République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Ain Temouchent Belhadj Bouchaib

Faculté des sciences et technologies
Département des Mathématiques et de l'Informatique

Mémoire

En vue de l'obtention du Diplôme de Master en Informatique

Option :

Réseaux et Ingénierie des Données (RID)

Présenté par :

Mr.NEGGAZ Hichem

Segmentation thématique des textes Anglais

Encadrant :

Mme BERRAKEM Fatima Zahra
Maitre Assistante "A" à U.A.T.B.B.

Devant le jury composé de :

Président :	*Dr BELGRANA* Fatima Zohra (M.C.A)	U.A.T.B.B.
Examinateurs :	*M^r MERAD BOUDIA* Djalal (M.A.A)	U.A.T.B.B.

Année universitaire 2020/2021

Remerciement

C'est un grand plaisir de remercier les nombreuses personnes qui m'ont soutenu durant tout mon parcours d'études.

Je tiens tout d'abord à exprimer mes sincères remerciements à mon encadrant Mme BERRAKEM Fatima Zahra pour son soutien tout au long de ce projet, c'est grâce à son aide et conseils que j'ai pu aller jusqu'au bout.

Je remercie chaleureusement les membres du jury Mme BELGRANA Fatima Zohra et Mr MERAD BOUDIA Djalal qui ont accepté d'évaluer et d'examiner ce travail.

Un grand merci à mes amis qui m'ont toujours soutenu et avec qui j'ai passé un agréable moment.

Bien sûr, je ne peux pas conclure sans remercier chaleureusement ma famille qui m'a toujours soutenu et encouragé, comme d'habitude.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Avec l'avènement de l'informatique, l'humanité a résolu le problème du stockage de grande quantité d'informations, il est maintenant possible de stocker des milliards de documents dans un espace réduit mais l'exploitation de cette grande quantité d'informations doit être optimisée. Pour résoudre ce problème, différentes solutions ont été proposées parmi elles la création de méthode de structuration de l'information comme les bases de données.

Mais il y a eu une explosion d'informations textuelles provenant d'un large éventail de sources. Les données textuelles sont aussi des ressource porteuses de beaucoup d'informations qu'elles doivent être exploitées. Dans ce contexte, les données textuelles fait référence à des données non structurées.

Le fouille des texte est un type d'analyse de données qui vise à récupérer des informations à partir d'informations textuelles. La segmentation thématique des textes est l'une des tâches les plus importantes dans ce domaine.

La segmentation thématique des textes (STT) est l'opération qui a pour but de trouver la structure thématique d'un texte, c'est à dire, elle permet de décomposer le texte par thème.

La STT est une tâche qui peut être incluse dans les systèmes de nombreuses applications liées au traitement automatique du langage naturel (TALN), telle que la recherche d'information (RI), où la tâche de STT consiste à extraire des segments appropriés pour répondre à une requête au lieu de retourner un texte complet.
La STT est aussi incluse dans d'autres applications comme pour le résumé automatique des textes en fusionnant les segments les plus pertinents ou en appliquant l'algorithme de résumé aux différents segments constituant le document, et pour les systèmes questions-réponses, segmenter thématiquement les documents avant l'indexation améliore la précision de ces système.

Après l'analyse des travaux déjà réalisés, on remarque que les méthodes de la STT principalement utilisées sont probabilistes (statistiques) qui font que les systèmes qui les utilisent sont parfois mis en échec par l'ambiguïté sémantique des mots et l'impossibilité d'identifier des relations sémantiques entre eux.

L'objectif principal de ce travail est de développer un segmenteur thématique des textes Anglais basé sur une représentation conceptuelle par l'intégration de l'ontologie WordNet Anglais qui va nous permettre d'identifier les relations sémantiques entre les mots.

Ce mémoire se compose de trois chapitres :

- Le premier chapitre présente les notions de base et les principales propriétés que la segmentation thématique doit prendre en considération, il comporte aussi les différentes mesures d'évaluation de la qualité des résultats obtenus en utilisant un corpus d'évaluation.
- Le deuxième chapitre présente les différentes représentations des textes (statistique et conceptuelle). Il dresse aussi un état de l'art sur le concept d'ontologie : les circonstances de leur apparition, différentes définitions, les principales ontologies pour TALN. Il se termine par l'utilisation des ontologies dans la tâche de la segmentation thématique des textes.
- Le dernier chapitre présente les caractéristiques de la méthode l'implémentation de notre système, les outils utilisés, l'architecture de notre système ainsi que les différentes étapes du développement de ce système et à la fin de ce chapitre on va présenter les différents résultats de l'évaluation de notre système.

Dans la conclusion générale, nous présentons les principaux points abordés dans notre travail et nous dégageons quelques pistes pour la poursuite de notre travail.

Chapitre 1

Généralités sur la segmentation thématique des textes (STT)

1.1 Introduction

Les machines peuvent comprendre la forme structurée des données comme les tableurs et les tableaux des bases de données, mais le langage humain et les textes forment une catégorie de données non structurées, et il devient difficile pour la machine de les comprendre. Et là se pose le besoin de traitement automatique des langues naturelles.

Le traitement automatique des langues naturelles TALN (*Natural Language Processing (NLP)*) est un domaine de l'intelligence artificielle qui porte essentiellement sur la compréhension, la manipulation et la génération du langage naturel par les machines. Le TALN est une discipline qui met l'accent sur l'interaction entre la science des données et le langage humain. Il est réellement à l'interface entre la science informatique et la linguistique. Il porte donc sur la capacité de la machine à interagir directement avec l'humain.

Le TALN est un terme assez générique qui recouvre un champ d'application très vaste, il peut aider dans de nombreuses tâches dans des différents domaines d'application :

— **Traduction automatique :** La traduction automatique est l'une des plus grandes applications du TALN. Aujourd'hui les textes sont traduits d'une manière différente grâce au développement d'algorithmes de traduction automatique. De nombreuses applications, comme **Google Translator**, ont la capacité de traduire des textes entiers sans avoir besoin d'une intervention humaine.
Le langage naturel étant par nature ambigu et variable, ces applications ne reposent pas sur un travail de remplacement mot à mot, mais ils utilisent la traduction automatique statistique (*Statistical Machine Translation en Anglais*) qui nécessite une véritable analyse et modélisation de texte.

— **Chatbot :** Le service client est la chose la plus importante pour toute entreprise.

Cela peut aider les entreprises à améliorer leurs produits et à satisfaire les clients. Mais interagir manuellement avec chaque client et résoudre les problèmes peut être une tâche fastidieuse. C'est là que les chatbots entrent en scène. Les chatbots aident les entreprises à atteindre l'objectif d'une expérience client fluide.
Aujourd'hui, de nombreuses entreprises utilisent des chatbots pour leurs applications et leurs sites Web, qui résolvent les requêtes de base d'un client. Cela facilite non seulement le processus pour les entreprises, mais évite également aux clients la frustration d'attendre.

D'autres applications du TALN telles que la classification des textes, la recherche d'informations et le résumé automatique nécessitent une étape qui joue un rôle très important dans leur développement, cette étape est la segmentation thématique des textes.

La segmentation thématique des textes est la division d'un texte sous forme des portions dont chaque portion parle d'un thème different que l'autre.

1.2 Définitions sur la segmentation thématique des textes

Le fractionnement d'un flux de texte en segments cohérents et significatifs est considéré comme la tâche de segmentation thématique. Dans la segmentation du texte, nous sommes à la recherche des points dans le texte au cours de lesquelles se fait le changement d'un thème à un autre [1].

La segmentation thématique de texte est l'opération qui a pour but de trouver la structure thématique d'un texte et d'en proposer une décomposition par thème.

Si la plupart des textes traitent un sujet unique, ils abordent en général plusieurs thèmes en leur sein. Plus le texte est volumineux, plus il est probable que ses thèmes, ou sous-thèmes d'un sujet donné, soient nombreux. Fondamentalement, la segmentation thématique de texte recherche au sein d'un texte le début et la fin des thèmes. Si l'on considère que la segmentation thématique doit diviser le document en plusieurs segments cohérents et distincts sur le plan thématique, alors chaque segment ne doit idéalement traiter qu'un seul thème. Mais un thème est une unité complexe sur le plan rhétorique, qui nécessite souvent des digressions, des exemples et des argumentations [2].

1.2.1 La notion de thème

Dans la littérature, nous trouverons plusieurs définitions de la notion *«thème»*.

Le terme thème vient du grec *thema* qui signifie ce qui est proposé. Si l'on ouvre un dictionnaire [3], la définition que l'on lira sera : *« Sujet, idée qu'on développe (dans un discours, un ouvrage) »*. En linguistique, il est définit comme : *« l'élément d'un énoncé qui est réputé connu par les participants à la communication »*. Le thème est l'information centrale sur laquelle s'articule un acte de communication. Plus simplement, la définition retenue de la notion thème sera : *« Ce dont on parle »*, l'information principale communiquée par l'auteur [2].

La segmentation thématique doit donc diviser le texte en portions dont chacune des phrases *« parle »* de la même chose que les autres.

1.2.2 Le segment thématique

Avant même de se poser la question *« Comment segmenter thématiquement un texte ?»*, on doit se demander ce qu'est un segment thématique. Si nous nous référons à la définition du thème que nous avons choisi d'utiliser dans ce travail, un segment thématique doit être une unité textuelle ne *« parlant »* que d'un seul et unique sujet. Toutefois, la notion de *« ne parler que d'un seul sujet »* est extrêmement vague et subjective.

Dans leurs travaux sur la structure du discours, [14] reconnaissent une imbrication successive de segments de discours. Dans une tâche automatique comme la notre, il nous faut tout de même choisir une échelle standard pour travailler, même si nous ne définissons pas numériquement et explicitement la taille d'un segment thématique *« typique »*. Nous nous retrouvons donc face à un problème de granularité. Quelle taille doit faire un segment thématique ? [15].

1.2.2.1 Le document/le texte dans son intégralité

Dans la mesure où nous souhaitons segmenter un texte en sous-unités textuelles, il est peu difficile de faire plus gros en terme de segment thématique que prendre l'intégralité du texte. En effet, un document traite, la plupart du temps, d'un sujet précis, le retourner dans son intégralité n'est donc pas une erreur, c'est d'ailleurs ce que font la plupart des moteurs de recherche en réponse à une requête.

Toutefois l'objectif de la segmentation thématique étant de diviser le document en sous parties plus facilement exploitables, on peut se poser la question de savoir si une telle échelle est justifiée, donc le document n'est pas un segment thématique [15].

1.2.2.2 Le chapitre

Dans le contexte de documents très volumineux (tels que des livres), un chapitre est une sous-unité thématique, et il est parfaitement acceptable en tant que segment thématique. En fait, par définition, un chapitre doit développer un sujet précis.
Cependant, nous sommes dans le cas d'une grande unité de texte. Ce chapitre peut avoir un fil conducteur unique, mais le sujet peut être général, il inclura donc plusieurs sous-sujets plus spécifiques. Par exemple, le chapitre que vous lisez implique plusieurs parties tels que l'utilisation de la segmentation thématique de texte, le prétraitement, etc.
La présente partie traite plus spécifiquement la notion de segment thématique. Cette notion plus spécifique peut être considérée comme un thème distinct. Par conséquent, la longueur du chapitre est encore trop grande pour être une partie thématique. [15].

1.2.2.3 La partie/la section

Encore une étape en dessous du chapitre, la partie (ou la section). Toutefois, elle reste une unité textuelle de taille importante, contenant plusieurs paragraphes ou sous-parties. De fait, elle peut très bien traiter plusieurs thèmes différents [15].

1.2.2.4 Le paragraphe

La plus petite unité du texte avant la phrase. Le paragraphe semble être le segment thématique *« idéal »*. Cependant, il est concevable qu'un paragraphe traite plusieurs thème, ou qu'un thème soit traité sur plusieurs paragraphes successifs. De plus, le choix de paragraphes comme segment thématique type peut provoquer des erreurs [15].

1.2.2.5 Typographie et segment thématique

Le choix d'une de ces unités de texte comme taille de référence pour un segment, revient à affirmer que la structure thématique d'un texte dépend de sa typographie. Il est indéniable que la typographie d'un texte et sa structure thématique sont liées. En tant qu'outils permettant de structurer de l'information textuelle, les indicateurs typographiques servent de jalons tout au long du texte indiquant pauses, ruptures et transition. Mais ces indications ne sont pas absolues. Elles peuvent parfois être mal utilisées (le langage étant le résultat d'un processus humain, il suit rarement les règles à la lettre), voir être détournées de leur fonction initiale volontairement (pour un effet de style, pour réduire la taille d'une portion de texte jugée trop longue, trop lourde, etc.) [15].

On considère alors le texte comme un bloc uniforme, faisant abstraction de toute information de type typographique. Si ça nous prive d'une information utile, il permet de bénéficier d'une plus grande liberté dans la définition d'un segment thématique.

1.2.3 Définition du segment thématique

On peut définir le segment thématique comme étant : *« The smallest text unit which sentences are thematically consistent and thematically distinct from sentences of the previous and next segments. »* [6].
« La plus petite unité textuelle thématiquement cohérente en son sein et thématiquement distincte des unités textuelles précédentes et suivantes.».

L'unité atomique standard du segment thématique est la phrase [15].

1.3 L'utilisation de la segmentation thématique des textes

1.3.1 La STT pour les systèmes de questions-réponses

Les systèmes de questions-réponses est un système informatique dans les domaines du TALN et la recherche d'information qui cherchent à répondre automatiquement des questions posées par les utilisateurs dans un langage naturel.

Un système de questions-réponses a pour but de répondre aux questions des utilisateurs en trouvant des petits segments textuels sur le Web ou sur une collection des documents. Dans la phase du traitement de la question, un certain nombre d'informations sont extraites de la question, le type de réponse spécifie le type d'entité dont la réponse se compose (personne, localisation, etc.). La requête spécifie les mots clés à utiliser pour le système de la RI pour la recherche des documents. Le système ne va pas rechercher dans tous les documents, mais il va filtrer la collection des documents selon les mot clés obtenus à travers la requête pour obtenir des documents qui peuvent porter une réponse à la question posée par l'utilisateur car les documents sont déjà thématiquement segmentés.

1.3.2 La STT pour le résumé automatique

Le résumé des textes est la tâche de condenser un morceau de texte en une version plus courte, en réduisant la taille du texte initial tout en préservant les éléments d'information clés et le sens du contenu. Comme le résumé des textes manuel est une tâche coûteuse en temps et généralement laborieuse, l'automatisation de cette tâche est devenu populaire. Il existe des applications importantes pour le résumé automatique des textes dans des diverses tâches liées au TALN tels que les systèmes de questions-réponses. La génération d'un résumé peut être intégrée dans ces systèmes comme une étape intermédiaire qui permet de réduire la longueur du document.
Le résumé automatique des textes nécessite de segmenter les textes et de reconnaître leurs sujets connexes. Alors, un système de segmentation thématique des textes est nécessaire [35].

1.3.3 La STT pour la recherche d'information

Comme il avait une augmentation dans le besoin de beaucoup d'informations, la construction des structures de données devient une nécessité pour avoir un accès rapide aux informations. L'indexation est la structure de données pour une recherche d'information plus rapide. Au fil des siècles, une catégorisation manuelle des hiérarchies a été effectuée. Les bibliothèques ont été les premiers à adopter les systèmes pour la recherche d'information. Au début, il s'agit de l'automatisation des technologies précédentes, et la recherche était basée sur le nom de l'auteur et le titre. Ensuite, la recherche par mots clés, titre de sujet, etc. a été inclus.

