MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE
LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et de l'Informatique
Département de Mathématiques et d'Informatique
Filière : Informatique

MÉMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master en Informatique
Option : **Ingénierie des Systèmes d'Information**

THÈME :

Les Résumés Automatiques des Documents Textuels

Etudiant(e) : **Boudraf Khadidja**

Encadrant(e) : Mme : Maghni Sandid Zoulikha

Année Universitaire 2016/2017

Résumé

La forte augmentation de texte disponible en format numérique a fait ressortir la nécessité de concevoir et de développer des outils de résumé performants dans le but de repérer et extraire l'information pertinente sous une forme abrégée.

Ce mémoire propose une méthode de production de résumés pour les documents textuels.

Notre démarche méthodologique consistait à étudier : les caractéristiques du résumé classique, les étapes les techniques utilisées dans le résumé automatique, la présentation de l'environnement de développement en détaillant les différents outils utilisés et expliquons notre approche proposé, ainsi la présentation de l'architecture de notre application, l'explication du déroulement de l'application, puis l'évaluation de notre système avec d'autres systèmes de référence.

L'objectif de cette étude fut de produire le bon et le mieux résumé automatiquement.

Mot clés : Résumé, Résumé automatique, Résumé mono-document, Résumé multi-document, Evaluation du Résumé automatique.

Abstract

The large increase of text available in digital format has highlighted the need to design and develop effective summary tools in order to locate and extract relevant information in an abbreviated form.

This paper proposes a method of producing abstracts for textual documents.

Our methodological approach consisted of studying: the characteristics of the classical abstract, the techniques used in the automatic abstract, the presentation of the development environment, detailing the different tools used and explaining our proposed approach, and the presentation of the architecture Of our application, the explanation of the implementation of the application, and then the evaluation of our system with other reference systems.

The objective of this study was to produce the right and best summarized automatically.

Keywords: Summary, Automatic abstract, Summary mono-document, Summary multi-document, evaluation of the automatic abstract.

Dédicaces

Il m'est agréable de profiter de cette occasion, pour rendre un hommage particulièrement sincère à travers ce modeste travail, à tous ceux qui me sont chers, à tous ceux qui m'ont soutenu moralement et matériellement.

Je dédie donc ce modeste travail:

A mes très chers et honorables parents ainsi que toute ma famille.

À tous mes enseignants de la faculté

À tous mes chères amies :

"Leila, Asmaà, Karima, Hafsa, Meriem, Kheira,

Nour el Imane, Naima"

Et à tous personne qui a contribué à la réussite de ce projet que ce soit de près ou de loin.

B.Khadidja

Remerciement :

Je tiens à remercier avant tout, Dieu de nous a prodiguée la force morale et physique et nous a permis d'achever ce travail.

J'adresse mes remerciements aux personnes qui m'ont aidé dans la réalisation de ce mémoire.

En premier lieu, je remercie **Mme. Maghni Sandid Zoulikha** *En tant que encadrante de mémoire, il m'a guidé dans mon travail et m'a aidé à trouver des solutions pour avancer.*

J'adresse mes plus sincères remerciements aux membres du jury d'avoir accepté d'examiner ce modeste travail.

Mes remerciements vont à tous les enseignants du département d'informatique que nous respectons beaucoup.

Enfin, je souhaiterais adresser des remerciements plus particuliers à toute ma famille.

Sommaire

CHAPITRE II Résumé automatique : état de l'art

Chapitre III : Conception et Approche Proposée

Chapitre IV : Implémentation et mise en œuvre

Liste des tableaux

Liste des figures

Abréviation

CDDL: Common Development and Distribution License

DUC: Document **U**nderstanding **C**onferences

RI: Recherche d'**I**nformation

ROUGE: Recall-**O**riented **U**nderstudy for **G**isting **E**valuation

ROUGEN: Recall-**O**riented **U**nderstudy for **G**isting **E**valuation de **N6-**gammes

SRI : Système de **R**echerche d'**I**nformation

TAL : Traitement **A**utomatique des **L**angues

TF.IDF: Terme **F**requency and **I**nverse **D**ocument **F**requency

TAC: Text **A**nalysis **C**onference

Introduction Générale

L'information textuelle sous forme de document numérique s'accumule rapidement et en très grandes quantités. L'immense quantité des documents est, dans la plupart de cas, non structurée : elle ne se trouve pas sous forme de base de données classique, mais sous un format de texte libre. Ces documents sont ainsi traités d'une façon très sommaire. En conséquence les taches d'analyse automatique de document deviennent extrêmement difficiles à mettre en œuvre. Le résumé automatique de documents, en condensant les textes de façon pertinente, peut aider à traiter cette masse grandissante d'information difficile à absorber. [1]

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations à travers la construction d'index. L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information. On peut aujourd'hui dire que la recherche d'information est un champ transdisciplinaire qui peut être étudié par plusieurs disciplines utilisant des approches qui devraient permettre de trouver des solutions pour améliorer son efficacité.

L'opérationnalisation de la **RI** est réalisée par des outils informatiques appelés Systèmes de Recherche d'Informations (**SRI**), ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation du contenu des documents (fiche ou enregistrement) au moyen d'une fonction de comparaison (ou de correspondance). [2]

Le traitement automatique du langage (**TAL**) est un domaine à la fois scientifique et technologique en plein essor qui débouche sur des applications très diverses : correction automatique des erreurs, analyse de textes, génération automatique de résumés, aide à la traduction, etc. En outre, la nécessité d'application de **TAL** s'avère de plus en plus indispensable avec l'explosion d'Internet où le langage humain reste un vecteur d'information prépondérant. [3]

Le résumé est la forme la plus utilisable de réduction d'un document. Dans le cas d'un livre, le sommaire et sa table de matières sont d'autres représentations condensées de ce document. Mais, qu'est-ce exactement qu'un résumé de texte ? Il est possible de trouver plusieurs définitions de sujet dans la littérature. Une d'entre elle dit que le résumé d'un document est une représentation réduite mais exacte qui cherche à rendre une idée précise de son contenu. Il a comme objectif principal celui de renseigner et de fournir un accès privilégié aux documents source. Le résumé devient automatique s'il est généré par un logiciel ou un système informatique.

Les premiers travaux sur les résumés automatiques de textes datent des années 50. Pour extraire les phrases pertinentes nécessaires à la construction d'un résumé, considère des caractéristiques comme la fréquence d'occurrence des termes, des mots de titres la longueur et la position de la phrase.

Avec l'avènement de l'Internet et de moteurs de recherche de plus en plus performants, l'importance d'informations condensées du type résumé est devenue nécessaire pour faire ressortir l'information pertinente. De ce fait le résumé automatique a inspiré de nouvelles orientations, plusieurs nouvelles approches ont commencé à être explorées en linguistique (basée sur l'analyse du discours et de sa structure) et en statistique (basée sur la distribution des occurrences des mots). [4]

Le text mining est un processus d'extraction de structures (connaissances) inconnues, valides et potentiellement exploitables dans les documents textuels, à travers la mise en œuvre de techniques statistiques ou de machine learning. Mais d'autres applications spécifiques aux textes sont possibles : résumé automatique, extraction d'information, etc. [5]

L'extraction d'information consiste à recherche des champs prédéfinis dans un texte plus ou moins rédigé en langage naturel. On s'appuie plus sur l'analyse lexicale et morphosyntaxique pour identifier les zones d'intérêts. [5]

Résumé d'un document. Recherche des phrases les plus représentatives dans un document. Résumé d'un corpus. Recherche du document le plus représentatif dans un corpus ou recherche des phrases représentatives à partir de plusieurs documents. [5]

Notre travail étape s'appuie sur l'extraction des phrases les plus pertinents à partir du calcul du score de chaque phrase et le trier de score le plus élevé au score le plus bas et retourner le résultat final suivant le choix du nombre des phrases extraites par rapport au nombre de phrases contenues dans le document.

Notre rapport est organisé en quatre chapitres incluant une introduction et une conclusion générale.

Dans le premier chapitre, nous présentons les mécanismes de la **RI**, les mécanismes du **TAL**, et une petite description de la relation entre le **TAL** et le résumé automatique et consistons à présenter le résumé classique d'une façon générale.

Le deuxième chapitre, consiste à étudier l'état de l'art du résumé automatique des documents textuels tout en présentant un ensemble des méthodes et mesures pour développer un système du résumé automatique.

Le troisième chapitre, consiste à aborder les aspects de développement de notre solution ensuite la présentation de l'environnement de développement, en détaillant les différents outils et méthodes utilisés.

Le dernier chapitre, consiste à expliquer l'architecture, le fonctionnement et le déroulement de notre application en détaillant les différentes étapes de l'implémentation ainsi les résultats obtenus.

Chapitre I
La Construction du Système de Résumé Automatique

I.1. Introduction

La Recherche d'Information (**RI**) peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur. La **RI** est réalisée par des Systèmes de Recherche d'Information (**SRI**) qui ont un but pour retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager. [2]

Le **TAL** est l'ensemble des méthodes et des programmes qui permettent un traitement par l'ordinateur des données langagières, mais quand ce traitement tient compte des spécificités du langage humain. Il y a des traitements de données langagières (écritures sur fichiers, sauvegardes ou autres) qui ne font pas partie du traitement automatique des langues. [3]

Résumer un texte consiste à réduire ce texte en un nombre limité de mots. Le texte ainsi réduit doit rester fidèle aux informations et idées du texte original, et dans la mesure du possible rendre compte du style et de l'intention de l'auteur. Cette discipline, quoique très ancienne, est mal formalisée. Le processus de résumé est en effet dépendant à la fois du type de texte à résumer et de l'utilisation qui en sera faite. [4]

Nous présentons dans ce chapitre les concepts clés de la recherche d'information et du **TAL**, la relation entre le **TAL** et le résumé automatique ainsi une représentation générale du résumé classique.

I.2. Recherche d'information

La recherche d'information (**RI**) suscite depuis fort longtemps l'attention de la communauté scientifique, elle se définit généralement par l'identification de documents qui satisfont le mieux le besoin de l'utilisateur. Ces documents doivent être trouvés parmi une large collection de documents non structurés. [6]

2.1. Définitions :

Plusieurs définitions de la recherche d'information ont vu le jour dans ces dernières années, nous citons dans ce contexte les trois définitions suivantes : [2]

Définition 1 : La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations.

Définition 2 : La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information.

Définition 3 : La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information.

2.2. Système de recherche d'information:

Un système de recherche d'information **(SRI)** est défini par un langage de représentation des documents (qui peut s'appliquer à différents corpus de documents) et des requêtes qui expriment un besoin de l'utilisateur (sous forme de mots-clés par exemple), et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information .[2]

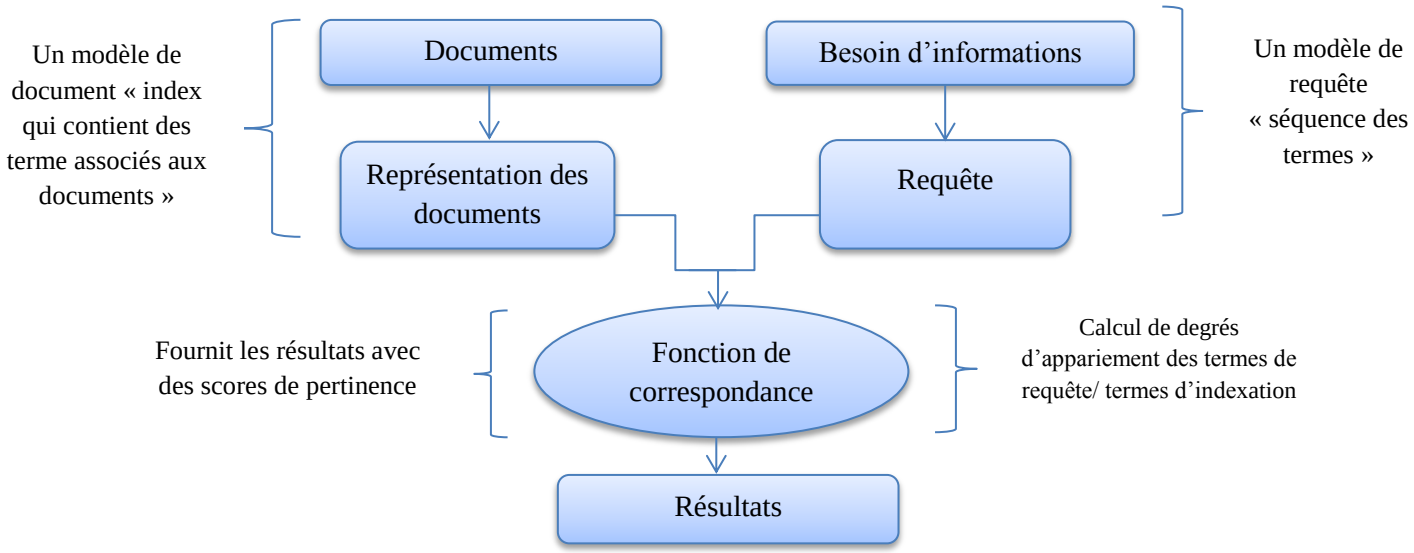

Figure 1.1 : Système de recherche d'information [2]

