
rules.sort_values("lift",ascending=False)
```

Out[60]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedents_length	consequents_length
1673	(medullary, thyroid)	(carcinoma, calcitonin)	0.280179	0.129478	0.124026	0.442669	3.418865	0.087749	1.561946	2	2
1672	(carcinoma, calcitonin)	(medullary, thyroid)	0.129478	0.280179	0.124026	0.957895	3.418865	0.087749	17.095746	2	2
1668	(carcinoma, calcitonin, thyroid)	(medullary)	0.128894	0.282321	0.124026	0.962236	3.408305	0.087637	19.004143	3	1
1677	(medullary)	(carcinoma, calcitonin, thyroid)	0.282321	0.128894	0.124026	0.439310	3.408305	0.087637	1.553633	1	3
591	(carcinoma, calcitonin)	(medullary)	0.129478	0.282321	0.124416	0.960902	3.403582	0.087861	18.356024	2	1
...	...	...	...	...	...	...	...	...	...	...	...
1845	(cell)	(patient, carcinoma, thyroid)	0.572625	0.279206	0.124611	0.217613	0.779401	-0.035269	0.921276	1	3
714	(cell)	(patient, carcinoma)	0.572625	0.282516	0.125974	0.219993	0.778694	-0.035802	0.919844	1	2
711	(patient, carcinoma)	(cell)	0.282516	0.572625	0.125974	0.445899	0.778694	-0.035802	0.771296	2	1
189	(cell)	(patient)	0.572625	0.424065	0.188084	0.328460	0.774550	-0.054746	0.857632	1	1
188	(patient)	(cell)	0.424065	0.572625	0.188084	0.443526	0.774550	-0.054746	0.768006	1	1

2100 rows × 11 columns

L'algorithme Apriori vous permet d'extraire des ensembles d'itemset fréquents et apprend les règles d'association entre les items sur les données des bases de données relationnelles (ensembles de données volumineux). L'algorithme identifie les items fréquents dans la base de données. Il les étend à des ensembles d'itemsets de plus en plus grands tant que ces ensembles d'itemsets apparaissent suffisamment souvent dans la base de données.

Conclusion

Lors de ce chapitre, nous avons proposé une architecture d'extraction de règles d'associations, les différents outils et langages de programmations que nous avons utilisés dans notre travail. Nous avons ensuite montré étape par étape le déroulement de notre démarche jusqu'à l'obtention du résultat escompté. Par contre, il y a une chose à laquelle il faut faire attention lors de l'utilisation d'Apriori sur de grands ensembles de données est le choix du seuil de prise en charge minimum. Si vous ne faites pas attention, vous pouvez rapidement manquer de mémoire avec un nombre potentiellement énorme d'itemsets de taille 2.

Conclusion générale et Perspectives

Les travaux présentés dans ce mémoire ont porté sur l'extraction de connaissances à partir des textes. Au cours de ce travail on a tout d'abord présenté un état de l'art qui explique brièvementle le concept de Fouille de textes (Text Mining) en précisant le processus, les techniques et les domaines d'application de ce dernier. On a choisi l'une des méthodes qui est l'extraction des règles d'association.

Puis, on a présenté le concept de notre travail en expliquant le fonctionnement des algorithmes qui permettent la recherche des motifs fréquents et la génération des règles d'association valides entre les différentes entités biologiques, présentes dans le corpus textuel étiqueté, à l'aide de l'algorithme Apriori.

Finalement, on a présenté l'application en exécutant notre code python de l'algorithme APRIORI, en le modifiant un peu pour qu'il puisse s'exécuter sur un corpus textuel. On a aussi fait des expérimentations pour extraire les règles d'association à partir des bases textuelles dans le domaine médical.

Malheureusement, le temps attribué à ce travail n'a pas suffi, d'où il était difficile d'arriver à l'étape finale qui est la classification, pour ne s'intéresser qu'au cancer de la Thyroïde. Aussi, on aurait aimé enrichir notre travail et comparer notre travail avec d'autres approches et algorithmes. Nous proposons comme perspectives :

- Faire de la classification pour ne s'intéresser qu'au cancer de la Thyroïde,
- Appliquer d'autres méthodes de l'extraction de connaissances à partir des textes.
- Tester d'autres algorithmes de génération des règles d'association tels que : FP-GROWTH, ECLAT et CLOSE ensuite comparer les résultats avec APRIORI.

Annexes

Appendix A

Code Python pour la suppression des mots vides

Entrée [31]:
```
import io
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
eng_stopwords = stopwords.words('english')
```

Entrée [33]:
```
data = data.astype(dtype={'full_text': 'string'})
data['clean_text'] =data['full_text'].apply(lambda x: ' '.join([item for item in re.split(r'\W+', x)
                                                                if item not in eng_stopwords and item != '']))