Non seulement les bibliothécaires s'engagent dans l'activité de recherche d'information, mais aujourd'hui des centaines de millions de personnes s'engagent dans la recherche d'information chaque jour à travers l'utilisation des moteurs de recherche Web.Le système de la recherche d'information aide les utilisateurs à trouver les informations qu'ils ont besoin mais il ne renvoie pas explicitement les réponses à la question. Il informe de

l'existence et de la localisation des documents pouvant contenir les informations souhaitées.

En informatique, spécialement en apprentissage automatique (Machine Learning), La recherche d'information (RI) peut être définie comme un programme qui traite l'organisation, le stockage, la recherche et l'évaluation de l'information à partir d'un document référentiel, des informations textuelles en particulier. La RI est l'activité consistant à obtenir des informations qui peuvent généralement être documentés d'une manière non structurée (des textes généralement) qui répond à un besoin d'information à partir d'une large collection (stockée dans des machines généralement).

L'indexation est la partie la plus vitale de tout système de recherche d'informations. Elle est le processus dans lequel les documents requis par les utilisateurs sont transformés en structures de données consultables. L'indexation peut également être définie comme le processus d'extraction plutot que d'analyse d'un contenu particulier. Elle crée une fonctionnalité de base du processus de la RI puisqu'il s'agit de la première étape de la RI et aide à une récupération efficace des informations.

L'indexation du document nécessite une étape préliminaire consiste à le segmenter en unités. La segmentation des textes est traitée sous plusieurs angles selon la finalité visée. Dans le domaine de la recherche d'information, on distingue différentes méthodes de segmentation :

— Segmentation en une suite de mots ;
— Segmentation en phrases ;
— Segmentation en paragraphes ;
— Segmentation thématique ;
— Segmentation en unités logiques répercutées dans le sommaire.

Ces différentes méthodes sont présentées en détail dans [4]. La Segmentation en une suite de mots est un découpage arbitraire, elle ne prend pas en considération les aspects syntaxiques et sémantiques du texte. Par conséquent, elle va produire du bruit lors de la réponse à la requête de l'utilisateur. La segmentation en phrases n'est pas fiable lorsqu'on attend en réponse une partie de texte ne nécessitant pas de travail d'inférence de la part de l'utilisateur, sachant que la phrase ne présente pas de garantie de complétude syntaxique. De la même façon que la segmentation en phrases, celle en paragraphes n'est pas non plus suffisamment fiable, du fait de la difficulté à interpréter un paragraphe dans des contextes dans lesquels il est rattaché à une unité qui le précède ou bien qui lui succède. Les méthodes de segmentation thématique procèdent a l'identification des différents thèmes véhiculés par le texte, pour le segmenter en unités homogènes formant des blocs thématiques [5].

La segmentation thématique du texte (Topical Text Segmentation (TTS)) devient une issue importante dans les systèmes de la recherche d'information. Elle a pour but de diviser le textes en segments, chacun d'eux correspondant à un thème différent. L'ap-

plication directe de la STT va mener à extraire des segments qui répondent aux besoins de l'utilisateur au lieu de lui retourner des textes complets, dans lesquels l'utilisateur ne trouve pas facilement les quelques phrases satisfaisantes ses besoins spécifiques [6].

Il y a des considérations pratiques qui sont importantes dans les systèmes de RI. Il est avantageux de maintenir l'architecture du système aussi simple et clair que possible, trois propriétés importantes de la segmentation du texte pour les systèmes de RI peuvent être identifiées :

1. **L'indépendance de domaine :** Comme peu de connaissances externes que possible sur le contenu du document est nécessaire pour la segmentation. Pour les grandes collections de documents, même les techniques semi-automatiques (par exemple, des techniques qui exigent une certaine sorte d'intervention de l'homme au cours du processus de segmentation) sont problématiques.
2. **L'indépendance de la langue :** Bien qu'il existe des techniques de détection automatique de la langue, l'utilisation des algorithmes séparés pour différents types de données d'entrée, rend difficile l'opération de maintenance. Un algorithme doit idéalement travailler en basant uniquement sur le contenu du document.
3. **Granularité :** Souhaitable, est une option qui permet de définir un niveau personnalisable de granularité de ce que constitue un changement «suffisant» de thème. Cela permet à l'administrateur du système de définir la stratégie de segmentation, basée sur la granularité attendue de la requête. Il est inutile d'engorger le système avec de nombreux petits segments qui sont conceptuellement identiques avec respect des intérêts des utilisateurs.

En plus de ces propriétés, d'autres aspects techniques telle que l'efficacité sont très importants. La segmentation des documents volumineux doit être effectuée en un temps raisonnable. Cela va pousser de choisir des algorithmes qui sont basés sur la cohésion du texte et la répétition lexicale[7].

1.3.3.1 La cohésion lexicale

La cohésion lexicale est un élément d'un vaste dispositif linguistique appelé *la cohérence*[1], La cohésion tient au fait que les éléments grammaticaux d'un texte aillent ensemble. Elle correspond au niveau grammatical et textuel.
Voici quelques formes de la cohésion lexicale suivie par des exemples :

— **Répétition :** Se produit quand une forme de mot est répétée à nouveau dans une autre section du texte.

«Every tongue brings in a several **tale**, And every **tale** condemns me for a villain.»

1. la cohérence : C'est la liaison, le rapport étroit d'idées qui s'accordent entre elles, c'est l'absence de contradiction. Elle correspond au niveau sémantique et informationnel [3].

— **Répétition à travers la synonymie :** Se produit quand les mots partagent le même sens, mais ils ont deux formes syntaxiques différentes.
«During the second world war, countless **lives** were taken and many **souls** were lost.»
— **à travers la spécialisation/généralisation :** Se produit quand une forme spécialisée/généralisée (hyperonyme/hyponyme) d'un précédent mot est utilisée.
«She was in a **jewelry** shop and she bought a **ring**. »
— **à travers les relations d'association partie de/ensemble de :** Se produit quand une relation partie de/ensemble de existe entre deux mots.
«The police caught the criminal and took him to the **jail**, he will spend his whole life behind the **bars**. »
— **Les associations statistiques entre les mots :** Ces types de relations se produisent lorsque la nature de l'association entre deux mots ne peut pas être définie en termes des types de relations qui précèdent. Ces relations sont le plus souvent trouvées par les statistiques de cooccurrence du mot [8] [9] [10], Par exemple :
«**Adolf Hitler**» «**Nazi Party**»

1.3.3.2 Les chaines lexicales

la cohésion lexicale se réfère à la connectivité entre deux mots en termes de relations de mot [11].

Bien qu'il n'y ait de nombreux types de relations entre les mots, la simple forme est les répétitions de mots. Les liens entre les mots captent la distribution d'un mot par des répétitions et sont souvent modélisés par chainage lexical qui garde la trace de toutes les positions d'occurrences d'un mot dans un document. Le chainage lexical est généralement représenté par des répétitions de mots au niveau de phrase [12], ou la distribution inclut toutes les occurrences de phrases d'un mot dans le texte.
La cohésion lexicale s'établit non seulement entre deux termes, mais entre des séries de mots reliés, à des distances variées dans le texte [8].
La tâche de la segmentation thématique de textes joue un rôle crucial dans le développement de nombreuses applications comme celles qu'on a cité au-dessus. Comme le nombre de mots présents dans un document textuel peut être très grand et possède des informations inutiles, alors une étape de prétraitement préliminaire à la segmentation est nécessaire.

1.4 Le prétraitement

Dans le traitement automatique de la langue naturelle, tout texte brut doit être soigneusement prétraité avant que l'algorithme puisse le digérer. Le prétraitement se

compose généralement de plusieurs étapes qui dépendent d'une tâche donnée, mais il peut être grossièrement classé en cinq étapes :

- **La segmentation :** L'analyse lexicale ou la tokenisation est le processus qui divise des chaînes de texte plus longues en morceaux plus petits, ou des tokens. Des morceaux de texte peuvent être segmentés en phrases, les phrases peuvent être segmentées en mots, etc [36].
- **Le nettoyage :** Consiste à se débarrasser des parties de texte les moins utiles à travers l'élimination des mots vides (ou stop words), traiter avec la capitalisation et les caractères et d'autres détails [36].
- **La normalisation :** La réduction linguistique des termes à travers la racinisation (ou stemming en Anglais) ou la lemmatisation.
- **L'annotation :** Consiste à appliquer un schéma aux textes, l'annotation est généralement l'extraction de la nature des mots (ou part-of-speech tagging).
- **L'analyse :** Extraire les relations entre les mots à partir d'un ensemble de données.

1.4.1 La segmentation

La segmentation est utilisée pour désigner la décomposition d'un texte en morceaux plus gros que des mots, tels que des paragraphes et des phrases, tandis que la tokenisation est réservée au processus de décomposition qui aboutit exclusivement à des mots.

Cela peut sembler un processus simple, mais la machine contrairement à l'être humain ne peut pas distinguer la ponctuation de fin de phrase des autres ponctuation.
«Shall we call **Mr.** Brown ?»
La machine peut facilement déviser cette phrase en deux phrases si l'abréviation n'est pas prise en compte, car le point est le même que ce soit à la fin d'une abréviation ou d'une phrase pour la machine.

Ce problème ne se pose pas seulement avec les phrases, mais aussi avec les mots, par exemple l'apostrophe dans «**he's**» va la faire un mot ou deux mots. Ensuite, il existe des stratégies de défi comme conserver la ponctuation avec une partie du mot ou l'éliminer complètement.

1.4.2 Le nettoyage

Le processus de nettoyage permet de mettre tous les textes sur un même pied d'égalité, impliquant des idées relativement simples de substitution ou de suppression :

- Mettre tous les caractères en minuscules.
- Suppression du bruit, y compris la suppression des chiffres et la ponctuation.
- Suppression des mots vides.

1.4.2.1 Mettre les caractères en minuscules

Le texte a souvent une une variété de majuscules reflétant le début des phrases ou des noms propres. L'approche courante consiste à tout réduire en minuscules pour plus de simplicité. Il faut prendre en considération que certains mots peuvent changer de sens lorsqu'ils sont réduits en minuscules, par exemple :
Le mot «**US**» (United States) qui signifie **Les états Unis**, mais le même mot en minuscule devient «**us**» qui veut dire **nous**.

1.4.2.2 Suppression du bruit

La suppression du bruit fait référence à la suppression des caractères et des morceaux de texte qui peuvent interférer avec l'analyse du texte. Il existe différentes manières de supprimer le bruit, notamment la suppression de la ponctuation, les caractères spéciaux, les chiffres, le formatage HTML, etc.

1.4.2.3 Suppression des mots vides

Les mots vides sont un ensemble de mots couramment utilisés dans une certaine langue comme **"a"**, **"the"**, **"is"**, **"are"**, **"and"**, etc. en Anglais, ces mots n'ont pas de sens important. L'intuition derrière cette approche est qu'en supprimant les mots avec faibles informations du texte nous pouvons nous concentrer seulement sur les mots importants à la place. En outre, cela réduit le nombre de fonctionnalités prises en compte, ce qui permet de garder les modèles dans une meilleure taille. La suppression des mots vides est souvent utilisée dans les systèmes de la RI, les applications de classification des textes, l'extraction des sujets, etc.

1.4.3 La normalisation

La normalisation met tous les mots sur un même pied d'égalité, et permet au traitement de se dérouler de manière uniforme. Elle est très liée au nettoyage, mais apporte au processus un pas en avant en mettant tous les mot sur un même pied d'égalité grace à la racinisation ou la lemmatisation.

1.4.3.1 La racinisation

La racinisation (ou Stemming en Anglais) est le processus d'élimination des affixes (suffixes, prefixes, infixes, circonfixes) d'un mot pour obtenir sa racine [36]. Il existe de nombreux modèles de racinisation tels que Porter et Snowball.
Il existe deux erreurs majeurs dans la racinisation :

- OverStemming
- UnderStemming

OverStemming : Est quand deux mots avec des différentes racines vont avoir une même forme tronquée. Ceci est également connu comme un faux positif :

— universal
— university
— universe

Les trois mots ci-dessus vont avoir la même racine "**univers**".

UnderStemming : Est quand deux mots vont avoir des différentes racines alors qu'en réalité ils doivent avoir la même. Ceci est également connu comme un faux négatif :

— alumnus –> alumnus
— alumni –> alumni
— alumnae –> alumna

Les mots au-dessus ont des racines différentes alors qu'elles doivent avoir une seule racine. La racinisation de ces mots est faite avec le Stemmer « *Snowball* ».

En outre, la racinisation peut donner des mots qui n'ont aucun sens par exemple : la racine de mot "**troubles**" va être "**troubl**" [39].

1.4.3.2 La lemmatisation

La lemmatisation est liée au racinisation, mais elle est capable de capturer des formes canoniques basées sur le lemme d'un mot en déterminant la nature des mots en utilisant des outils spéciaux, tels que La base de données lexicale de l'Anglais de WordNet ou de SpaCy [39].

1.4.3.3 Comparaison entre la racinisation et la lemmatisation

Les deux tableaux ci-dessous représentent une explication du processus de la racinisation et de la lemmatisation avec des exemples :

Forme	Suffixe	Racine
Studies	-es	Studi
Studying	-ing	Study

Table 1.1: La racinisation

Forme	informations morphologiques	Lemme
Studies	Third person, singular number, present tense of the verb "**study**"	Study
Studying	Gerund of the verb "**study**"	Study

Table 1.2: La lemmatisation

1.4.4 L'annotation

L'annotation est un processus fournissant du texte avec des balises pertinentes. La pratique la plus courante est d'ajouter l'étiquetage morpho-syntaxique qui est le processus qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, etc.

L'étiquetage morpho-syntaxique fournit des informations plus granulaires sur les mots. Par exemple, dans un problème de classification de documents, l'apparition du mot "**book**" comme étant un **nom** pourrait entraîner une classification différente de celle du mot "**book**" comme un **verbe**. L'étiquetage morpho-syntaxique a pour but d'attribuer les informations grammaticales (comme nom, verbe, adjectif, etc.) pour chaque mot d'un texte donné en fonction de sa définition et de son contexte [37].

1.4.5 L'analyse

L'analyse est la dernière étape du prétraitement, cette phase a pour but d'extraire des fonctionnalités qui peuvent être utilisées dans le développement du modèle.

1.4.5.1 Le comptage

Le comptage c'est l'ajout des informations statistiques tels que le nombre des mots et le nombre des phrases.

1.4.5.2 L'analyse syntaxique de surface

L'analyse syntaxique de surface (Chunking ou Shallow Parsing en Anglais) est le processus qui identifie les éléments constitutifs des phrases, tels que les nom, les verbes, les adjectifs, etc. et les relie à des unités d'ordre supérieur qui ont des significations grammaticales discrètes (groupe nominal, verbe, etc.) [38].

1.4.5.3 L'extraction de collocation

Les collocations sont des combinaisons de mots plus ou moins stables, tels que "**keep in mind**", "**get ready**", "**free time**", etc. Comme ils véhiculent généralement un sens

précis et établi, il vaut la peine de les extraire avant l'analyse.