2.3. Processus de recherche d'information :

Un système de recherche d'information **(SRI)** manipule un corpus de documents qu'il transpose à l'aide d'une fonction d'indexation en un corpus indexé. Ce corpus lui permet de résoudre des requêtes traduites à partir de besoins utilisateur. Un tel système repose sur la définition d'un modèle de recherche d'information qui effectue ces deux transpositions et qui fait correspondre les documents aux requêtes. La transposition d'un document en un document indexé repose sur un modèle de document. De même, la transformation du besoin utilisateur en requête repose sur un modèle de requête. Enfin, la correspondance entre une requête et des documents s'établit par une relation de pertinence, la figure 1.2, présente les différentes étapes d'un processus de recherche d'information. [2]

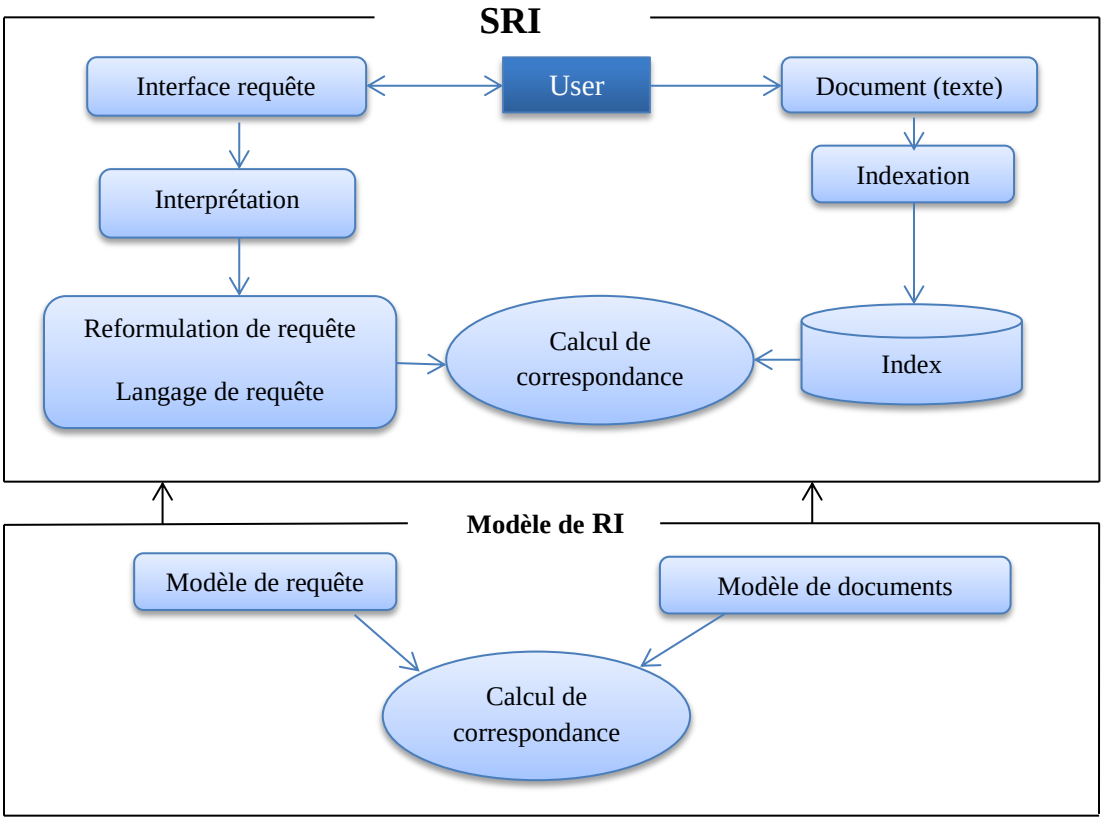

Figure 1.2 : Processus de recherche d'information [2]

2.3.1. Indexation :

L'indexation consiste à extraire des documents les mots les plus discriminants encore appelés index. Cette première tâche est généralement effectuée en marge du processus de recherche car, la construction des index peut être assez longue en fonction du nombre de documents de la collection ainsi que de la taille des documents. Les index ont un caractère réducteur car tous les termes d'un document ne sont pas importants à prendre en compte pour la recherche. L'indexation peut se faire de trois manières différentes : manuellement (faite par un humain), de manière semi-automatique (créée par un humain assisté d'un programme proposant des termes), ou de manière automatique (créée par un programme informatique). [2]

2.3.2. Interrogation :

Il s'agit de l'expression du besoin d'information de l'utilisateur dans la forme imposée par le système, la recherche dans le corpus, et la présentation des résultats. Cette phase nécessite un modèle de représentation du besoin de l'utilisateur, appelé modèle de requêtes, ainsi qu'une fonction de correspondance qui doit évaluer la pertinence des documents par rapport à la requête. La réponse du système est un ensemble de références à des documents qui obtiennent une valeur de correspondance élevée. [2]

2.3.3. Fonction de correspondance :

Tout système de recherche d'information s'appuie sur un modèle de recherche d'information. Ce modèle se base sur une fonction de correspondance qui met en relation les termes d'un document avec ceux d'une requête en établissant une relation d'égalité entre ces termes. Cette relation d'égalité représente la base de la fonction de correspondance et, par la même, du système de recherche d'information. [2]

2.4. Evaluation des SRI :

Les systèmes de **RI** sont toujours évalués en fonction de la pertinence des documents retrouvés. Afin de procéder à des évaluations automatiques, nous avons besoin de corpus de test « standard ». Chaque corpus contient :

- L'ensemble de documents.
- L'ensemble de requêtes de test sur l'ensemble de documents du même corpus.
- La liste de documents pertinents pour chaque requête.

Un système de **RI** quelconque peut utiliser ce corpus pour trouver des documents pour les requêtes données, et nous pouvons comparer ces documents retrouvés avec la liste de documents pertinents pour évaluer la qualité du système. [7]

Les deux principales mesures utilisées pour évaluer un système de RI sont la précision et le rappel. [7]

Précision = *nombre total de documents pertinents retrouvés par le système / nombre total de documents retrouvés par le système.*

Rappel = *nombre total de documents pertinents retrouvés par le système / nombre total de documents pertinents dans le corpus*

2.5. Les modèles de recherche d'information :

Il existe un certain nombre de modèles théoriques dans la littérature les plus connus étant le « Modèle Booléen», le «Modèle Vectoriel», et le «Modèle Probabiliste ». Dans le modèle booléen, les requêtes sont représentées sous forme de termes reliés par des opérateurs booléens (ET, OU, NON, . . .). Le modèle vectoriel considère les documents et les requêtes comme des vecteurs pondérés, chaque élément du vecteur représentant le poids d'un terme dans la requête ou le document. Le modèle probabiliste tente d'estimer la probabilité qu'un document donné soit pertinent pour une requête donnée. [2]

2.5.1. Définition :

Un modèle de **RI** a pour rôle de fournir une formalisation du processus de **RI** et un cadre théorique pour la modélisation de la mesure de pertinence. Ce modèle a en commun le vocabulaire d'indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d'appariement requête-document. Le vocabulaire d'indexation $V = \{ti\}, i \in \{1, ..., n\}$ est constitué de n mots ou racines de mots qui apparaissent dans les documents.

Un modèle de **RI** est défini par un quadruplet (D, Q, F, R (q, d)) : où

– D est l'ensemble de documents.

– Q est l'ensemble de requêtes.

– F est le schéma du modèle théorique de représentation des documents et des requêtes.

– R (q, d) est la fonction de pertinence du document d à la requête q. Parmi les modèles de la **RI**, nous présentons le modèle vectoriel [2], alors :

2.5.2. Le modèle vectoriel :

Dans ce modèle, la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel représente les documents et les requêtes par des vecteurs d'un espace à *n* dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation. L'index d'un document *dj* est le vecteur $\vec{d} = (w_{1,j}, w_{2,j}, w_{3,j}, .., w_{n,j})$, où $w_{k,j} \in [0\ 1]$ dénote le poids du terme t_k dans le document *dj*. Une requête est également représentée par un vecteur $\vec{q} = (w_{1,q}, w_{2,q}, w_{3,q}, .., w_{n,q})$,où $w_{k,q}$ est le poids du terme t_k dans la requête q. La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents. Une mesure classique utilisée dans le modèle vectoriel est le cosinus de l'angle formé par les deux vecteurs :

$$\mathbf{RSV\ (q, d) = COS\ \vec{q}\vec{d}}$$

Plus les vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. La fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante. [2]

2.5.3. La mesure TF-IDF :

La TF-IDF suit la logique du Cosinus de SALTON. On cherche à accorder une pertinence lexicale à un terme au sein d'un document. En ce qui concerne **TF-IDF**, on applique une relation entre un document, et un ensemble de documents partageant des similarités en matière de mots clés. On recherche en quelque sorte une relation de quantité / qualité lexicale à travers un ensemble de documents.

Pour une requête avec un terme X, un document a plus de chances d'être pertinent comme réponse à la requête, si ce document possède une certaine occurrence de ce terme en son sein, et que ce terme possède une rareté dans d'autres documents reliés au premier. [7]

$Tf_{i,j}$: le nombre du terme i dans le document d_j.

Df_i : le nombre des documents dans le corpus du terme t_i.

Idf_j : l'inverse de la fréquence documentaire mesure l'importance d'un terme dans l'ensemble de la collection.

TF-IDF

Poids du terme ti dans le document dj → $W_{ij} = tf_{ij} * idf_j$

Poids du terme ti dans la requête qk → $W_{ik} = idf_i$

Où :

$Idf_j = \log (N/df_i)$ avec **N** : le nombre total des documents

$$RSV(D_j, Q) = \cos(D_j, Q) = \frac{\vec{D}\,\vec{Q}}{\|D\|\,\|Q\|}$$

On a : $\vec{D}\,\vec{Q} = \sum w_{ij} * w_{ik}$ tel que **i = 1..N**

Et : $\|D\|\,\|Q\| = \sqrt{((\sum w_{ij})^2 * (\sum w_{ik})^2)}$ **tel** que **k = 1..N**

$$RSV(Dj, Q) = \frac{\sum w_{ij} * w_{ik}}{\sqrt{((\sum w_{ij})^2 * (\sum w_{ik})^2)}}$$

I.3. Traitement automatique des langues

Le Traitements Automatique des Langues (**TAL**) est une discipline qui associe étroitement linguistes et informaticiens. Il repose sur la linguistique, les formalismes (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'informatique. Le **TAL** a pour objectif de développer des logiciels ou des programmes informatiques capables de traiter de façon automatique des données linguistiques. Pour traiter automatiquement ces données, il faut d'abord expliciter les règles de la langue puis les représenter dans des formalismes opératoires et calculables et enfin les implémenter à l'aide de programmes informatiques. [3]

I.4. Résumé automatique des textes et TAL

L'élaboration de systèmes plus performants passe donc par le détour de recherches fondamentales, en matière notamment de **compréhension de textes** et de **génération de textes**, le traitement de la langue porte non seulement sur les formes, mais aussi sur le

contenu ; il doit mettre en œuvre des connaissances linguistiques très complètes ainsi que des connaissances d'univers. [8]

- **Compréhension automatique des textes :**

Comprendre un texte, c'est en effet, par-delà le simple décodage du contenu littéral de ce qui et dit phrase après phrase, être capable de relier les phrases entre elles de façon à reconstruire un tout signifiant et cohérent, et être capable d'interpréter le message reçu par rapport à la situation et aux conditions d'énonciation.

