République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITE FERHAT ABBAS : SETIF 1 FACULTE DES SCIENCES

DEPARTEMENT : INFORMATIQUE

MEMOIRE DE MASTER

DOMAINE .**Mathématiques et Informatique**

FILIERE : **Informatique**

SPECIALITE : **Informatique Fondamentale et Intelligence Artificielle**

Thème

Extraction des éléments pertinents dans un texte

Présenté par :

Elkolli Soulef

Dirigé par :

Dr.Douar Amel

Promotion : 2021/2022

dédicaces

، إلى أبي

الغائب بجسده الذي لن يعود

الحاضر بروحه في كل الوجود

إلى أمي

أطال الله في عمرها و جعلها نورا لنا

و عوضا

إلى زوجي منجي .

حفظه الله لي و جعله سندا

إلى أولادي. هارون،أحمد و ماس.

إلى إخوتي و أخواتي.

إلى كل من علمني حرفا في هذه الحياة.

أهدي هذا العمل المتواضع

Remerciments

أتقدم بالشكر أولا و قبل كل شيء لله عز و جل على منحه لي القوة

.و العزيمة و الصبر و عدم اليأس لإنجاز هذا العمل

كما أتقدم بالشكر لكل من

زوجي على ما بدله من مجهود لمساعدتي في هذا العمل

أولادي على صبرهم معي و تحمل كل الإهمال و التقصير الذي كان مني

الأستاذة المشرفة على العمل: دوار أمال

على تشجيعها و مساندتها

كل من ساندني و شجعني من زملاء و زميلات و خاصة

ليندة زروال

الملخص

يتقاطع عملنا مع المجالات التالية. معالجة النصوص، معالجة اللغة الطبيعية،استخراج المعلومات و التعليم عن بعد.

لهذا ، إقترحنا طريقة عمل تدمج تقنيات معالجة النص و طرق تلخيص ورقة الأعمال التطبيقية. من أجل إنشاء ملخص تلقائي لهذه الورقة باستعمال طرق التلخيص المختلفة لاستخراج العناصر ذات الصلة.

الكلمات الرئيسية. استخراج البيانات ، معالجة اللغة الطبيعية ، الملخص ،العناصر ذات الصلة.

Résumé

Notre travail est a l'intersection des domaines suivant : le Text Minig, le traitement du langage naturel, l'extraction des informations et l'apprentissage des travaux pratique à distance.

L'objectif de notre travail est de générer un résumé automatique d'une feuille de TP. Pour cela on a proposé un processus de génération automatique qui intègre les techniques de Text Minig, les méthodes de Summurization du langage naturel et les différentes méthodes d'extraction de l'information.

Mots-clés : l'exploration de données, traitement du langage naturel NLP, un résumé, informations pertinentes.

Abstract

Our work is at the intersection of the following areas: Text minig, Natural Language Processing, information extraction and remote practical work learning.

The objective of our work is to generate an automatic summary of a TP sheet. For this , we have proposed an automatic generation process that integrates text minig techniques, natural language summurization methods and different extraction methods of extraction.

Keywords: Data mining, Natural Language Processing(NLP), summary, relevant information.

Table de matière

Chapiter1. Généralités

Chapitre2 . Le résumé automatique du texte et l'extraction des informations

Chapitre4 . Contribution

.

Table des figures

Liste des abréviations :

NLP : Natural Language Process

NLG : la génération du langage naturel

NLU : la compréhension du langage naturel.

TP : Travaux pratiques.

POS : Part of speach

ML : Machine Learning

TLN : traitement du langage naturel

NLTK : Natural Language Tools Kits

Tk : Tkinter

API : interface de programmation applicative

Liste des équations :

Liste des tableaux :

Introduction générale

L'émergence du Big Data a apporté de nouveaux défis aux équipes informatiques spécialisées dans l'analyse de données, et spécialement les données textuelles non structurées.

Toutefois, la difficulté majeure rencontrée est que ces informations ne peuvent pas être utilisées pour tout traitement d'extraction d'informations pertinentes. Ceci est du à cause de leurs formats, ce qui pose des problèmes au niveau de la recherche, de la découverte, de l'extraction et de la modélisation des connaissances.

Pour cela, ces données impliquent de nombreuses applications pour une telle taille de données tels que les mails, les pages web, les enregistrements audio, les articles de recherche, etc.

Notre objectif et de réaliser une adaptation ou une optimisation des techniques de Text Mining pour les feuilles de TP, et montrer le lien entre le Text Mining et le traitement du langage naturel et les documents textuels, en les appliquant sur cette feuille de la spécialité « chimie », afin d'extraire les termes ou les informations les plus pertinentes, et cela va nous conduire au point de rencontre de plusieurs discipline : l'analyse des données, le traitement du langage naturel et l'extraction des concepts.

Organisation du mémoire :

Après, la présentation de notre objectif et l'introduction générale, ce mémoire sera organisé en cinq chapitres de la façon suivante.

Chapitre1. Généralités

Dans ce chapitre, nous allons donner des généralités sur le traitement

automatique du texte avec toutes ses étapes(Nettoyage du texte(Suppression des mots vides et quelques signes de ponctuation), Tokennization(découper le texte en mots ou séquence) appelés Tokens, Lemmatisation et Stemming) ,ainsi que le traitement du langage naturel NLP et leur application sur une feuille de TP .Le traitement du texte est la phase la plus importante qui abouti à avoir un texte nettoyé prêt à l'emploi.

Chapitre2. Résumé automatique du texte et extraction des éléments

Dans ce chapitre nous avons vu deux étapes très importantes de notre travail : ***Le résumé automatique du texte*** avec ses types et les méthodes de réalisation possibles pour leurs modèles d'automatisation, on a choisit **Méthode basée sur la classification des éléments, car** c'est cette méthode seule, qui nous a aider à aboutir notre but.

Et les différentes méthodes d'***extraction des éléments pertinents***, dans notre cas d'étude, on a choisis la méthode : **Dictionnaire** qui se base sur la comparaison directe des éléments du texte résumé et un dictionnaire généré manuellement à partir de la documentation sur le web, dans le domaine choisis : **chimie.**

Chapitre3. Contribution

Ce chapitre est consacré au travail théorique précédent avec toutes ses étapes avec une vue global du travail et des propositions pour les problèmes rencontrés durant le travail.

Chapitre4. Implémentation

Dans ce chapitre, nous allons présenter la phase de réalisation et d'implémentation du notre système avec exemples de code et captures d'écran.

CHAPITRE 1

Généralités

1. Introduction :

Depuis plusieurs années, le contenu d'informations électroniques se trouvent sous plusieurs formes tirées de plusieurs sources (les news groups, les courriers électroniques, les forums de discussions, les réseaux sociaux, les documents électroniques…etc.). Plus de 80% de ce flux d'informations est stocké sous une forme textuelle (non structurée ou semi structurée). Dans ce travail on s'intéresse aux feuilles de Travaux Pratiques (TP) numériques se trouvant sur les plateformes d'enseignement à distance.

Les sources de connaissances nécessitent une formalisation de ces informations. Pour remédier à cela, plusieurs outils et techniques (langage naturel, extraction d'information, recherche d'information), qui rentrent dans le cadre de la fouille de texte ou Text Mining, ont été conçus.

Donc, dans ce chapitre on va présenter des généralités sur le ***Text Mining***, ***le langage naturel*** ainsi le domaine de notre travail qui est les ***travaux pratiques a distance.***

Ce domaine des travaux pratiques à distance, a réapparut après l'apparition du pandémie du Covid 19, cette pandémie qui a conduit la moitié de l'humanité au confinement, tout le monde est à l'arrêt, et parmi les secteurs impactés, il y a l'éducation, avec prés de 1.7 milliards d'enfants et de jeunes dans le monde qui ne vont ni à l'école, ni à l'université.

Cette pandémie a provoqué la plus grande révolution de l'éducation que le monde n'a jamais vu : une accélération de l'apprentissage en ligne à distance, pour assurer à la fois la protection sanitaire des étudiants, le maintient de la qualité pédagogique, la continuité de la formation professionnelle et une préparation efficace des cours et Tps à distance qui présente notre domaine de travail.

2 .Text Mining :

2.1. Définition :

Le Text Mining est une branche de l'intelligence artificielle qui se spécialise dans le traitement de corpus de textes pour en analyser le contenu puis en extraire des connaissances.

le Text Mining est défini comme suit : « Le Text Mining est un traitement permettant de passer en revue un ensemble de données textuelles grâce à des techniques statistiques et linguistique pour en déduire de l'information utile par rapport au but fixé par un utilisateur lui évitant une lecture séquentielle de l'information » [1].

2.2. Domaines d'application du Text Mining :

L'exploration de texte est un domaine interdisciplinaire qui intègre des domaines tels que la recherche d'informations, l'extraction d'informations, l'exploration de données, la linguistique informatique et le traitement du langage naturel. Les domaines de l'exploration de texte sont illustrés dans la figure [1] ci-dessous.

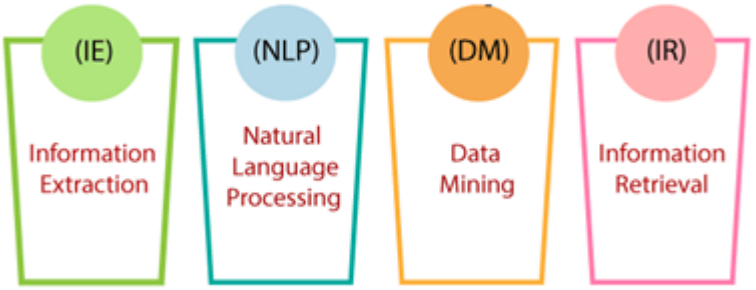


Figure 1 : Les domaines du Text Mining [2]

2.3. Les technique du Text Mining :

Il existe plusieurs techniques et outils pour le Text Mining, parmi ces techniques (voir figure [2]) on cite les suivant :

- *Traitement automatique du langage naturel et apprentissage automatique :*

La plupart des outils utilisent le traitement automatique du langage naturel et/ou des techniques de Machine Learning pour Le Text Mining.

- *Les méthodes statistiques* : telles qu'elles sont utilisées pour l'exploration de données, sont également appliquées pour l'exploration de texte. En fait, la plupart des outils utilisent les méthodes statistiques en conjonction avec d'autres méthodes.
- *Intelligence artificielle* : techniques telles que les réseaux de neurones sont également utilisés dans de nombreux outils du Text Mining.

- *Les techniques de classification*: sont également utilisées pour catégoriser le texte et les documents. Ces techniques doivent être capables de traiter des données non structurées.
- *Apprentissage linguistique, analyse sémantique et modélisation* prédictive des techniques sont également employées pour extraire du texte. [3]

  La figure ci-dessous illustre quelques techniques du Text Mining :

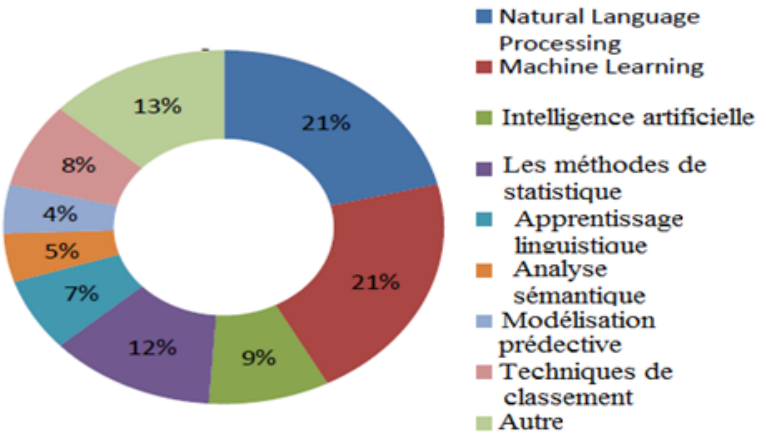


Figure 2 : Les techniques les plus populaires du Text Mining [3]

2.4. Les étapes du processus du Text Mining :

Le processus du Text Mining implique une série d'activités à effectuer pour extraire l'information. [4] .Ces activités sont :

- Le prétraitement du texte.
- La transformation du texte.
- La sélection des caractéristiques.
- Data Mining.
- L'évaluation
- L'application.

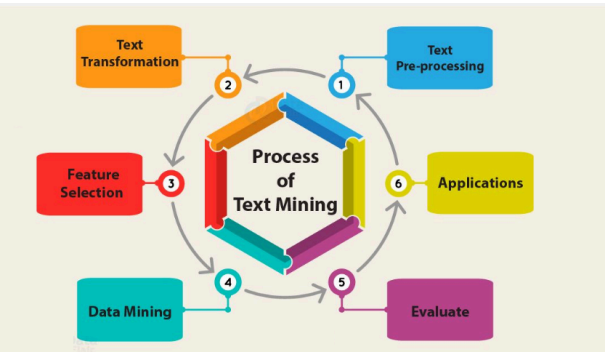


Figure 3 .Activités effectuées par le processus du Text Mining

L'étape la plus importante est le prétraitement, ou la préparation des données, c'est une tâche très importante qui consomme le plus de temps dans le processus de Text Mining (figure [3]). Cette étape inclut tous les traitements, les processus et les méthodes nécessaires pour la préparation des données pour les opérations de bases de la découverte de connaissance du système Text Mining. L'étape du prétraitement en général converti les informations de leur source originale en un format intermédiaire [5].

Le prétraitement dans le Text Mining consiste à identifier le nettoyage qui devra être effectué pour permettre notre analyse. Le nettoyage fait référence aux étapes prises pour normaliser le texte et pour supprimer le texte et les caractères qui ne sont pas pertinents. Après avoir effectué ces étapes, on se retrouve avec un ensemble de données textuelles "propres" prêt à être analysé. [6]. Dans cette étape en procède comme suit :

- Transformer toutes les lettres en minuscule**.**
- Faire la Tokénisation : mots ou les phrases.
- Supprimer les mots vides (Stopwords).
- Supprimer la ponctuation.
- Stemming.
- Lemmatisation.
- La normalisation.
- Créer un nouveau corpus propre prêt à l'emploi.

3. **Le langage naturel :**

3.1. Définition :

C'est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle. Il vise à créer des outils de traitement de la langue naturelle pour diverses applications. Il ne doit pas être confondu avec la linguistique informatique, qui vise à comprendre les langues au moyen d'outils informatiques [7]

3.2 .Traitement du langage naturel (NLP)

Le traitement du langage naturel est le problème le plus difficile dans le domaine de l'intelligence artificielle. C'est l'étude du langage humain afin que les ordinateurs puissent comprendre des langages naturels similaires à ceux des humains. Le traitement du langage naturel concerne la génération du langage naturel (NLG) et la compréhension du langage naturel (NLU).

NLG s'assure que le texte généré est grammaticalement correct et fluide. La plupart des systèmes NLG incluent un réalisateur syntaxique pour s'assurer que les règles grammaticales telles que l'accord du verbe sujet sont respectées et un planificateur de texte pour décider comment organiser les phrases, les paragraphes et les autres parties de manière cohérente. La meilleure application NLG connue est la traduction automatique.

NLU se compose d'au moins l'un des composants suivants : a tokenizer, analyseur lexical, analyseur de syntaxe et analyseur sémantique.