1.4.5.4 Word Embedding/Vectorisation de mots

Word Embedding est la façon moderne de représenter les mots par des vecteurs de nombres réels. Les mots apparaissant dans des contextes similaires possèdent des vecteurs correspondants qui sont relativement proches. En d'autres termes, Word Embedding représente les mots par des cordonnées vectorielles X et Y où les mots liés, basés sur un corpus de relations, sont placés plus près les uns des autres.

1.5 Les méthodes de calcul de poids

Une fois la liste des termes est déterminée à travers une étape dite la représentation du texte qu'on va la voir dans le chapitre suivant, il faut maintenant attribuer à chaque terme un poids. Il existe plusieurs facons d'associer un poids à un terme. Il peut être tout simplement binaire (1 si le mot est présent dans le texte, 0 sinon). Il peut également indiquer le nombre de fois qu'un mot apparaît dans le texte.

1.5.1 Le modèle vectoriel

Ce modèle est introduit par Salton [13], représente chaque unité textuelle atomique par un vecteur et calcule le poids de chaque terme. De nombreuses solutions ont été proposées pour coder les composants des vecteurs. Parmi les mesures les plus utilisées, la mesure TF-IDF.

1.5.1.1 TF-IDF

TF-IDF (term frequency-inverse document frequency) TF-IDF (terme fréquence-fréquence inverse du document) est une mesure statistique utilisée pour évaluer l'importance de chaque terme [2] d'un document par rapport à une collection ou un corpus. Elle se calcule en multipliant deux métriques : la fréquence du terme dans le document (TF) et la fréquence du terme dans le corpus (IDF).

TF-IDF a été inventé pour la recherche d'information, elle fonctionne en augmentant proportionnellement au nombre de fois qu'un mot apparaît dans un document, mais il est compensé par le nombre de documents qui contiennent le mot. Alors les mots qui sont communs dans chaque document tels que "**this**", "**and**", "**if**", etc. sont classés moins importants même s'ils peuvent apparaître plusieurs fois car ils ne signifient pas grand-chose pour ce document en particulier.

2. En général que les termes dits « utiles » sont considérés et donc les mots comme certains pronoms ou déterminants seront ignorés. De plus, les mots sont lemmatisés pour éviter que les formes fléchies d'un mot ne soient considérées comme des termes différents.

— La fréquence du terme (TF) est le nombre d'occurrences de ce terme dans le document considéré. Il existe plusieurs façons de calculer cette fréquence, la plus simple étant le nombre brut d'occurrences d'un mot dans un document. Ensuite, il existe des moyens d'ajuster la fréquence : par la longueur d'un document ou par la fréquence brute du mot le plus fréquent dans un document.

— La fréquence inverse de document (IDF) est une mesure de l'importance du terme dans l'ensemble du corpus. Cette métrique est calculée en prenant le nombre total de documents, en le divisant par le nombre de documents qui contiennent le mot et en calculant le logarithme.
si le mot est très courant et apparaît dans de nombreux documents, ce nombre se converge vers 0. Sinon, il se convergera vers 1.
En multipliant ces deux métriques on obtient le score TF-IDF d'un mot dans un document. Plus le score est élevé, plus ce mot est pertinent dans ce document particulier.
Pour le mettre en terme mathématique plus formel, le score TF-IDF pour un mot t dans un document d d'un corpus D est calculé comme suit :

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \qquad (1.1)$$

$$\text{Où : } tf(t, d) = \frac{freq(t,d)}{|d|} \qquad (1.2)$$

$$\text{Et : } idf(t, D) = log(\frac{|D|}{|d \in D : t \in d|}) \qquad (1.3)$$

1.5.1.2 Le modèle probabiliste

Concernant le modèle probabiliste, on considère que les documents qui sont générés par tirage aléatoire des différents termes qui les composent. Le calcul de probabilité s'appuie sur les calculs de la probabilité qu'un terme soit présent sachant que le document est pertinent et la probabilité qu'un terme soit présent sachant que le document est non pertinent.

1.6 Les approches de segmentation thématique de texte

Nous distinguons deux approches pour la segmentation thématique de texte, une avec un aspect actif et l'autre avec un aspect passif.
Les méthodes « passives » ne cherchent pas à retrouver les frontières thématiques d'un texte, mais à regrouper les phrases en segment. Les frontières apparaissant ainsi par « défaut ». La méthode active, contrairement à la méthode passive tentent d'identifier spécifiquement les propriétés des frontières thématiques, les segments thématiques devenant l'espace qu'il y a entre deux frontières, nous les appelons méthodes à détection active des frontières [15].

1.6.1 Les approches passives

Les méthodes passives tentent de regrouper les phrases en segments thématiques, ensuite détecter les frontières à travers ce regroupement [15]. Deux approches passives sont communes : les méthodes graphiques et les méthodes par calcul de distance ou de similarité.

1.6.1.1 Les méthodes graphique

Dans le but de faciliter la visualisation de la répartition des termes dans un document, une représentation graphique a été proposée. Helfman dans [16] a présenté la méthode du nuage de points pour la recherche d'information.Le principe est de positionner sur un graphique chaque occurrence des termes du document.
Dans cette représentation, un terme apparaissant à une position i et une position j du texte, sera représenté par les 4 couples $(i,\ i)$, $(i,\ j)$, $(j,\ i)$ et $(j,\ j)$. les zones de forte concentration de points dans le graphique représentent Les portions du document où les répétitions de termes sont nombreuses.
Reynar dans [17] a adapté cette approche graphique dans son algorithme *DotPlotting*. L'idée de cet algorithme est d'identifier les frontières des segments qui sont thématiquement cohérents sur le graphique en cherchant les limites des zones les plus denses. La densité d'une région du graphique est calculée en divisant le nombre de points présents dans la région par l'aire de cette dernière.
L'approche passive est fortement incarnée dans les méthodes graphiques par la transformation du texte en un ensemble de points, pour ensuite retrouver les segments thématiques en regroupant ces points en nuages, par contre, la transformation d'un texte en une représentation graphique de ces termes mène à perdre toute notion de compréhension et on se contente alors de compter des mots [15].

1.6.1.2 Les méthodes par calcul de distance/similarité

Ces méthodes considèrent chaque portion du document à traiter comme autant de vecteurs. Après avoir éliminer les termes inutiles du textes, ces vecteurs sont composés, dans la plupart des cas, des fréquences d'apparition des termes dans une portion de texte. Ces fréquences peuvent être pondérées par un IDF pour renforcer l'importance des mots. un algorithme C99 dans [18] a été proposé par Choi qui utilise une mesure de similarité calculée pour chaque paire de phrases du texte en utilisant chaque mot commun entre les phrases, l'idée de base de cette méthode est que les mesures de similarité entre des segments de texte court sont statistiquement insignifiantes, et que donc seuls les classement locaux sont à considérer pour ensuite appliquer un algorithme de catégorisation sur la matrice de similarité.
Lamprier et al. dans [19] ont utilisé un algorithme de clustering évolutif pour segmenter thématiquement des textes dans leurs algorithme *ClassStruggle*. Le clustering peut être utilisé pour la segmentation thématiques de texte car il existe une grande similarité entre le regroupement des éléments d'un ensemble en classe et le regroupement des phrases en segments thématiques . La seule contrainte supplémentaire est que les phrases d'un segment thématique doivent se suivre dans le texte.
Les méthodes par calcul de distance ou de similarité sont typiquement des méthodes à détection passive des frontières, mais il existe certaines méthodes utilisant des distances ou des similarités qui peuvent être classées dans les approches actives [15].

1.6.2 Les approches actives

Contrairement aux approches passives, les approches actives cherchent à détecter les frontières thématiques en identifiant les propriétés des phrases frontières.
L'identification des propriétés des frontières thématiques utilise, dans la plupart des cas, des ressources externes, ces méthodes d'identification ont été définies sous le nom «approches exogènes» selon Ferret dans [20]. Comme il existe d'autres méthodes tentent de retrouver les frontières thématiques, sans s'appuyer sur des données externes qui ont été définies toujours selon Ferret dans [20] sous le nom «approches endogènes» [15].

1.6.2.1 Les méthodes supervisées

Les méthodes supervisées servent à construire un modèle permettant de retrouver les frontières thématiques en les déduisant à partir d'une bases de données d'apprentissage. cette base de données est un corpus composé de textes qui ont été déjà thématiquement segmentés.
Le fonctionnement de ces méthodes est limité car elles ne sont approuvées seulement sur des textes proches de leurs corpus. En outre, les ressources d'apprentissage sont très

difficiles à trouver ce qui rend la phase d'apprentissage lourde et couteuse [15].

1.6.2.2 Les méthodes par chaînes lexicales

La méthode *Segmenter* de Kan [23], effectue une segmentation linéaire basée sur les chaînes lexicales présentes dans le texte. Ces chaines relient les occurrences des termes dans les phrases, une chaine est rompue si le nombre de phrases séparant deux occurrences est très important. Ce nombre dépend de la catégorie syntaxique du terme considéré, une fois tous les liens établis, un poids leur est assigné en fonction de la catégorie syntaxique des termes en jeu et de la longueur du lien. Un score est ensuite donné à chaque paragraphe en fonction des poids et des origines des liens qui le traversent ou qui y sont créés. Les marques de segmentation sont alors apposées au début des paragraphes ayant les scores maximaux.
Sitbon dans ses travaux [24] a fusionné les approches à base des chaînes lexicales avec celles de similarité [15].

1.6.2.3 Les méthodes par calcul de distance/similarité

Ces méthodes se basent sur le calcul de la similarité ou la distance. Contrairement aux approches passives qui utilisent beaucoup plus la similarité, les approches actives se basent surtout sur les distances.
Lamprier et al. dans [25] introduisent une approche originale en utilisant un algorithme génétique pour optimiser deux critères : la cohérence interne des segments thématiques et la dissimilarité avec les segments adjacents [15].

1.7 L'évaluation de la segmentation thématique de texte

La segmentation thématique de texte est une tâche plutôt subjective, alors choisir un ensemble de critères à évaluer et quantifier les résultats peuvent se révéler difficiles. Les principales mesures existantes évaluent la qualité d'une segmentation de texte en la comparant à une segmentation de référence.
Il existe deux difficultés majeures associées aux algorithmes d'évaluation de la segmentation de texte. La première est la difficulté de trouver des textes pré-segmentés.
La deuxième difficulté avec ces algorithmes d'évaluation est que pour différentes applications de la segmentation thématique de textes différents types d'erreurs deviennent importants. Par exemple, pour la recherche d'information, un décalage de quelques phrases peut être acceptable. Comme il existe d'autre application ou le placement précis des frontières thématiques est crucial.

1.7.1 La création de corpus de référence

Dans toutes les tâches du traitement automatique du langage naturel, la création d'un corpus pour l'évaluation a été toujours un soucis. Spécifiquement dans la segmentation thématique de texte, on a besoin des textes qui ont été déjà pré-segmentés. Ce type de ressources est difficile à trouver.
Labadie dans [15] a proposé certaines alternatives pour la création de corpus de référence qu'on va les voir.

1.7.1.1 La création de corpus par la concaténation de textes courts

Cette solution repose sur la création des corpus de référence nommés « artificiels ». Un document de référence est le résultat d'une concaténation de plusieurs textes courts traitant de sujets différents. Alors chaque texte est considéré comme un segment thématique.
Cette méthode présente le double avantage d'être simple et peu couteuse à mettre en œuvre. Le problème de cette méthode c'est qu'elle ne fournit pas un cadre de test pour une tâche de segmentation thématique car les textes utilisés pour la création du document de référence n'ont, dans la plupart des cas, rien à voir les un avec les autres. Donc cette méthode ne va pas vraiment évaluer la segmentation thématique, car la segmentation thématique consiste à diviser un texte qui parle d'un seule thème général en sous-unités textuelles dont chaque unité parle d'un sujet précis [15].

1.7.1.2 La création de corpus par la référence de l'expert

Lorsqu'on veut avoir des textes pré-segmentés pour les utiliser comme une référence pour l'évaluation de la segmentation thématique, on peut faire appel à un expert pour faire la segmentation. Cet expert peut être un linguiste ou l'auteur même du texte.
Les résultats de l'évaluation obtenus à partir ces corpus sont plus fiables, comme ils sont faciles à justifier devant la communauté.
Cependant, le caractère subjectif de cette tâche que nous avons à évaluer est considéré comme une faille. Si dans certains domaines l'avis de l'expert est difficilement contestable, ce n'est pas le cas de la segmentation thématique. En outre, les résultats de la segmentation des experts peuvent se diffèrent d'un expert à l'autre. Le problème se pose alors de savoir lequel des résultats choisir. Doit-on n'en prendre qu'un et considérer les autres comme faux ? Sont-ils tous justes et dans ce cas comment évaluer les résultats d'une méthode automatique sur la base de ces références ? [15].

1.7.1.3 La création de corpus par la référence consensuelle

Cette méthode s'appuie sur des textes segmentés par « consensus ». En d'autres termes, chaque texte sera présenté pour un groupe de juges, chaque juge va segmenter le texte individuellement. Les frontières acceptables sont celles qui vont être désignées par la majorité des juges.
L'avantage d'un corpus consensuel, c'est qu'il représente une « moyenne » de ce que l'utilisateur s'attend à recevoir comme solution [15].

1.7.2 Métriques d'évaluation

La mesure utilisée est un aspect très important dans l'évaluation de la segmentation thématique. Plusieurs solutions ont été proposées.

1.7.2.1 Le rappel et la précision

Le rappel et la précision sont deux standards qui ont été utilisés pour évaluer les algorithmes de segmentation [26]. Ces deux mesures sont inversement proportionnelles, une diminution de l'un entraîne généralement une augmentation de l'autre. Ils évaluent si les frontières obtenues sont à leurs places par rapport aux celles de référence.
Le problème de ces deux mesures c'est qu'elles ne permettent pas de différencier une erreur faible (un décalage d'une phrase) d'une erreur grave (un oubli de frontière).
La précision s'exprime ainsi :

$$Precision = \frac{Nombre\ de\ frontires\ justes\ ramenes}{Nombre\ de\ frontires\ ramenes} \qquad (1.4)$$

De même que l'on écrirait le rappel de la sorte :

$$Rappel = \frac{Nombre\ de\ frontires\ justes\ ramenes}{Nombre\ de\ frontires\ attendues} \qquad (1.5)$$

1.7.2.2 La mesure Pk

La mesure Pk a été proposée par Beeferman dans [28], cette mesure prend en compte la distance entre la frontière ramenée par l'algorithme et la frontière attendue. Alors, elle évalue la probabilité d'erreur sur la segmentation en tenant compte la probabilité pour deux phrases éloignées d'une distance k [3] dans les mêmes segments du document

3. Cette distance peut être un nombre de mots, phrases, paragraphes ou de toute autre unité textuelle. Dans notre cas, cette distance est le nombre des phrases.

de référence(*ref*) et du document produit (*hyp*). La formule pour un corpus de longueur n étant la suivante :

$$Pk = \sum_{1 \leq i \leq j \leq n} D_k(i,j)\delta_{ref}(i,j) \oplus \delta_{hyp}(i,j) \qquad (1.6)$$

Dans cette formule δ_{ref}(i, j) et δ_{hyp}(i, j) sont des fonctions prenant la valeur 1 si les deux indices i et j appartiennent au même segment dans leurs corpus respectifs (ref et hyp) et 0 si ce n'est pas le cas [15].
Le score Pk trouve alors la pénalité entre toutes les paires de phrases (i,i+k) :

$$Penality = (\delta_{ref}(\text{i, j}) \oplus \delta_{hyp}(\text{i, j}) \qquad (1.7)$$

Un score plus élevé (par exemple pk = 1) signifie que le système a une moins bonne qualité, sinon (par exemple pk = 0) il a une meilleure qualité. L'opérateur $\oplus$ est l'opérateur XNOR qui suit la table 1.3.