L'élaboration de systèmes de compréhension automatique de textes écrits se heurte à deux problèmes liés, d'une part aux relations inter-phrastiques, et d'autre part au contexte, mais cela n'empêche pas, qu'aujourd'hui on est capable de traiter des aspects limités du sens d'un texte quelconque, de manière à pouvoir par exemple l'indexer correctement dans une base de données. Par ailleurs, on peut aller plus loin pour obtenir une compréhension plus profonde, mais cela serait relatif à un domaine très précis. [8]

- **Génération automatique de textes :**

Ce n'est qu'au début des années 80 que le problème de génération de textes, dans un acte de communication donné, est abordé. Dans ce cadre, le texte généré par la machine doit satisfaire des exigences : d'une part indiquer à l'utilisateur les informations qu'il désire, et d'autre part offrir une formulation de ces informations dans une langue correcte.

Il s'en suit que le processus de génération comporte deux composants : le premier (système expert de raisonnement) traite la question "quoi dire ?" (Détermination du contenu informatif), le second (module de génération linguistique) traite la question "comment le dire?" (Formulation du contenu informatif dans une langue correcte). [8]

I.5. Le résumé

Avant de parler de résumé automatique, on peut peut-être déjà identifier ce qu'est le résumé de manière classique. Au départ, le résumé, c'est quand un résumeur professionnel - donc un humain - prend un texte et en dégage les idées essentielles pour en faire un texte plus court. Le résumé est donc un autre texte, plus court et censé dégager les idées saillantes qui étaient présentes dans le texte initial.

5.1.Définitions :

Définition 1 : Présentation abrégée, orale ou écrite, qui rend compte de l'essentiel.

Définition 2 : Consiste à réduire un texte en gardant les idées essentielles.

Définition 3 : Texte court qui récapitule à des fins didactiques ce qu'il fait retenir d'un sujet, d'un livre, d'un auteur.

Définition 4 : C'est une réduction d'un document.

Définition 5 : Présentation réduite mais exacte et précise.

5.2. Pourquoi des résumés?

A mon avis, on fait le résumé car nous n'avons pas le temps de tout lire, et pour se rappeler que l'essentiel.

5.3.Stratégie du résumé : [9]

On lit l'original attentivement, avec un dictionnaire pour vérifier la compréhension, et on fait une première hypothèse sur le thème, l'idée principale du texte.

- Si on veut, on peut faire alors une carte sémantique, qui représente l'idée centrale au milieu, avec d'autres idées réparties autour, regroupées selon leurs affinités et leur importance.
- On marque les divisions en parties et les rapports entre ces parties. (Le rapport peut être que la deuxième partie est un exemple, ou qu'elle s'oppose à la première partie, ou que telle partie est une conséquence d'une idée exprimée, ou simplement un deuxième argument ou une deuxième idée complémentaire.)
- On met les exemples entre crochets, pour se rappeler qu'ils sont secondaires.
- On souligne les **mots clés**, ceux qui portent les **idées clés**.
- On **reformule** les idées clés (en utilisant des **synonymes,** des **simplifications** et des **réductions**).
- Pour chaque paragraphe ou division, on prépare une phrase qui énonce l'idée principale en utilisant les reformulations.
- On se relit, en vérifiant que chaque élément du résumé est bien dans le texte.
- On se relit, en vérifiant l'orthographe, les constructions, les accords.
- Éventuellement, pour un texte un peu long, on reformule l'idée principale du texte qu'on utilise pour lui donner un titre ou une phrase introductive.
- On compte les mots du résumé et on met le total à la fin.

5.4. Les types du résumé :

Il existe plusieurs types de résumé selon leur longueur, leur style et leur subjectivité :

- **Résumé indicatif (indicative abstract) :**
  - Il fait moins de 100 mots.
  - Il signale le ou les thèmes du document.
  - Il est utilisé en particulier pour les documents trop courts (abstract), trop détaillé (thèse,...) ou impropre au résumé informatif (synthèse bibliographique, dictionnaire,...)
  - Son style peut être télégraphique.
- **Résumé informatif (informative abstract) :**
  - Il fait entre 100 et 250 mots.

- Il renseigne sur les informations quantitatives et qualitatives essentielles contenues dans le document.
- Les informations sont présentées dans l'ordre du document, mais leur importance relative peut différer de celle du document (les informations originales étant bien sur plus développées).

o **Résumé sélectif (selective abstract) :**

Ce résumé ne retient du document que les éléments nécessaires à une catégorie particulière d'utilisateurs (ceux du centre de documentation, ou de la base).

o **Résumé critique (critical abstract) :**
- Il s'agit d'un résumé descriptif, assorti d'une critique originale du document.
- Il ne peut être rédigé que par un spécialiste de la question. [10]

Il est souvent utile, pour bien assimiler le sens d'un texte, de préparer un **résumé synthétique** après avoir fait le **résumé ou le compte rendu analytique** dont il s'agit ci-dessus : [9]

- **Résumé analytique :**
  - On suit le texte pas à pas.
  - On inclut chaque idée importante.
  - On ne juge pas.
  - On réduit à un quart de la longueur.
- **Résumé synthétique :**
  - On caractérise le texte globalement, réorganisant les données si nécessaire.
  - On peut émettre des jugements de valeur.
  - On rassemble les idées du résumé analytique, pour n'en retenir que les caractères généraux.
  - On réduit à une ou deux phrases normalement.

I.6. Conclusion

Résumer nécessite une connaissance complète du contenu du document, donc une lecture attentive et longue. Le processus de synthèse et de reformulation demande également du temps. C'est donc une provision que la plupart des services ne peuvent se permettre.

Grâce aux progrès de la scannérisation et de la reconnaissance de caractère, ainsi qu'à la généralisation des logiciels d'indexation et de recherche, le texte intégral vient concurrencer désormais le résumé.

Chapitre II
Le Résumé Automatique : état de l'Art

II.1. Introduction

Le résumé automatique se propose de faire une extraction de l'information jugée importante d'un texte d'entrée pour construire, à partir de cette information, un nouveau texte de sortie, condensé. Ce nouveau texte permet d'éviter la lecture en entier du document source.

Nous présentons dans ce chapitre l'état de l'art du résumé automatique ainsi une présentation de leurs différentes étapes et techniques.

II.2. Le Résumé Automatique

Le résumé automatique c'est de faire par une machine la tache faite par un humain résumeur, et aussi c'est un résumé qui est généré par un logiciel ou un système d'information, à partir d'un texte source on produit un texte court.

Le résumé automatique de document est un processus de compréhension avec perte d'informations, à la différence des méthodes et logiciels de compression du texte.

2.1. Pourquoi le résumé automatique ?

Le but d'un résumé automatique de texte est de produire une représentation abrégée d'un ou de plusieurs documents. Il peut aider à traiter de façon efficace cette masse grandissante d'informations que les personnes s'avèrent tout simplement incapables d'absorber.

2.2. Approches du résumé automatique :

Il existe deux approches en matière de résumé automatique : l'approche par compréhension et l'approche par extraction : [11]

- **Approche par compréhension :**

Elle repose sur des modèles fondés sur des concepts de psychologie cognitive et sur le paradigme de l'intelligence artificielle. À partir d'un texte source, elle permet de générer un nouveau texte, avec de nouvelles phrases et de nouvelles constructions syntaxiques. Pour obtenir un résumé pertinent, il faut coder un grand nombre de connaissances qui ne figurent pas toujours explicitement dans le texte originel.

- **Approche par extraction :**

Elle est utilisée dans des produits commerciaux et dans certains laboratoires, est inspirée du postulat : « Pour résumer, il suffit d'extraire ». Elle repose sur des algorithmes de repérage d'unités textuelles pertinentes. Le résumé respecte la linéarité et la structure du texte source.

2.3. Les étapes du résumé automatique :

Dans le résumé automatique de documents, on peut identifier trois différentes étapes. Ces étapes sont : l'identification du thème, l'interprétation, et la génération du résumé. La plupart des systèmes aujourd'hui utilisent la première étape seulement.

L'identification des thèmes produit un résumé simple ; une fois le système repère les unités importantes, il les présente comme un extrait. Ensuite, **l'interprétation** qui comporte la

fusion des concepts, l'évaluation, ou autres procédures qui utilisent une connaissance autre que le (les) document(s) d'entrée. Le résultat de l'interprétation est un abstrait non lisible, ou un extrait incohérent. Donc, l'étape de **génération** sert à produire un texte (document) lisible par l'humain, et dans le cas de l'extrait cette étape peut être considérée comme étape de "lissage" pour rendre le résumé plus cohérent. [12]

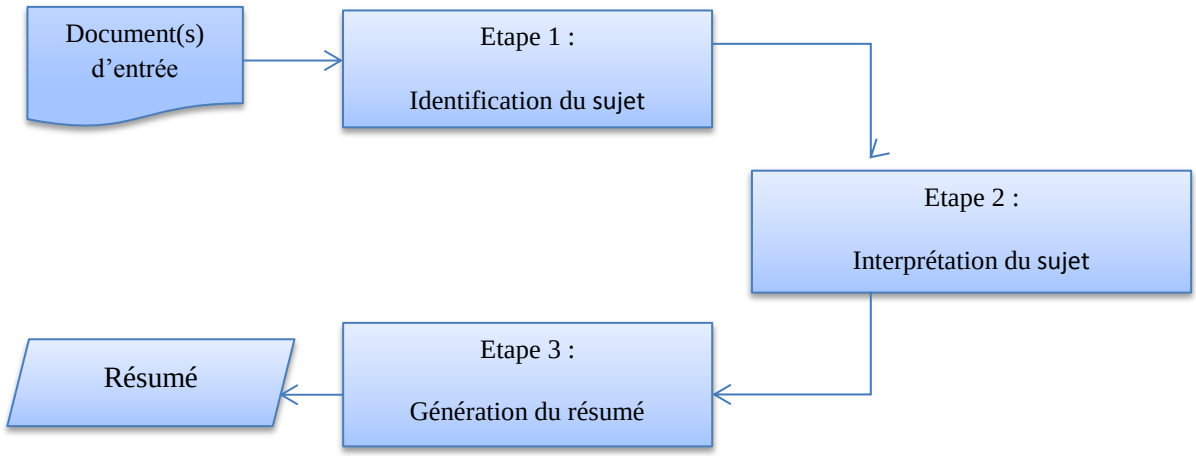

Figure 2.1. Les étapes du résumé automatique [12]

<u>Etape 1</u> : Identification des thèmes :

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

Chaque module attribue un score aux unités d'entrée (mot, phrase ou passage plus long), puis un module de combinaison, combine les scores pour chaque unité afin d'attribuer un score unique.

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

<u>Etape 2</u> : Interprétation des thèmes :

Dans l'interprétation, le but est de faire un compactage en réinterprétant et en fusionnant les thèmes extraits pour avoir des thèmes plus brefs. Ceci est indispensable du moment que les abstraits sont généralement plus courts que les extraits équivalents. Cette deuxième phase de résumé automatique (passage de l'extrait vers l'abstrait) est naturellement plus complexe que la première. Pour compléter cette phase, le système a besoin de connaissances sur le monde (par exemple, les anthologies), puisque sans connaissance aucun système ne peut fusionner les sujets extraits pour produire des sujets moins nombreux afin de former une abstraction.

Lors de l'interprétation, les thèmes identifiés comme importants sont fusionnés, représentée en des termes nouveaux, et exprimé en utilisant une nouvelle formulation, en utilisant des concepts ou des mots qui n'existent pas dans le document original.

Etape 3 : Génération du résumé :

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

2.4. Les types du résumé automatique :

- **Le résumé mono-document :** C'est le résumé d'un seul document isolé.
- **Le résumé multi-document :** C'est le résumé d'un groupe de documents, pas forcément hétérogènes, portant souvent sur une thématique bien précise.

2.4.1. Résumé automatique multi-documents :

Un système de résumé multi-documents permet de produire un résumé d'une collection de textes en rendant compte de ses idées principales. Cependant, certaines sont plus adaptées que d'autres au résumé multi-documents. Par exemple, les méthodes fondées sur la programmation linéaire ont montré plus de succès que les méthodes fondées sur les graphes. [13]

Les trois problèmes principaux pour le résumé multi-documents sont :

• la reconnaissance des unités saillantes redondantes.