data.head()
```

Appendix B

Code Python pour le nettoyage du corpus

Entrée [7]:
```
def span_handling(text):
    cleanr=re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', text)
    return cleantext
tag = '<span class=""misc"" mc=""Mediastinal"">mediastinal</span> goiter : a surgical challenge . <span class=""misc"" mc=""Media
span_handling(tag)
```

Out[7]:
```
' mediastinal  goiter : a surgical challenge .  Mediastinal '
```

Entrée [11]:
```
s = ' st"rinè_-àç&&&&&&&&&&g. With. Punctuation!!!!!" '
def remove_punct(s):
    exclude = set(string.punctuation)
    #table = bytes.maketrans("","")
    regex = re.compile('[%s]' % re.escape(string.punctuation))
    return regex.sub('', s)   # From Vinko's solution, with fix.
remove_punct(s)
```

Out[11]:
```
' strinèàçg With Punctuation '
```

Entrée [12]:
```
def extract_class(text):
    m = re.findall('class=(.+?)\s', text)
    return m

def extract_mc(text):
    m = re.findall('mc=(.+?)"', text)
    return m


def extract_span_content(text):
        m = re.findall('>(\w.+?)<', text)
        return m

def spliter(sentence):
    m=sentence.split()
    return m
x='Life threatening giant mediastinal goiter  '
spliter(x)
```

Out[12]:
```
['Life', 'threatening', 'giant', 'mediastinal', 'goiter']
```

```
In [13]: data['classes'] = data['text'].apply(lambda x: extract_class(x))
         data['mcs'] = data['text'].apply(lambda x: extract_mc(x))
         #data['span_content'] = data['text'].apply(lambda x: extract_span_content(x))

In [14]: data['text'] = data['text'].apply(lambda x: span_handling(x))

In [15]: #TypeError: expected string or bytes-like object Handling
         data['text'] =data['text'].astype(str)
         data['mcs'] =data['mcs'].astype(str)
         #data['span_content'] =data['span_content'].astype(str)
         data['classes'] =data['classes'].astype(str)

In [16]: data['mcs'] = data['mcs'].apply(lambda x: remove_punct(x))
         #data['span_content'] = data['span_content'].apply(lambda x: remove_punct(x))
         data['text'] = data['text'].apply(lambda x: remove_punct(x))
         data['classes'] = data['classes'].apply(lambda x: remove_punct(x))
         data.head()
```

Appendix C

Code Python pour la tokenization

```
Entrée [54]: data["new"] = data["new"].apply(nltk.word_tokenize)

Entrée [55]: data.head()
```

Appendix D

Code Python pour le stemming et la lemmatization

Entrée [56]:
```
import time
from nltk.stem.snowball import SnowballStemmer
from nltk.tokenize import TweetTokenizer
stemmer = SnowballStemmer("english")


start = time.time()
data['stemmed']=data['new'].apply(lambda x: [stemmer.stem(y) for y in x])
print ("stemming.apply duration :"), (time.time() - start)
```