1	XNOR	1	=1
1	XNOR	0	=0
0	XNOR	1	=0
0	XNOR	0	=1

Table 1.3: La table de vérité de l'opérateur XOR

1.7.2.3 La mesure WindowDiff

Pevzner et Hearst dans [29] ont évolué la mesure Pk proposée par Beeferman, et ils montraient l'existence des failles dans cette méthode. La signification de la mesure n'est pas claire et l'ajout des petits segments n'est pas pris en considération.
La mesure WindowDiff est une mesure inspirée de la mesure Pk mais qui pallie certains de ses défauts. En fait cette nouvelle mesure prend en compte le nombre de frontières séparant deux phrases espacées d'une distance k.

$$WindowDiff = \frac{1}{N-K} \sum_{i=1}^{N-K} (|b(ref_i, ref_{i+k})|) - (|b(hyp_i, hyp_{i+k})|) \qquad (1.8)$$

ou N est le nombre de phrases du texte et b(x_i, x_j) une fonction donnant le nombre de frontières du texte x entre les phrases i et j.
Ils ont montré que cette mesure est d'une grande stabilité face aux variations des tailles des segments, et qu'elle est aussi sévère avec les ajouts qu'avec les oublis de frontières thématiques. Cependant elle peut être supérieure à 1, et ne peut donc plus être assimilée à un taux d'erreur. Il est désormais évident qu'elle n'est qu'un élément de comparaison de la fiabilité des méthodes, et non pas un indice absolu de leur qualité [27].

1.8 Conclusion

Nous avons présenté dans ce chapitre la segmentation thématique de textes en essayant d'éclairer les notions de thème et de segment thématique. Ensuite on a cité quelques applications utilisant la segmentation thématique de textes, ce qui nous a amené à la notion de cohésion lexicale. Après, on a présenté le phase du prétraitement qui est une phase nécessaire pour la segmentation thématiques prétraitement que toute tâche du traitement automatique du langage naturel l'utilise.

Finalement, nous avons cité les principales méthodes de segmentation thématique des textes, suivi par les différentes mesures d'évaluation de la qualité du résultat de la tâche vis à vis du corpus de référence.

Afin de prendre en compte les relations sémantiques entre les mots, les ontologie peuvent être une solution.

Nous développerons ci-après le concept d' « ontologie » dont l'utilisation est récente dans le domaine de la segmentation thématique des textes.

Chapitre 2

La représentation conceptuelle des textes et les ontologies

2.1 Introduction

Les méthodes de représentation de texte jouent un rôle crucial dans la tâche de la segmentation thématique de texte, la méthode du représentation choisie peut affecter d'une manière significative les résultats obtenus. Parmi les méthodes de la segmentation thématique de texte, celles qui sont basées sur les statistiques (La représentation statistique), et les autres qui sont basées sur la similarité sémantique (La représentation conceptuelle). Les méthodes basées sur les statistiques, qui se concentrent sur les mêmes mots qui apparaissent dans le texte. Ces méthodes se conforment à l'hypothèse que des parties similaires d'un texte doivent avoir de nombreux mots en commun, mais elles ignorent la sémantique.
Par exemple, bien que les deux phrases « Donald Trump invites the winning team to the White House » et « The 45th president of the USA has dinner with the champions in his home », ont aucun mot en commun, elles véhiculent presque la même sémantique.

C'est pour cela, des efforts ont été faits pour créer une représentation basée sur la similarité sémantique. Ce type de représentation utilise des ressources sémantiques (thesaurus, ontologies, etc.) qui ont un apport considérable pour le traitement des documents. c'est pour cette raison qu'on va étudier dans ce chapitre : en premier lieu, la signification d'une ontologie, ces constituants et les plus grands projets réalisés pour le TALN. Parmi ces projets on trouve l'ontologie lexicale WordNet qu'on va la détailler en définissant ces concepts de base et les relations sémantiques entre concepts.

2.2 La représentation du texte

Puisque les documents à traiter sont des données non structurées, ils nécessitent un codage pour être utilisés par les approches de segmentation. C'est pourquoi une étape préliminaire dite de représentation de termes est nécessaire.

2.2.1 La représentation statistique

Cette étape consiste en la transformation de chaque unité textuelle (unité atomique [1]) par un vecteur dont chaque composante représente un terme, alors toute unité u_j est représentée par un vecteur $u_j = (w_{1j}, w_{2j}, w_{3j}, ...w_{Tj})$, ou T est l'ensemble des termes qui apparaissent au moins une fois dans l'unité textuelle, le poids w_{kj} est le nombre d'occurrence d'un terme donné t_k dans une unité textuelle u_j.

2.2.1.1 Représentation par sac de mots (Bag of Words)

La représentation des textes la plus simple est celle qui porte le nom de « sac de mots », elle consiste à transformer les textes en vecteurs dont chaque composante représente un mot. D'autre auteurs parlent d' « ensemble de mots » lorsque les poids associés sont binaires, avec cette approche, seule la présence ou l'absence du terme est porteuse d'information.

2.2.1.2 Représentation du texte par des phrases

Particulièrement, d'autres auteurs utilisent une phrase à la place d'un mot. Cette représentation permet d'effectuer une sélection des phrases (pas l'entité grammaticale « phrase » que l'entend habituellement). Une phrase est composée d'un ensemble ou d'une séquence de mots se suivant dans le texte et qui porte un sens important.
D'après les résultats des travaux, cette approche n'est pas encouragée, car le grand nombre de combinaisons (séquences) possible même à des fréquences faibles et un problème de taille (pour n mots il y a probablement n^k séquences, avec k est la longueur de la séquence).

2.2.1.3 Représentation par des stems ou des lemmes

Dans la représentation précédente (en sac de mots), la taille du vecteur sera très grande de sorte que chaque mot est considéré comme un terme. Si on a l'ensemble de mots suivant dans un texte donné : "**Consult**", "**Consultant**", "**Consulting**", "**Consultative**", "**Consultants**" sont considérés comme 5 termes dans un vecteur, alors qu'on peut les remplacer par l'unique racine (ou stem) «**consult**».

1. Notre tâche consistant à découper le texte en plusieurs segments, cela nous interdit d'office l'usage du texte comme unité atomique.

Les algorithmes d'extraction des racines sont plus simples que ceux d'extraction des lemmes, ils extraient la racine seulement à partir de la forme du mot, par contre la lemmatisation prend en considération le sens du mot, cela signifie qu'après avoir appliqué la lemmatisation, nous obtiendrons toujours un mot valide.

2.2.2 La représentation conceptuelle

La représentation conceptuelle se base aussi sur le formalisme vectoriel pour représenter les documents, mais elle reste fondamentalement différente de la représentation statistique. Les dimensions de l'espace ne sont pas associées à des termes mais à des concepts. Pour permettre une telle représentation des documents, il est nécessaire de pouvoir projeter n'importe quelle lexie du dictionnaire sur l'espace généré par l'ensemble des concepts prédéfinis. Comme espace de concepts, nous citons le thesaurus WordNet.

L'avantage de la représentation conceptuelle est en particulier, de réduire les effets synonymiques du vocabulaire.

2.3 Les ontologies

2.3.1 Définitions

— En philosophie, le terme Ontologie est utilisé depuis le 17^{me} siècle. Il est construit à partir des racines grecques ontos (ce qui existe, l'existant) et logos (le discours ou l'étude) c'est à dire l'étude de ce qui existe. Selon ARISTOTE l'Ontologie est une partie de la métaphysique qui étudie l'être en tant qu'être, étude des propriétés générales de ce qui existe [30].

— L'ontologie est employée en informatique et en science de l'information au début des années 90. Elle est définit comme un ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métas données d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine concerné [30].

— Une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance [30]. Le terme conceptualisation, dans la définition fait référence à un système de concepts, autrement dit à un ensemble structuré de concepts. La spécification signifie pour sa part que la conceptualisation est représentée dans un langage (un artefact informatique). Ce langage peut être une langue naturelle (ex : francais, Anglais, arabe), ou un langage formel (ex : logique du 1^er ordre) [30].

— Une ontologie est une structure de données hiérarchique qui comprend toutes les entités du domaine que l'on tente de décrire ainsi que les relations sémantiques

qui existent entre ces différentes entités. Mais une ontologie doit être plus qu'une simple taxonomie. Une ontologie informatique doit absolument contenir des relations ajoutant de l'information sémantique aux concepts.

— Une ontologie O est un système formel qui est constitué d' :
Un ensemble C de concepts organisés en une hiérarchie H^C ou les concepts sont reliés par une relation directe, acyclique, transitive et réflexive : $H^C(C_1, C_2)$ signifie que C_1 est un sous concept de C_2.

Pour conclure cette partie, on peut donc affirmer que les définitions du terme ontologie traitent les connaissances, leurs définitions et leurs manipulations.

2.3.2 Les objectifs de l'ontologie

Une ontologie nous aide à :

— Analyser la connaissance du domaine : Les ontologies sont les outils pour fournir une description complète du domaine d'intérêt par rapport aux besoins des utilisateurs.

— Partager une compréhension commune de la structure de l'information entre les personnes ou les agents logiciels : Par exemple, des informations d'un domaine précis sont publiées sur plusieurs sites Web différents, si ces sites Web partagent la même ontologie, les agents informatiques peuvent ensuite extraire et agréger des informations de ces différents sites et les utiliser pour répondre aux requêtes des utilisateurs.

— Réutiliser les connaissances du domaine :Une grande ontologie peut être construite par l'intégration de plusieurs ontologies existantes décrivant des portions du grand domaine. Et au contraire, nous pouvons réutiliser une ontologie générale pour décrire notre domaine d'intérêt.

2.3.3 Les concepts et les relations dans une ontologie

2.3.3.1 Concept

Un concept peut représenter un objet matériel, une notion, une idée. Il peut être divisé en trois parties ; un terme (ou plusieurs), une notion et un ensemble d'objets.
L'intension d'un concept est sa définition, contient la sémantique du concept exprimée en termes de propriétés et d'attributs, de règles et de contraintes.

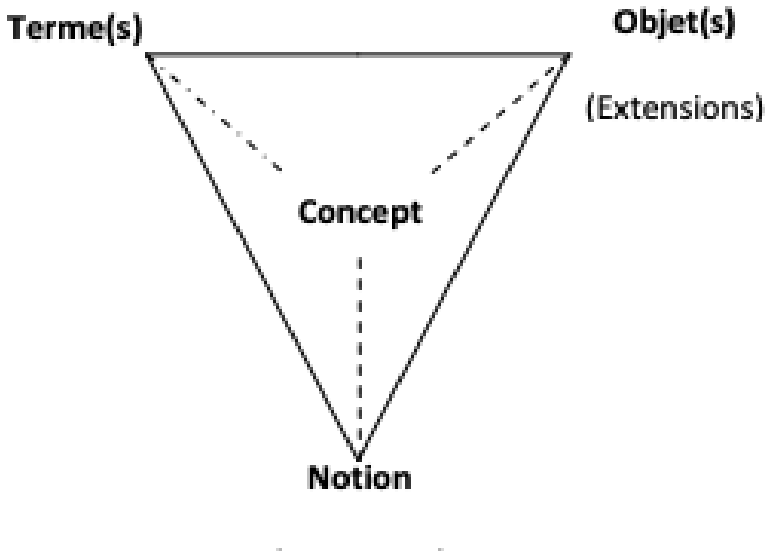

Figure 2.1: Le triangle sémantique [31].

Les objets, aussi appelés extension du concept, servent à regrouper les objets manipulés à travers le concept, ces objets sont les instances d'un concept. Par exemple le terme « **table** » renvoie à la fois à la notion de table comme objet de type « **meuble** » possédant « **une planche plate** » et « **des tréteaux** » , et à l'ensemble des objets de ce type. Un concept ayant une extension vide est appelé concept générique, ces concepts génériques correspondent généralement à des notions abstraites, par exemple, « **tout ce qui est vrai** » et « **vérité** ». Les concepts manipulés dans un domaine de connaissance sont organisés au sein d'un réseau de concepts liés par des propriétés conceptuelles. Les propriétés portant sur des concepts sont :

— La généricité : un concept est générique s'il n'admet pas d'extension.
— L'abstraction : un concept est abstrait si toute instance de ce concept est aussi instance d'un de ses concepts fils. Par exemple, dans une hiérarchie comportant les concepts « **mère** » et « **père** », fils du concept « **parent** »
— L'identité : Guarino dans [32] propose cette propriété et indique qu'un concept porte une identité si cette propriété permet de conclure quant à l'identité de deux instances de ce concept. Cette propriété peut porter sur des attributs du concept ou sur d'autres concepts. Par exemple, le concept « **étudiant** » porte une propriété d'identité liée au numéro d'étudiant, deux étudiants étant identiques s'ils ont le même numéro.
— La rigidité : Proposée par Guarino dans [32], un concept est rigide si toute extension de concept en reste extension dans toutes les connaissances du monde possibles. Par exemple, « **l'homme** » est un concept rigide, « **étudiant** » est un

concept non rigide.
- L'équivalence : deux concepts sont équivalents s'ils ont la même extension.
- La disjonction (incompatibilité) : deux concepts sont disjoints si leurs extensions sont disjointes. Par exemple « **un homme** » et « **une femme** ».
- La dépendance : proposée par Guarino, est définie comme suit : Un concept C1 est dépendant d'un concept C2, si pour toute instance de C1 il existe une instance de C2 qui ne soit ni partie ni constituant de l'instance C2. Exemple : « **père** » est un concept dépendant de « **fils** » (et vice-versa).

2.3.3.2 Relation

Les relations ont pour but de lier des instances d'un concept, ou des concepts génériques. Elles sont caractérisées par un ou plusieurs termes et une signature qui définit le nombre d'instances de concepts liées par cette relation, leur type et l'ordre des concepts (la manière qu'un concept doit être lu). Par exemple une relation nommée « **lire** » qui lie une instance du concept « **un homme** » avec une instance du concept « **texte** » dans cet ordre.
Les concepts peuvent être spécifiés par des propriétés, alors c'est aussi le cas pour les relations :
- Les propriétés algébriques : symétrie, réflexivité, transitivité ,etc.
- La cardinalité : nombre possible de relations de ce type entre les mêmes concepts (ou instances de concept). Les relations portant une cardinalité représentent souvent des attributs.
  Exemple : « **un homme** » « **a ente 0 et 2** » « **main** »

On a vu les propriétés des relations liant des instances des concepts. Il existe aussi des propriétés qui lient deux relations :
- L'incompatibilité : deux relations sont incompatibles si elles ne peuvent pas lier les mêmes instances de concepts simultanément.
  Exemple : les relation « **est chaud** » et « **est froid** » sont incompatibles.
- L'inverse : deux relations binaires sont inverses l'une à l'autre si, quand l'une lie deux instances I1 et I2, l'autre lie I2 et I1. Par exemple : les relations « **père** » et « **fils** » sont inverses l'une de l'autre.