• l'identification des différences entre les documents ;

• la cohérence du résumé même quand le contenu vient de différents documents sources. [13]

2.5. Les méthodes de résumé automatique :

Dans cette partie, nous présentons brièvement différentes méthodes employées pour l'extraction de phrases clefs, elles sont basées essentiellement sur le calcul d'un score associé à chaque phrase afin d'estimer son importance dans le texte. Le résumé final ne gardera que les phrases avec les meilleurs scores. [4]

2.5.1. Méthodes à base de mots clés :

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

a. Mots-clés prédéfinis :

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

Score $_{\text{mot-clé}}$(S) = a (t) * F (t)
F(t) est la fréquence du terme t dans la phrase S

$$a(t) = \begin{cases} \text{A si t} \in \text{liste de mots - clés (A > 1)} \\ \text{1 sinon} \end{cases}$$

La liste de mots-clés peut être introduite par l'utilisateur (domaine d'intérêt) ou composée des mots-clés établis par l'auteur. L'importance du poids du terme t est donné par $A \times F(t)$, avec $A>1$.

b. Titre :

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

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

Score $_{\text{titre}}$(S) = b (t) * F (t)
F(t) est la fréquence du terme t dans la phrase S

$$b(t) = \begin{cases} \text{A si t} \in \text{liste de mots du titre (A>1)} \\ \text{1 sinon} \end{cases}$$

c. Distribution des termes :

L'idée de cette méthode est de considérer comme importantes les phrases qui contiennent des mots importants du texte. Un mot est considéré important s'il est employé assez fréquemment dans le texte. Le même principe que la mesure **TF-IDF.**

2.5.2. Méthode à base de position :

Cette méthode suppose que la position d'une phrase dans un texte indique son importance dans le contexte. Les premières et les dernières phrases d'un paragraphe, par exemple, peuvent transmettre l'idée principale et devraient donc faire partie du résumé. Comme variante de cette méthode on peut citer la méthode Lead ; c'est une méthode qui détermine les phrases importantes en extrayant celles qui sont en tête. Cette méthode est efficace pour résumer les articles de journaux, puisque les phrases importantes ont tendance à apparaître dans les premières phrases de l'article.

On définit le score d'une phrase S à la position i comme suit :

Score $_{lead}$ (S_i) = β_i

$$\beta_i = \begin{cases} B > 0 \text{ si } i > 0 \\ 0 \text{ si } i >= N \end{cases}$$

β_i est une fonction rectangulaire qui modélise la distribution de phrases importantes selon leur position dans l'article.

Dans le cas où les dernières phrases auraient une certaine importance, il suffit d'introduire un nouvel intervalle pour la valeur de i. L'inconvénient de cette méthode est qu′elle dépend de la nature du texte à résumer ainsi que du style de l'auteur.

2.5.3. Méthode dépendant de la longueur de phrase :

Cette méthode attribue un poids à une phrase en fonction du nombre de mots dans la phrase.

Deux techniques peuvent être employées pour le calcul du score :

• longueur de chaque phrase (Li) par rapport à la longueur maximale de la phrase. Score-long **(Si) = Li/L$_{max}$**.

• affecte un score nul à une phrase plus courte qu'une certaine longueur (L minimale).

2.5.4. Méthode à base d'expressions indicatives (cuemethods) :

Cette méthode choisit des unités de texte avec des indications spécifiques ou des expressions spécifiques. Par exemple, pour les textes scientifiques, on a comme expressions *le but de ce travail …, ce papier présente …, les résultats* et *des conclusions* sont de bons candidats pour indiquer les phrases à inclure dans un résumé. Des textes de types différents peuvent avoir des expressions indicatives différentes. On peut déduire un score pour une phrase d'un texte quelconque à analyser en fonction de la ressemblance qu'elle présente, pour le trait donné.

On pourrait définir le score d'une phrase *S* correspondant à un certain motif comme:

$$\textbf{Score-}_{cue}\textbf{ (S)} = \begin{cases} \textbf{1 si S correspond à un motif} \\ \textbf{0 sinon} \end{cases}$$

2.5.5. Méthode basée sur les relations (cohésion lexicale) :

L'exploitation des fréquences de mots est un bon moyen pour faire ressortir les termes clés dans un texte mais elle ne prend pas en compte les relations entre les mots dans les différentes parties du texte. L'extraction de phrases basée sur la fréquence de mots cause souvent un manque de cohésion. Pour pallier ce problème, les déployés dans ce domaine ont développé une approche basée sur la cohésion grammaticale (c'est-à-dire, la référence, la substitution, la conjonction) et la cohésion lexicale (c'est-à-dire, des mots liés sémantiquement).

Cette méthode suggère que plus une phrase est liée à une autre dans un texte, plus elle est appropriée dans ce contexte c'est-à-dire qu'elle exprime le même sujet. Ainsi, des phrases

liées doivent être choisies ensemble pour composer un résumé. L'omission de certaines phrases fortement corrélées pourrait produire des textes incohérents. L'identification de telles corrélations est basée normalement sur un thésaurus ou lexique informatisé qui permet de déterminer les relations entre les mots. On construit des chaînes lexicales à partir des mots candidats du texte, ces chaînes regroupent des mots liés par des relations obtenues à partir du thésaurus. Les phrases qui sont les plus connectées aux chaînes lexicales sont extraites.

II.3. Evaluation du résumé automatique

Évaluer les résumés automatiques est une tâche difficile à laquelle la communauté à des réponses partielles. En effet, une évaluation automatique demande de disposer d'un système capable de générer des résumés de qualité humaine, afin qu'il soit capable de juger. Des solutions pragmatiques peuvent être envisagées. Un des objectifs des conférences NIST (Document Understanding Conferences **DUC** devenu Text Analysis Conference (**TAC**)), consiste à utiliser la métrique **ROUGE** (Recall Oriented Understudy for Gisting Evaluation). Cette métrique mesure la couverture entre les N-gramme produits automatiquement par une machine à ceux contenus dans des résumés écrits par un certain nombre de juges humains. Un haut niveau en **ROUGE** implique empiriquement un niveau de corrélation avec les résumés humains.

3.1. La mesure ROUGE:

Les mesures produites par **ROUGE** (Recall Oriented Understudy for Gisting Evaluation) sont des mesures automatiques dont le calcul s'appuie sur la comparaison du résumé système avec plusieurs résumés de référence. Cette comparaison se base sur la cooccurrence des n-grammes.

La formule, proposée pour les métriques générées par **ROUGE**, utilise une moyenne pondérée des n-grammes à longueurs variables et extraits à partir des résumés systèmes et un ensemble de résumés de référence.

3.2. Les mesures de rappel et de précision :

Le rappel et la précision présentent des mesures de similarité classiques de recherche d'information.

Ces mesures issues de cette discipline, ont pour objectif d'indiquer à quel point un système obtient des performances proches à celles obtenues manuellement par des humains.

Ces deux mesures ont trois paramètres suivants :

– P : nombre de phrases non pertinentes fournies par le système,

– Q : nombre de phrases pertinentes fournies par le système,

– R : nombre de phrases pertinentes présentes dans un fond textuel et non fournies par le système.

Le rappel et la précision sont alors calculés comme suit :

Rappel = Q/ (Q + R)

Précision = Q/ (Q + P)

II.4. Travaux liés au résumé automatique

Plusieurs travaux ont été réalisés dans le but de développer des systèmes de résumé automatique. Le premier objectif est d'obtenir des textes plus clairs et précis, faciles à comprendre et bien structurés. Tandis que le deuxième objectif est d'aider à éviter la difficulté de lire des textes trop longs.

Des chercheurs du RALI, sous la direction de Guy Lapalme, travaillent dans le domaine du résumé automatique depuis plusieurs années.

D'ailleurs depuis 2002, le RALI a systématiquement participé à toutes les compétitions de Document Understanding Conference (**DUC**) et plus récemment de Text Analysis Conference (**TAC**). [14]

Les principaux travaux au domaine du résumé automatique :

- **SumUM :**

SumUM a été développé par Horacio Saggion dans le cadre de sa thèse de doctorat (1997-2000). SumUM génère de courts résumés (10-15 lignes) de longs documents (15-20 pages) scientifiques et techniques. Il produit le résumé en deux étapes: l'utilisateur reçoit d'abord un résumé *indicatif*, qui identifie les sujets importants du document et le système génère ensuite un résumé *informatif* qui élabore quelques sujets choisis par l'utilisateur.

L'entrée du système est un article scientifique en anglais, contenant les éléments structuraux suivants: titre de l'article, auteur et affiliation, introduction, sections principales, conclusion, bibliographie et remerciement. La sortie du système est un court résumé indicatif composé de phrases complètes. Ce résumé n'est pas qu'un simple extrait de phrases du texte original, il est régénéré à partir des informations trouvées. Il est de qualité comparable à celle des résumés d'auteur. Il est ensuite possible d'obtenir des informations supplémentaires sur des sujets identifiés par l'usager.

- **CATS :**

CATS (Cats is an Automatic Text Summarizer) a été développé par Atefeh Farzindar et Frédérik Rozon au cours de l'été 2005 pour participer à la compétition Document Understanding Conferences 2005 (DUC2005). La tâche consistait à résumer, en moins de 250 mots, des groupes d'une vingtaine d'articles de journaux traitant du même sujet. Le résumé devait traiter d'un aspect particulier identifié par une question de quelques lignes. La performance de CATS, décrite dans cet article, a été excellente par rapport à l'ensemble de la trentaine de systèmes qui ont participé à la compétition.

- **Lakhas :**

Fouad Douzidia a développé Lakhas (signifiant *résumer* en arabe), un système de résumé de textes journalistiques arabes basé sur la combinaison de méthodes d'extractions utilisées jusqu'ici en anglais.

Lakhas a également été utilisé pour produire de très courts résumés en arabe lors de l'évaluation à DUC2004. Ces résumés étaient ensuite traduits en anglais avec un système de traduction automatique. Malgré le fait d'avoir suivi un chemin différent des autres compétiteurs, les résultats de l'évaluation ont été excellents et même les meilleurs lorsqu'on disposait du même système de traduction automatique que celui qui avait été utilisé pour traduire les autres textes. Ces résultats sont décrits dans cet article et dans son mémoire de maîtrise.

- **LetSUM (Legal text Summarizer) :**

En collaboration avec le groupe LexUM, qui faisait alors partie du Centre de recherche en droit public de la Faculté de Droit de l'Université de Montréal, Atefeh Farzindar a étudié la problématique des résumés de textes juridiques, plus particulièrement les jugements. La méthodologie repose sur l'exploitation de la structure thématique des décisions juridiques afin de constituer automatiquement une fiche de résumé augmentant la cohérence et la lisibilité du résumé. LetSUM permet aux juristes de consulter rapidement les idées clés d'un jugement pour trouver les jurisprudences pertinentes.

- **Automatic summarization of Legal Text (ASLI):**

Entre juillet 2007 et juin 2008, NLP Technologies et le RALI ont collaboré au projet ASLI (Automatic Summarization of Legal Information) financé dans le cadre du programme Alliance de Precarn.

La collaboration avec le RALI portait sur deux points:

Développement de règles sémantiques du domaine juridique permettent de segmenter un jugement, d'en identifier les domaines, d'y choisir les phrases pertinentes et d'identifier la catégorie du jugement et de déterminer les citations. Il faudra en développer un modèle général de ces diverses règles en tenant compte des aspects de maintenance, de facilité de modification et d'amélioration de la performance.

II.5. Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur le résumé automatique. Dans un premier temps, nous avons présenté quelques notions pour le résumé automatique, afin de comprendre ce domaine. Il peut appartient à plusieurs classes ou types passant par des étapes, et utilisant des différentes méthodes. Donc le système de résumé automatique est un domaine très important pour faciliter l'accès à un résumé bien compris et bien précis rapidement.

Chapitre III
Conception et Approche Proposée

III.1. Introduction

Dans ce chapitre, nous abordons les aspects de développement de notre solution. Nous y décrivons l'architecture et l'approche de développement du notre système de résumé automatique des documents textuels. Nous commençons tout d'abord par la présentation de l'environnement de développement, en détaillant les différents outils utilisés et expliquons notre approche proposée.