En d'autres termes, la NLP est un composant de l'exploration de texte qui effectue un type particulier d'analyse linguistique qui aide essentiellement une machine à "lire" du texte. Il utilise une méthodologie différente pour déchiffrer les ambiguïtés du langage humain, notamment les éléments suivants : résumé automatique, étiquetage des parties du discours, segmentation, ainsi que la compréhension et reconnaissance du langage naturel. Nous verrons tous les processus étape par étape en utilisant Python. [8]

3.3. Les méthodes du NLP :

Globalement, nous pouvons distinguer deux aspects essentiels à tout problème de NLP :

- La partie **« linguistique »**, qui consiste à prétraiter et transformer les informations en entrée en un jeu de données exploitable.
- La partie **« apprentissage automatique »** ou « Data Science », qui porte sur l'application de modèles de Machine Learning ou Deep Learning à ce jeu de données. [8].

3.4. Les avantage du NLP :

Le principal avantage du NLP est qu'il améliore la façon dont les humains et les ordinateurs communiquent entre eux. Pour manipuler un programme informatique, le moyen le plus direct passe par code, le langage de l'ordinateur. En autorisant les machines à comprendre le langage humain, l'interaction avec celles-ci devient beaucoup plus intuitive pour les humains. Les autres avantages sont :

- Amélioration de la précision et de l'efficacité de la documentation.
- Réalisation automatique d'un résumé lisible d'un texte original plus vaste et plus complexe.
- Utilisation d'assistants personnels et de chabots.
- Analyse des sentiments facile à réaliser.
- Mise à disposition d'information avancée à partir d'analyse qui n'était pas disponible auparavant en raison du volume de données. [9]

4. Le NLP et le Text Mining :

Le domaine de la NLP concerne l'analyse des informations textuelles et a été appliqué récemment dans d'autres domaines comme le contexte des sciences expérimentales (la biologie moléculaire, la chimie pharmaceutiques...etc.).

L'approche d'exploration de texte implique l'analyse et l'extraction des collections d'informations de données textuelles libres en utilisant des systèmes.

En général, les applications d'exploration de texte tirent parti d'une gamme de méthodes indépendantes du domaine telles que la partie du discours (POS) taggers, qui étiquettent chaque mot avec son correspondant partie du discours (par exemple nom, verbe ou adjectif), ou les stemmers, qui sont des algorithmes qui retournent la morphologie racine d'une forme de mot. En outre, des outils et des ressources spécifiques à un domaine tels que les marqueurs de protéines et les ontologies sont utilisés [10].

4. 1. Comparaison entre Text Mining et NLP :

Le terme «*Text Mining*» est utilisé pour l'apprentissage automatique automatisé et les méthodes statistiques utilisées à cet effet. Il est utilisé pour extraire des informations de haute qualité à partir de texte non structuré et structuré.

Les informations peuvent être structurées en texte ou en structure correspondante, mais la sémantique du texte n'est pas prise en compte. *Le langage naturel* est ce qu'on

utilise pour la communication. Les techniques de traitement de ces données pour comprendre la signification sous-jacente sont collectivement appelées traitement du langage naturel (NLP). Les données peuvent être de la parole, du texte ou même une image et l'approche implique l'application de techniques d'apprentissage automatique (ML) sur les données pour créer des applications impliquant la classification, l'extraction de la structure, la synthèse et la traduction des données structurées, analyse des sentiments, etc. [11]

Le tableau suivant résume la différence entre les deux termes :

	Text Mining	Traitement du langage naturel
1.	Il traite de la conversion du contenu textuel en données qui font l'objet d'une analyse plus approfondie.	Son objectif est que les systèmes informatiques puissent comprendre les langues ou les textes humains.
2.	Pour traiter les données, il utilise différents types d'outils et de langages.	Il utilise des modèles d'apprentissage automatique de haut niveau pour traiter les données et produire des résultats.
3.	Pour effectuer des tâches, il ne prend pas en compte l'analyse sémantique.	Il prend en compte l'analyse syntaxique et l'analyse sémantique pour effectuer des tâches.
4.	La principale source de données dans l'exploration de texte comprend des documents volumineux.	En cela, il peut y avoir plusieurs sources de données telles que des panneaux, des discours, etc.
5.	En cela, nous pouvons mesurer facilement les performances du système et sa précision par rapport à la PNL.	En cela, mesurer les performances du système est assez difficile par rapport au Text Mining.
6.	Il ne nécessite pas d'intervention humaine.	Pour traiter des données, cela nécessite parfois une intervention humaine.
7.	Il produit le modèle et la fréquence des mots.	Il produit une structure comme une structure grammaticale.
8.	Il peut être utilisé pour surveiller les médias sociaux.	Il peut être utilisé dans la traduction de sites Web.

Tableau1 : Comparaison entre Text Mining et NLP [12]

5. Les travaux pratiques :

5.1. Notions de Travaux Pratiques :

Le progrès des toutes les sciences est fondu essentiellement sur la comparaison entre les lois **théoriques** et les résultats des **expériences et l**es lois théoriques sont censés décrire objectivement la réalité scientifiques ainsi les résultats des expériences supposés être représentatifs.

Les deux démarches, théoriques et expérimentales, sont naturellement accompagnées de la notion d'incertitude, ce qui rend l'une indispensable pour compléter l'autre.

Les travaux pratiques constituent un type d'enseignement fondé sur l'apprentissage pratique avec en particulier la réalisation d'expériences permettant de vérifier et compléter les connaissances dispensées dans les cours théoriques.

Les travaux pratiques concernent généralement les sciences expérimentales. Contrairement aux autres types de cours qui se passent exclusivement à l'oral ou à l'écrit, les séances de travaux pratiques nécessitent souvent un matériel spécifique (verrerie et produits chimiques, circuits électriques, ordinateurs...). La salle de classe, de type laboratoire, affectée à ces travaux est généralement appelé *Salle de travaux pratiques* ou *salle de TP*.

Les travaux pratiques sont une mise en application (et une mesure de la maîtrise par les étudiants) de la méthode scientifique, basée sur la pose d'hypothèse, la conception d'un protocole expérimental, l'expérimentation, l'interprétation des résultats et le raffinement des hypothèses initiales.

Les travaux pratiques permettent de mettre en évidence les transferts et les techniques mis en œuvre dans les ateliers d'application et d'adapter les supports pédagogiques en fonction des techniques étudiées. [13]

5 .2 .Objectifs :

Les travaux pratiques peuvent offrir des avantages innombrables aux scientifiques, citons quelques-uns :

- Acquérir une connaissance de base des phénomènes scientifique (chimie, physique, informatique…….) .

- Apprendre à représenter et interpréter les résultats obtenus et en tirer les conclusions.
- Apprendre quelques règles pour estimer les incertitudes expérimentales et valoriser les mesures effectuées au laboratoire.
- Renforcer la compréhension du cours.
- Apprendre certains comportement et pratiques expérimentaux, tels que le bon usage des outils et le respect des méthodologies de recherche.

5.3 .Objectifs pédagogiques et typologie de travaux pratiques :

Les travaux pratiques et spécialement en sciences expérimentales mettent en œuvre des supports matériels qui peuvent être classés selon deux types principaux :

• Des mécanismes réels et authentiques : ceux-ci sont éventuellement appareillés pour analyser des solutions constructives, mesurer des caractéristiques et, plus généralement, permettre des observations relatives à leur comportement en situation réelle,

• Des supports destinés à l'étude et la vérification des phénomènes.

Ces travaux pratiques permettent :

• De mettre en relation le réel et ses représentations ; d'apprécier les écarts entre les résultats obtenus à partir du modèle d'étude et les résultats mesurés sur le réel,

• De mettre en lumière le choix des solutions constructives adoptées en regard des fonctions techniques à réaliser.

• De découvrir ou d'illustrer les lois fondamentales. [14]

5.4 La feuille de TP :

La feuille de TP est un document riche en information scientifique, il est rédigé par des chercheurs dans un domaine donné, il fait généralement référence à une feuille de papier avec des questions ou des exercices que les étudiants doivent accomplir.

5.4.1. Description d'une feuille de TP :

Le format général d'une feuille de TP est généralement présenté sous plusieurs parties, la figure ci-dessous illustre ces parties :

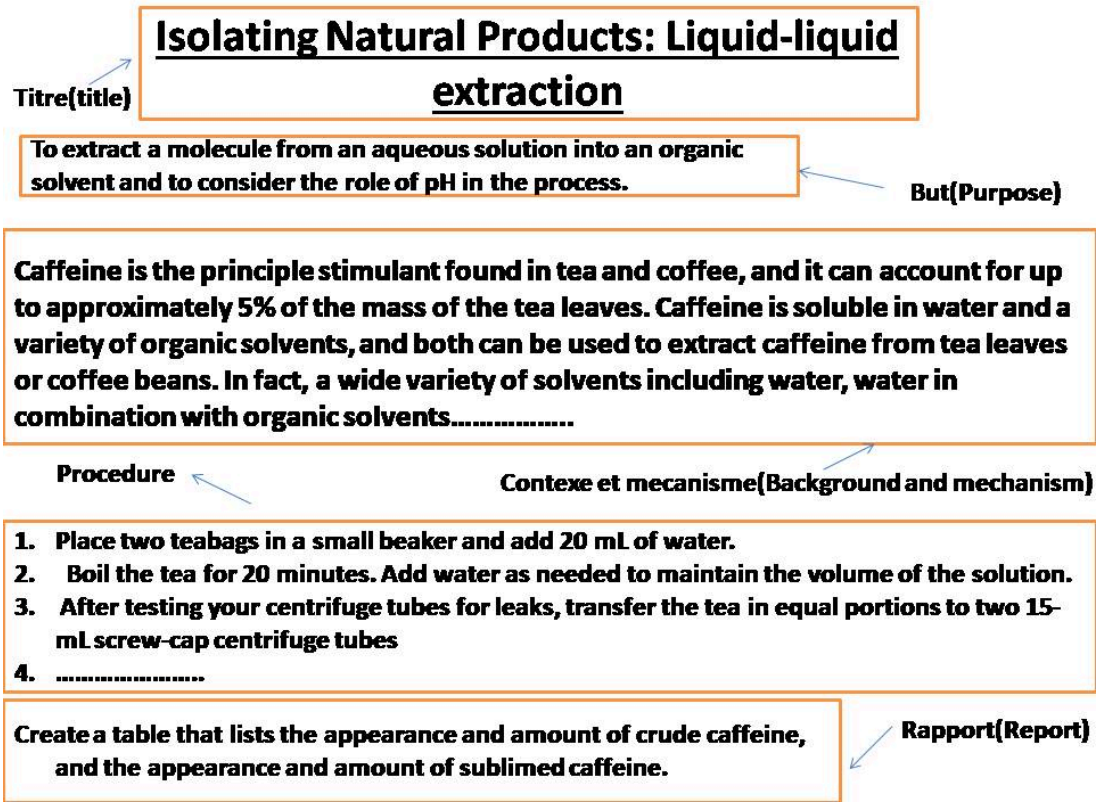


Figure 4 : Format générale de la feuille de TP

6- Conclusion :

Dans ce chapitre, nous avons présenté des généralités sur les différents axes de notre travail : Text Mining, NLP et travaux pratiques ainsi la feuille de TP qui représente le noyau de notre travail ; c'est le document source qu'on va étudier durant ce mémoire.

CHAPITRE 2

Le résumé automatique du texte

Et

L'extraction des informations

Etat de l’art

Ⅰ .résumé automatique du texte : Première partie :

Ⅰ.1. Introduction :

Le résumé automatique de texte est l'une des applications du traitement du langage naturel (TLN) qui a forcément un impact énorme sur nos vies. Avec la croissance des médias numériques et l'édition toujours croissante : qui a le temps de parcourir des articles, des documents ou même des livres entiers pour décider s'ils sont utiles ou non ? Heureusement, cette technologie est déjà là.

Le résumé automatique de texte est l'un des problèmes les plus difficiles et les plus intéressants dans le domaine du traitement automatique du langage naturel (TLN).

Le résumé de texte a attiré l'attention dès les années 1950. Un article de recherche, publié par Hans Peter Luhn à la fin des années 1950, intitulé "La création automatique de résumés de littérature", utilisait des caractéristiques telles que la fréquence des mots et la fréquence des phrases pour extraire des phrases importantes du texte à des fins de résumé.

Il s'agit d'un processus de génération d'un résumé concis et significatif du texte à partir de plusieurs ressources textuelles telles que des livres, des articles de presse, des articles de blog, des documents de recherche, des mails et des tweets.

La demande de résumé automatique de texte augmente de nos jours grâce à la disponibilité de grandes quantités de données textuelles. [15]

Ⅰ.2 .Définition :

Le résumé de texte est un processus permettant d'exprimer le contenu d'un document sous une forme condensée qui répond aux besoins de l'utilisateur. Il sert d'outil qui aide l'utilisateur à trouver efficacement des informations utiles à partir d'une immense quantité d’informations. Le résumé permet la réduction de la taille du document original tout en préservant son contenu informatif et représente moins de la moitié du texte principal. [15]

Ⅰ.3. Les différentes approches de résumé automatique de texte :

Il existe quatre approches principales pour générer des résumés de texte, ces approches sont classées par rapport au document de sortie **:**

Ⅰ.3.1. Approche d'extraction :

Consiste à choisir des extraits appropriés (des phrases, des paragraphes, etc.) du texte original et à les enchainer dans une forme plus courte. Le texte résumé est extrait du texte sur base statistique ou en employant des méthodes heuristiques ou une combinaison de deux.

Souvent, on extrait du texte source les phrases complètes jugées les plus importantes. Cette approche a l'avantage d'être facile à réaliser mais elle risque d'introduire une certaine incohérence dans les résumés.

L'objectif de cette approche est de pouvoir fournir rapidement, sans analyse en profondeur du texte, un résumé à l'utilisateur. On repère et extrait les segments textuels (phrases ou paragraphe) les plus pertinents du texte afin de construire un sous-ensemble d'extraits textuels que l'on considère comme un résumé.

L'avantage de la méthode par extraction est de ne pas passer par une analyse en profondeur du texte, et de pouvoir fournir un résumé de façon plus simple sans devoir générer du texte.

Le résumé par extraction évite la génération de texte. Ceci permet de se concentrer sur la sélection du contenu pertinent et d'autre part, d'obtenir un résumé lisible et linguistiquement correct

Ⅰ.3.2. Approche d'abstraction :

Le texte résumé est une interprétation du texte original avec un processus de production par réécriture du texte source en une version plus courte par le remplacement de certains concepts.sa mise en œuvre exige l'utilisation de grammaire et le lexique pour l'analyse syntaxique et la génération, en plus d'une modélisation de la compréhension humaine des textes. Cette approche est la plus difficile.

Ⅰ.3.3. Approche statistique :

Les approches statistiques peuvent résumer un document en utilisant des caractéristiques statistiques de la phrase telles que le titre, l'emplacement, la fréquence des

termes, en attribuant des poids aux mots clés, puis en calculant le score de la phrase et en sélectionnant la phrase la mieux notée dans le résumé.

Ⅰ.3.4. Approche linguistique :

La linguistique est une étude scientifique du langage qui comprend l'étude de la sémantique et de la pragmatique. L'étude de la sémantique signifie comment le sens est déduit des mots et des concepts et l'étude de la pragmatique comprend la façon dont le sens est déduit du contexte

Ces approches peuvent être combinées en vue d'obtenir de meilleurs résumés.

Ⅰ.4. Les domaines d'application de résumés automatiques :

Nombreux sont les domaines d'outils du résumé automatiques, en effet, on peut recenser les usages suivants :

- Pour résumer les nouvelles au SMS pour les téléphones portables.
- Pour laisser un ordinateur synthétique lu le texte résumé. Le texte écrit peut être long et ennuyeux pour le lire.
- Dans des moteurs de recherche pour présenter les descriptions compressées des résultats de recherche.
- Pour chercher dans des langues étrangères et obtiennent un résumé automatiquement traduit du texte automatiquement résumé.