2.4 Ontologies et segmentation thématique de textes

2.4.1 Les ontologies les plus connus en TALN

2.4.1.1 Corelex [16]

La base de données sémantique (CoreLex) de 126 types sémantiques, couvre près de 40.000 noms et définie un grand nombre de classes polysémiques systématiques qui sont dérivés d'une analyse minutieuse des distributions de sens dans WordNet. Les types sémantiques sont des représentations sous-spécifiées basées sur la théorie lexique générative. La base de données CoreLex est librement disponible à des fins de recherche, ou commerciales.
La base de données se compose de trois fichiers, chacun a un certain format qui les lie ensemble comme une base de données relationnelle.

2.4.1.2 GUM (Generalized Upper Model)

Le Generalized Upper Model est une ontologie linguistique générale, indépendante de tout domaine et de tout type de tâche qui fournit une sémantique pour le langage naturel. GUM essaye d'être multilingue le plus loin possible. Cette ontologie peut être utilisée pour créer d'autre ontologie pour des langues spécifiques, telles que l'Anglais, l'allemand, etc.

2.4.1.3 WordNet

WordNet est une base de données lexicale des relations sémantiques entre les mots dans plus de 200 langues. Les synonymes sont regroupés en *synsets* en Anglais avec de courtes définitions et des exemples d'utilisation. Dans notre travail, on va utiliser WordNet pour l'Anglais qui sera abordée avec plus de détails.

2.4.2 Quelle ontologie choisir ?

Les ontologies choisis à être utiles dans la segmentation thématique des textes doivent être adaptées à la tâche de la segmentation thématique considérée et plus particulièrement elles doivent comprendre de la connaissance suffisante pour intégrer la prise en compte des relations sémantiques entre les mots. WordNet utilise des relations sémantiques clairement représentées dans un réseau lexico-sémantique.

2.5 WordNet

Une création d'un petit filet de 45 noms sur un PC d'IBM est le commencement de WordNet en 1984 par George Miller. Ce produit a été montré plus tard à IBM et

à Bellcore. Le début effectif du projet WordNet était en 1985 après une collaboration entre Bellcore et l'Université de Princeton. WordNet a été vu comme entrée, plutot que similitudes alphabétiques. En 1986, la liste Fred Chang de mots a été employée et s'est ajoutée comme entrée.
En 1993, le lexique de COMPLEX se compose de 39143 mots. La liste des mot a été entrain de se développer progressivement ce qui a mené à la nécessité de division de la base de données. Une division par catégorie syntaxique des mots était la première division de la base de données.
WordNet avais un souci avec la morphologie flexionnelle, il n'était pas capable d'identifier les pluriels. En 1989, c'était le développement d'un programme appelé Word Filter qui a employé une liste d'exceptions pour déterminer des mots avec une morphologie régulière.
En 1991, une interface appelée ConText a été développée. ConText prétraite le texte par tokenisation, étiquetage des mots, et lemmatisation avant de produire le mot cible avec la signification pour l'entreprise dans WordNet.
La première version de WordNet qui était publiquement disponible était la version 1.0 de 1991, la version 3.1 est la plus récente à ce jour. Cependant, WordNet est toujours en train d'être travaillé dessus.

2.5.1 Définition

WordNet est un thesaurus pour la langue Anglaise basé sur des études psycholinguistiques et développé à l'université de Princeton par G.Miller en 1985 [33]. Il a pour but de répertorier, classifier et de mettre en relation le contenu sémantique et lexical de la langue Anglaise (Il existe d'autres versions de WordNet pour d'autres langues mais la langue Anglais est la plus complète).
WordNet est une base de données électronique téléchargeable sur internet. Il est distribué avec une licence libre qui permet l'utilisation commerciale ou à des termes de recherche.
Ce système repose sur un composant atomique appelé le *synset* qui veut dire *l'ensemble de synonymes* (synonymes set en Anglais) et des relations sémantiques.
Il y a trois principes auxquels le processus de construction de synset doit adhérer :

- **Minimalité :** Ce principe insiste sur la capture d'un ensemble minimal de mots dans le synset qui identifie de manière unique le concept. Par exemple l'ensemble $\{family, house\}$ peut identifier un même concept, exemple : « He is from the **house**/**family** of the Jacksons. ».
- **Couverture :** Ce principe met l'accent sur l'achèvement du synset, c'est à dire la capture de tous les mots qui représentent le concept exprimé par le synset. Au sein du synset, les mots doivent être classés en fonction de leur fréquence dans le corpus.
- **Remplaçabilité :** Au sein du synset, les mots doivent être ordonnés en fonction

de leur fréquence dans le corpus. La remplaçabilité exige que les mots les plus courants dans le synset c'est à dire un mot d'un synstet doit pouvoir remplacer d'autres mots du même synset dans les phrases des exemples associées au synset.

WordNet est différent des autres dictionnaires traditionnels crées avant. WordNet est un thesaurus qui consiste à décrire la langue Anglaise.

2.5.2 Les concepts de base WordNet

La division de la base de données en fonction des catégories est la différence majeure entre WordNet et les autres dictionnaires traditionnels. WordNet sépare les données en quatre catégories : catégorie de noms, de verbes, d'adjectifs et d'adverbes.

2.5.2.1 La base des noms

Les noms sont ainsi classés en un système de catégories complet et précis comprenant plusieurs niveaux d'imbrications (on retrouve notamment certaines sections de cette ontologie ou la profondeur dépasse 10 niveaux) .Certains synsets ne sont couverts par aucun autre synset, chacun d'eux constitue une hiérarchie séparée correspondant à une distinction relative des champs sémantiques. On retrouve 25 types de synsets [34] organisés en une hiérarchie.

2.5.2.2 La base des verbes

On retrouve un système de classification beaucoup moins élaboré pour les verbes, qui sont organisés en un système hiérarchiques beaucoup plus plat (moins de niveaux d'imbrication), ou on passe très rapidement d'un concept spécialisé (le sens *operate*, *run* du verbe *running*, par exemple) à un concept très général (*control*, *command*).
Au départ, les verbes ont été regroupés sur des critères généraux en 17 grandes familles comme le mouvement, la possession, la perception. Le contact, la communication, le changement, l'apprentissage, la conception, la création, l'émotion, le soin du corps et la vitalité, les relations et les interactions sociales et la météorologie. Ces classes ont été subdivisées jusqu'à l'obtention d'un synset.
à ce jour, les adjectifs et les adverbes ne sont pas encore hiérarchiquement catégorisés.

2.5.3 Les relations dans WordNet

WordNet répertorie une grande variété de relations sémantiques permettant d'organiser le sens des mots, Nous allons présenter ci-dessous ces différentes relations.

— **Synonymie :** la relation de synonymie est la base de la structure de WordNet. Les unités lexicales sont regroupées en ensembles de synonymies (synsets). On a donc dans un synset tous les termes utilisés pour dénoter le concept. La définition

de synonymie utilisée dans WordNet par Miller [34] est la suivante :
« Deux expressions sont synonymes dans un contexte linguistique C, si la substitution de l'une pour l'autre en C ne modifie pas la valeur de vérité de la phrase dans laquelle la substitution est faite »

— **Antonymie :** C'est une relation lexicale qui indique l'opposé. Elle provient principalement d'adjectifs descriptifs. De plus, chaque membre d'une paire d'antonymes directs est associé à des adjectifs sémantiquement similaires. Par exemple : « *fat* » est l'opposé de « *thin* », « *obese* » est aussi l'opposé de « *thin* » car « *fat* » et « *obese* » sont dans le même synset.
— **Gradiation :** C'est une relation lexicale qui représente des états intermédiaires possibles entre deux antonymes. Par exemple : « morning, noon, evening ».
— **Hypernymy and Hyponymy :** représentent des relations lexicale entre un terme général et des instances spécifiques de celui-ci. Elles construisent un arbre hiérarchique avec des concepts de plus en plus concrets/particuliers issus de la racine abstraite.
— **Méronymie :** Elle exprime la partie de la relation. Les synsets désignant des parties, des composants ou des membres de synsets.
A est un méronyme de B implique que A est une partie de B(ou B est construit de A).
— **Holonymie :** Est l'inverse de la méronymie.
A est un holonyme de B implique que A est fait de B (A contient B).
— **Implication :** Est une relation sémantique entre deux verbes. Un verbe A implique un verbe B, si le sens du verbe B est logiquement et strictement inclus dans le sens du verbe A. Cette relation est unidirectionnelle. Par exemple : « *snoring* » implique « *sleeping* », par contre « *sleeping* » n'implique pas « *snoring* ».
— **Troponymie :**Il s'agit d'une relation sémantique entre deux verbes lorsque l'un est une élaboration « de manière » spécifique d'un autre.

2.6 L'utilisation des ontologies dans la segmentation thématique des textes

Dans les méthodes de segmentation thématique qui sont basées sur la représentation conceptuelle de texte (considérant un concept est signé par un ensemble des mots), on utilise des réseaux lexico-sémantiques comme des ressources sémantiques, tels que l'ontologie WordNet.

2.7 Conclusion :

Durant ce chapitre, nous avons vu les différentes définitions d'ontologie, ses objectifs, nous avons aussi souligné les différents types d'ontologies et les plus connus pour le traitement automatique des langues naturelles TALN.

Le choix des ontologies est une tâche cruciale afin de résoudre le problème de variation sémantique dans les textes à segmenter thématiquement.

Dans le chapitre suivant, nous allons présenter notre approche de segmentation thématique des textes Anglais à l'aide de l'ontologie WordNet Anglais comme nous allons évaluer notre système en terme de quelques mesures en utilisant un corpus de référence.

Chapitre 3

L'implémentation d'un segmenteur thématique pour les textes Anglais

3.1 Introduction

Beaucoup de travaux qui sont réalisés sur la segmentation thématique de textes utilisent des méthodes qui sont probabilistes (statistiques) avec leur nature lexicale et leur absence de structuration thématique explicite qui font que les systèmes qui les utilisent sont parfois mis en échec par l'ambiguité sémantique des mots ou par l'impossibilité d'identifier des relations comme spécifiquement sémantiques.
L'intégration de la prise en compte des relations sémantiques devient nécessaire par l'utilisation des relations sémantiques clairement identifiées dans des thesaurus ou des réseaux lexico-sémantique tel que WordNet.
L'objectif de notre travail est de développer un système de segmentation thématique des textes Anglais. On essaye de résoudre le problème de variation sémantique par l'intégration de l'ontologie WordNet pour une représentation conceptuelle des textes, afin d'obtenir des résultats satisfaisants.

3.2 La méthode implémentée

Nous avons implémenté une méthode passive non supervisée qui est basée sur le calcul de similarité en utilisant la mesure TF-IDF après qu'on a fait une représentation conceptuelle à l'aide de WordNet.
Pour l'évaluation, on a utilisé un corpus qui contient des textes parlant de différents sujets.

3.3 Le langage et les outils utilisés

3.3.1 Python

Python est un langage de programmation orientée objet à typage dynamique et il peut être utilisé pour la programmation fonctionnelle aussi.
Python est le langage le plus populaire comme il existe un nombre de communautés et de ressources disponibles pour soutenir les développeurs.
Python est livré avec une énorme quantité de bibliothèques intégrées. De nombreuse bibliothèques sont destinées à l'intelligence artificielle et à l'apprentissage automatique.
Il y a beaucoup de choses à propos de Python qui en font un très bon choix de langage de programmation pour un projet de TALN. La syntaxe simple et la sémantique transparente de ce langage en font un excellent choix pour les projets qui incluent des tâches de traitement du langage naturel. Il fournit aux développeurs une vaste collection d'outils et de bibliothèques de TALN tel que « *SpaCy*»

3.3.2 SpaCy

SpaCy est une bibliothèque open source pour effectuer des tâches simples à avancées de traitement du langage naturel tels que la tokenisation, lemmatisation, etc.

3.3.3 WordNet Anglais

WordNet est une base de données électronique téléchargeable sur internet. Il est distribué avec une licence libre qui permet l'utilisation commerciale ou à des termes de recherche. Elle est organisée autours de la structure des synsets.
Cette base de données contient 175 979 synstets (voir chapitre 2).

3.3.4 The Jupyter Notebook

The Jupyter Notebook est une application Web open source qui nous permet de créer et de partager des documents contenant du code en direct, des équations et des visualisations. Les utilisations incluent : le nettoyage et la transformation des données, la simulation

numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore.

3.3.5 Pandas

Pandas est une bibliothèque pour le Python, elle nous permet la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques

3.4 L'architecture de notre système

La figure ci-dessous (figure 3.1) représente l'architecture générale de notre système de segmentation thématique des textes Anglais :

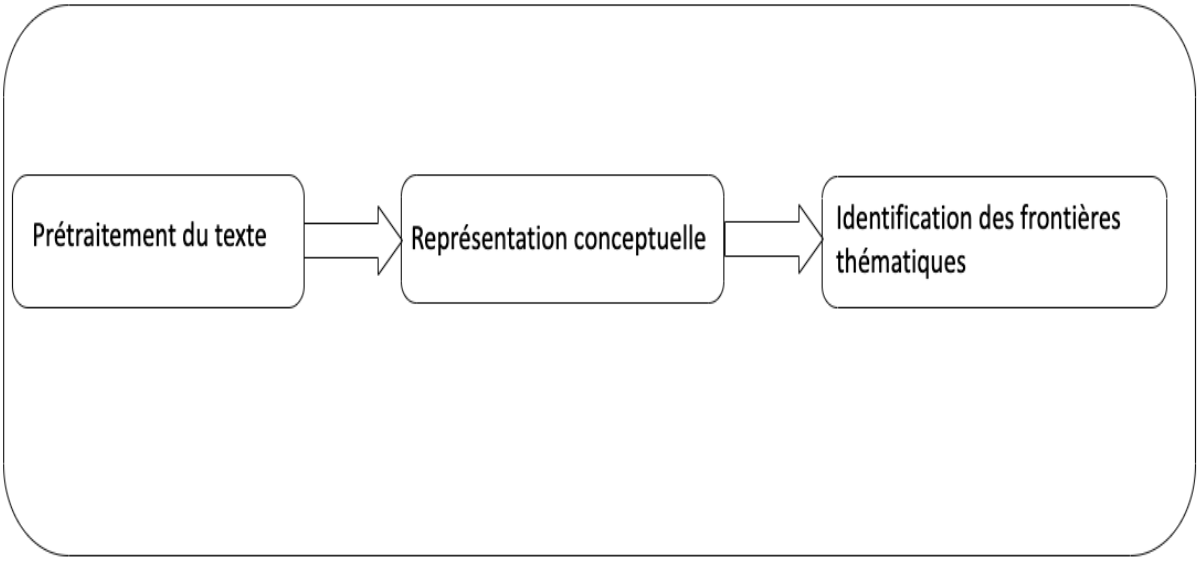

Figure 3.1: L'architecture générale de notre système de segmentation thématique de textes Anglais.

Le développement de notre système se compose de trois étapes principales qu'on va les détailler :

1. **Le prétraitement du texte :** Cette étape consiste à diviser le texte en entrée en phrases, ensuite on va nettoyer les phrases extraites en lemmatisant les mots de chaque phrase après qu'on a éliminé les mots vides et mettre toute phrase en minuscule. Finalement nous allons créer le sac des mots qu'il ne va contenir aucun mot vide et que des lemmes.

2. **La représentation conceptuelle :** La deuxième étape, consiste à regrouper les mots qui sont sémantiquement similaires sous un seul ensemble que nous allons l'appeler « *un concept* ».
3. **L'identification des frontières thématiques :** La dernière étape, elle permet de définir les segments thématiques.