III.2. Environnement de l'application

L'implémentation et les tests de notre application ont été réalisés dans l'environnement matériel et logiciel suivant :

- Processeur : Intel ® Core ™i3-2328M CPU @ 2.20 GHz
- Mémoire installée (RAM) : 4.00 GO
- Windows : Windows 7 Professionnel
- Type de système : système d'exploitation 32 bits, processeur x64.
- Java sous l'environnement Netbeans 7.4

2.1. Langage d'application :

Java est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts. [15]

2.1.1. IDE Netbeans 7.4 [1]:

L'EDI NetBeans est un environnement de développement - un outil pour les programmeurs pour écrire, compiler, déboguer et déployer des programmes. Il est écrit en Java - mais peut supporter n'importe quel langage de programmation. Il y a également un grand nombre de modules pour étendre l'EDI NetBeans.

L'EDI NetBeans est un produit gratuit, sans aucune restriction quant à son usage.

Également disponible, La Plateforme NetBeans; une fondation modulable et extensible utilisée comme brique logicielle pour la création d'applications bureautiques. Les partenaires privilégiés fournissent des modules à valeurs rajoutées qui s'intègrent facilement à la plateforme et peuvent être utilisés pour développer ses propres outils et solutions.

Les deux produits sont open source et gratuits pour un usage commercial et non-commercial. Le code source est disponible pour réutilisation sous la Common Development and Distribution License (**CDDL**). [16]

[1] https://netbeans.org/downloads/7.4/?pagelang=pt_BR

III.3. Intégration des ressources logicielles

Notre application s'appuie sur les ressources logicielles et linguistiques suivantes :

Intellexer Summarizer	Système de Résumé automatique
Lucene	**Moteur de recherche open source**

Tableau 1 : Ressources logicielles intégrées dans notre application

3.1. Intellexer Summarizer :

Document Summarizer est une solution sémantique qui analyse le document, extrait ses idées principales et les met dans un bref résumé ou crée une annotation. Vous pouvez résumer un document, un courrier électronique ou une page Web directement à partir de votre application préférée ou générer des annotations. La fonctionnalité unique d'Intellexer Summarizer est de créer des orientations théoriques (par exemple, la politique, l'économie), orientées structurellement (par exemple article scientifique, brevet) et des résumés axés sur les concepts. [17]

Nous avons utilisé ce logiciel pour évaluer notre système par la comparaison de notre résultat des résumés et le résultat obtenu par Intellexer Summarizer[2].

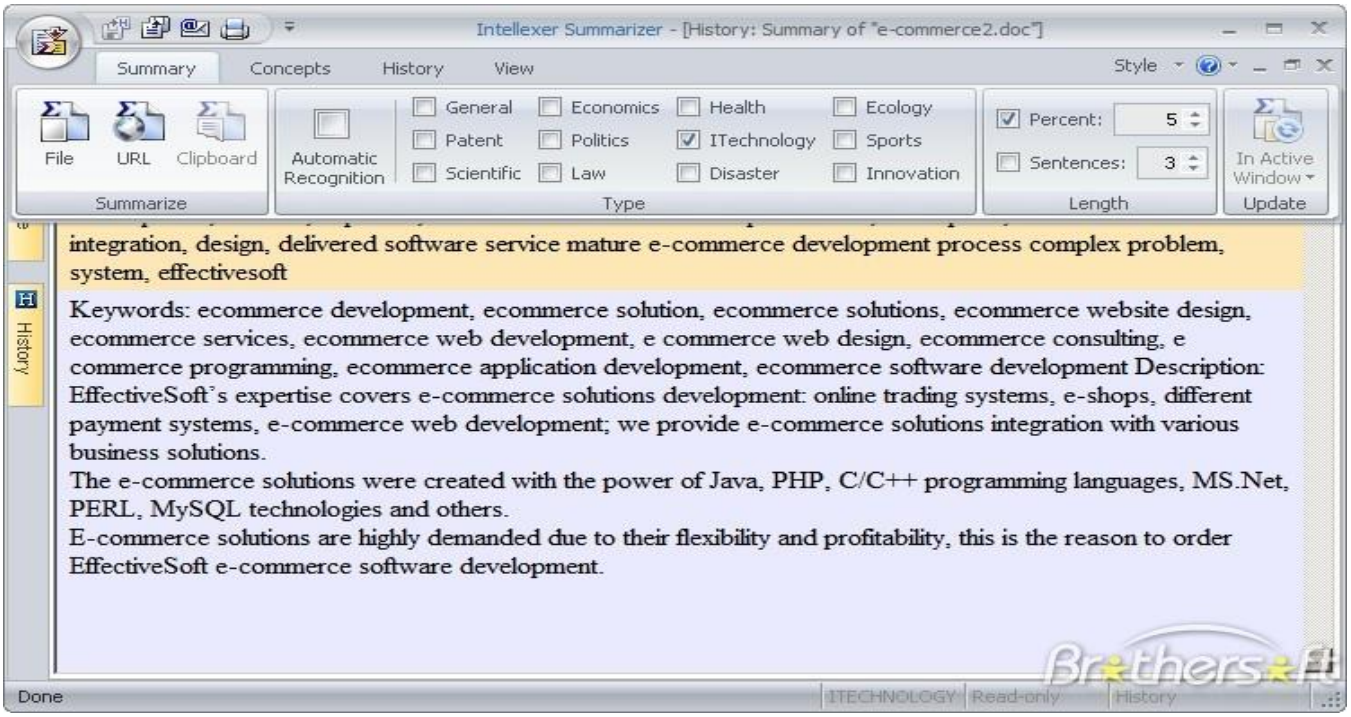

Figure 3.1 : Intellexer Summarizer [17]

[2] http://summarizer.intellexer.com/downloads.html

3.2. Fonctionnement de Lucene :

Lucene[3] est une bibliothèque de recherche basée sur java open source. C'est une bibliothèque très populaire et rapide utilisée dans une application basée sur Java pour ajouter une capacité de recherche de documents à tout type d'application de manière très simple et efficace.

Lucene généralement permet de créer un **IndexWriter** utilisé pour écrire le fichier d'index en choisissant un **Analyseur** compatible avec ce dernier, mais dans notre cas nous avons utilisé lucene pour analyser nos documents pour supprimer les mots inutiles (stop-words). Nous allons discuter de divers types d'objets Analyzer et d'autres objets pertinents qui sont utilisés lors du processus d'analyse.

- **Analyser :**

Il s'agit d'un ensemble de classes qui ont pour but le découpage du texte en « token » (mot) et l'analyse du texte. Les principaux analyzer fournis sont :

StandardTokenizer	**découpe le texte en mot et le converti en minuscule.**
StopFilter	**découpe le texte en mot et le converti en minuscule et supprime les mots vides (the, of, and, into, an, such, …).**
StandardAnalyzer	**Combine les deux analyzers précédents.**

Tableau 2 : les principaux analyzers de la normalisation d'un texte sur Lucene

3.3. Regroupement des phrases :

Le regroupement des phrases consiste à regrouper automatiquement ces phrases en clusters en fonction de leur similarité de contenu pour éviter le cas de la redondance des mêmes phrases. Le problème du regroupement de texte peut être défini comme suit. Étant donné un ensemble de n phrases notés PS et un nombre de cluster pré défini K (généralement défini par les utilisateurs), PS est regroupé dans K clusters de phrases PS1, PS2, ..., DSk, afin que les phrases d'un même cluster soient similaires les uns aux autres alors que les phrases provenant de différents clusters sont dissemblables.

[3] https://archive.apache.org/dist/lucene/java/3.6.2/

3.4. Prétraitement de nos documents :

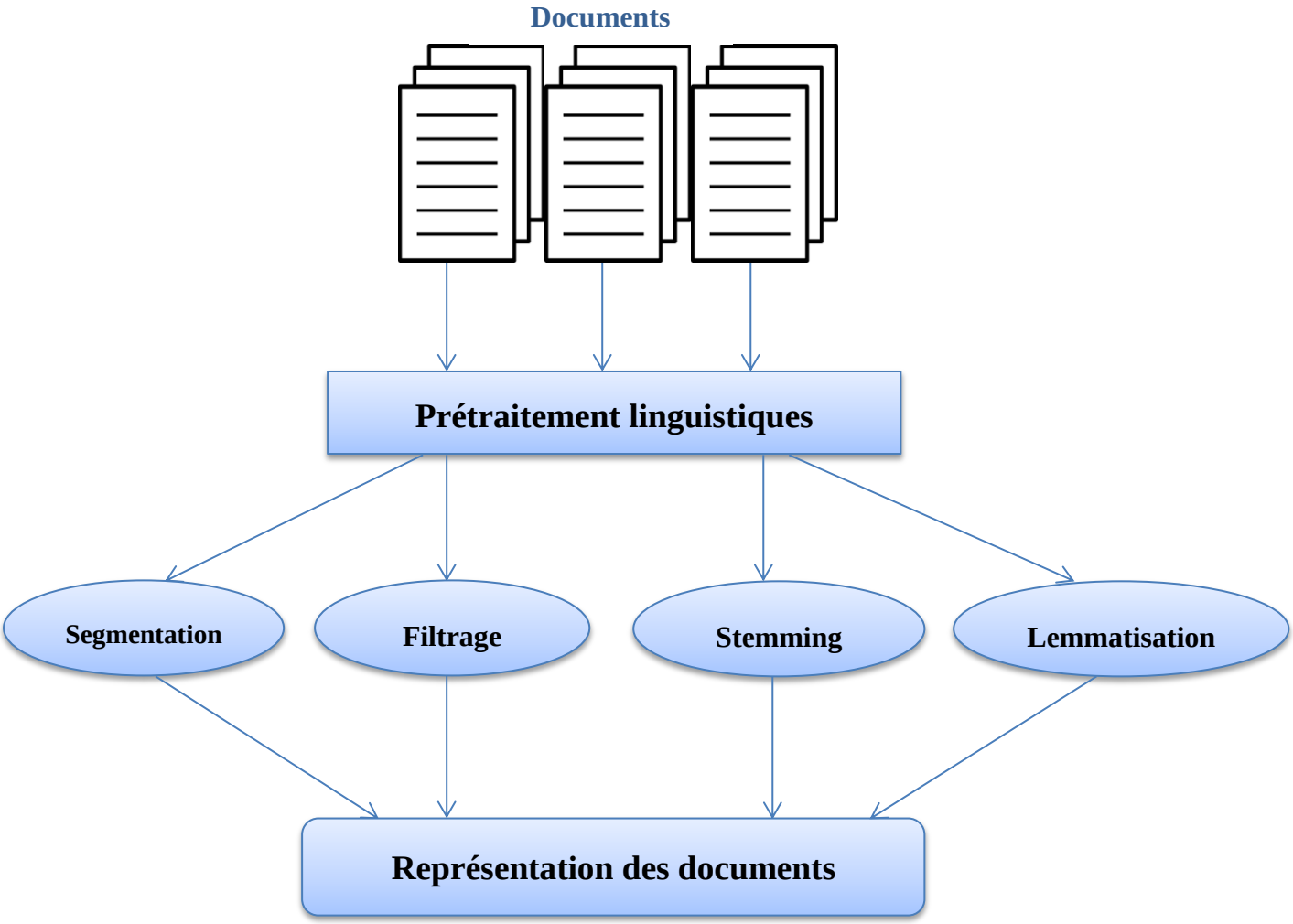

Figure 3.2 : La phase de prétraitement du document

- **Segmentation :**

C'est la conversion du document en un ensemble des phrases dépendant de la ponctuation.

- **Filtrage :**

C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs. Ces mots peuvent être dépendants d'un domaine ou pas. L'ensemble des mots éliminés est conservé dans un anti-dictionnaire sous lucene appelé « STOP_WORDS_SET ».

Par exemple: dans les documents en englais, "a, an, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, such, that, the, their, then, there, these, they, this, to, was, will", sont des mots non informatifs qu'ils doivent supprimés.

- **Stemming:**

C'est le principe qui basé sur la suppression des préfixes et suffixes des mots du document pour obtenir la forme originale ou canonique de ce mot.

C'est l'utilisation d'une forme canonique pour représenter les variantes morphologiques d'un mot :

Par exemple : les mots « dynamic, dynamics, dynamically » seront représentés par un même terme qui est « dynamic ».

- **Lemmatisation :**

C'est l'analyse qui permettant de trouver la forme présente dans les dictionnaires**.** Permet de convertir les verbes en l'infinitif et les mots pluriels aux singuliers.

Par exemple : - « **petit, petite, petits** et **petites** », la forme canonique de tous ces mots est « **petit** ».