Sans oublier les domaines de l'archivage, des bibliothèques et du journalisme

Ⅰ.5. Les types de résumé automatique du texte :

L'objectif de résumé automatique de textes présente le texte source dans une version plus courte avec la sémantique.

Les résumés automatiques se diffèrent par rapport à leurs objectifs ou buts : **Indicatif** ou **informatif**.

Ⅰ.5.1. Le résume indicatif :

Le résumé indicatif a pour fonction de fournir au lecteur suffisamment d'information pour qu'il puisse juger s'il doit consulter ou non le texte source. Pour un document correspondant à ce que recherche un lecteur, le résume indicatif doit pouvoir acheminer correctement ce lecteur vers celui-ci à travers la lecture de son contenu qui doit pouvoir le

faire décider ou non de la consultation du document. Ce type de résumé contient seulement des éléments partiels par rapport au résume informatif, mais surtout des éléments en vue de répondre à sa fonction. De ce fait, s'il consiste à diriger le lecteur vers le document initial, il ne se substitue pas à la lecture de ce dernier.

I .5.2. Le résume informatif :

Le résumé informatif fournit un ensemble d'informations permettant de donner un large panorama du contenu d'un texte. Pour cela, l'ensemble des principaux sujets doit être rapporté.

De plus, le résume informatif tend à conserver l'organisation générale du texte source. Ainsi, les sujets principaux qui sont rappelés dans le résume sont répartis de manière fidèle par rapport à l'organisation initiale afin de donner un juste aperçu de texte source.

I .6. Etapes de résumé de texte :

Dans le résumé automatique de texte, on peut identifier trois différentes étapes illustrées dans la figure ci-dessous. La plupart des systèmes aujourd'hui utilisent la première étape seulement. [7]

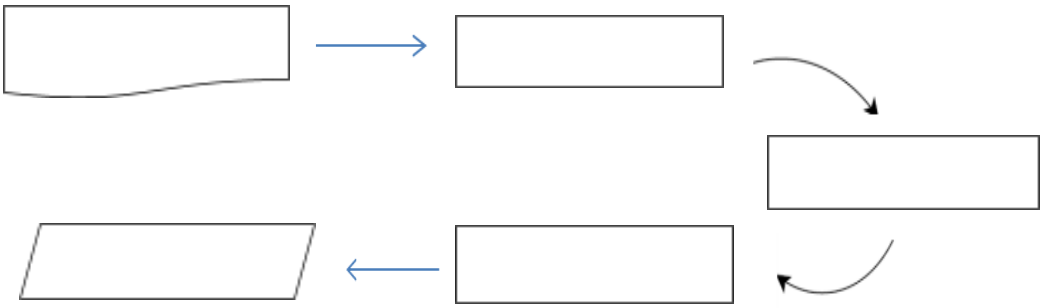


Figure 4. Etapes de résumé de texte

I .6.1 .Etape1 : Identification des thèmes

Elle sert à produire un résumé simple (extrait) en détectant les unités importantes dans le document (mot, phrase, paragraphes, etc.). Les systèmes de résumé qui utilisent seulement l'étape d'identification du thème, produisent un résumé extractif. Ceci se fait par filtrage du fichier d'entrée pour obtenir seulement les thèmes les plus importants. Une fois ces thèmes identifiés, ils sont présentés sous forme d'un extrait.

Pour effectuer cette étape, presque tous les systèmes utilisent plusieurs modules indépendants. Chaque module attribue un score aux unités d'entrée (mot, phrase ou passage plus long), puis un module de combinaison combine les scores pour chaque unité

afin d'attribuer un score unique. Enfin, le système renvoie les unités du plus haut en score, en fonction de la longueur du résumé, demandé par l'utilisateur ou fixé préalablement par le système.

I .6.2 .Etape2 : L'interprétation :

Dans l'interprétation, le but est de faire un compactage en réinterprétant et en fusionnant les thèmes extraits pour avoir des thèmes plus brefs. Ceci est indispensable du moment que les abstraits sont généralement plus courts que les extraits équivalents. Cette deuxième phase de résumé automatique (passage de l'extrait vers l'abstrait) est naturellement plus complexe que la première. Pour compléter cette phase, le système a besoin de connaissances sur le monde (par exemple, les anthologies), puisque sans connaissance aucun système ne peut fusionner les sujets extraits pour produire des sujets moins nombreux afin de former une abstraction. Lors de l'interprétation, les thèmes identifiés comme importants sont fusionnés, représentée en des termes nouveaux, et exprimé en utilisant une nouvelle formulation, en utilisant des concepts ou des mots qui n'existent pas dans le document original.

I .6.3 .Etape3 : Génération du résumé :

Le résultat de l'interprétation est un ensemble de représentations souvent non lisibles, c'est le cas du résumé par abstraction. Pour le résumé extractif, le résultat est un extrait rarement cohérent, à cause des références coupées, la négligence des liens entre les phrases, et la redondance ou la négligence de quelques matériels. De ce fait, les systèmes incluent une étape de génération du résumé afin de produire un texte cohérent et lisible par l'humain.

Dans cette étape, le système a besoin d'utiliser des techniques de la génération de langage naturel qui, selon [**7**], a besoin de deux modules : le micro-planeur et le générateur des phrases. Le micro-planeur, dans le contexte du résumé automatique, a comme fonction d'assurer que l'information sélectionnée par les deux étapes précédentes (identification et interprétation du thème) est rédigée d'une manière compacte et brève autant que possible en restant dans un état grammatical. Il peut être construit pour mener son travail sur deux niveaux : le niveau textuel et le niveau représentatif. Dans le premier niveau, l'entrée est une liste de phrases ou fragments de phrases, et la sortie est une liste compacte de phrases.

Dans le deuxième niveau, l'entrée est exprimée sous une notation abstraite, et la sortie est une spécification abstraite et syntaxique pour chaque phrase. La sortie de ce niveau n'est pas lisible par l'humain c'est pour ça il faut passer par la génération des phrases. Un des micro-planeurs destinés pour le résumé automatique est celui de **[16]**.

Ⅰ.7. Évaluation du résumé :

Le facteur le plus important qui doit être défini à la fin de toute l'expérience dans le résumé du texte est de savoir si le résumé généré est valide ou non. Par conséquent, l'évaluation du résumé du texte doit être soigneusement testée et évaluée. Il n'est pas non plus possible d'évaluer le résumé par une opération manuelle générée à partir de plusieurs documents. L'évaluation du résumé n'est pas une tâche facile car même un expert de la langue peut ignorer les informations lors de la rédaction du résumé. Le résumé du texte peut être évalué de deux manières :

Ⅰ.7.1.Évaluation intrinsèque : l'évaluation intrinsèque consiste à évaluer le système de résumé en interne. Elle s'occupe surtout de l'évaluation de cohérence et le contenu informatif des résumés produits.

Ⅰ.7.2.Évaluation extrinsèque : l'évaluation extrinsèque consiste à tester l'impact de résumé sur les tâches comme l'évaluation de pertinence, la compréhension en lecture, etc. [8]

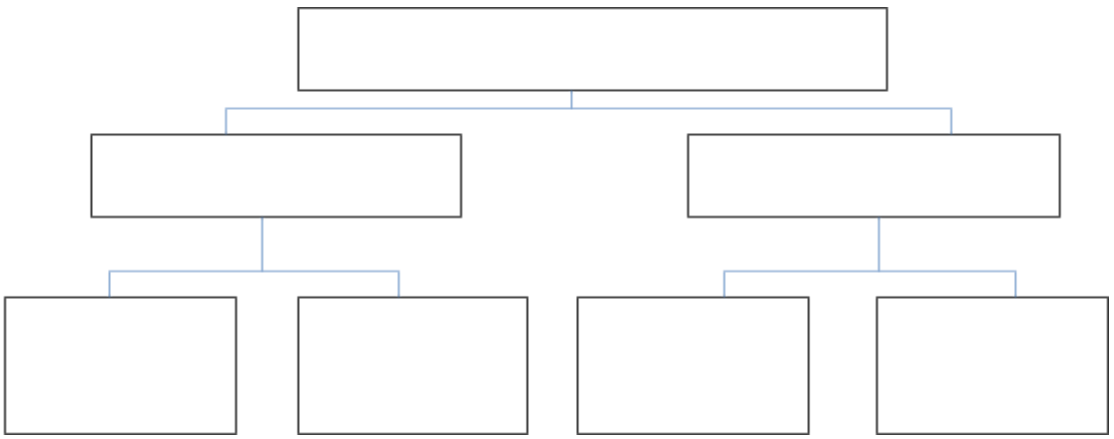


Figure 5. Les approches d'évaluation des systèmes de résumé automatique

Ⅰ.8. Méthodes de résumé :

L'objectif des méthodes d'extraction des phrases est de repérer dans le texte source les phrases les plus importantes. Le résultat obtenu est alors un extrait du texte source [9].

I .8.1. Méthodes à base de mots clés :

Cette méthode est basée sur le fait que l'auteur se sert (pour exprimer ses idées principales) de quelques mots-clés qui ont tendance à être récurrents dans le texte [10]. Le résumé automatique est alors produit en recherchant dans le texte source les unités de texte minimales réunissant ses mots-clés. Ce principe est souvent appliqué en différentes variantes présentées dans les sous-sections qui suivent :

I .8.1.1. Mots-clés prédéfinis

Pour calculer le score de chaque phrase S selon les mots-clés qu'elle contient, on peut calculer le score suivant :

$$Scoremot - clés(S) = \sum a(w) * F(w) \qquad [\textbf{\textit{Equation1}}] \qquad , w \in S$$

$$a(w) = \{A \; si \; w \in liste \; de \; mots\_cles(A > 1) \; 1 \; sinon$$

et F(w) est la fréquence du terme w dans la phrase S

La liste de mots-clés peut être introduite par l'utilisateur (domaine d'intérêt) ou composée des mots- clés établis par l'auteur. L'importance du poids du terme w est donné par A * F(w), avec A > 1.

I . 8.1.2 .Titre

Etant donné que le titre est l'expression la plus significative et qui résume le mieux un document en quelques mots, on peut dire que la phrase qui ressemble le plus au titre est la plus marquante du document. Par conséquent, on peut attribuer à chaque phrase un poids en fonction de sa ressemblance avec le titre. [17]

Dans ce cas on considère les mots du titre du texte comme des mots-clés et on produit le résumé en sélectionnant les phrases qui couvrent certains mots apparaissant dans un titre.[12].

$$Scoretitre(S) = \sum wes \; b(w) * F(w) \qquad [\textbf{\textit{Equation2}}]$$

A si w ∈ liste de mot titre (A<1)

B(w) = 1 sinon

I .8.1.3 .Méthode de distribution de terme(DT) :

L'idée de cette méthode est de considérer comme «importantes» les phrases qui contiennent des mots «importants» du texte. Un mot est considéré important s'il est employé assez fréquemment dans le texte. En premier lieu, le texte source est traité pour calculer la fréquence de chaque mot de «contenu» du texte et en second lieu les fréquences sont utilisées pour associer un poids à chaque phrase.

On considère comme un même mot les mots dérivés de la même racine (par exemple, résumé, résumés et résumer). Une fois la fréquence de chaque mot calculée, une liste triée par fréquence est obtenue, il s'agit de la liste de distribution de termes. Pour mesurer le poids d'une phrase, on utilise le texte source et la liste de distribution de termes.

Ensuite, on choisit les phrases les plus « pesantes». La sélection peut être faite en termes d'un pourcentage du texte original, en nombre de phrases ou en nombre de mots. La génération du résumé consiste à juxtaposer les unités sélectionnées en ordre d'apparition dans le texte source. Les avantages de la méthode sont sa robustesse (n'importe quel texte aura un résumé) et sa facilité d'implantation. Les limitations sont toutefois nombreuses. Comme on ne prend pas en considération les relations entre les différents éléments du texte, le résultat risque d'être incohérent et même d'omettre de l'information importante.

I .8.2 .Méthodes de la position(P) :

Cette méthode a été introduite par Edmunson. [18] pour compléter la méthode de distribution de termes. Elle est utilisée en combinaison avec d'autres méthodes d'attribution de poids pour faire augmenter ou diminuer le poids d'une phrase. La méthode de la position considère que les premières et dernières phrases de chaque paragraphe sont importantes car elles sont considérés comme thématiques, c'est-à-dire elles «résument» le contenu du paragraphe, donc ces phrases auront leurs poids augmentés (cette affirmation est appuyée par les expériences de Baxendale[19].

La méthode considère aussi des phrases positionnées dans certaines sections conceptuelles importantes, par exemple dans «Introduction» et «Conclusion».On définit le score d'une phrase S à la position i comme suit :[20]

$$Score\ \ lead\ (Si) = \ \beta i \qquad \textbf{[Equation3]}$$

$\beta i\{B > 0\ si\ i < N\ 0 \qquad si\ i < N$

βi est une fonction rectangulaire qui modélise la distribution de phrases importantes selon leur position

L'inconvénient de cette méthode est qu'elle dépend de la nature du texte à résumer ainsi que du style de l'auteur.

I .8.3. Méthode à base d'expressions indicatives :

Cette méthode a été introduite par Paice.[21] ,elle choisit des unités de textes avec des indications spécifiques ou des expression spécifiques. Par exemple, pour les textes scientifiques, on a comme expression : le but de ce travail…, ce papier représente les résultats et les conclusions sont de bons candidats pour indiquer les phrases à inclure dans un résumé. De textes de types différents peuvent avoir des expressions indicatives différentes.

On peut déduire un score pour une phrase d'un texte quelconque à analyser en fonction de la ressemblance qu'elle présente, pour le trait donné.

On pourrait définir le score d'une phrase S correspondant à un certain motif comme :[22].

$$Score\ cue(S) = \ \{1\ si\ S\ correspond\ \text{à}\ un\ motif\ 0 \qquad sinon$$

[Equation 4]

I .8.4 .Méthode basée sur la classification des éléments :

Dans les textes de science et de technique il y a des phrases qui font référence à des catégories conceptuelles telles que: Connaissances Antérieures, Contenu, Méthode et Résultat, on peut également constater que dans les résumés de science et technique des informations relatives à ces catégories sont souvent retenues pour le résumé. Cette

approche essaie de classifier sémantiquement les phrases d'un texte tout en oubliant le contexte [23].

I .8.5. Méthode basée sur les approches hybrides :

Les méthodes présentées dans les sections précédentes utilisent des traits (fréquence, position, expression indicative, etc.) qui ne peuvent isolément garantir des résultats optimaux. On combine souvent ces trains par exemple avec l'équation suivante :

$$Score\ hybride\ (S) = a_1 * \text{Score}_{DT}(S) + a_2 * \text{Score}_{P}(S) + a_3 * \text{Score}_{exp.ind}(S) + a_4 * \text{Score}_{Titre}(S)$$

[***Equation5***]

Les poids a_i peuvent être fixés arbitrairement ou déterminés de manière expérimentale (par apprentissage par exemple). Dans le cas de textes journalistiques, Strzalkowski et al [24] ont combiné les méthodes de distribution de termes, du titre et de la position, en considérant la spécificité du texte. Pour garder la cohérence du texte, le résumé est composé d'une sélection de paragraphes pertinents.