3.5 L'implémentation de notre méthode

3.5.1 Le prétraitement

Le prétraitement est une étape qui se compose de trois parties essentielles.
La figure 3.2 représente les différentes étapes dans notre module de prétraitement :

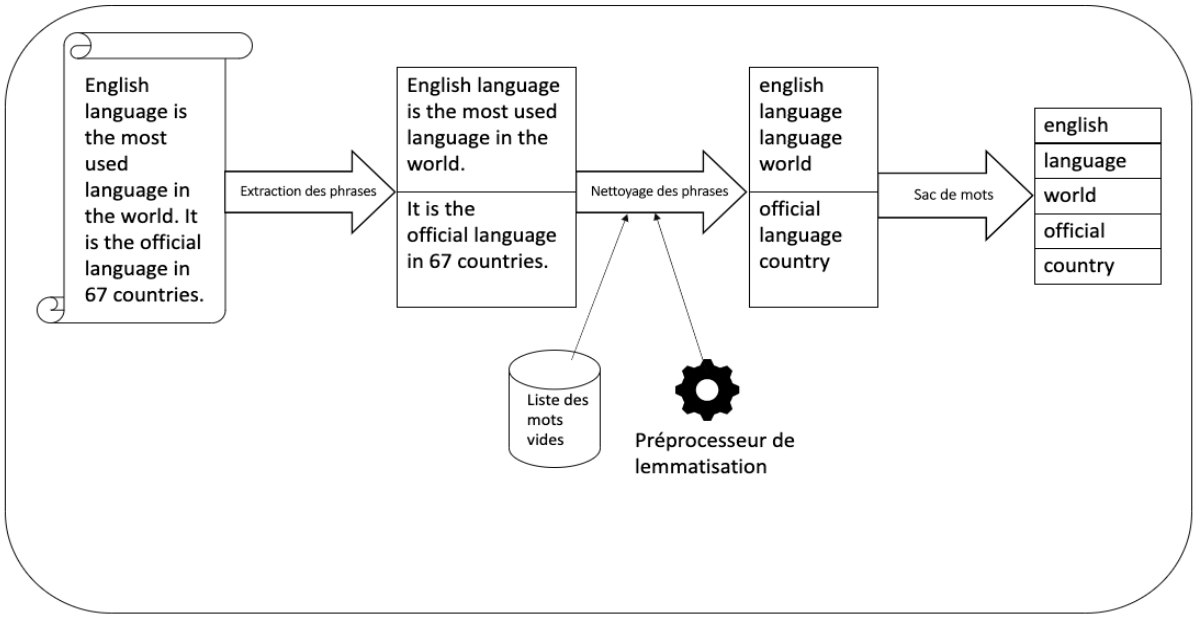

Figure 3.2: Le module de prétraitement de notre système de la STT.

3.5.1.1 L'extraction des phrases

Cette étape consiste à déterminer les phrases du texte. à l'aide de la bibliothèque « *SpaCy* », nous allons extraire les phrases et leur ordre dans le texte. à chaque fois que le processus trouve un symbole indiquant la ponctuation («., ?, !, ...»), il va considérer la portion qui est entre deux symboles de ponctuation (La fin de la phrase précédente et la fin de la phrase actuelle) comme étant une phrase dans le cas général. (La première phrase et celle qui est située entre le début du texte et le premier symbole de ponctuation).

La figure ci-dessous (figure 3.3) représente la liste des phrases obtenues avec leurs positions d'un texte donné :

	Order	Sentence
0	0	Chemical reaction process substances reactants converted different substances products
1	1	Substances chemical elements compounds
2	2	chemical reaction rearranges constituent atoms reactants create different substances products
3	3	Chemical reactions integral technology culture life
4	4	Burning fuels smelting iron making glass pottery brewing beer making wine cheese examples activities incorporating chemical reactions known thousands years
5	5	Chemical reactions abound geology Earth atmosphere oceans vast array complicated processes occur living systems
6	6	Chemical reactions distinguished physical changes
7	7	Physical changes include changes state ice melting water water evaporating vapour
8	8	physical change occurs physical properties substance change chemical identity remain
9	9	matter physical state water H2O compound molecule composed atoms hydrogen atom oxygen
10	10	water ice liquid vapour encounters sodium metal Na atoms redistributed new substances molecular hydrogen H2 sodium hydroxide NaOH
11	11	know chemical change reaction occurred

Figure 3.3: Liste des phrases avec leurs positions.

3.5.1.2 Le nettoyage des phrases

La liste des phrases obtenue contient des mots vides et des symboles qui ne seront pas utiles pour notre tâche, ce qui nécessite l'élimination de ce bruit. Cette étape a pour but de remplacer chaque mot par son lemme, d'éliminer les mots vides, les caractères inutiles tels que le caractère signifiant le saut de ligne ($\backslash n$) et de rendre tous les mots dans chaque phrase en minuscule. Pour cela, nous avons utilisé le lemmatiseur de SpaCy, la liste des mots vides de SpaCy qui contient 326 mots en Anglais par défaut et quelques fonctions prédéfinies sur python tels que la fonction *lower()* qui a pour but de mettre les phrases en minuscule.

La figure 3.4 représente la liste des phrases obtenues avec leurs positions d'un texte donné après l'étape de nettoyage :

	Order	Sentence
0	0	chemical reaction process substance reactant convert different substance product
1	1	substance chemical element compound
2	2	chemical reaction rearrange constituent atom reactant create different substance product
3	3	chemical reaction integral technology culture life
4	4	burn fuel smelt iron make glass pottery brew beer make wine cheese example activity incorporate chemical reaction know thousand year
5	5	chemical reaction abound geology Earth atmosphere ocean vast array complicated process occur live system
6	6	chemical reaction distinguish physical change
7	7	physical change include change state ice melting water water evaporating vapour
8	8	physical change occur physical property substance change chemical identity remain
9	9	matter physical state water h2o compound molecule compose atom hydrogen atom oxygen
10	10	water ice liquid vapour encounter sodium metal Na atom redistribute new substance molecular hydrogen H2 sodium hydroxide NaOH
11	11	know chemical change reaction occur

Figure 3.4: Liste des phrases nettoyées avec leurs positions.

3.5.1.3 La création du sac des mots

La création du sac des mots (lemmes) est la dernière étape du prétraitement, dans cette étape nous allons créer le sac des mots, c'est à dire d'extraire les mots à partir des phrases nettoyées que nous avons déjà obtenu, alors le sac des mots ne va contenir aucun mot vide et que des mots en minuscule. On va aussi éliminer tous les mots dupliqués dans notre sac des mots.

Le sac des mots d'un texte crée avec notre système de la STT est représenté dans la figure suivante :

	Words
0	chemical
1	reaction
2	process
3	substance
4	reactant
...	...
71	new
72	molecular
73	h2
74	hydroxide
75	naoh

Figure 3.5: Le sac des mots d'un texte crée par notre système de la STT.

3.5.1.4 Le résultat du module de prétraitement

Dans cette phase, nous avons décortiqué le texte en entrée sous forme des phrases, ensuite nous avons nettoyé les phrases obtenues en éliminant les mots vides, les caractères inutiles et nous avons mis toute phrase en minuscule. Enfin nous avons extrait les mots de chaque phrase nettoyée et nous avons éliminé la duplication des mots, et comme ça, nous avons crée notre sac des mots.

3.5.2 La représentation conceptuelle

La représentation conceptuelle est le deuxième module de note tâche.

La figure 3.6 représente une une brève démonstration de la représentation conceptuelle de notre système :

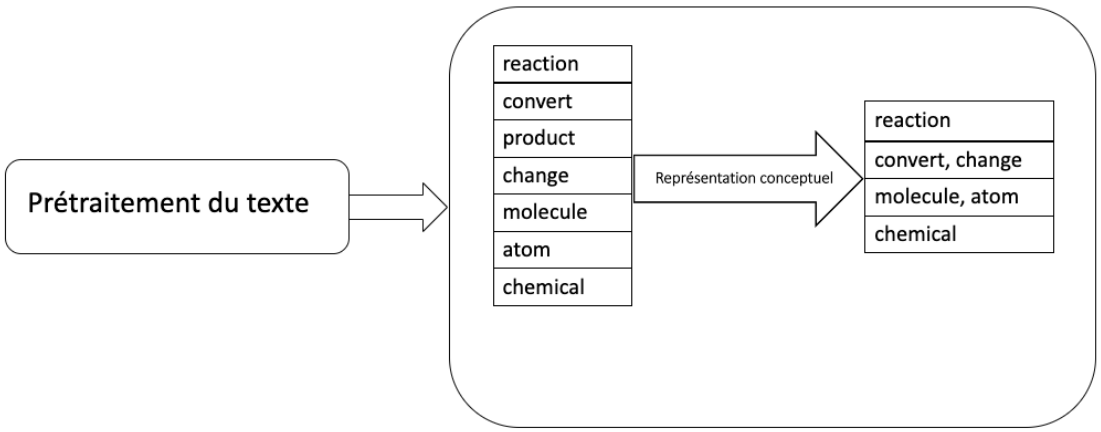

Figure 3.6: La représentation conceptuelle à partir d'un sac des mots.

La représentation conceptuelle joue un rôle crucial pour obtenir des résultats satisfaisants dans la segmentation thématique des textes. Dans cette étape, nous avons utilisé l'ontologie WordNet pour créer « *des concepts* ». Ces concepts sont crées en regroupant les termes qui ont une similarité sémantique.

Pour chaque mot A dans le sac des mots, on va vérifier s'il existe un autre mot B qui est sémantiquement similaire à A à l'aide de WordNet. Si oui alors on regroupe les deux mots A et B dans une liste et on continue la vérification.

Si un mot appartient déjà à un concept, nous allons pas le traiter encore une fois pour éviter la duplication. Par exemple, dans le traitement d'un mot A, nous avons trouvé un mot B sémantiquement similaire à A, alors nous allons mettre le mot B dans une liste qui contient les mots qui sont déjà faits partie d'un concept. Alors notre programme ne va pas faire le traitement de recherche des mots similaires pour les mots qui sont dans la liste contenante les mots qui appartiennent à un concept.

La figure 3.7 ci-dessous représente les résultats qu'on a obtenu après la représentation conceptuelle à l'aide de WordNet :

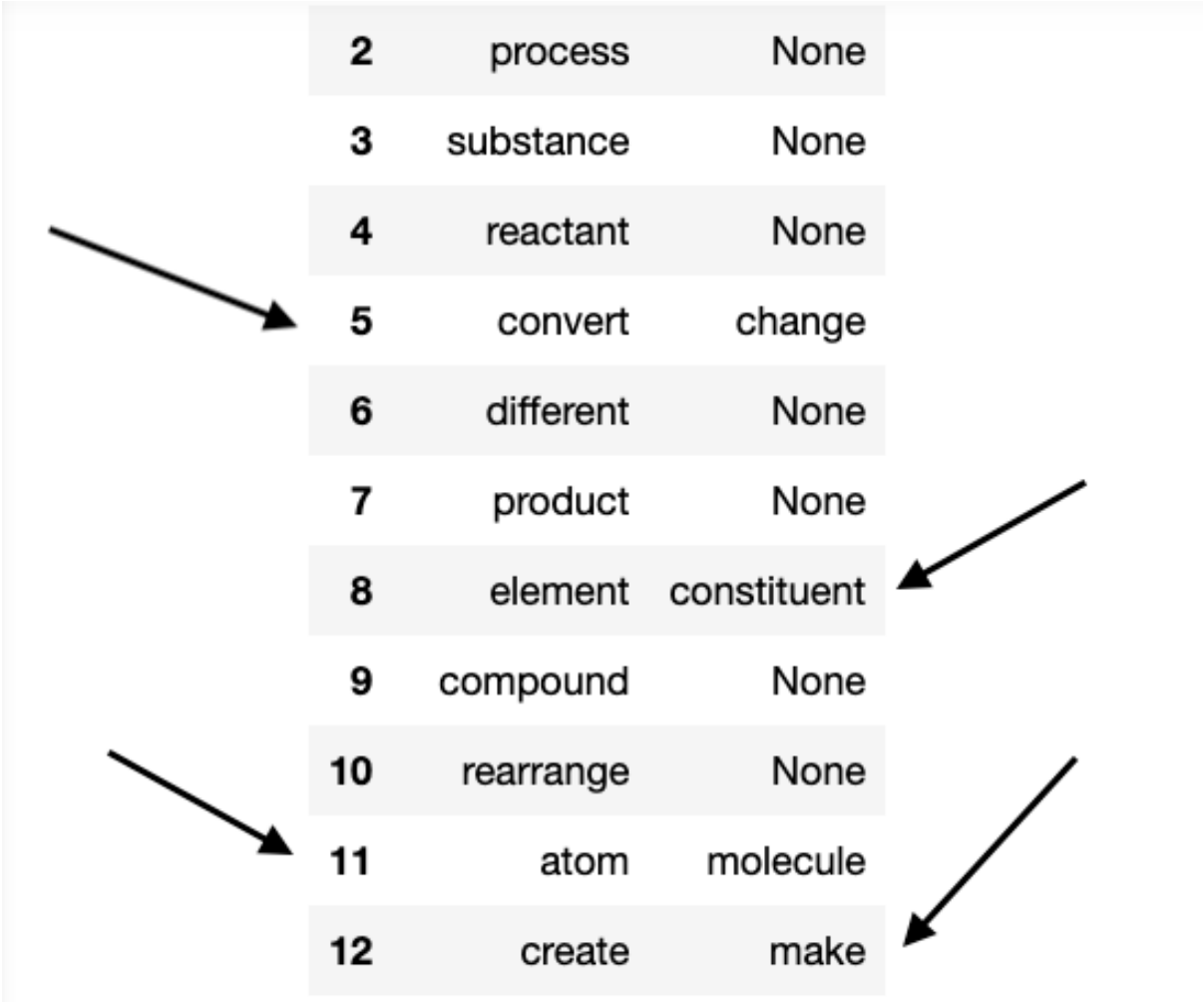

Figure 3.7: Un échantillon de la représentation conceptuelle d'un texte.

3.5.2.1 Le résultat de la représentation conceptuelle

Dans cette étape, nous avons créé des concepts. Ces concepts vont nous permettre de traiter les termes qui ont une même signification comme étant un seul élément. Et comme ça nous avons pris en compte les relations sémantiques entre les mots.

3.5.3 L'identification des frontières thématiques

L'identification des frontières thématiques est la phase finale dans le processus de segmentation, elle utilise les informations précédemment crées pour la détection des frontières thématiques. Notre méthode d'extraction des segments thématiques est basée sur le calcul du TF-IDF. Cette phase se compose de trois étapes principales.

1. La création des suites successives des phrases.
2. Le calcul du TF-IDF pour chaque concept dans chaque suite des phrases.

3. L'extraction des segments.

3.5.3.1 La création des suites successives des phrases

La mesure TF-IDF est une mesure qui vise à refléter l'importance d'un mot dans un ensemble des documents. Dans notre cas, cette mesure va refléter l'importance d'un concept dans un ensemble des phrases. Pour cela, nous allons traiter chaque phrase comme étant un document et chaque suite successive des phrases que nous allons créer est l'ensemble des documents.
Dans un texte T qui contient N phrases, nous allons avoir N suites des phrases. Chaque suite S_n va contenir les phrases de K à n tel que K est initialement égale à 0.
à chaque fois que notre système identifie un segment thématique, K va devenir égale à l'indice de la phrase qui suit la fin du dernier segment obtenu.