- "il **jouera**", "nous **jouons**", "ils **ont joué**" la forme canonique de tous ces verbes est « **jouer** ».

Pour les documents en anglais on utilise la technique des systèmes itératifs à base de règles simples par exemple Porter stemming : on établit une liste de suffixes et de préfixes qui sont éliminés itérativement.

« Porter Stemming » : L'algorithme de déroulement de Porter (ou 'Porter stemmer') est un processus pour éliminer les terminaisons morphologiques des mots en anglais. Son utilisation principale est dans le cadre d'un processus de normalisation des termes qui se fait habituellement lors de la mise en place des systèmes de récupération d'informations.

III.4. Notre approche proposée

Notre approche consiste à implémenter une approche capable de produire des résumés à partir de la segmentation d'un texte ou un document en phrases, ensuite l'analyse du texte à partir de la suppression des mots inutiles, non informatif existants dans le document, l'obtention de la forme originale de chaque mot et enfin le regroupement des phrases similaires en classe sémantique afin d'obtenir un résultat convenable par les méthodes du résumé automatique.

4.1. Les méthodes utilisées pour générer notre résumé :

4.1.1. Méthode dépendant de la longueur de phrase :

Cette méthode attribue un poids à une phrase en fonction du nombre de mots dans la phrase.

Les techniques qui peuvent être employées pour le calcul du score sont :

• On a calculé la longueur de chaque phrase (Li) qui est le nombre de mots dans chaque phrase identifiée par ligne.

• On a défini la longueur maximale (L_{max}) du document.

• Si la longueur de chaque phrase (Li) est supérieur à la longueur maximale du document, le score est : score_phrase **= L_i/L_{max}.**

Exemple :

Texte :

« C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs. Ces mots peuvent être dépendants d'un domaine ou pas. L'ensemble des mots éliminés est conservé dans un anti-dictionnaire sous lucene ».

Phrase1 = « C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs ».

Phrase2 = « Ces mots peuvent être dépendants d'un domaine ou pas ».

Phrase3 = « L'ensemble des mots éliminés est conservé dans un anti-dictionnaire sous lucene ».

✓ **Le calcul de la longueur de chaque phrase :**

Longueur(Phrase1) = nombre de mots qu'elle contient Phrase1 = 14.

Longueur(Phrase2) = nombre de mots qu'elle contient Phrase2 = 9.

Longueur(Phrase3) = nombre de mots qu'elle contient Phrase1 = 11.

✓ **Le calcul du score de chaque phrase :**

Si on choisit la technique de la longueur maximale :

D'après le calcul de la longueur de chaque phrase, on déduit que la longueur maximale du texte = 14.

Score(Phrase1) = Longueur(Phrase1) / longueur maximale = 14/14 = 1.

Score(Phrase2) = Longueur(Phrase2) / longueur maximale = 9/14= 0.64

Score(Phrase3) = Longueur(Phrase3) / longueur maximale = 11/14 = 0.78

Après le calcul du score de chaque phrase, on trie ce score par ordre décroissant :

1-------- Score(Phrase1)

2-------- Score(Phrase3)

3-------- Score(Phrase2)

Si on veut créer le résumé à partir de l'extraction des deux premières phrases qui ont le score le plus élevé, **le résumé** est : « C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs. L'ensemble des mots éliminés est conservé dans un anti-dictionnaire sous lucene».

4.1.2. Méthode à base de position de la phrase :

Cette méthode attribue un poids à une phrase en fonction de la position de cette phrase dans le document. Elle suppose que la position d'une phrase dans un texte indique son importance dans le contexte. Les premières et les dernières phrases d'un paragraphe, par exemple, peuvent transmettre l'idée principale et devraient donc faire partie du résumé.

Score (phrase) = β_i

$$\beta_i = \begin{cases} \mathbf{B > 0 \quad si\ i < N} \\ \mathbf{0 \quad si\ i >= N} \end{cases}$$

N : le nombre total des phrases.

i : la position de la phrase dans le document.

Exemple :

Texte :

« C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs. Ces mots peuvent être dépendants d'un domaine ou pas. L'ensemble des mots éliminés est conservé dans un anti-dictionnaire sous lucene ».

D'après ce texte on trouve que :

N = 3.

i = 1 = « C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs ».

i = 2 = « Ces mots peuvent être dépendants d'un domaine ou pas ».

i = 3 = « L'ensemble des mots éliminés est conservé dans un anti-dictionnaire sous lucene ».

On a choisi que **B = 2,** donc :

Pour la première phrase : i = 1 < N = 3

Score (phrase1) = 2

Pour la deuxième phrase : i = 2 < N = 3

Score (phrase2) = 2

Pour la troisième phrase : i = 3 = N = 3

Score (phrase3) = 0

Après le calcul du score de chaque phrase, on trie ce score par ordre décroissant :

1-------- Score(Phrase1)

2-------- Score(Phrase2)

Si on veut créer le résumé à partir de l'extraction des deux premières phrases qui ont le score le plus élevé, **le résumé** est : « C'est le principe qui basé sur la suppression des mots inutiles ou non informatifs. Ces mots peuvent être dépendants d'un domaine ou pas».

Dans le cas où les dernières phrases auraient une certaine importance, il suffit d'introduire un nouvel intervalle pour la valeur de i de la manière suivante :

Score (phrase) = β_i

$$\beta_i = \begin{cases} B > 0 & si\ i > N \\ 0 & si\ i <= N \end{cases}$$

N : le nombre total des phrases.

i : la position de la phrase dans le document.

4.1.3. Méthode à base de mots-clés du document :

On extraire du document les mots les plus fréquents c'est-à-dire les mots les plus répétés.

Cette méthode attribue un poids à une phrase selon les mots-clés qu'elle contient, on peut calculer le score de chaque phrase comme suit :

Score (phrase) = $\sum$ (A (mot_i) * F (mot_i)), i = 1..n

n : nombre total des mots dans la phrase

F (mot) : La fréquence de ce mot dans la phrase.

A (mot) = a > 1 si ce mot appartient à la liste des mots clés. (On a choisi a=3)

= 1 sinon.

Exemple :

Texte :

« Master2 Master2 informatique.Master2 informatique.Master2 isi. »

Liste des mots clés (phrase1) = {Master2, informatique}.

Phrase1 **=** Master2 Master2 informatique.

Phrase2 = Master2 informatique.

Phrase3 = Master2 isi.

✓ **Le calcul de la fréquence de chaque mot dans la phrase :**

Pour la phrase1 :

F(Master2) = 2.

F (informatique) = 1.

***Pour la phrase2* :**

F(Master2) = 1.

F (informatique) = 1.

***Pour la phrase3* :**

F(Master2) = 1. F (isi) = 1.

✓ **Le calcul de l'appartenance de chaque mot dans la phrase à la liste des mots clés :**

Pour la phrase1 :

A (Master2) = 3 car le mot « Master2 » appartient à la liste des mots clés.

A (informatique) = 3 car le mot « informatique » appartient à la liste des mots clés.

***Pour la phrase2* :**

A(Master2) = 3 car le mot « Master2 » appartient à la liste des mots clés.

A (informatique) = 3 car le mot « informatique » appartient à la liste des mots clés.

***Pour la phrase3* :**

A(Master2) = 3 car le mot « Master2 » appartient à la liste des mots clés.

A (isi) = 1 car le mot « isi » n'appartient pas à la liste des mots clés.

✓ **Le calcul du score de chaque phrase :**

Score (phrase1) = (2*3) + (1*3) = 9.

Score (phrase2) = (1*3) + (1*3) = 6.

Score (phrase3) = (1*3) + (1*1) = 4.

Après le calcul du score de chaque phrase, on trie ce score par ordre décroissant :

1-------- Score(Phrase1)

2-------- S.c.ore(Phrase2)

3-------- Score(Phrase3)

Si on veut cr0.

éer le résumé à partir de l'extraction des deux premières phrases qui ont le score le plus élevé, **le résumé** est : « Master2 Master2 informatique.Master2 informatique ».

4.1.4. Méthode à base de mots clés de la première phrase du document :

Cette méthode s'applique sur le même principe de la méthode à base de mots clés du titre, nous avons choisi les mots de la première phrase du document comme des mots clés pour cette méthode parce que la première phrase est en générale l'expression la plus significative et qui résume le mieux un document en quelques mots, on peut dire que la phrase qui ressemble le plus à la première phrase est la plus marquante du document. Par conséquent, on peut attribuer à chaque phrase un poids en fonction de sa ressemblance avec la première phrase.

Cette méthode attribue un poids à une phrase en fonction de la fréquence des mots de cette phrase et l'appartenance de ces mots à la liste des mots clés.

Score (phrase) = $\sum$ (B (mot_i) * F (mot_i)), i = 1..n

n : nombre total des mots dans la phrase

F (mot) : La fréquence de ce mot dans la phrase.

B (mot) = a > 1 si ce mot appartient à la liste des mots clés. (On a choisi a=3)

= 1 sinon.

Exemple :

Texte :

« Résumé des textes Résumé. Résumé Automatique des textes. Résumé Automatique des documents.»

Phrase1 = Résumé des textes Résumé.

Phrase2 = Résumé Automatique des textes.

Phrase3 = Résumé Automatique des documents.

Liste des mots clés (phrase1) = {Résumé, des, textes}.

✓ **Le calcul de la fréquence de chaque mot dans la phrase :**

***Pour la phrase1* :**

F(Résumé) = 2.

F (des) = 1.

F (textes) = 1.

***Pour la phrase2* :**

F(Résumé) = 1.

F (Automatique) = 1.

F (des) = 1.

F (textes) = 1.

***Pour la phrase3* :**

F(Résumé) = 1.

F (Automatique) = 1.

F (des) = 1.

F (documents) = 1.

✓ **Le calcul de l'appartenance de chaque mot dans la phrase à la liste des mots clés :**

***Pour la phrase1* :**

B (Résumé) = 3 car le mot « Résumé » appartient à la liste des mots clés.

B (des) = 3 car le mot « des » appartient à la liste des mots clés.

B (textes) = 3 car le mot « textes » appartient à la liste des mots clés.

Pour la phrase2 **:**

B (Résumé) = 3 car le mot « Résumé » appartient à la liste des mots clés.

B (Automatique) = 1 car le mot « Automatique » appartient à la liste des mots clés.

B (des) = 3 car le mot « des » appartient à la liste des mots clés.

B (textes) = 3 car le mot « textes » appartient à la liste des mots clés.

Pour la phrase3 **:**

B (Résumé) = 3 car le mot « Résumé » appartient à la liste des mots clés.

B (Automatique) = 1 car le mot « Automatique » appartient à la liste des mots clés.

B (des) = 3 car le mot « des » appartient à la liste des mots clés.

B (documents) = 1 car le mot « textes » appartient à la liste des mots clés.

✓ **Le calcul du score de chaque phrase :**

Score (phrase1) = ((2*3) + (1*3) + (1*3) = 12.

Score (phrase2) = (1*3) + (1*1) + (1*3) + (1*3) = 10.

Score (phrase3) = (1*3) + (1*1) + (1*3) + (1*1) = 8.

Après le calcul du score de chaque phrase, on trie ce score par ordre décroissant :

1-------- Score(Phrase1)

2-------- Score(Phrase2)

3-------- Score(Phrase3)

Si on veut créer le résumé à partir de l'extraction des deux premières phrases qui ont le score le plus élevé, **le résumé** est : « Résumé des textes Résumé. Résumé Automatique des textes».

4.2. Corpus de test :

Dans ce mémoire, on a utilisé deux corpus, l'un c'est le corpus de test qui contient 32 documents et l'autre c'est un corpus de jugement qui est le corpus de l'ensemble des résumés générés par le système intégré « **Intellexer Summarizer »**, Ce dernier nous permet d'accomplir l'évaluation par le calcul de la mesure « Rouge ».

III.5. Conclusion

Dans ce chapitre nous avons présenté les différents outils et logiciels utilisés afin de concevoir notre système de résumé, le prétraitement utilisé pour nos documents, l'explication des méthodes exploitables, ainsi que les corpus utilisés pour tester la validité de notre système.

Chapitre IV
Implémentation et mise en œuvre

IV.1. Introduction

Ce chapitre vise à expliquer les différentes étapes d'implémentation et d'expérimentation de notre système ainsi que les résultats obtenus. Nous commençons tout d'abord par la présentation de l'architecture de notre application, puis nous expliquons le déroulement de l'application, et enfin nous interprétons et commentons les résultats obtenus.