- Les différentes méthodes de résumé automatique du texte présentées ci-dessus, offrent beaucoup d'avantages. Certaines méthodes semblent offrir de meilleurs résultats que d'autres, cela est du en grandes partie à la nature de texte et au style de l'auteur.

D'après les différentes méthodes, on peut constater que les méthodes d'extraction offrent certains avantages : simplicité de mise en œuvre.

Ⅱ.Extractions des informations :Deuxième partie :

Ⅱ.1. Introduction :

De nos jours nous disposons d'une quantité énorme d'écrits sur de très nombreux domaines scientifiques, techniques, littéraires, journalistiques, historiques, De plus en plus, ces écrits sont disponibles ou existent quelque part sous format électronique. Nous sommes de plus en plus submergés d'information. Il nous est cependant de moins en moins possible de lire l'ensemble des textes qui nous intéressent. Malgré cela nous cherchons à tirer l'information essentielle afin de prendre des décisions éclairées et ne pas refaire ce qui a déjà été fait, d'éviter des erreurs. Est-il possible, sans lire un ensemble de textes, d'en extraire l'information essentielle, les tendances, les associations entre les idées ? Peut-on découvrir des connaissances/informations non triviales, implicites, non connues, potentiellement utiles et compréhensibles à partir d'une grande masse de données textuelles ? C'est ce qui invoque l'opération d'extraction des informations dans le domaine du Text Mining.

Dans cette deuxième partie, nous allons présenter une définition de l'extraction, l'objectif de l'extraction ainsi une recherche bibliographique sur les différentes méthodes d'extraction existantes.

Ⅱ.2. Définition:

Il existe deux grandes directions pour l'extraction de l'information cachée dans les données textuelles :

(1) – Développer des méthodes de fouilles pour des objets **non structurés** (extraction de concepts, constructions d'ontologies), axe suivi par les linguistes, les analystes du langage naturel et de la sémantique.

(2) – Utiliser les outils de fouille (Data Mining) qui fonctionnent avec des données **structurées**, après avoir traité l'information qui est contenue dans les textes pour la mettre sous une forme adéquate (on parle alors de méta données, qui elles seront structurées), axe privilégié par les coutumiers du Data Mining traditionnel. Nous nous placerons dans le cadre du second axe.

Ⅱ.3 .But :

La fouille de données textuelles (Text Mining) a pour but d'extraire des patrons intéressants à partir d'un ensemble de données textuelles. Malheureusement les données textuelles sont moins structurées que les habituelles bases de données, ou fichiers Excel. Les informations ne signalent pas leur présence et peuvent être cachées à différents endroits. Aussi le même concept peut être énoncé de différentes manières avec du vocabulaire différent.

La fouille de données textuelles s'effectue par une étape essentielle qui vise à extraire de chaque texte une information pertinente. Il s'agit principalement de filtrer le texte (suppression des termes vides de sens (Mots vides : le, la, un, une …) et de compter les termes identiques (avec conjugaison, synonymes, …) afin de construire un vecteur pondéré qui contient seulement les mots « pertinents » du texte.

Ce vecteur est exploité pour :

- Classer les documents en fonction de leur contenu
- Regrouper les documents de contenu proche et organiser les documents en hiérarchies
- Retrouver un document à partir de son contenu
- Dresser des relations entre les personnes/lieux/organisations/concepts
- Extraire des caractéristiques dans les textes

Ces points permettent de donner une vue sur les possibilités offertes par la fouille de données textuelles.

Ⅱ.4. Méthodes d'extraction des données :

Ⅱ.4.1. **Méthode d'extraction d'information à partir de textes :**

La méthode d'extraction d'information à partir de textes est en trois étapes. La première étape vise à extraire des métas donnés sur chaque texte. Ces métas donnés, sous forme de vecteur, auront pour but de représenter le texte.

Ensuite ces métas donnés seront exploitées avec des outils de fouilles structurés. Losiewicz et al. [25] considèrent trois étapes dans la recherche d'information. [26]

La première étape : (Data collection) à pour but de rassembler les sources documentaires pertinentes (Source selection) et de prétraiter les textes pour retirer les informations non utilisables pour ne garder que l'essentiel du document, là où est sensée être l'information recherchée (Feature extraction). La sélection d'un corpus de textes est cruciale, autant le corpus de textes est adapté, autant les réponses seront pertinentes. Aussi les documents doivent être datés, car une information, une relation évolue en fonction du temps. [27]

La seconde étape : actuellement la plus cruciale et la plus difficile est aussi la plus sujette à la critique. Il s'agit d'extraire de chaque texte l'ensemble de son contenu (Meta data assignement) et de le représenter sous une forme adaptée à la future fouille (Data Storage). En pratique l'ensemble des mots du texte est supposé apporter de l'information. On extrait donc chaque mot du texte que l'on gardera dans une base de données. Cependant quelques transformations peuvent être effectuées pour enrichir cette étape. Toutes les transformations ne sont systématiquement pas employées, cela dépend des textes utilisés et de ce que l'on cherche à prédire. Notons les transformations suivantes : retirer les mots peu significatifs (le, la, les, du, des, …), transformer les majuscules en minuscules, conjuguer les verbes à l'infinitif, regrouper les mots de même racine, regrouper les synonymes … Des codes informatiques sont disponibles sur internet pour chacune de ces tâches, mais dépendent de la langue considérée. Toutes les méthodes ont pour but de réduire le nombre de mots utilisés pour représenter un texte, simultanément la fréquence des mots qui restent augmente.

La troisième étape : l'exploitation des données (Data exploitation) se divise en deux étapes : la fouille (Data Mining) va permettre de construire des modèles, de faires ressortir des liens entre les « données » ; les résultats de la fouille ne seront souvent utilisables qu'avec une présentation adéquate à l'utilisateur (Présentation). [28]

Ⅱ.4.2. Méthode Format des documents :

Les documents à explorer peuvent avoir plusieurs formats, il peut s'agir par exemple d'un format propriétaire plus ou moins facile à exploiter, d'un fichier ASCII, ou d'une image si le texte est simplement scanné sans autre traitements. La plupart des travaux supposent que les documents sont formatés selon la convention XML. Ce format est couramment employé, notamment sur le web. Un document XML contient des données de structure telles que <TITLE>, <AUTHOR>, <ASTRACT>, <TEXT> … qui serviront à délimiter les différents éléments du texte. Par la suite ces différents éléments pourront être traités différemment en fonction de ce que l'on cherche à produire. Notamment les mots apparaissant dans les différentes sections ne seront pas pondérés de la même manière

Ⅱ.4.3. Méthode de représentation d'un texte par un vecteur !

Chaque texte sera représenté par un vecteur. Ce vecteur devrait contenir la totalité (ou un maximum) de l'information pertinente contenue dans le texte. Un certain nombre d'informations peuvent être extraits avec peu ou pas de perte d'information : l'auteur, l'institution à laquelle appartient l'auteur, le format du document (article, livre, résumé, page web, …), la date, les mots clefs, la liste des références utilisées par l'auteur, la revue, les numéros de page…

Il reste cependant encore à extraire l'information contenue dans le reste du texte (la plus grande partie du document!). En pratique, on procède par l'extraction de tous les mots contenus dans le texte. Les termes vides de sens sont retirés (par exemple : le, la, les, du, des, …). Toutes les majuscules deviennent des minuscules. Toutes les conjugaisons sont converties à l'infinitif. Les mots de même racine et les synonymes peuvent être traités. Ensuite chaque texte est représenté par un vecteur qui décrit l'ensemble des mots contenus dans le texte (après traitement).

Le vecteur est construit dans le but de représenter le contenu du texte. Chaque coordonné du vecteur représente un mot du texte :

$d = (w_1, w_2, w_3, \dots w_n)$ [28]
w_i est le poids du $i^{ème}$ terme
w_i montre l'importance du terme ***i*** dans le texte ***d.***

Ⅱ.4.4. Méthode dictionnaires :

Faire appel à un dictionnaire des synonymes permet de regrouper les termes qui ont la même signification. Cependant il faut être prudent, dépendamment du contexte le même mot peut dramatiquement changer de sens. Par exemple le mot « bâtiment » peut se référer à de nombreux types de construction d'habitation, de bureaux, d'usine, d'hôpitaux … pour un architecte, ou bien il peut s'agir de différents types de bateaux de guerre pour un militaire.

On utilisera donc plutôt des dictionnaires thématiques que des dictionnaires des synonymes à vocation générale. Par exemple, dans notre cas, on peut utiliser un dictionnaire de termes chimiques, cela permet à l'ensemble des utilisateurs de partager (d'apprendre, d'imposer) le même vocabulaire. Tout le monde est donc sensé trouver l'information pertinente vis-à-vis du domaine étudié.

Ces dictionnaires sont réalisés par apprentissage. L'utilisateur (ou le groupe d'utilisateur) définit lui-même en fonction de son contexte (de vie, de travail, de culture …) son dictionnaire des synonymes.

Ces dictionnaires électroniques sont des algorithmes spéciaux qui donnent la priorité aux éléments supposés les plus importants du texte. Ce sont des bases de données de mots et de concepts de mots conçus entre autre pour faciliter la reconnaissance automatique de parole. **Cette méthode utilise ce dictionnaire afin de comparer son contenu avec les différents mots du résumé.** [28]

II .4.5. Méthode Notions de distance :

Afin de traiter les informations de similarité entre documents il a fallu définir des notions de métrique. Ces notions pourront être utilisées pour comparer des documents, évaluer la pertinence d'un document vis-à-vis d'un ensemble de mots clés, regrouper les documents similaires, classifier un document nouveau …

Différentes notions de distances sont définies.

- La distance entre deux textes est donnée par la mesure suivante :

$$Cos(d1, d2) = \frac{d1*d2}{||d1||*||d2||} \qquad [\textbf{\textit{Equation 6}}]$$

Ceci représente le produit scalaire de d_1 avec d_2, divisé par le produit des normes des vecteurs. [30]

- La distance d'un texte à un ensemble de texte peut se calculer de différentes manières : [29]

- Plus proche voisin :

$$D = min(d(Pi, Pj), Pi \in G1, Pj \in G2) \qquad \textbf{\textit{[Equation 7]}}$$

- Voisin le plus éloigné :

$$D = max(d(Pi, Pj), Pi \in G1, Pj \in G2) \qquad [\textbf{\textit{Equation 8}}]$$

- Distance moyenne :

$$D = \sum_{J}^{n2} \frac{d(P1,Pj)}{n2} \qquad [\textbf{\textit{Equation 9}}]$$

- Distance de centres de gravité :

$$D = d(\theta 1, \theta 2) \qquad [\textbf{\textit{Equation 10}}]$$

Toutes ces mesures sont appliquées et permettent des regroupements différents

Ⅲ. Conclusion :

Dans ce chapitre nous avons vu deux étapes très importantes de notre travail : ***Le résumé automatique du texte*** avec ses types et les méthodes de réalisation possibles pour leurs modèles d'automatisation, on a choisit **Méthode basée sur la classification des éléments, car** c'est cette méthode seule, qui nous a aider à aboutir notre but.

Et les différentes méthodes d'***extraction des éléments pertinents***, dans notre cas d'étude, on a choisis la méthode : **Dictionnaire** qui se base sur la comparaison directe des éléments du texte résumé et un dictionnaire généré manuellement à partir de la documentation sur le web, dans le domaine choisis : **chimie.**

CHAPITRE 3

Contribution

1. Introduction :

La génération du résumé du texte est un travail fait depuis longtemps, mais générer un résumé du feuille de TP qui contient juste les informations nécessaires et suffisantes

pour comprendre le principe du TP et les différentes étapes de son déroulement, c'est le plus grand défit qui nous est présenté.

Problème : Pour faire ce résumé, on doit prendre en considération la pertinence de ces informations.

2. Objectif :

L'objectif de ce mémoire est de résumer une feuille de TP sous forme textuelle dans le domaine de chimie, et d'extraire les informations pertinentes dans ce domaine

1- Automatisation de la feuille de TP : Appliquer quelques procédés afin de rendre cette feuille prête pour un ensemble de traitement.
2- Résumer la feuille de TP : Vise à réduire la taille de la feuille de TP afin de minimiser le nombre des mots contenants cette feuille tout en gardant le sens du contenu.
3- Extractions des informations pertinentes : Ce sont les mots qui ont une relation directe avec le sujet de la feuille de TP, et dans notre cas d'étude est le domaine de chimie.

3. Problématique :

Lors de l'analyse du thème de la génération automatique du résumé de la feuille de TP on déduit les points problématiques suivants :

1- La spécificité de la feuille de TP : il ne s'agit pas d'un simple texte, mais il contient plusieurs autres éléments tel que : les images, les équations, terminologies,….
2- Résumer un texte peut entrainer la perte d'information ce qui mène à une mauvaise compréhension du contenu de la feuille de TP ; provoquant des erreurs de déroulement du TP (Mauvais apprentissage)
3- Les méthodes existantes ne correspondent pas à notre contexte .Dans ce cas, on doit adapter les techniques existantes pour avoir des bons résultats.

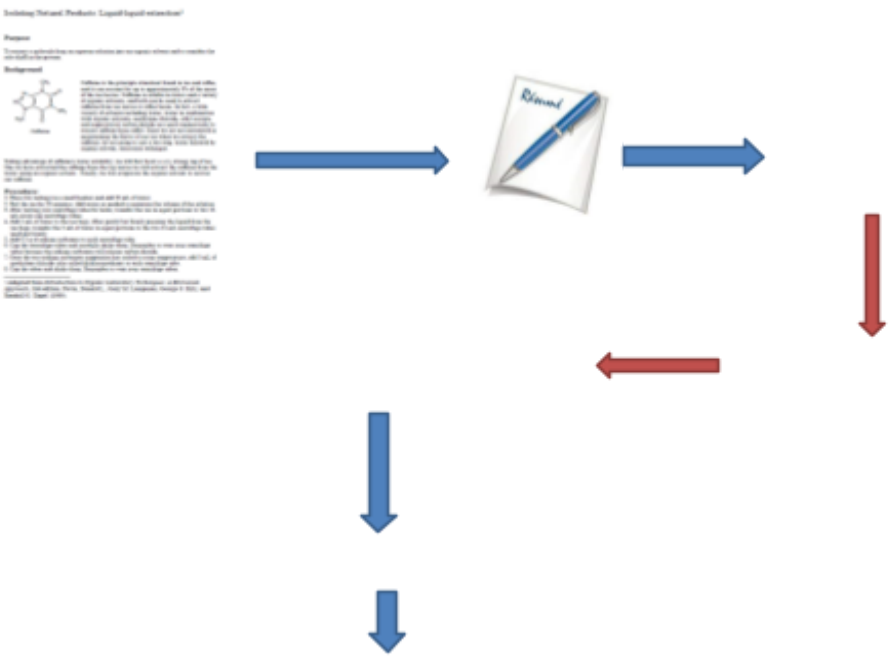


Figure.6 .Problématique

4. Contribution :

4.1 Contexte

Notre travail est basé sur une feuille de TP et précisément en spécialité de chimie. Ce choix n'a pas été fait aléatoirement, mais après une discussion avec des enseignants spécialisés en chimie, et après avoir vu le déroulement des différentes étapes du TP, on a choisi 3 feuilles de TP :

1- « Preparation of benzoic acid » : Module : Chimie organique, 2eme année Master, spécialité : Génie des procédés.
2- « liquid-liquid_extraction(1) » : Module : Génie de l'environnement, 1ere année Master.
3- « Synthesis-of-Aspirin » : Module : Chimie organique, 2eme année Licence.