La figure 3.8 représente les suites des phrases crée pour un texte de 12 phrases :

	0	1	2	3	4	5	6	7	8	9	10	11
0	0	None	None	None	None	None	None	None	None	None	None	None
1	0	1.0	None	None	None	None	None	None	None	None	None	None
2	0	1.0	2.0	None	None	None	None	None	None	None	None	None
3	0	1.0	2.0	3.0	None	None	None	None	None	None	None	None
4	0	1.0	2.0	3.0	4.0	None	None	None	None	None	None	None
5	0	1.0	2.0	3.0	4.0	5.0	None	None	None	None	None	None
6	0	1.0	2.0	3.0	4.0	5.0	6.0	None	None	None	None	None
7	0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	None	None	None	None
8	0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	None	None	None
9	0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	None	None
10	0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	None
11	0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0

Figure 3.8: Un exemple des suites des phrases crées à partir d'un texte contenant 12 (douze) phrases.

Cette étape consiste à diviser le texte en parties (les suites des phrases) pour les utiliser ultérieurement pour savoir l'importance de chaque concept dans chaque portion à travers le calcul du TF-IDF.
Pour calculer TF-IDF dans chaque suite des phrases, nous avons besoin de savoir le nombre des termes existants dans chaque suite (cette donnée est nécessaire pour le calcul de la mesure TF).

La figure ci-dessous (figure 3.9) représente les suites des phrases obtenues suivies avec le nombre des termes dans chaque suite :

	Suite	NB_termes
0	[0]	9
1	[0, 1]	13
2	[0, 1, 2]	23
3	[0, 1, 2, 3]	29
4	[0, 1, 2, 3, 4]	49
5	[0, 1, 2, 3, 4, 5]	63
6	[0, 1, 2, 3, 4, 5, 6]	68
7	[0, 1, 2, 3, 4, 5, 6, 7]	79
8	[0, 1, 2, 3, 4, 5, 6, 7, 8]	89
9	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]	101
10	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	119
11	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]	124

Figure 3.9: Les suites des phrases avec le nombre des termes dans chaque suite.

3.5.3.2 Le calcul du TF-IDF

Après que nous avons représenté le texte en suites des phrases, il faut alors calculer le TF-IDF pour chaque concept dans chaque suite des phrases.
Le calcul du TF-IDF d'un concept se passe en plusieurs étapes.

1. **Le calcul du TF :** Term Frequency est la fréquence d'un mot dans un texte divisée sur le nombre total des mots du texte. Dans notre cas nous allons calculer le TF-IDF pour des concepts et non pas pour des mots. Sachant qu'un concept est composé d'un ensemble des mots, pour chaque concept, dans chaque suite des phrases, notre système va parcourir toutes les phrases, à chaque fois qu'un mot qui appartient au concept apparait la fréquence du concept va être augmentée par 1 jusqu'a la fin de cette suite. Ensuite, on va diviser la fréquence du concept sur le nombre total des termes. Et comme ça on a obtenu la mesure TF de chaque concept dans chaque suite.

2. **Le calcul du IDF :** Inverse Data Frequency est le *log* de nombre des phrases totale dans une suite des phrases (N) divisé sur le nombre des phrases dans lesquelles au moins un terme d'un concept apparait (df_t). Donc pour chaque concept, notre système va parcourir toutes les phrases d'une suite S, s'il trouve dans une phrase P un terme qui appartient au concept alors il augmente le df_t par 1 et on passe à la phrase suivante, Sinon il passe directement à la phrase suivante jusqu'a la fin de la suite.
3. **Le calcul du TF-IDF :** Le TF-IDF est tout simplement la métrique TF multipliée par la métrique IDF.

3.5.3.3 L'extraction des segments

L'extraction des segments est la dernière tâche de notre système. L'extraction du premier segment est faite par l'identification du concept qui a la plus grande valeur du TF-IDF. Cette identification va nous retourner trois informations importantes :
le concept C qui a la plus grande valeur du TF-IDF, son score TF-IDF et la suite des phrases S dans laquelle il a obtenu ce score. Et comme ça, on a obtenu le premier segment qui est l'ensemble des phrases composant la suite S et qui a un concept dominant C.
Après l'identification du premier segment, une modification dans la liste des concepts et dans la liste des suites des phrases est nécessaire pour identifier les autres segments.
il faut éliminer le concept C de la liste des concept puisqu'il est déjà associé à un segment. Comme il faut modifier la liste des suites des phrases, le nombre des suites va être égale au nombre des phrases composant le texte moins le nombre des phrases composant les segments identifiés. Chaque suite des phrases va commencer par la phrase qui suit la dernière phrase du dernier segment identifié, et on doit recalculer le nombre des termes dans chaque nouvelle suite des phrases et refaire les opérations précédentes jusqu'a l'obtention du dernier segment.

La figure ci-dessous (figure 3.10) représente l'identification des segments thématiques. Dans cet exemple notre programme nous a donné trois segments. Chaque segment est identifié respectivement par : les phrases composant le segment, le concept dominant dans le segment et le score TF-IDF du concept dans le segment obtenu :

	0
0	[[0, 1], [element, constituent], 5.331901388922657]
1	[[2, 3], [life, living], 4.332169878499658]
2	[[4, 5, 6, 7, 8, 9, 10, 11], [water], 3.097355535826504]

Figure 3.10: L'identification des segments thématiques.

Voir la figure 3.11 ci-dessous pour un exemple d'une segmentation thématique d'un texte avec notre système éxécuté sur Jupyter Notebook :

Chemical reaction, a process in which one or more substances, the reactants, are converted to one or more different substances, the products. Substances are either chemical elements or co mpounds. A chemical reaction rearranges the constituent atoms of the reactants to create diff erent substances as products.

Chemical reactions are an integral part of technology, of culture, and indeed of life itself. Burning fuels, smelting iron, making glass and pottery, brewing beer, and making wine and che ese are among many examples of activities incorporating chemical reactions that have been kno wn and used for thousands of years. Chemical reactions abound in the geology of Earth, in the atmosphere and oceans, and in a vast array of complicated processes that occur in all living systems.

Chemical reactions must be distinguished from physical changes. Physical changes include chan ges of state, such as ice melting to water and water evaporating to vapour. If a physical cha nge occurs, the physical properties of a substance will change, but its chemical identity wil l remain the same. No matter what its physical state, water (H2O) is the same compound, with each molecule composed of two atoms of hydrogen and one atom of oxygen. However, if water, as ice, liquid, or vapour, encounters sodium metal (Na), the atoms will be redistributed to give the new substances molecular hydrogen (H2) and sodium hydroxide (NaOH). By this, we know that a chemical change or reaction has occurred.

segment number 1 :

Chemical reaction, a process in which one or more substances, the reactants, are converted to one or more different substances, the products. Substances are either chemical elements or compounds.

segment number 2 :

A chemical reaction rearranges the constituent atoms of the reactants to create different substances as products.

Chemical reactions are an integral part of technology, of culture, and indeed of life itself. Burning fuels, smeltin g iron, making glass and pottery, brewing beer, and making wine and cheese are among many examples of activities inco rporating chemical reactions that have been known and used for thousands of years. Chemical reactions abound in the g eology of Earth, in the atmosphere and oceans, and in a vast array of complicated processes that occur in all living systems.

Chemical reactions must be distinguished from physical changes. Physical changes include changes of state, such as i ce melting to water and water evaporating to vapour. If a physical change occurs, the physical properties of a substa nce will change, but its chemical identity will remain the same. No matter what its physical state, water (H2O) is th e same compound, with each molecule composed of two atoms of hydrogen and one atom of oxygen.

segment number 3 :

However, if water, as ice, liquid, or vapour, encounters sodium metal (Na), the atoms will be redistributed to give the new substances molecular hydrogen (H2) and sodium hydroxide (NaOH). By this, we know that a chemical change or re action has occurred.

Figure 3.11: Le résultat d'une segmentation thématique d'un texte en utilisant notre système.

3.6 L'interface de l'application

L'interface de notre application se compose de deux champs de textes, un pour le texte en entrée et l'autre pour le résultat de la segmentation ; et un bouton «**Browse**» pour charger le texte qu'on souhaite le segmenter.

La figure ci-dessous représente l'interface de l'application :

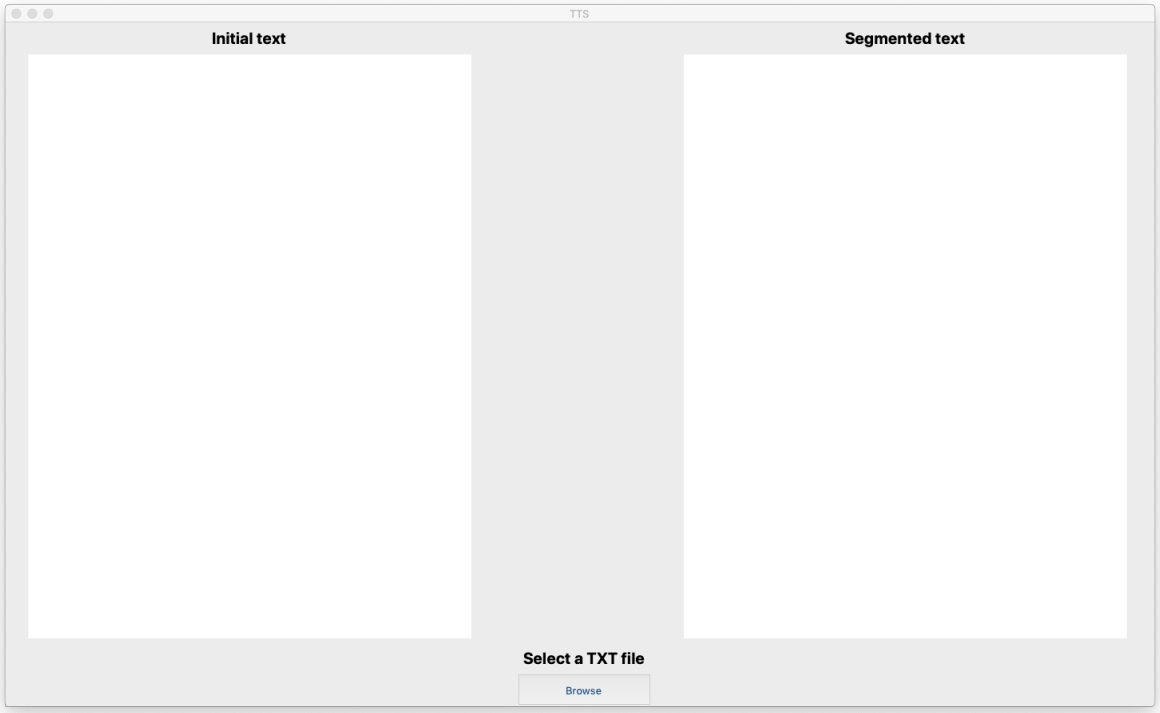

Figure 3.12: L'interface de notre application.

En cliquant sur le bouton «**Browse**» l'application va vous demander de choisir un fichier .txt pour le segmenter.

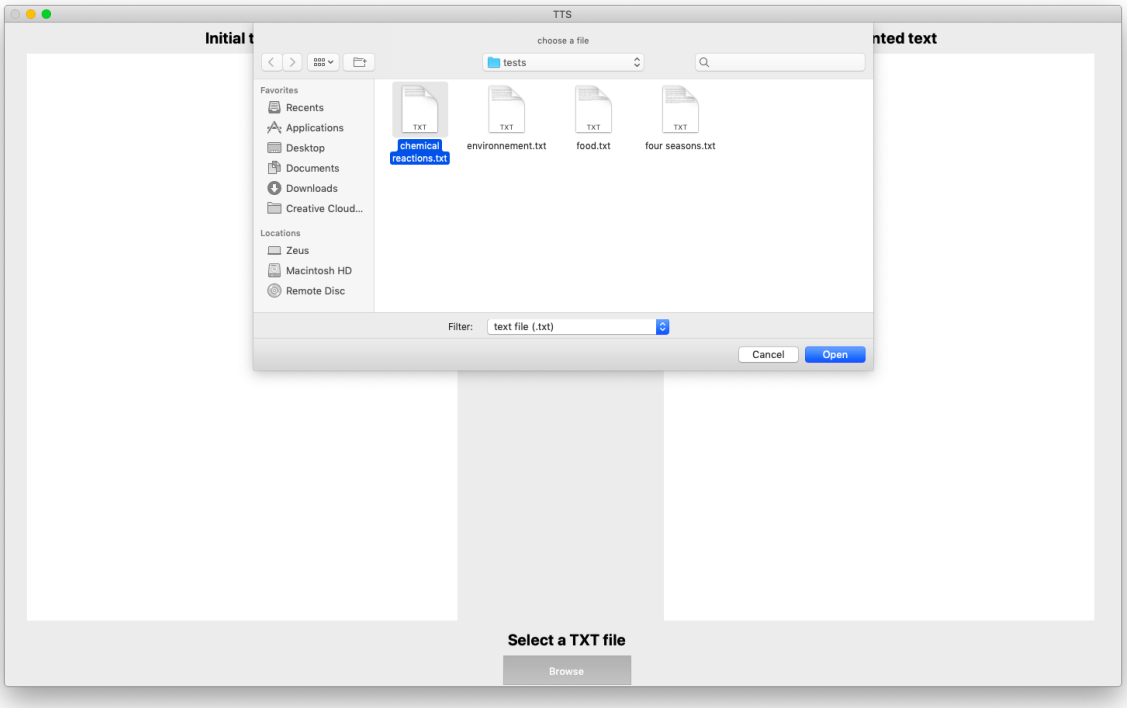

Figure 3.13: L'ouverture d'un fichier .txt sur notre application.

Après que le texte est chargé, il va s'afficher sur le champ du texte à gauche et sa segmentation va s'afficher sur le champ du texte à droite.

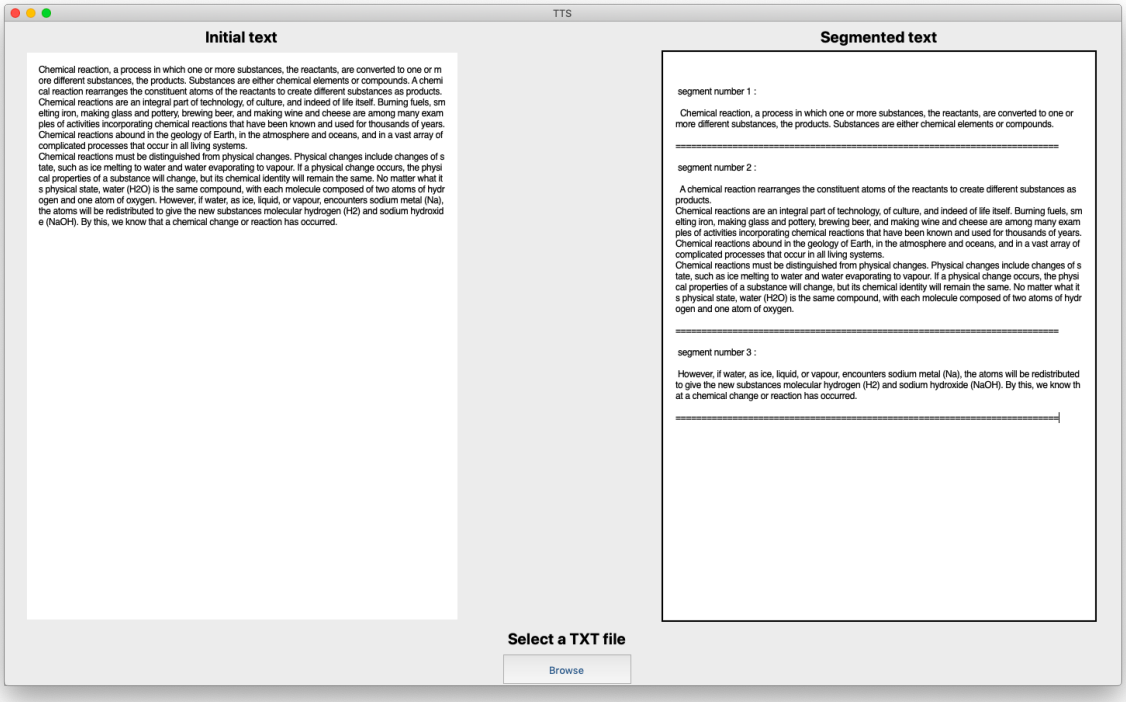

Figure 3.14: Le résultat d'une segmentation avec notre application sur l'interface.