```
stemming.apply duration :
```

Out[56]:
```
(None, 14.033479928970337)
```

Entrée [57]:
```
from nltk.stem import WordNetLemmatizer
def lemmatize_text(text):
    lemmatizer = WordNetLemmatizer()
    return [lemmatizer.lemmatize(w) for w in text]
```

Entrée [58]:
```
import time
start = time.time()
data['lemmed'] = data['new'].apply(lemmatize_text)
print ("Lemmed.apply duration :"), (time.time() - start)
```

```
Lemmed.apply duration :
```

Out[58]:
```
(None, 5.666422367095947)
```

Appendix E

Code Python pour convertir du texte en matrice d'éléments

Entrée [84]:
```
text_matrix = data['text'].apply(lambda x: pd.Series(x.split(' '))) #  convert text into matrix of items

# remove duplicate items PER ROW as we don't need them

text_matrix=(pd.DataFrame(text_matrix.apply(pd.Series.unique, axis=1).tolist()))
text_matrix.head()
```

Bibliographie

[0] Royaute J., François C., Zasadzinski A., Besagni D., Dessen P., Maunoury M-T., Le Minor S. (2003). "Mining corpora of texts on genes involved in thyroid cancers : a bioinformatic text mining and clustering process". ECCB'2003, Poster Session, September 2003. Paris

[1] Toussaint, Yannick : « Extraction de connaissances à partir de textes structurés », Vol.8, p. 11-34, 2004/3.

[2] Kantardzic M: "Data mining concepts, models, methods and algorithms". Press ,Piscataway, NJ, USA, 2003

[3] R. Lefébure, G.Venturi : « Le Data Mining » Edition EYROLLES, deuxième tirage 1998

[4] A.Taibi, H.Lazreg : «Utilisation des algorithmes d'apprentissage dans la catégorisation automatique thématique de documents Etude de cas : les algorithmes K_PPV, Naïve Bayes», Mémoire de Licence, Université de M'sila, 2011-2012

[5] S.Raheel : « L'Apprentissage Artificiel pour la Fouille de Données Multilingues: Application à la Classification Automatique des Documents Arabes », Thèse de doctorat en Sciences de l'Information et de la Communication, Université Lumière Lyon 2, 2010

[6] R. Agrawal, T. Imielinski, A. N. Swami: "Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference"*, volume 22, pages 207–216,Washington, DC, 1993.

[7] Blanchard Julien, Kuntz Pascale, Guillet Fabrice, Gras Régis : "Mesure de la qualité des règles d'association par l'intensité d'implication entropique", IRIN - École polytechnique de l'université de Nantes, 2002.

[8] H. Cherfi : « Etude et réalisation des d'un système d'extraction de connaissance à partir de textes », 2006

[9] Abderraouf Nouasria : « Extraction D'associations Lexicales Fortes Dans Les Commentaires », L'université Du Québec À Trois Rivières, Juin 2016.

[10] Achouri Abdelghani : "Extraction de relations d'association maximales dans les textes : représentation graphique", Université du Québec à Trois-Rivières, 2012.

[11] Pagé, Christian : "Bases de règles multi-niveaux", Université du Québec à Montréal, Fevrier 2008.

[12] Shashikumar G. Totad, Geeta R. B,Prassad Reedy: "Batch Processing for Incrementing FP-tree Construction", international Journal of Computer Applications, 2010.

[13] Fuchs. Patrick, Poulain. Pierre : "Cours Python", Université de Paris France, 2020

[14] R. Agrawal and R. Srikant: “Fast algorithms for mining association rules”. In *Proceedings of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.

[15] Dawid Weiss: «Descriptive Clustering as a Method for Exploring Text Collections», PhD thesis, Institute of Computing Science, Poznàn, Poland. 2006.

[16] Mohamed Mahdi MALIK, « Mise en évidence de relations entre entités biologiques au moyen de structures prédicatives à partir d’un corpus de textes indexés »