Ces feuilles de TP présentent une source pour notre travail.

5. Description de la feuille de TP :

La feuille de TP fait généralement référence à une feuille de papier avec des questions ou des exercices que les étudiants doivent accomplir.

5.1 .Contenu

5.1.1. Des verbes d'actions : Les verbes d'action sont des verbes faisant partie d'une des classes des verbes. Ils expriment une action faite ou subie par le sujet. [30]

- Exemples : to extract, to add, to to wait, to mix, to heat, to notice….etc

5.1.2- Les produits chimiques utilisés dans les laboratoires d'enseignement

- Exemples : Vitamins, Chelating Agents ,Amino Acids, Surfactants, Emulsifiers, Thickeners, Acidulants, Binders, Antioxidants, Defoamers, Minerals, Proteins, Nutraceuticals, Fillers, Lubricants, Solvents, Preservatives, Pigments, Oils, Bases, Acids ,et plus.

5.1.3- Le matériels

-Exemples :

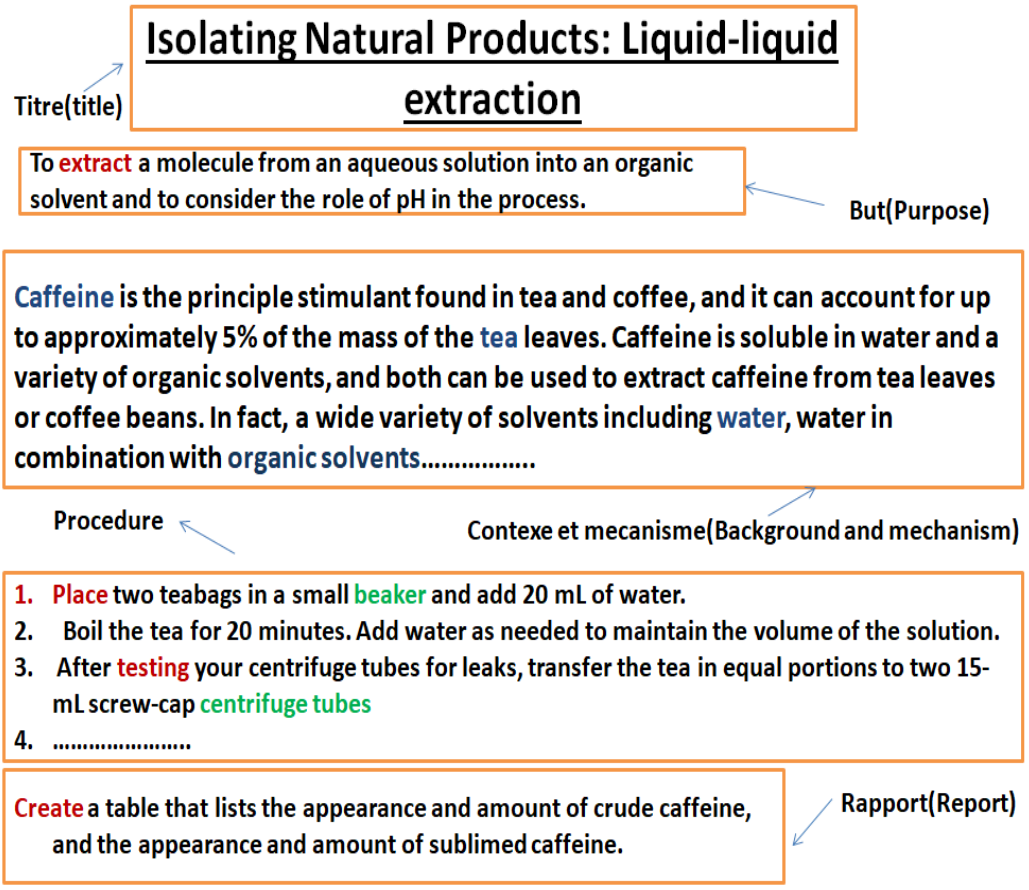
Beaker

round bottom flask

wash bottle …etc

Figure 7.Exemples de materiels

Voici des exemples dans le schéma précédant :



Rouge=verbe,Bleu=produits,Vert=materiels

Figure 8.Exemple de résultats d'extraction

5.2. Définition des mots pertinents dans une feuille de TP :

Les mots pertinents sont des mots qui ont une signification précise ou logique à la matière, et dans une feuille de TP, ce sont les mots appartenant au domaine chimique.

6 : Processus de génération automatique du résume de la feuille de TP :

La figure suivante, illustre toutes les étapes de notre processus de génération

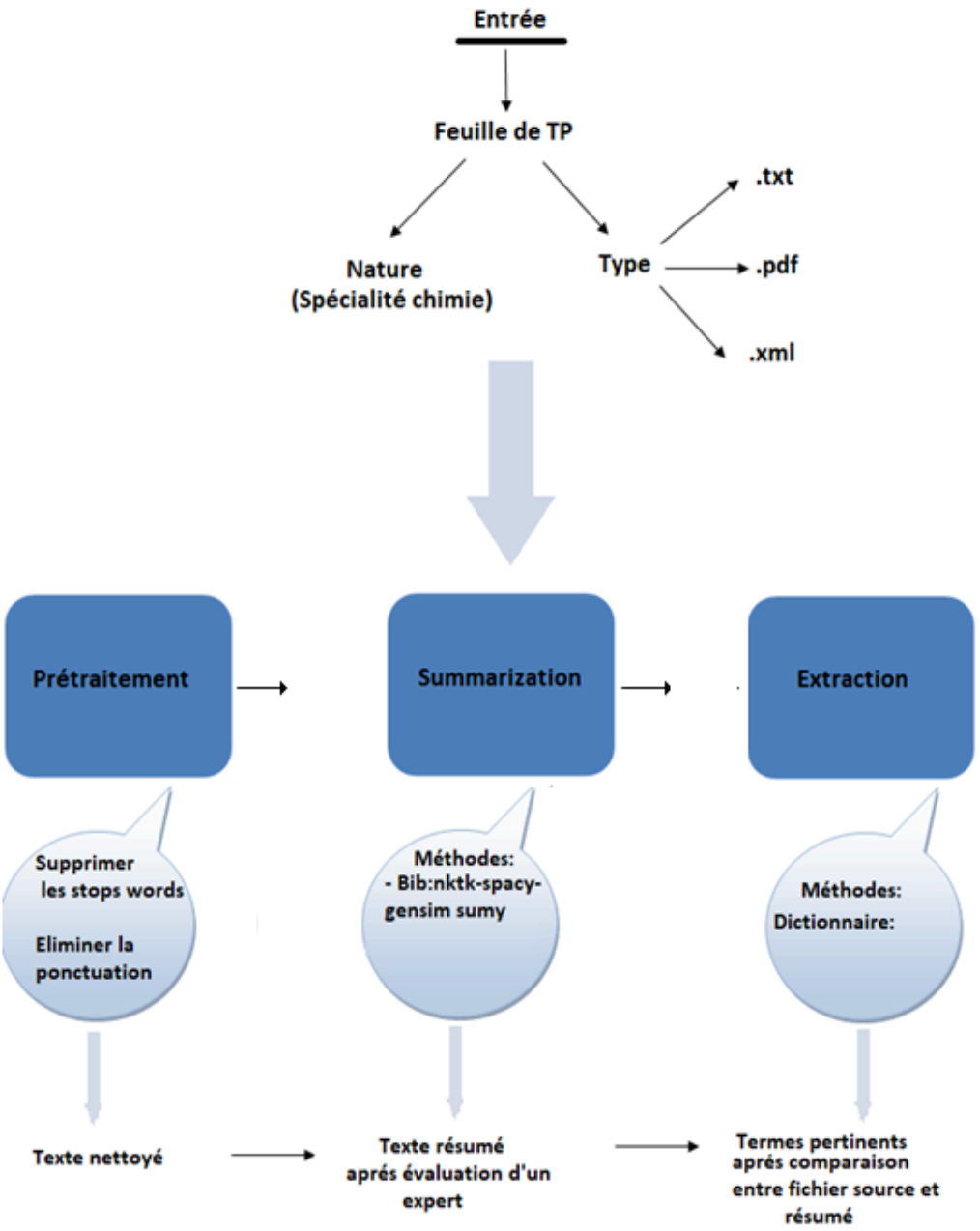


Figure 9. Architecture générale du processus de résumé extraction

6.1 .Description des Etapes :

6.1.1. L'acquisition ou la sélection : c'est la source de donnée présentée sous forme de fichier textuel obtenu à partir du texte original : .PDF qui représente notre feuille de TP.

6.1.2. Le prétraitement :

6.1.2.1.. Nettoyage : Consiste à éliminer les Stopwords (mots vides) et quelques signes de ponctuation afin de garder juste les mots qui ont un sens dans le domaine

6.1.2..2. Tokenization : Consiste à découper notre corpus ou texte en plusieurs pièces appelés : Token

Exemple : « Vous trouverez en pièce jointe le document en question » ;

« Vous », « trouverez », « en pièce jointe », « le document », « en question » [31]

6.1.2..3. Stemming : On peut trouver les mêmes mots sous différentes formes en fonction du genre (masculin ou féminin), du nombre singulier ou pluriel. Le Stemming est généralement le processus brut qui consiste à découper la fin des mots dans afin de ne conserver que la racine du mot.

Exemple: trouverez ⟶ trouv

6.1.2..4. Lemmatization : On applique un traitement lexicale à notre texte.

Exemple : le lemme petit renvoie 4 formes : petit, petite, petits, petites.

6.1.2.5 .schéma :

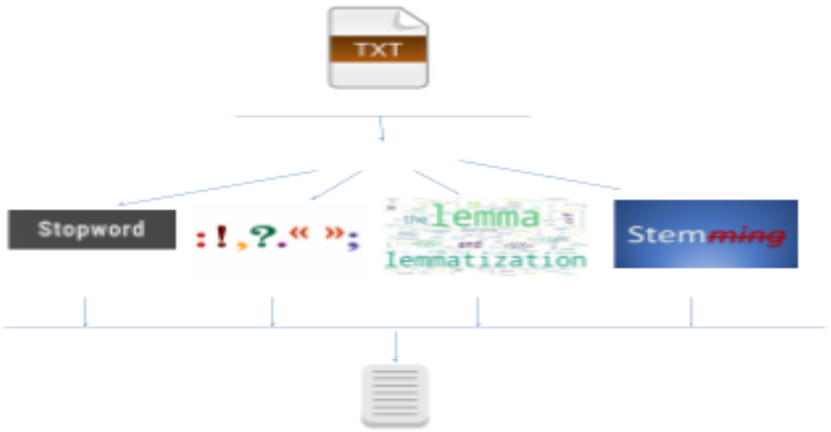


Figure 10. Le processus du prétraitement

6.1.2.6. algorithme :

```
Algorithme Prétraitement :
    Entrée : Fichier à traiter(F), Liste_mot_vide_en(Vide),
             Liste_ponctuation(P)
    Procédure : read (f)
                Fot in F
                    If (mot is in Vide) or (mot in P)
                        Delete (mot)
                    end If
                end For
    Sortie :  Fichier traite=Fichier à traiter
end
```

6.1.3.- résumé automatique du texte (Summarization):

6.1.3.1.étapes

a- Récupérer le corpus traité (de l'étape précédente)

b- Appliquer les méthodes de summarization

c- Choisir le résumé le plus accepté selon un expert

d- Sauvegarder le résultat sous forme de fichier texte nommé : résumé.txt

6.1.3.2. Schéma

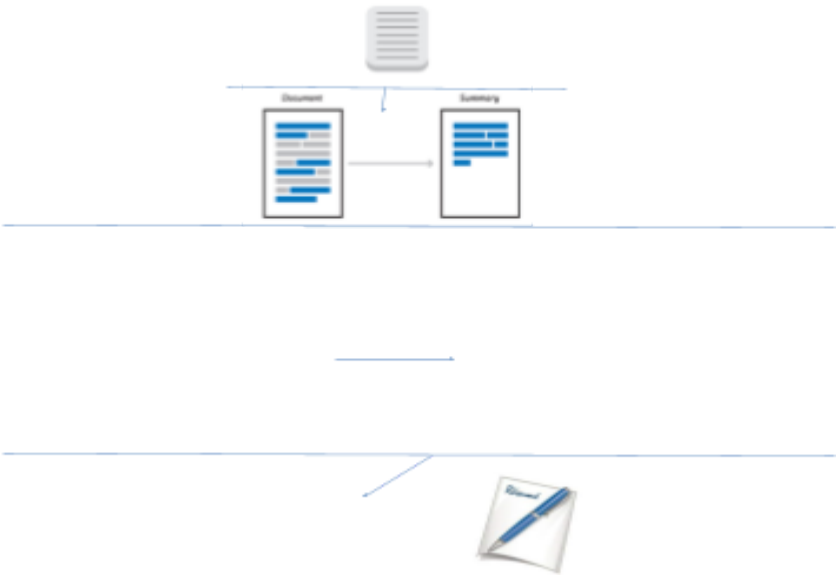


Figure 11. Le processus du résumé du texte(Summarization)

6.1.4. Extraction des informations:

6.1.4.1 :-étapes

1- Récupérer le fichier résumé (de l'étape précédente)

2- Appliquer la méthode d'extraction qui nous aide à aboutir notre but.

Dans notre cas, nous avons choisis la méthode de dictionnaire.

Cette méthode passe par les étapes suivantes :

1 : Récupérer le fichier résumé

2 : Construire un fichier source sous forme textuelle nommé : source.txt

3 : Comparer les deux fichiers en parcourant le fichier résumé mot par mot

Et en le comparant par le texte résumé.

4 : Construire un fichier contenant les termes extraits après la comparaison et qui présentent les termes ou les mots pertinents.