IV.2. Architecture de L'application

Notre système est un système de résumé automatique de texte en anglais basé principalement sur des techniques d'extraction. La mise en œuvre fonctionnelle de notre système est représentée à la **figure 4.1**. Elle repose sur une segmentation à différents niveaux afin de permettre la génération de résumé. Ce système est flexible et comporte plusieurs modules qui peuvent communiquer entre eux.

Nous présentons dans la (Figure **4.1**) l'architecture globale de notre application :

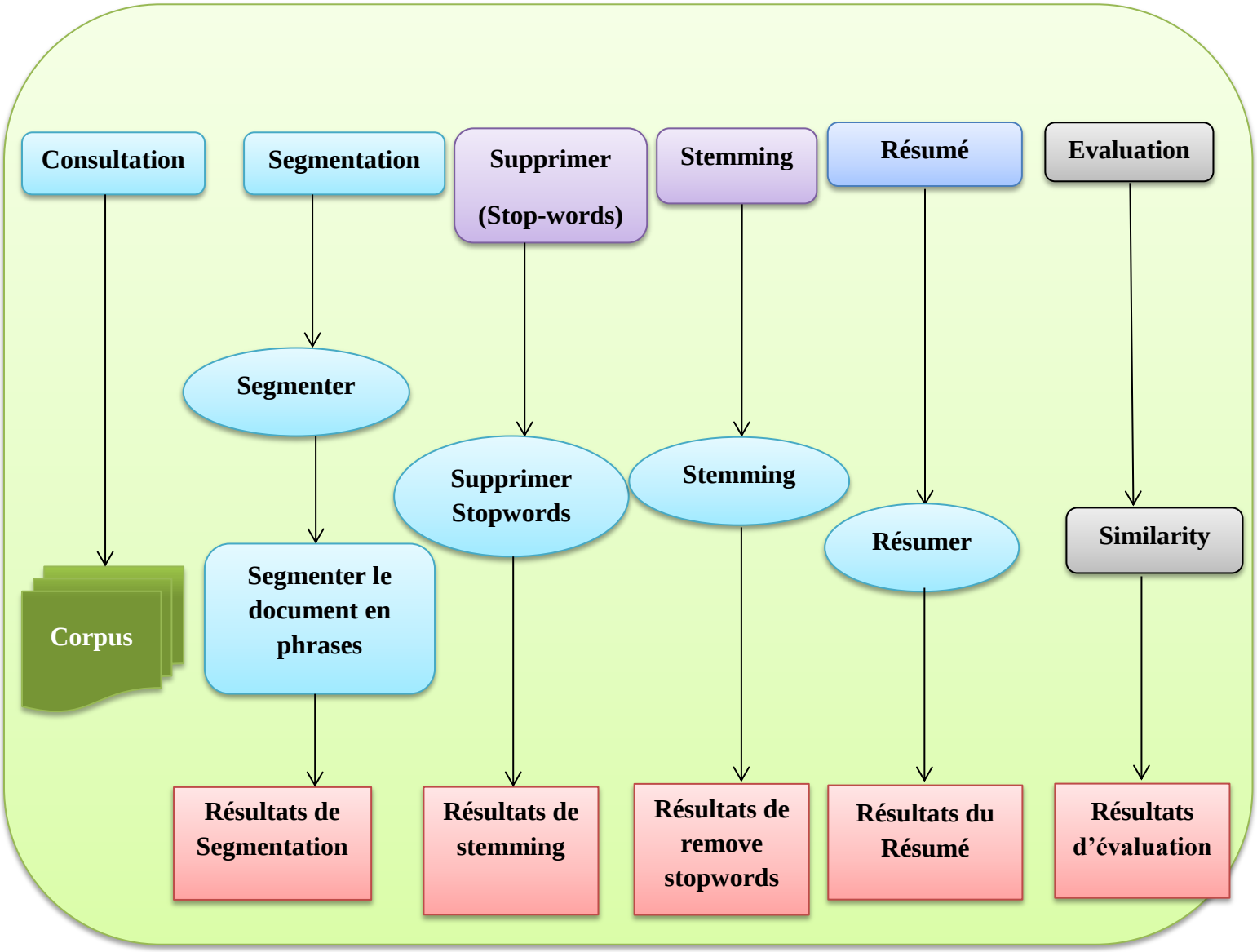

Figure 4.1 : Architecture globale de l'application

2.1. Description des principaux modules composant de l'architecture globale de l'application:

- **Consultation :** permet l'ouverture et la consultation des documents.
- **Segmentation :**
  Segmentation du document en phrases : identifie chaque phrase du document dans le corpus par ligne.
- **Supprimer Stop-words :** consiste à éliminer tous les mots non significatifs. Pour chaque mot reconnu, on le compare avec un des éléments dans l'anti-dictionnaire sous Lucene qui contient tous les mots non-significatifs. Si un mot en fait partie, il ne sera pas pris en considération pour le calcul de sa fréquence, aussi la phrase qui contient ces mots inutiles peut influencer le résumé généré.
- **Stemming :** permet d'obtenir la forme canonique et originale des mots, un mot peut être trouvé sous différentes formes dans le même document. Ces mots doivent être convertis sur leur forme originale pour simplifier. L'algorithme de dérivation est utilisé pour transformer les mots en leurs formes canoniques. Dans notre travail, on utilise le stemmer qui divise un mot dans sa forme racine.

Cette étape est nécessaire à cause des variations qui peuvent exister lors de l'écriture d'un même mot. L'extraction se fait à partir du document original ce qui permet de préserver l'intégralité de l'information.

- **Résumé:** pour faire le résumé du document.
- **Evaluation :** pour évaluer la qualité de notre système par apport à d'autres systèmes.

2.1. Fonctionnement du « Résumé » :

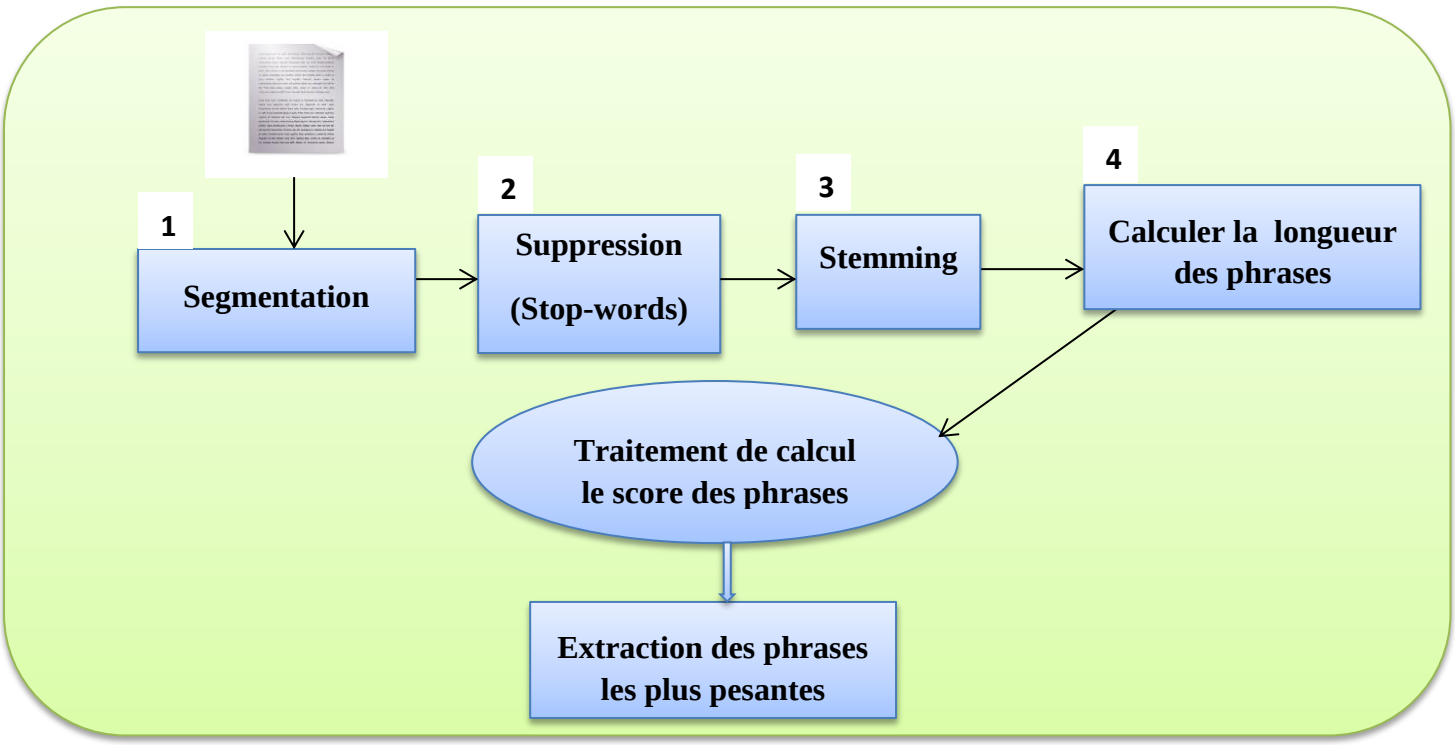

Figure 4.2 : Fonctionnement du « Résumé » (méthode1)

Description des principaux modules composant du « Résumé » (méthode1):

- **Segmentation :** c'est la segmentation du document en phrases.
- **Supprimer Stop-words :** consiste à éliminer tous les mots non informatifs dans le document.
- **Stemming :** permet d'obtenir la forme canonique et originale des mots du document.
- **Calculer la longueur des phrases :** consiste à calculer le nombre des mots dans chaque phrase.
- **Calcul le score des phrases :** Le score sera calculé par apport à la longueur du chaque phrase. Le résultat sera retourné sous forme de liste de phrases triée par score.
- **Extraction des phrases :** permet de retourner le résultat final suivant le choix du nombre de phrases extraites par rapport au nombre de phrases contenues dans le document.

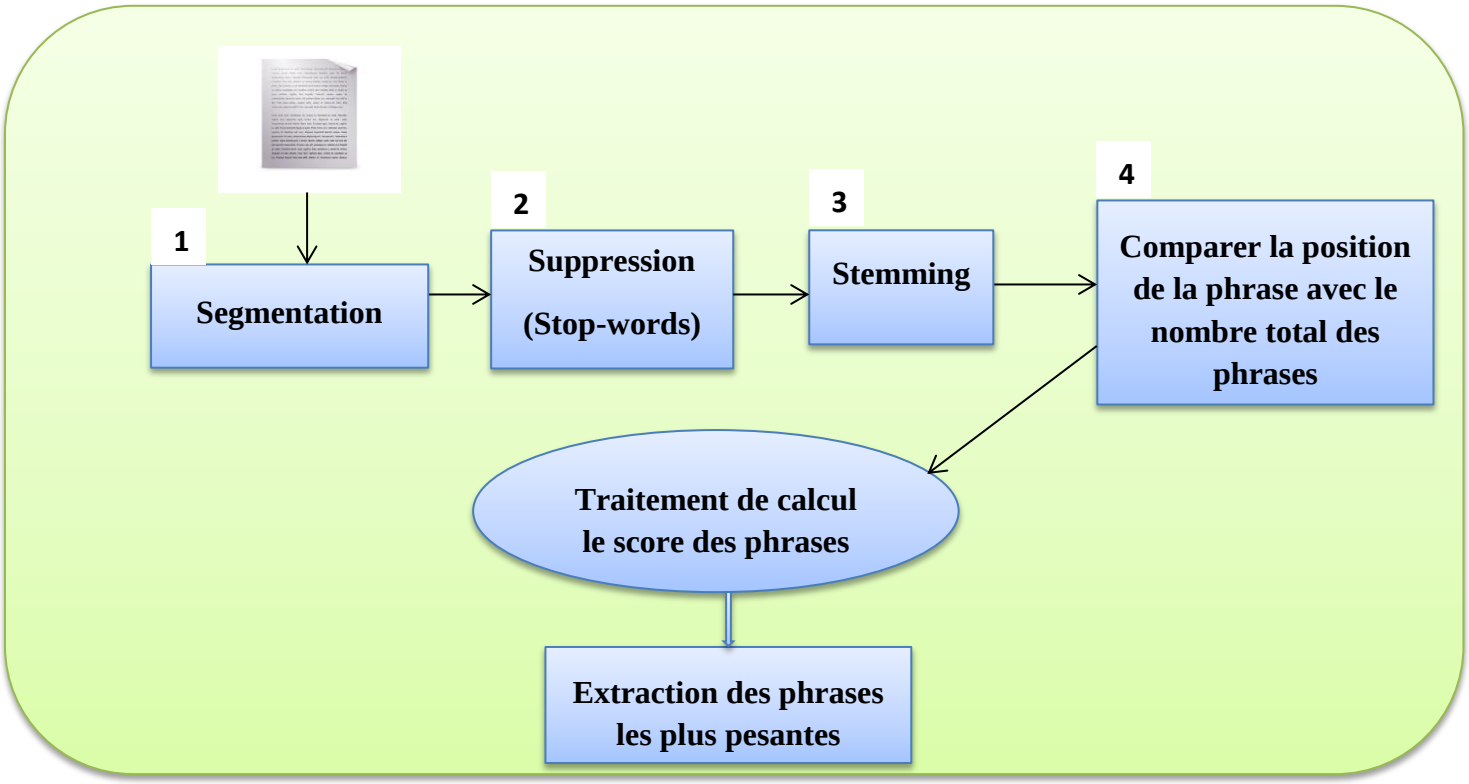

Figure 4.3 : Fonctionnement du « Résumé » (méthode2)

Description des principaux modules composant du « Résumé » (méthode2) :

- **Segmentation :** c'est la segmentation du document en phrases.
- **Supprimer Stop-words :** consiste à éliminer tous les mots non informatifs dans le document.
- **Stemming :** permet d'obtenir la forme canonique et originale des mots du document.
- **Comparer la position de la phrase avec le nombre total des phrases :** à partir du calcul du nombre total des phrases, on compare ce dernier avec la position de chaque phrase.