[17] STILOU, S., BAMIDIS, Panagiotis D., MAGLAVERAS, Nicos, et al. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. Studies in health technology and informatics, 2001, no 2, p. 1399-1403.

[18] TAN, Ying, YIN, Guo-Fu, LI, Gui-Bing, et al. Mining Compatibility Rules from Irregular Chinese Traditional Medicine Database by Apriori Agorithm. Journal of Southwest JiaoTong University, 2007, vol. 15, no 4, p. 288-293.

[19] ABDULLAH, Umair, AHMAD, Jamil, et AHMED, Aftab. Analysis of effectiveness of apriori algorithm in medical billing data mining. In:2008 Emerging Technologies. ICET 2008. 4th International Conference on. IEEE,2008. p. 327-331.

[20] NOMA, Nasir G. et GHANI, Mohd Khanapi Abd. Discovering pattern in medical audiology data with FP-growth algorithm. In: Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on. IEEE, 2012. p. 17-22.

[21] NAHAR, Jesmin, IMAM, Tasadduq, TICKLE, Kevin S., et al. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 2013, vol. 40, no 4, p. 1086-1093.

[22] RASHID, Mahmood A., HOQUE, Md Tamjidul, et SATTAR, Abdul. Association rules mining based clinical observations. arXiv preprint arXiv:1401.2571, 2014.

[23] YANG, H. S., XIE, Y. M., CHEN, C. Zhuang, Y., & Zhang, Y. Association rules analysis of Fufang Kushen injection in combination with modern medications in treating lung cancer: real- world study based on hospital information. Zhongguo Zhong yao za zhi= Zhongguo zhongyao zazhi= China journal of Chinese materia medica, 2018, vol. 43, no 8, p. 1708-1713.

[24] ALWIDIAN, Jaber, HAMMO, Bassam H., et OBEID, Nadim. WCBA: Weighted classification based on association rules algorithm for breast cancer disease. Applied Soft Computing, 2018, vol. 62, p. 536-549.

[25] HU, Ruijuan. Medical data mining based on association rules. Computer and Information Science, 2010, vol. 3, no 4, p. 104.

[26] CHEN, Wei, YANG, Jun, WANG, Hui-Ling, et al. Discovering Associations of Adverse Events with Pharmacotherapy in Patients with Non-Small Cell Lung Cancer Using Modified Apriori Algorithm. BioMed research international, 2018, vol. 2018.

[27] PASQUIER, Nicolas, BASTIDE, Yves, TAOUIL, Rafik, et al. Pruning closed itemset lattices for association rules. In : BDA'1998 international conference on Advanced Databases. 1998. p. 177-196.

[28] VERHEIN, Florian. Frequent pattern growth (FP-growth) algorithm. School of Information Studies, The University of Sydney, Australia, 2008, p. 1-16.

[29] DAHMANI, Djilali. Fouille des règles d'association guidée par des ontologies et des schémas de règles : Application au domaine de la production SONATRACH / AVAL. 2011 Thèse de magister. Université des Sciences et de la Technologie d'Oran « Mohamed Boudiaf ».

[30] BLASCHKE C., ANDRADE M.A., OUZOUNIS C. and VALENCIA A. (1999). "Automatic extraction of biological information from Scientific text: protein-protein interactions". ISMB, 7,60-67

Webographie

[31] https://www.cairn.info/revue-document-numerique-2004-3-page-11.htm?contenu=article

[32] https://fr.wikipedia.org/wiki/Natural_Language_Toolkit

[33] https://docs.python.org/fr/3.9/library/string.html#module-string

[34] https://fr.wikipedia.org/wiki/Nuage_de_mots-cl%C3%A9s

[35] https://www.geeksforgeeks.org/generating-word-cloud-python/

[36] https://towardsdatascience.com/mlxtend-a-python-library-with-interesting-tools-for-data-science-tasks-d54c723f89cd

[37] https://www.lemagit.fr/definition/Extraction-dinformation-EI

[38] https://www.cancer-environnement.fr/287-Cancer-de-la-thyroide.ce

[39] https://fluoptics.com/comment-evolue-le-cancer-de-la-thyroide/