6.1.4.2. Schéma

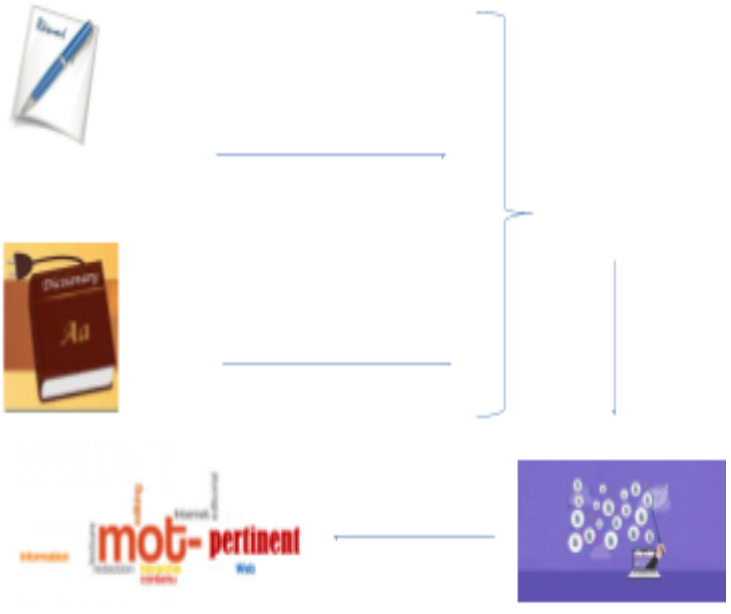


Figure 12. Le processus d'extraction des éléments

6.1.4.3.:Algorithme :

```
Algorithme Extraction :
    Entrée : Fichier source(S), Fichier résumé(R),
             Liste_mot_commun(MC)=[]
    Procédure : read (S),
                Read(R),
                For mot in R
                    If mot is S
                    Liste_mot_commun(MC)=
                    Liste_mot_commun(MC)+mot
                    end If
                end For
    Sortie :  Fichier_mots_pertinents= Liste_mot_commun(MC)
end
```

6.1.5. Création de fichier source :

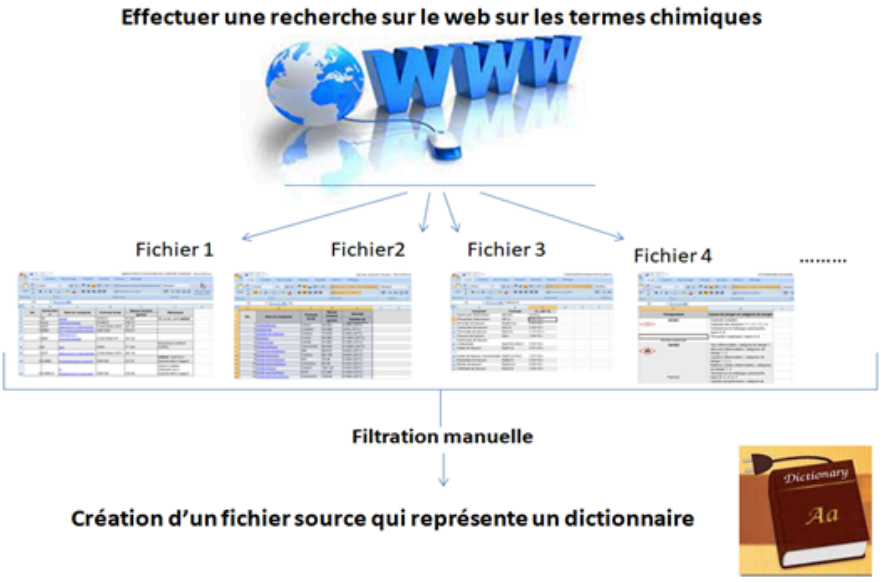


Figure 13.Le processus de la création du fichier source pour l'extraction

6.1.5.1.Description de l'algorithme :

Notre algorithme passe par trois étapes :

1- Après avoir faire le résumé du texte original, on le sauvegarde sous le nom : résumé final.txt
2- On crée un fichier texte nommé : Source.txt qui présente un dictionnaire ou une base référentielle [32] de tous les termes chimiques : verbes d'action, noms des produits utilisés dans les labos d'enseignement et nom des matériels utilisés dans les TP.
3- On parcourt les deux fichiers pour extraire les mots communs qui présentent des éléments pertinents dans notre cas.

7 : Conclusion

Dans le travail de l'extraction, on passe toujours par les mêmes étapes, afin d'obtenir les éléments pertinents, mais en voyons ces étapes plus profondément, on trouve que chaque étape amène à créer un résumé, c'est pour cela qu'il ne faut pas confondre entre la deuxième étape du travail général qui est la Summarization avec toutes ses méthodes, et l'extraction qui about à la fin d'avoir un résumé plus précis contenant seuls les éléments pertinents qui ont une relation directe avec le domaine choisis.

CHAPITRE 4

Implémentation et résultats

1. Introduction :

Après avoir faire l'étude conceptuelle de notre approche d'extraction les éléments pertinents, nous allons présenter dans ce chapitre la phase de réalisation et d'implémentation du notre système.

Ce chapitre a pour objectif de présenter l'aspect implémentation de notre application, il s'agit donc d'expliquer l'environnement matériel sur lequel notre système a été développé, les langages de programmation et les outils utilisés. Par la suite, nous allons présenter les interfaces graphiques en décrivant les différentes fonctionnalités de notre application et nous présenterons un exemple qui nous permettra d'illustrer les résultats obtenus lors de l'utilisation de notre approche.

2. Les Outils Utilisés:

Dans cette partie, nous allons présenter la définition du langage que nous allons utilisés dans l'implémentation et la réalisation de notre système.

2.1 .Le langage :

Python est un langage de programmation interprété multi-paradigme. Il favorise la programmation impérative structurée, et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Scheme, Small talk et Tcl. [33].

Le langage Python est placé sous une licence libre proche de la licence BSD et fonctionne sur la plupart des plates-formes informatiques, des supercalculateurs aux ordinateurs centraux, de Windows à Unix en passant par Linux et MacOs, avec Java ou encore .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. Il est également apprécié par les

pédagogues qui y trouvent un langage où la syntaxe, clairement séparée des mécanismes de bas niveau, permet une initiation plus aisée aux concepts de base de la programmation.

Python est un langage qui peut s'utiliser dans de nombreux contextes et s'adapter à tout type d'utilisation grâce à des bibliothèques spécialisées à chaque traitement. Il est cependant particulièrement utilisé comme langage de script pour automatiser des tâches simples.

Figure 14. PYTHON

Ce choix a été motivé par les raisons suivantes :

- L'une des principales langues parmi les langues appropriées pour la programmation de problèmes d'apprentissage profond.
- Il dispose un grand nombre de bibliothèques pour le traitement du langage naturel, telle que NLPnet, NLTK, …
- Un langage simple, productif et utilisable dans presque tous les domaines et systèmes

2.1.1. L'environnement de développement PyCharm :

Pycharm est un environnement de développement intégré (IDE) utilisé dans la programmation informatique, spécifiquement pour le langage Python. Il est développé par la société tchèque JetBrains. Il fournit l'analyse de code, un débogueur graphique, un testeur d'unité intégré, l'intégration avec des systèmes de contrôle de version et prend en charge le développement Web avec Django ainsi que la science des données avec Anaconda.

Figure .15. Pycharm

2.1.1.1. Tkinter :

Tkinter (de l'anglais Tools kit interface) est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. Elle vient d'une adaptation de la bibliothèque graphique Tk écrite pour Tcl. [34]

2.1.2 Les bibliothèques :

2.1.2.1 .NLTK (Natural Language Toolkit) :

Guide de l'outil de Traitement Naturel du Langage en Python

Le **NLTK, ou Natural Language Toolkit**, est une suite de bibliothèques logicielles et de programmes, dédiée au traitement naturel du langage ou Natural Language Processing.

Elle est conçue pour le traitement naturel symbolique et statistique du langage anglais en langage Python. C'est l'une des bibliothèques de traitement naturel du langage les plus puissantes.

Cette suite d'outils rassemble **les algorithmes les plus communs du traitement naturel du langage** comme le Ttokenizing, le part-of-speech Tagging, le Stemming, l'analyse de sentiment, la segmentation de topic ou la reconnaissance d'entité nommée.

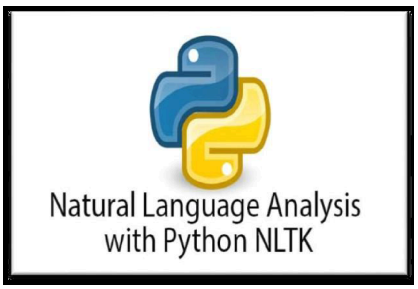


Figure 16.NLTK

2-1-2-2 Spacy :

spaCy est l'une des principales bibliothèques du langage Python pour le Traitement Naturel du Langage (NLP)

spaCy est **une bibliothèque Python gratuite et open source** publiée sous la licence MIT pour le traitement naturel du langage (Natural Language Processing ou NLP). Elle est écrite en Cython, et conçue pour l'usage en production grâce à une API concise et simple d'utilisation.

Cette bibliothèque est **initialement développée par Matt Honnibal de Explosion AI**. Pour les connaisseurs du langage Python, on peut considérer spaCy comme l'équivalent de numPy pour le NLP : une bibliothèque de bas niveau, mais intuitive et performante.

Grâce à cet outil, il est possible de créer des applications permettant **de traiter et de comprendre de larges volumes de texte**. Il peut être utilisé notamment pour développer des systèmes d'extraction d'information, de compréhension du langage naturel, ou encore pour prétraiter des textes pour le Deep Learning.

Figure.17. spaCy

2.1.2.3. TextBlob :

TextBlob module est une bibliothèque Python et offre une API simple pour accéder à ses méthodes et effectuer des tâches NLP de base. Il est construit sur le dessus du module NLTK.

Figure.18.TextBlob

3. Implémentation :

3.1 Démarrage de l'application : Après avoir lancer l'exécutable, notre application démarre. En voici la page de démarrage :

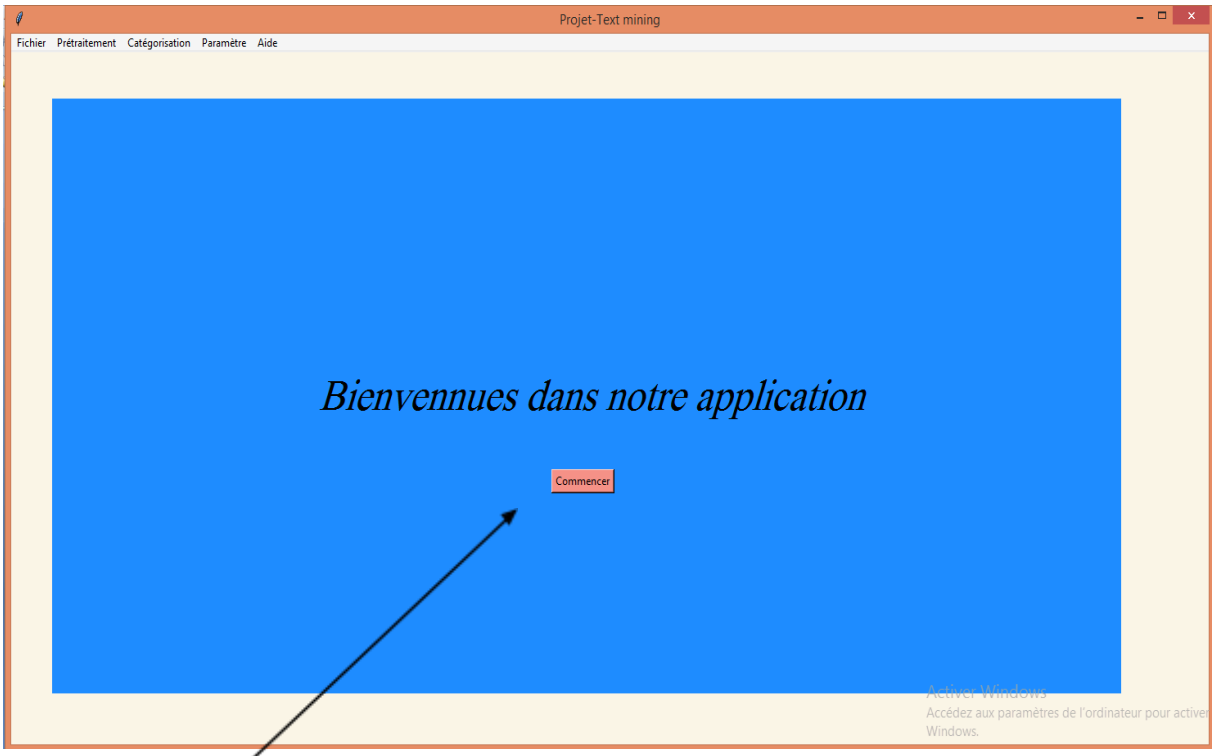
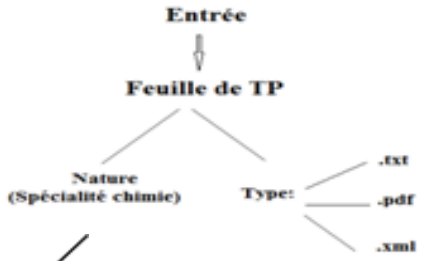


Figure19. La fenêtre de démarrage de notre application

Lancer l'application

3.2. Le prétraitement : C'est la première étape, en voici le processus :



Etape1

Supprimer
La ponctuation

```
def punk():
    pont =" [<a[^>]*>(.*?)</a>]"
    raw_text = str(frame2_display5.get('1.0', tk.END))
    tokens = raw_text.split()
    clean_tokens = [t for t in tokens if not t in pont]
    clean_text = " ".join(clean_tokens)
    print(raw_text, clean_text)
    frame2_display6.insert(tk.END, clean_text)
```

Eliminer les stop_words

Etape2

```
def stop():
    stopwords = ['your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she',
                 "she's" , 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
                 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this',
                 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be',
                 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing',
                 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until',
                 'while', 'of', 'at']
    raw_text = str(entry1.get('1.0', tk.END))
    tokens = raw_text.split()
    clean_tokens = [t for t in tokens if not t in stopwords]
    clean_text = " ".join(clean_tokens)
    print(raw_text, clean_text)
    frame2_display5.insert(tk.END, clean_text)
```

Etape4 **:Lemmatisation**

```
def nlpiffy_file():
    raw_text = str(frame2_display6.get('1.0', tk.END))
    docx = nlp(raw_text)
    final_text = [ (token.text,token.shape_,token.lemma_,token.pos_) for token in docx ]
    result = '\nSummary:{}'.format(final_text)
    frame2_display7.insert(tk.END,result)
```

Etape3 **: Tokenization**

```
def get_tokens():
    raw_text = str(entry1.get('1.0', tk.END))
    #raw_text = str(raw_entry.get())
    new_text = TextBlob(raw_text)
    final_text = list(str(new_text.words).split(" "))
    result = '\nTokens:{}'.format(final_text)
    frame2_display.insert(tk.END, result)
```

3-1 Le prétraitement présenté par notre application :

Figure.20. Le processus du prétraitement d'un document .TXT

Le prétraitement est fait à l'aide la bibliothèque NLTK.