3.7 L'évaluation de notre système de segmentation thématique

L'évaluation de la segmentation thématique est toujours un défi en soi. Pour une évaluation plus fiable, il faut prendre comme une référence pour l'évaluation un corpus de textes pré-segmentés. Cette ressource est difficile à trouver.

3.7.1 La création de corpus

La création d'un corpus pour évaluer la segmentation thématique des textes est un souci majeur à cause du manque des ressources nécessaires. Pour cette raison nous avons créé notre corpus de référence par consensus.
nous avons présenté 4 textes qui traitent des différents sujets sur 7 enseignants d'Anglais dans une école des langues à Ain Temouchent appelée « Odyssey Learning Center », chaque enseignant a segmenté les textes individuellement et nous avons pris la segmentation qui est désignée par la majorité des enseignants comme des textes de référence.

Le tableau 3.1 représente les textes du corpus d'évaluation :

N° texte	**Titre du texte**	**Nombre de mots pleins**	**Nombre des phrases**
1	Chemical reactions	123	12
2	environment	170	15
3	The four seasons	226	15
4	Food	186	11

Table 3.1: Les textes du corpus d'évaluation

3.7.2 Les résultats de l'évaluation de notre système

Nous avon utilisé comme des standards d'évaluation les trois mesure : le rappel, la précision et le score Pk.

Les résultats obtenus sont représentés sur le tableau 3.2 ci-dessous :

N° texte	**Segments de référence**	**Segments trouvés**	**Rappel** (%)	**Précision** (%)	**Score Pk** (%)
1	3, 6, 12	2, 10, 12	33	33	44
2	3, 6, 9, 15	9, 14, 15	25	33	27
3	1, 4, 8, 11, 15	6, 15	20	50	36
4	2, 6, 8, 11	5, 10, 11	25	33	35

Table 3.2: Les résultats de l'évaluation

Le tableau 3.2 représente les valeurs de rappel, précision et le score Pk pour les 4 textes segmentés.

On remarque que les valeurs du rappel et de la précision ne sont pas assez satisfaisantes et qu'elles varient entre 20% et 50%. Ces deux mesures ne sont pas assez fiables parce qu'elles s'intéressent seulement à la comparaison des frontières thématiques obtenues avec celles de la référence. Un ajout ou un manque d'une seule phrase dans un segment obtenu est considéré le même qu'un ajout ou un manque de plusieurs phrases, le segment va être considéré comme « *faux* » dans les deux cas.

Par contre, au niveau du score Pk, nous avons obtenu des résultats plus ou moins satisfaisants qui varient entre 27% (le meilleur score) et 44% (le pire score) comme une probabilité d'erreur. Cette mesure est plus fiable que le rappel et la précision parce qu'elle s'intéresse au contenu des segments en tenant compte de la probabilité pour deux phrases éloignées d'une distance k dans les mêmes segments du document de référence et du document produit.

L'évaluation de notre système reste toujours une tâche difficile à réaliser à cause de l'absence d'un corpus valide et bien conçu qui nous aide à réaliser une évaluation précise et fiable.

3.8 Conclusion

Dans ce chapitre nous avons présenté l'architecture de notre système de la segmentation thématique des textes Anglais ainsi que les différents modules que nous avons développé en détails qui sont le prétraitement, la représentation du texte dont nous avons utilisé l'ontologie WordNet Anglais pour une représentation conceptuelle et enfin l'identification des segments thématiques.

Enfin, nous avons évalué notre système en utilisant un corpus de référence crée par consensus. Les mesures utilisées sont le rappel, la précision et le score Pk.

Conclusion générale

Notre projet consiste à développer un système pour la segmentation thématique des textes Anglais en prenant en considération les relations sémantiques entre les mots.

Notre méthode implémentée est une méthode passive basée sur le calcul de la similarité entre les phrases en utilisant la mesure TF-IDF.

Notre travail est basé sur l'intégration des ontologies dans le processus de la segmentation thématique des textes Anglais afin d'établir une représentation conceptuelle.

La réalisation de notre projet se passe en trois parties principales. La première est consacrée pour la phase de prétraitement qui a pour but de mettre le texte à segmenter sous une forme analysable.

La deuxième partie est le représentation conceptuelle du texte qui est la tâche qui nous permet de créer des concepts en regroupant les mots du texte qui sont sémantiquement similaire dans un seul ensemble nommé « un concept » à l'aide de l'ontologie WordNet Anglais.

La troisième et la dernière partie est la partie d'identification des segments thématiques.

Ce mémoire nous a permis d'intégrer l'ontologie WordNet Anglais comme une ressource sémantique. Il nous a permis aussi d'étudier les caractéristiques, le contenu et l'exploitation de cette ontologie pour la segmentation thématiques des textes Anglais.

L'utilisation de la bibliothèque SpaCy qui est une bibliothèque Python dédiée pour le traitement automatique du langage naturel nous a aidé à realiser plusieurs tâches de prétraitement tels que la lemmatisation, la tokenisaion, l'élimination des mots vides, etc.

Enfin, on a évalué notre système en utilisant des différentes mesures d'évaluation qui sont le rappel, la précision et le score Pk.

Le système de la segmentation thématique des textes Anglais réalisé est plus ou moins satisfaisant, mais ce travail ouvre plusieurs perspectives comme :

-Reconnaitre les mots composés tel que « **United States** ».
-Utiliser un corpus de référence plus riche pour l'évaluation et qui est fait par des linguistes experts.
-Développer une interface graphique plus moderne.
-Ajouter d'autre fonctions comme le stockage de chaque segment dans un fichier ou dans une base de données pour une meilleur exploitation.

Bibliographie

[1] M. Shafiei et E. Evangelos, « Statistical Model for Topic Segmentation and Clustering », Canadian AI 2008, LNAI 5032, pp. 283–295, 2008.

[2] A. Labadié et V. Prince, « Comparaison de méthodes lexicales et syntaxico- sémantiques dans la segmentation thématique de texte non supervisée », TALN 2008, Avignon, juin 2008.

[3] M. CHAROLLES, "Les études sur la cohérence, la cohésion et la connexité textuelles depuis la fin des années 1960", Modèles Linguistiques, X, 2, 45-66. , 1988.

[4] T. Ouerfelli, « La segmentation des documents techniques composites dans une perspective d'indexation : vers la définition d'un modèle dans une optique d'automatisation », Thèse de Doctorat en sciences de l'information et de la communication, Université Stendhal de Grenoble, 2001.

[5] T. OUERFELLI, « La Segmentation des documents techniques en amont de l'indexation : définition d'un modèle », Institut Supérieur de Documentation Campus Universitaire de la Manouba Tunisie, 2010.

[6] A. Labadié, « Intended boundaries detection in topic change tracking for text segmentation », Int J Speech Technol, 2008.

[7] R. Radim, « Text Segmentation Using Context Overlap », EPIA, 2007.

[8] J. Morris et G. Hirst, « Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text », Computational Linguistics, 17(1), pp. 21–48, 1991.

[9] J. Reynar , « Topic Segmentation : Algorithms And Applications », Ph.D. thesis, Computer and Information Science, University of Pennsylvania, 1998.

[10] N. Stokes, et Al., «Segmenting Broadcast News Streams using Lexical Chains », In Proceedings of STAIRS, 2002.

[11] M. Halliday et R. Hasan, « Cohesion in English », Longman, New York, 1976.

[12] J. Vasak et F. Song, « Word Distribution Based Methods for Minimizing Segment Overlaps », LNAI 4629, pp. 147-154, 2007.

[13] G. Salton, A. Wong, et C. S. Yang, « Model for Automatic Indexing : A Vector Space », ACM, pp 613–620, 1975.

[14] J. Barbara et L. Sidner, «Attention, Intentions, And The Structure Of Discourse », Computational Linguistics,12(1986), pp. 175–204. July- September, 1986.

[15] A. LABADIE, « Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Francais », Thèse présentée à l'Université des Sciences et Techniques du Languedoc pour obtenir le diplome de doctorat, UMII 2009.

[16] P. Buitelaar, « CoreLex : Systematic Polysemy and Underspecification », PhD Thesis, Computer Science, Brandeis University, February, 1998.

[17] J. Reynar , « Topic Segmentation : Algorithms And Applications », Ph.D. thesis, Computer and Information Science, University of Pennsylvania, 1998.

[18] F. Choi, « Advances in domain independent linear text segmentation », NAACL'00, p. 26-33, 2000.

[19] Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, et Frederic Saubion. ń Using an Evolving Thematic Clustering in a Text Segmentation Process ż. Universal Computer Science, pp 178–192, 2008.

[20] O. Ferret, « How to thematically segment texts by using lexical cohesion ? », ACL-COLING'98, p. 1481-1483, 1998.

[21] Olivier Ferret, Brigitte Grau, Jean-Luc Minel, et Sylvie Porhiel. ń Repérage de structures thématiques dans des textes ż. TALN2001, 2001.

[22] H. Kozima, « Text segmentation based on similarity between words », 31th annual meeting of the ACL, pp 286–288, 1993.

[23] M. Kan et Al., « Linear Segmentation and Segment Significance », Proceedings of WVLC-6, pp 197–205, 1998.

[24] Laurianne Sitbon. ń Fusion d'approches non supervisées et génériques pour la segmentation thématique ż. 2004.

[25] Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, et Frederic Saubion. ń SegGen : a Genetic Algorithm for Linear Text Segmentation ż. Proceedings of IJCAI'07, 2007.

[26] R. Baeza-Yates et B. Ribeiro-Neto, « Modern Information Retrieval », Addison-Wesley, 1999.

[27] L. Sitbon et P. Bellot, «Adapting and comparing linear segmentation methods for French », In Proceedings RIAO'04, Avignon, France, 2004.

[28] D. Beeferman et Al., « Text segmentation using exponential models », In Proceedings of the 2nd conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Somerset, New Jersey, USA, 1997.

[29] L. Pevzner et M. Hearst, « A critique and improvement of an evaluation metric for text segmentation », Computational Linguistics, p. 19–36, 2002.

[30] T. Gruber, “A translation approch to portable ontology specifications”,Knowledge Acquisition, 199-220, 1993.

[31] C. OGDEN et I. RICHARDS, « The Meaning of Meaning : A Study of the Influence of Language upon Thought and of the Science of Symbolism », New York, Harcourt, 1946.

[32] N. Guarino, « Understanding, building and using ontologies». International J. Human-Computer Studies, pp 293-310, 1997.

[33] http ://wordnet.princeton.edu/wordnet/

[34] M. Nagata et Al., « Using Goi-Taikei as an Upper Ontology to Build a Large- Scale Japanese Ontology from Wikpedia », Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pages 11–18, Beijing, August, 2010.

[35] Synthèse automatique de texte avec Machine Learning - Un aperçu. (2020, 26 novembre). ICHI.PRO. https ://ichi.pro/fr/synthese-automatique-de-texte-avec-machine-learning-un-apercu-115306273012261

[36] Fabien, M. (2019, 3 novembre). Traitement Automatique du Langage Naturel en français (TAL / NLP). Stat4decision. https ://www.stat4decision.com/fr/traitement-langage-naturel-francais-tal-nlp/

[37] Pykes, K. (2020, 25 novembre). Part Of Speech Tagging for Beginners - Towards Data Science. Medium. https ://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba

[38] Bachani, N. (2021, 31 mai). Chunking in NLP : decoded - Towards Data Science. Medium. https ://towardsdatascience.com/chunking-in-nlp-decoded-b4a71b2b4e24

[39] Balodi, T. (2020, 15 juillet). What is Stemming and Lemmatization in NLP ? | Analytics Steps. Analytics Steps. https ://www.analyticssteps.com/blogs/what-stemming-and-lemmatization-nlp

Résumé

La segmentation thématique des textes (STT) est une tâche importante dans plusieurs applications du traitement automatique du langage naturel (TALN), tels que la recherche d'information (RI), le résumé automatique des textes et les système questions-réponses. La STT consiste à diviser les textes en segments, chaque segment correspond à un thème différent.
Les méthodes principalement utilisées sont des méthodes probabilistes (statistiques) ; Ce qui provoque parfois l'échec du système du STT par l'ambiguïté sémantique et l'impossibilité d'identifier les relations sémantiques entre les mots.
Notre travail à pour objectif de developper un segmenteur thématique des textes Anglais basé sur une représentation conceptuelle. Pour cela, on a intégré l'ontologie WordNet Anglais comme une ressource sémantique.

Mots clés : Intelligence artificielle, TALN Anglais, Fouille de texte Anglais, Segmentation thématique, Ontologies, WordNet Anglais, Python, SpaCy.

Abstract

The Topical Text Segmentation (TTS) is an important task in many natural language processing (NLP) applications, such as information retrieval (IR), automatic text summarization and question answering. It consists of dividing the texts into segments, each segment corresponds to a different topic.
The methods used are mainly probabilistic (statistical), which sometimes leads to the failure of the TTS system due to the semantic ambiguity and the impossibility to identify the semantic relations between the words.
Our task is to develop an english thematic text segmenter based on a conceptuel representation. That is why we have integrated the ontology WordNet English as a semantic resource.

Keywords : Artificial Intelligence, English NLP, English Text Mining, Topic Segmentation, Ontologies, WordNet English, Python, SpaCy.

ملخص

التجزئة الموضوعية للنصوص هامة في العديد من تطبيقات المعالجة الآلية للغات مثل البحث عن المعلومات، التلخيص الآلي للنصوص نظام سؤال جواب. التجزئة الموضوعية للنصوص تتمثل في تجزئة النصوص إلى أجزاء، كل جزء يتوافق مع موضوع مختلف.
الأساليب المعتمدة غالبا ما تكون مرتكزة أساسا على الاحتمالات (الاحصائيات)، هذا ما ينجم عنه فشل أنظمة التجزئة الموضوعية للنصوص بسبب الغموض الدلالي و استحالة تحديد العلاقات الدلالية بين الكلمات.
عملنا يهدف إلى تطوير نظام تجزئة موضوعية للنصوص الانجليزية على أساس التمثيل المفاهيمي. لهذا، دمجنا الأنطولوجيا ووردنيت للغة الانجليزية كمورد دلالي.

كلمات مفتاحية: الذكاء الاصطناعي، المعالجة الآلية للغة الانجليزية، تنقيب النصوص الانجليزية، التجزئة الموضوعية، الأنطولوجيات، ووردنيت للغة الانجليزية، بايثون، سبيسي.