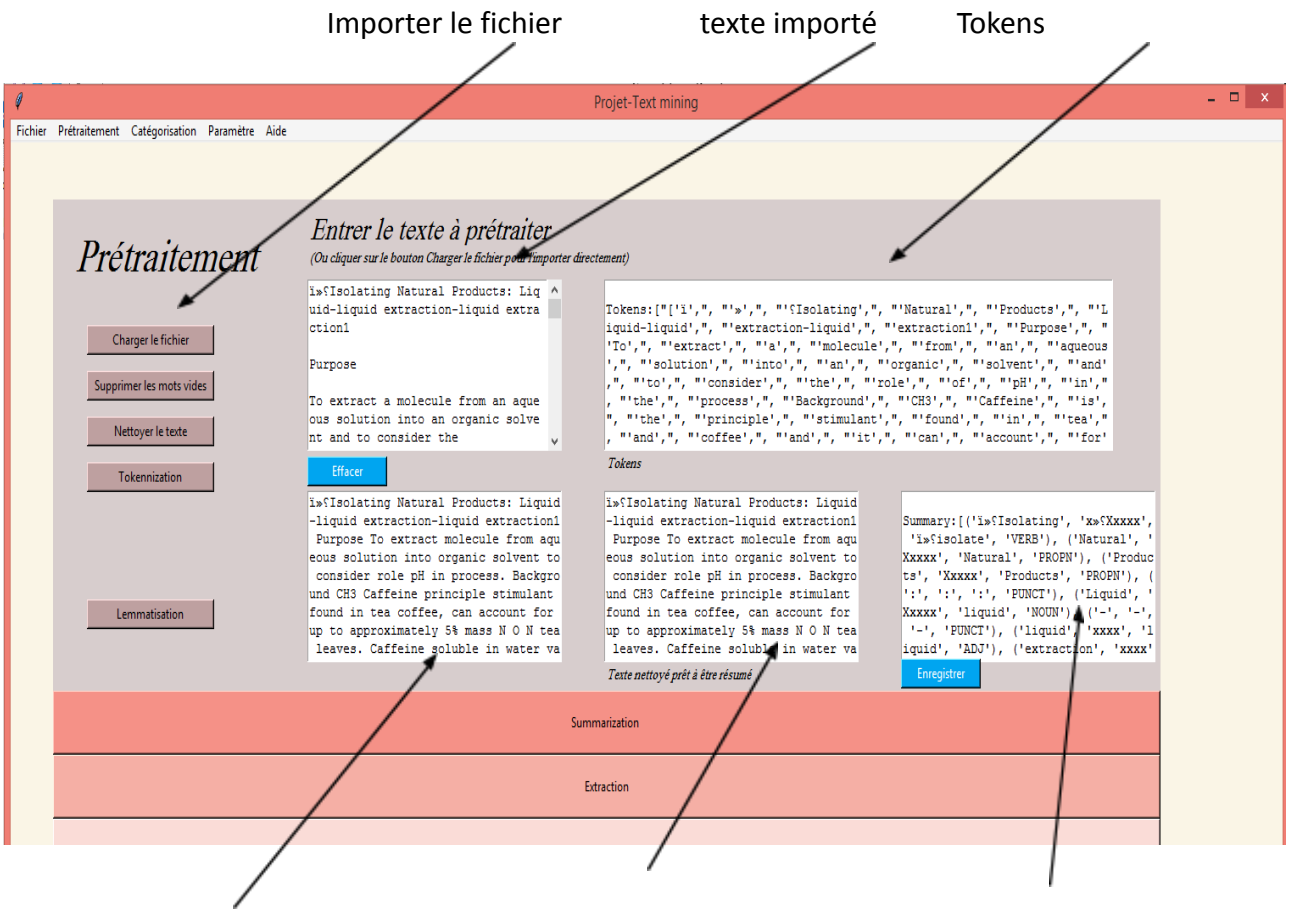


Figure.21. Le prétraitement présenté par notre application

3.3. Résumé automatique du texte (Summarization) : Consiste à faire un résumé du texte nettoyé de l'étape précédente, en voici le processus :

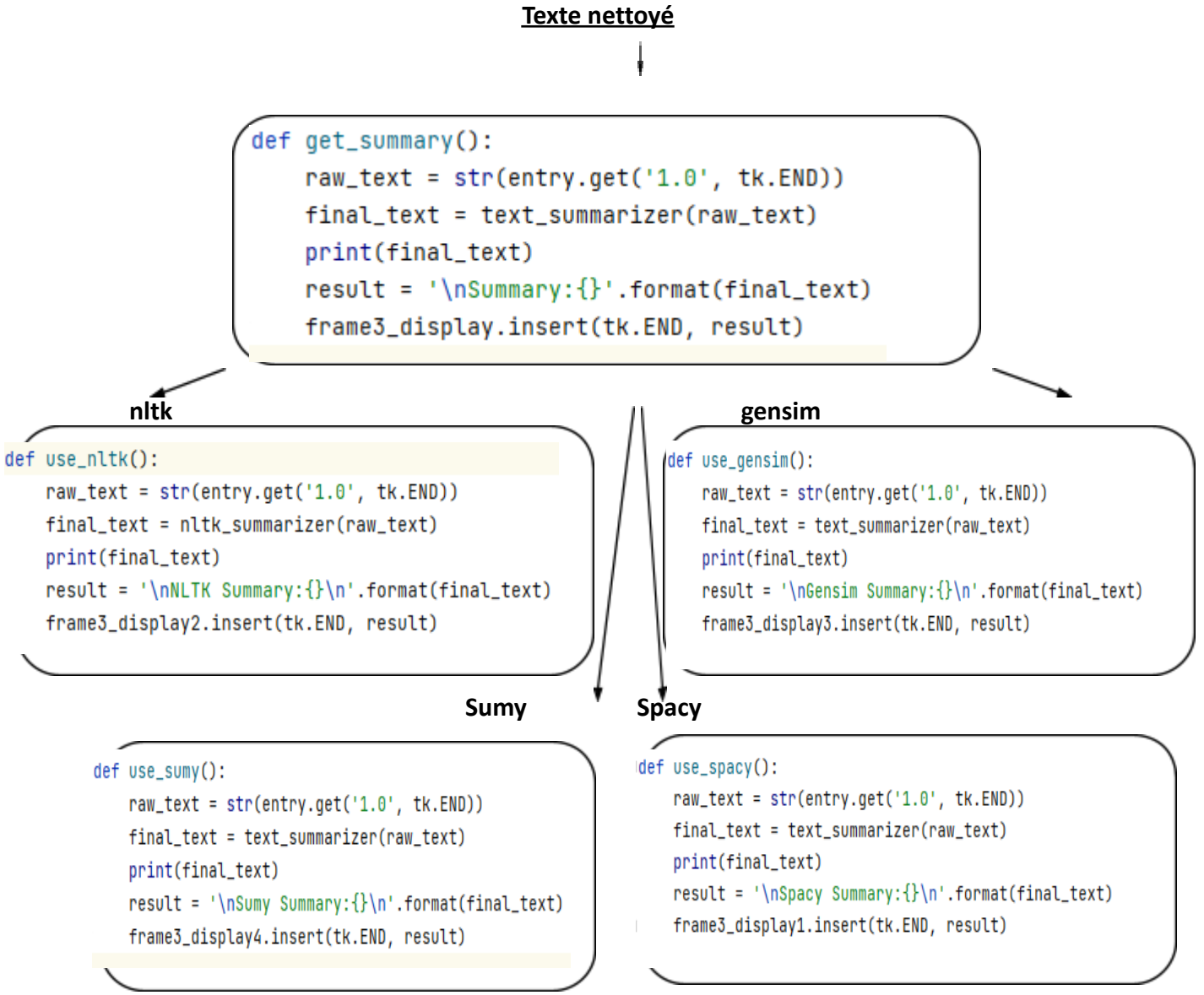


Figure.22. Le processus du résumé automatique du texte (Summarization)

Voici les résultats du résumé avec les quatre bibliothèques utilisée :

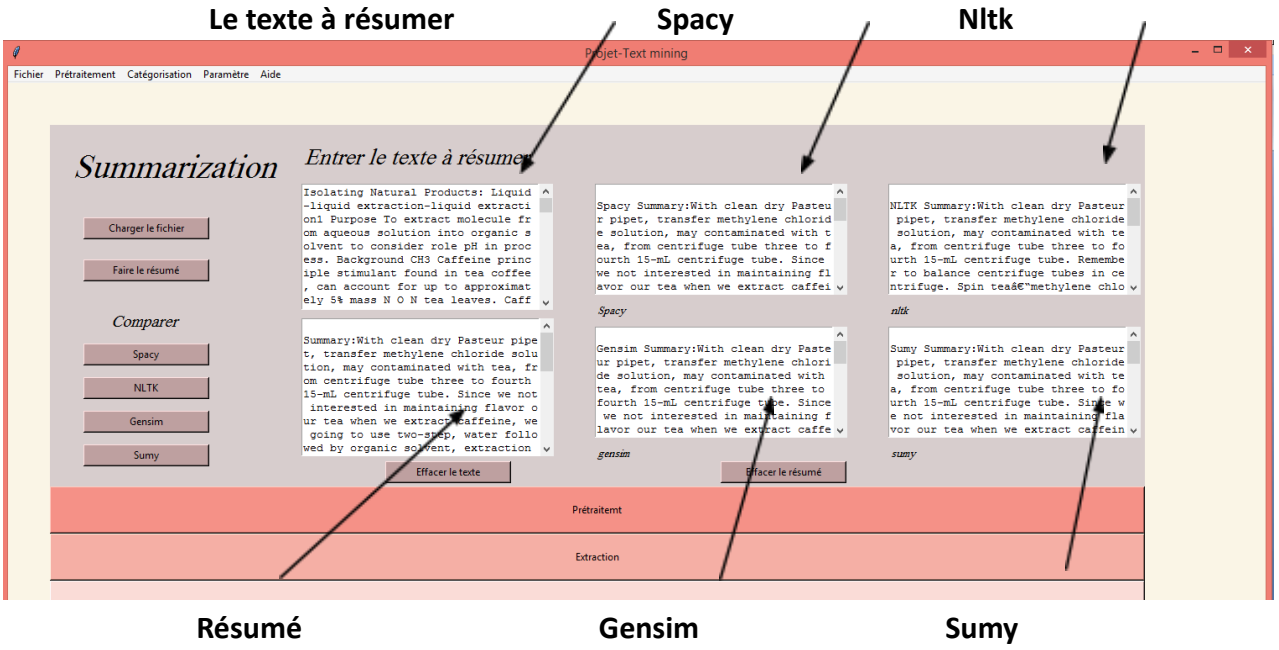


Figure.23. Le résumé automatique du texte présenté par notre application

3.3.1. Comparaison :

Toutes ces bibliothèques (Spacy, Nltk, Gensim et Sumy), sont des bibliothèques open-source gratuites du langage python, pour le traitement naturel du Langage(NLP).

Mais le choix d'utilisation de chaque bibliothèque diffère selon le but du travail et la nature du texte.

1. Spacy : est conçu spécifiquement pour une utilisation en production et vous aide à créer des applications qui traitent et « comprennent » de gros volumes de texte. Elle

peut être utilisée pour créer des systèmes d'extraction d'informations ou de compréhension du langage naturel, ou pour prétraiter du texte pour un apprentissage en profondeur.

Avantage :

- Le Framework PNL le plus rapide
- Utilise le réseau de neurones pour entraîner certains modèles.
- Intègre la fourniture des vecteurs de mots
- Traite des objets ; plus précisément orienté- objets, par rapport aux autres bibliothèques

Inconvénients :

-Manque de flexibilité par rapport au NLP
- La tolérisation des phrases est plus lente que NLTK.
- Ne prend pas en charge de nombreuses langues

1- **Gensim :** est présentée comme un package de traitement du langage naturel qui effectue la « modélisation de sujets pour les humains ». Mais c'est pratiquement beaucoup plus que cela. Il s'agit d'un package leader et à la pointe de la technologie pour le traitement de textes, le travail avec des modèles vectoriels de mots (tels que Word2Vec, FastText, etc.) et pour la création de modèles de sujets.

Avantages :

- Travaille avec de grands ensembles de données et traiter des flux de données.
- Soutient l'apprentissage en profondeur (Deep Learning).

Inconvénients :

- N'a pas assez d'outils pour fournir un pipeline NLP complet , elle doit donc être utilisée avec une autre bibliothèque (spacy ou nltk)
- Conçue principalement pour la modélisation de texte non supervisé.

2- **Sumy :** simple et utilitaire de ligne de commande pour extraire un résumé à partir de pages HTML ou de textes bruts. Le package contient également un cadre

d'évaluation simple pour les résumés de texte. Les méthodes de synthèse implémentées sont décrites dans la documentation.

3- NLTK : c'est la bibliothèque la plus populaire et la plus utilisée dans le traitement du langage naturel sous python car elle donne un résumé plus précis et plus le accepté par les experts.

Avantage :

- La bibliothèque la plus connue en PNL
- De nombreuses extensions
- Tokenisation rapide des phrases
- Prend en charge le plus grand nombre de langues par rapport aux autres bibliothèques

Inconvénients :

- compliqué à apprendre et à utiliser
- assez lent
- ne fournit pas de modèles de réseau neurone
- pas travail pas avec les vecteurs. [35]

- Après l'évaluation et la comparaison des différents résultats d'implémentation des différents algorithmes de Summarization, nous avons fait rencontre aux experts (quelques enseignants de la spécialité), et après discussion avec ces experts, on a abouti à choisir le résultat obtenu en appliquant l'algorithme qui intègre la bibliothèque NLTK et qui présente un résumé précis.

3.4. L'extraction des informations pertinentes : C'est l'étape finale de notre travail, en voici le processus :

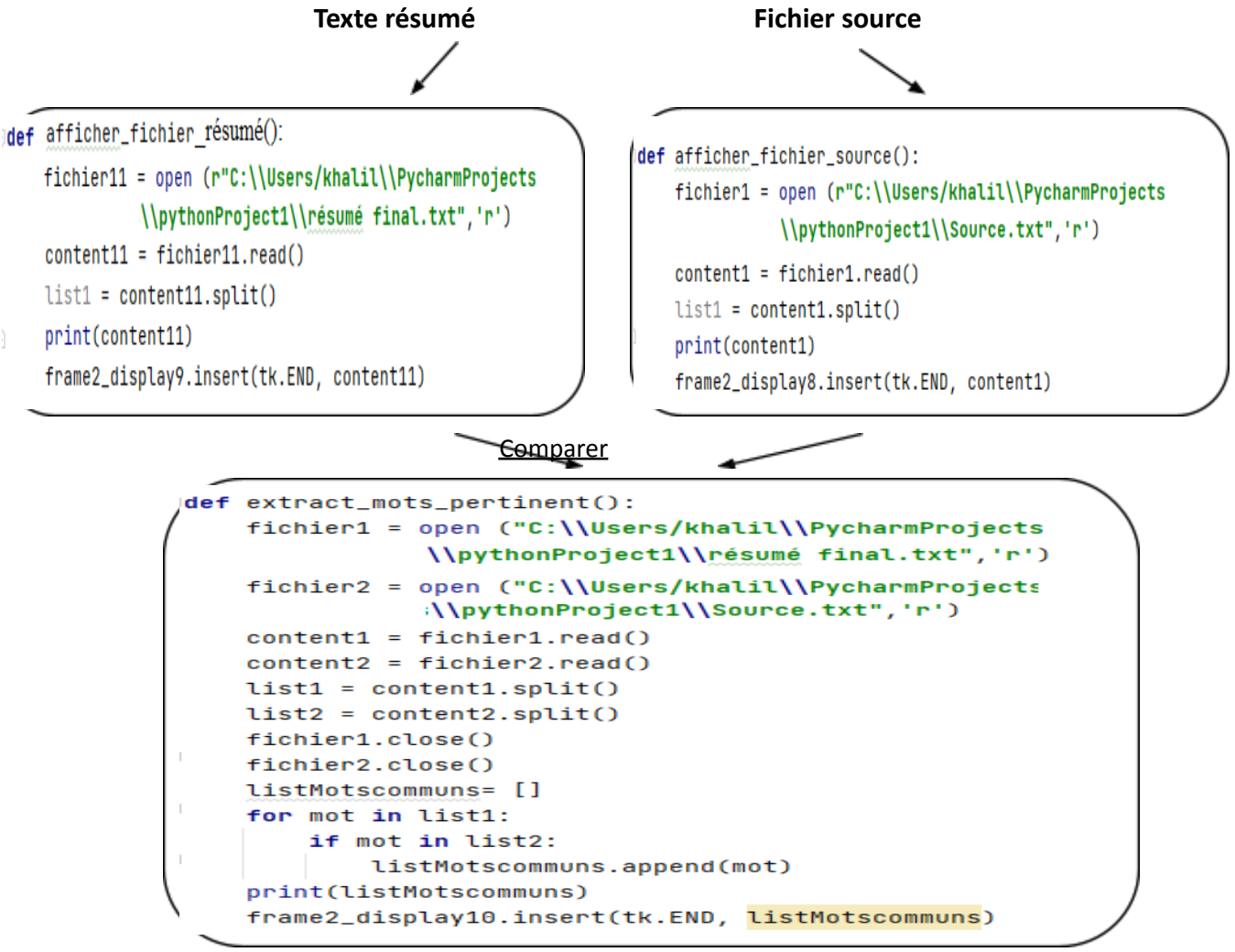


Figure.24.Le processus de l'extraction

Voici le résultat de l'extraction après comparaison des deux fichier :
Fichier source et Fichier résumé

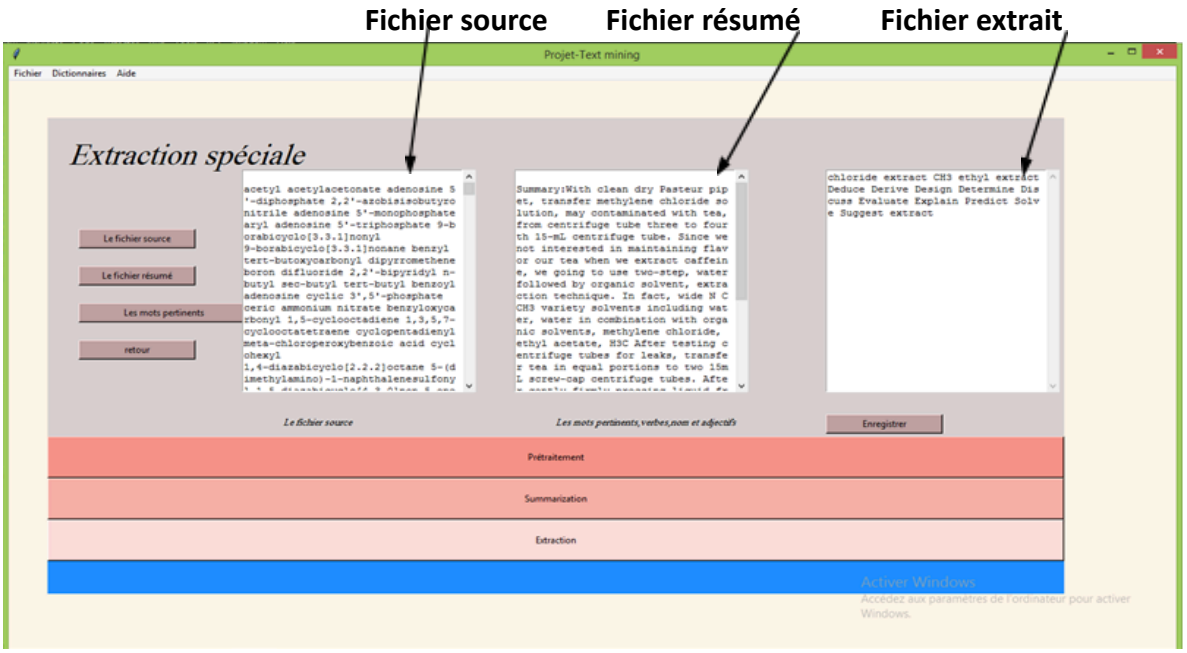


Figure 25. L'extraction présentée dans notre application

Et pour plus précisément, après avoir importer le fichier résumé (Figure 26 et 27), on peut remarquer l'existence de quelques mots (soulignés) dans la figure 28 qui présente le fichier final et ça se réalise après une comparaison avec le fichier résumé.

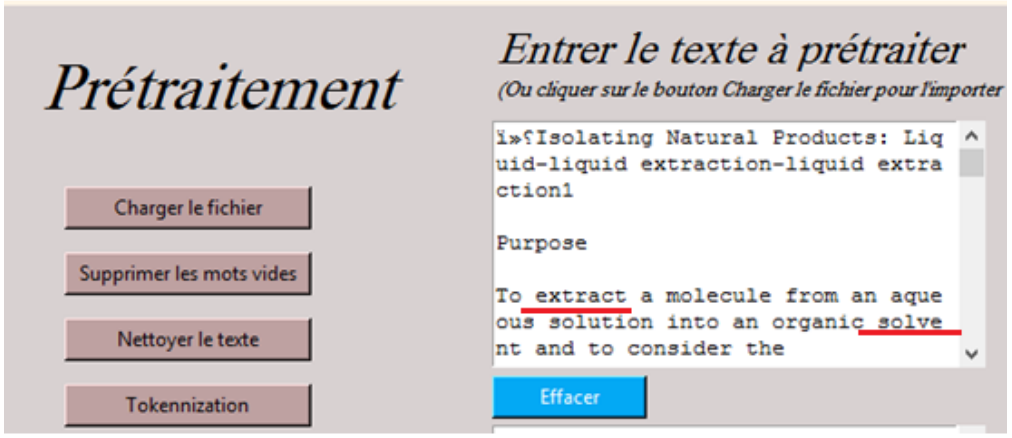


Figure 26. Exemple de mots pertinents1

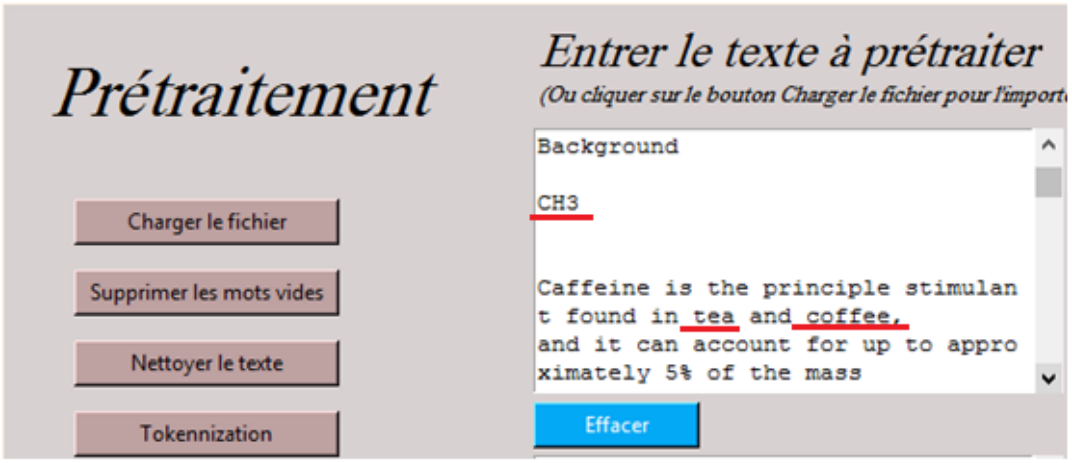


Figure 27. Exemple de mots pertinents2

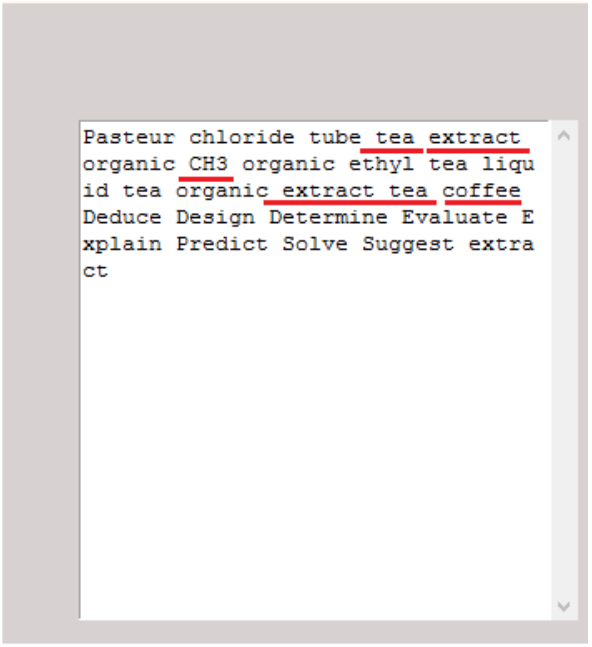


Figure 28. Aperçu final avec les mots pertinents

4. Conclusion :

Dans ce chapitre, nous avons présenté toutes les étapes de réalisation et d'implémentation du notre système avec le code source de chaque étape ainsi que des captures d'écran dans notre application.

1. Conclusion générale

Ce mémoire a passé en revue d'utilisation de l'exploration du texte « Text Mining » avec le traitement du langage naturel « PNL » pour extraire les termes les plus importants des documents textuels, présenté sous forme d'une feuille de TP dans un domaine expérimentale qui est « la chimie » .

Le Text Mining et le PNL sont deux domaines qui se développent rapidement ces dernières années. Par conséquent, les algorithmes, les méthodes et les approches sont aussi en évolution continue.

Face au problème rencontré dans notre travail lors de la conversion de la feuille de TP d'origine de type **PDF** au format textuel **Txt**, le changement Mentionné précédemment nous a mené à utiliser la méthode « ***Dictionnaire*** » pour l'extraction des information, avec un changement dans la méthode elle-même présenté par la comparaison des mots de notre corpus résumé avec le contenu d'un dictionnaire crée manuellement , contenant les termes les plus utilisés dans le domaine chimique, afin d'extraire les mots tout en respectant leur pertinence.

2. Perspectives :

Le résultat de notre travail peut être le noyau d'un futur travail, qui mène à une procédure de classification des termes pertinents extrais selon leur nature (action ou objet), ainsi que la génération final de la feuille de TP.

Bibliographie :

[1] : C. C. Aggarwal, Y. Zhao, P. S. Yu. On Text Clustering with Side

Information , ICDE Conference, 2012.

[2] : International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4 October 2013 .L.Sumathy -Research and Development Center -BharathiyarUniversity

Coimbatore,M.Chidambaram, Ph.d -Assistant Professor/CS –Rajah Serfoji Govt. College Thanjavur.

[3] :http ://datascientest.com/introduction-au-nlp-natural-language-processing

(Consulté le 12mars 2022).

[4] : Comparison of Text Mining Tools-Arvinder Kaur1, Deepti Chopra2-1,2

University School of Information and Communication Technology,

Guru Gobind Singh Indraprastha University, New Delhi, Delhi, India.

[5] :https://libguides.library.usyd.edu.au/text_data_mining/cleaning(Consulté le 13 janvier 2022).

[6] :https://www.analyticsvidhya.com /blog /2018/11/ introduction - text -summarization-textrank - python (Consulé le /22 fevrier 2022).

[7] :Hovy 1998, Eduard Hovy et Chin-Yew Lin. Automated text summarization and the SUMMA- RIST system. In Proceedings of a workshop on held at Baltimore,Maryland:October 13-15, 1998,pages 197–214.Association for ComputationalLinguistics, 1998.

[8] : ARIES Abdelkrime (26/06/2013) - Résume automatique de textes.

Thèse- l'Ecole Nationale Superieure d'Informatique (ESI)(specialite IRM)

[9] : Aidjouli Hocine -LES RESUMES ET LES RESUMES AUTOMATIQUES-Thèse-Centre universitaire de Tissemsilt.

[10] : T.A.S. Pardo, L.H.M. Rino, M.G.V. Nunes, Extractive summarization: how to Identify the gist of a text,-International.Bibliographie 63 Information

Technology Symposium – I2TS 2002, Florianópolis-SC, Brazil,pp.245-260,

01- 05 October 2002.

[11] : Ishikawa, K., Ando, S., Okumura, A.: Hybrid Text Summarization Method based on the TF Method and the-Lead Method. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese-Text Retrieval and Text Summarization.-Tokyo.Japan. pp.5-219-5-224, March 2001.

[12] : BASAGIR,R,KRUPIC,D., & SUZIC,B.(2009).Automatic text

Summarization . Information Searches and Retrieval,WS.

[13] : H. P. Edmundson: New methods in automatic abstracting, Journal of the Association for ComputingMachinery (ACM), vol. 16 N°2pp. 264-285,

April1969

[14] : 21- J. Baxendale, Man-made Index for Technical Litterature - an experiment. IBM J. Res.-Dev., 2(4) :354-361, 1985.

[15] : https://www.analyticsvidhya.com/blog/2018/11/introduction-text-

summarization-textrank-python(Consulté le 22 Fèvrier 2022).

[16] : Kathleen McKeown et Dragomir R. Radev. Generating summaries of multiple news articles. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '95, pages 74–82, New York, NY, USA, 1995. ACM. 10, 17, 20

[17] : Ishikawa, K., Ando, S., Okumura, A.: Hybrid Text Summarization Method based n the TF Method and the-Lead Method. Proceedings of the Second NTCIR orkshop Meeting on Evaluation of Chinese & Japanese-Text Retrieval and Text Summarization.- Tokyo.Japan. pp.5-219-5-224, March 2001.

[18] : H. P. Edmundson: New methods in automatic abstracting, Journal of the Association for ComputingMachinery (ACM), vol. 16 N°2pp. 264-285,

April 1969

[19] : 21-J Baxandale,Man made Index for Technical Littérature-an experiment.IBM J .Res-Dev..2(4) :354-361.1985.

[20] : GAHBICHE-BRAHAM.S .(2013).Am »Amélioration des systèmes de

traduction par analyse linguistique et thématique :application à la traduction

depuis l'arabe(Doctorat dissertation,université paris Sud-parisXI).

[21] : D. Paice, The Automatic Generation of Literary Abtracts : An Approach based On Identication of Selfindicating-Phrases. In Norman, O., Robertson, S., van Rijsbergen,- C., and Williams, P., editors, Information Retrieval Research, Butterworth, London 1981.

[22] : LEBARBIER,E,& MARY-HUARD,T(2008).classification non supervisée

[23] : Larkey, L. S., Ballesteros, L. and Connell M., improving Stemming for Arabia-Information Retrieval: Light Stemming and co-occurrence Analysis, In proceeding of the 25th annual-International conference on Research and

development in information Retrieval (SIGIR 2002), Tampere,Finland, August 2002.

[24] : T. Strzalkowski, J. Wang and B. Wise, Summarization-based Query Expansion In Information Retrieval,Proceedings of 36th Annual Meeting of the ACL, V. 2, pp. 1258- 1264, Montreal 1998.

[25] : R. Feldman, H. Hirsh, (1998), Mining Text Using Keyword Distributions,*Journal of Intelligent Information Systems*, vol 10, pp. 281-300

[26] : P. Losiewicz, D. Oard, R. Kostoff, (2000), Textual Data Mining to Support Science and Technology Management, Journal of Intelligent Information Systems, vol 15, pp. 99-119

[27] : Ben-Dov, R. Feldman (2005), Chapter 38: Text Mining and Information Extraction, *in Oded Maimon and Lior Rokach (Ed.), The Data Mining and Knowledge Discovery*

[28] :https://www.futura-sciences.com/tech/actualites/tech-logiciel-cree-

automatiquement- bons-resumes-texte-6024/(Consulté le 13 janvier 2022)

[29] : A.J.C. Trappey, C.V.Trappey, B.H.S. Kao (2006), Automated Patent Document Summarization for R&D Intellectual Property Management, *10th International Conference on Computer Supported Cooperative Work in Design*, pp.1-6.

[30] :https:////junior.universalis.fr/encyclopedie/verbes- d- etat- et- verbes –d –action (Consulté le 25 Février 2022

[31] : - https://dspace.univ-bba.dz/xmlui/handle/123456789/189925/02/2022

[32] :https://www.dicocitations.com /dico -mot- definition /100487/ pertinent.php (Consulté le 25 Février 2022)

[33] :https://www.techno-science.net/glossaire-definition/Python-langage. Html (Consulté le 20 Mars 2022)

[34] : https://wikimonde.com/article/Tkinter (Consulté le 20 Mars 2022)

[35] :https://medium.com/activewizards-machine-learning-company/ comparison - of-top-6-python-nlp-libraries-c4ce160237eb (Consulté le 01 Février 2022)