- **Calcul le score des phrases :** Le score sera calculé par apport à la position de chaque phrase. Le résultat sera retourné sous forme de liste de phrases triée par score.
- **Extraction des phrases :** permet de retourner le résultat final suivant le choix du nombre de phrases extraites par rapport au nombre de phrases contenues dans le document.

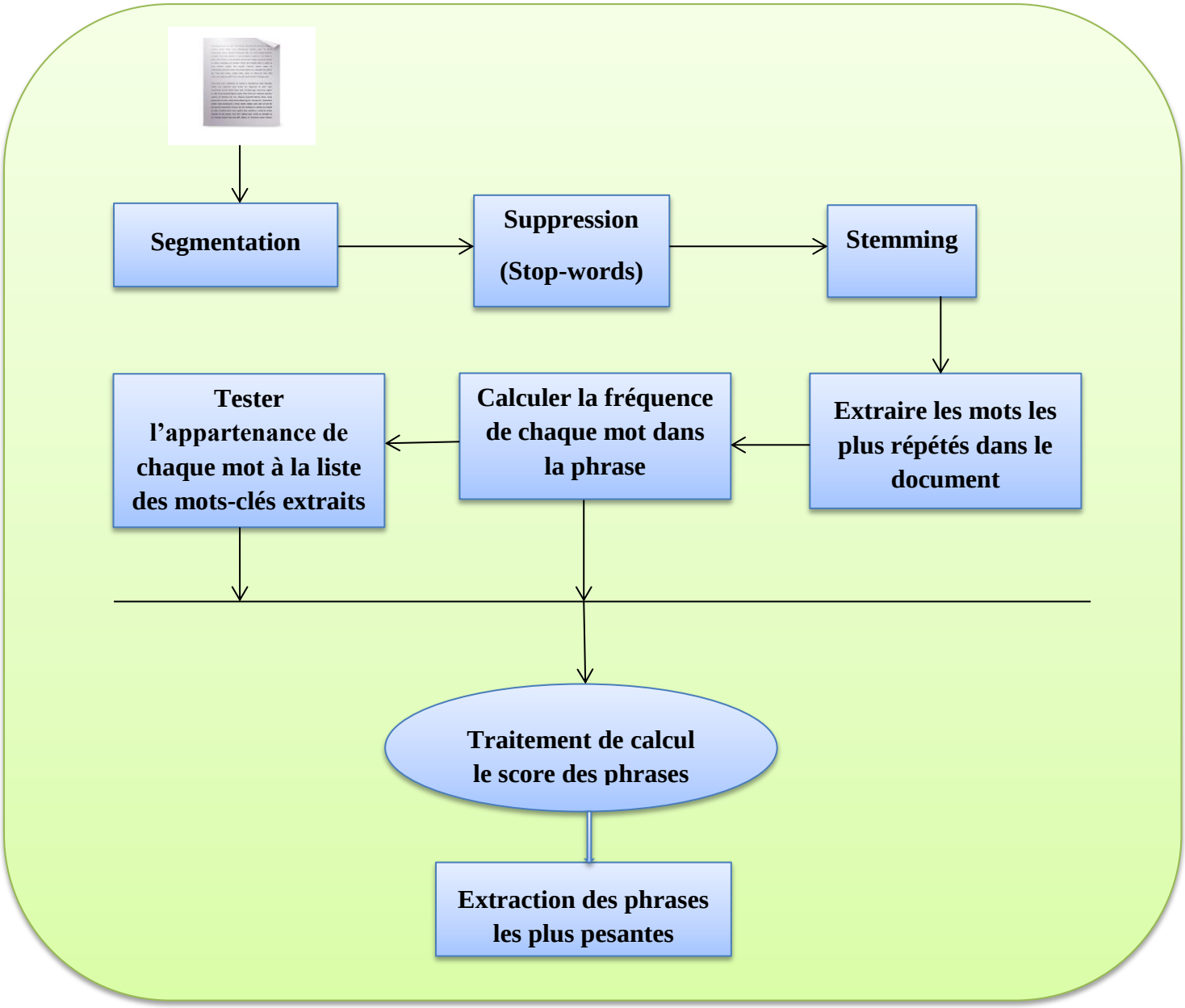

Figure 4.4 : Fonctionnement du « Résumé » (méthode3)

Description des principaux modules composant du « Résumé » (méthode3) :

- **Segmentation :** c'est la segmentation du document en phrases.
- **Supprimer Stop-words :** consiste à éliminer tous les mots non informatifs dans le document.
- **Stemming :** permet d'obtenir la forme canonique et originale des mots du document.

- **Extraire les mots les plus répétés dans le document :** après le calcul de la fréquence de chaque mot dans le document, on extraire les mots les plus fréquents et les mettez dans une liste comme des mots-clés.
- **Calculer la fréquence de chaque mot dans la phrase :** consiste à calculer le nombre d'occurrences de chaque mot dans la phrase.
- **Tester l'appartenance de chaque mot à la liste des mots-clés extraits :** c'est-à-dire on teste si chaque mot dans phrase appartient à la liste des mots-clés extraits ou pas.
- **Calcul le score des phrases :** Le score sera calculé par apport à la fréquence de chaque mot dans la phrase et l'appartenance de ce mot à la liste des mots-clés. Le résultat sera retourné sous forme de liste de phrases triée par score.
- **Extraction des phrases :** permet de retourner le résultat final suivant le choix du nombre de phrases extraites par rapport au nombre de phrases contenues dans le document.

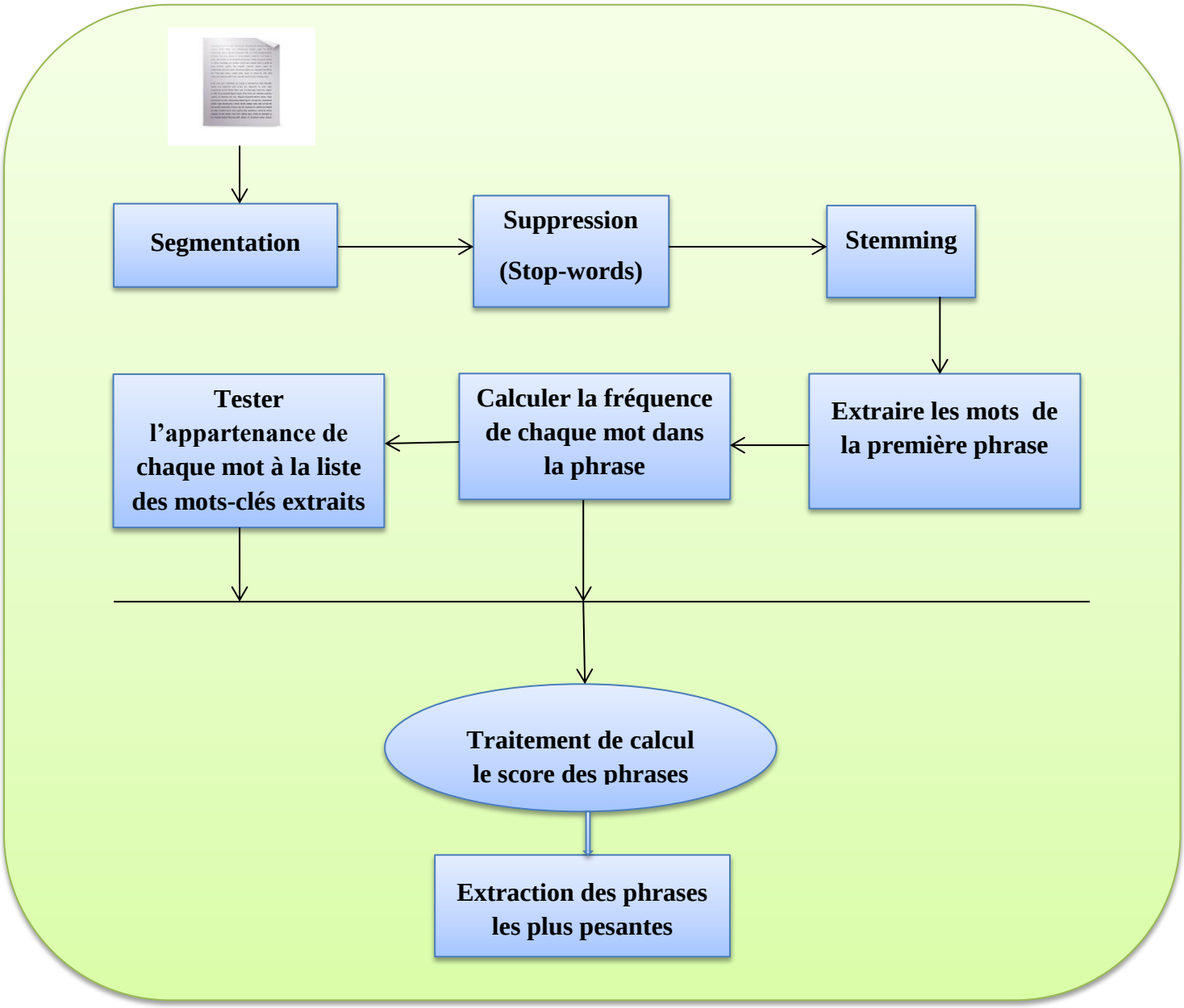

Figure 4.5 : Fonctionnement du « Résumé » (méthode4)

Description des principaux modules composant du « Résumé » (méthode4) :

- **Segmentation :** c'est la segmentation du document en phrases.
- **Supprimer Stop-words :** consiste à éliminer tous les mots non informatifs dans le document.
- **Stemming :** permet d'obtenir la forme canonique et originale des mots du document.
- **Extraire les mots de la première phrase :** permet l'extraction des mots de la première phrase et les mettez dans une liste comme des mots-clés.
- **Calculer la fréquence de chaque mot dans la phrase :** consiste à calculer le nombre d'occurrences de chaque mot dans la phrase.
- **Tester l'appartenance de chaque mot à la liste des mots-clés extraits :** c'est-à-dire on teste si chaque mot dans la phrase appartient à la liste des mots-clés extraits ou pas.
- **Calcul le score des phrases :** Le score sera calculé par apport à la fréquence de chaque mot dans la phrase et l'appartenance de ce mot à la liste des mots-clés. Le résultat sera retourné sous forme de liste de phrases triée par score.
- **Extraction des phrases :** permet de retourner le résultat final suivant le choix du nombre de phrases extraites par rapport au nombre de phrases contenues dans le document.

IV.3. Mise en œuvre

Dans cette partie on va décrire les différentes parties de notre application coté interface graphique et les différentes opérations de chaque bouton et menu.

La figure suivante représente l'interface principale de notre application :

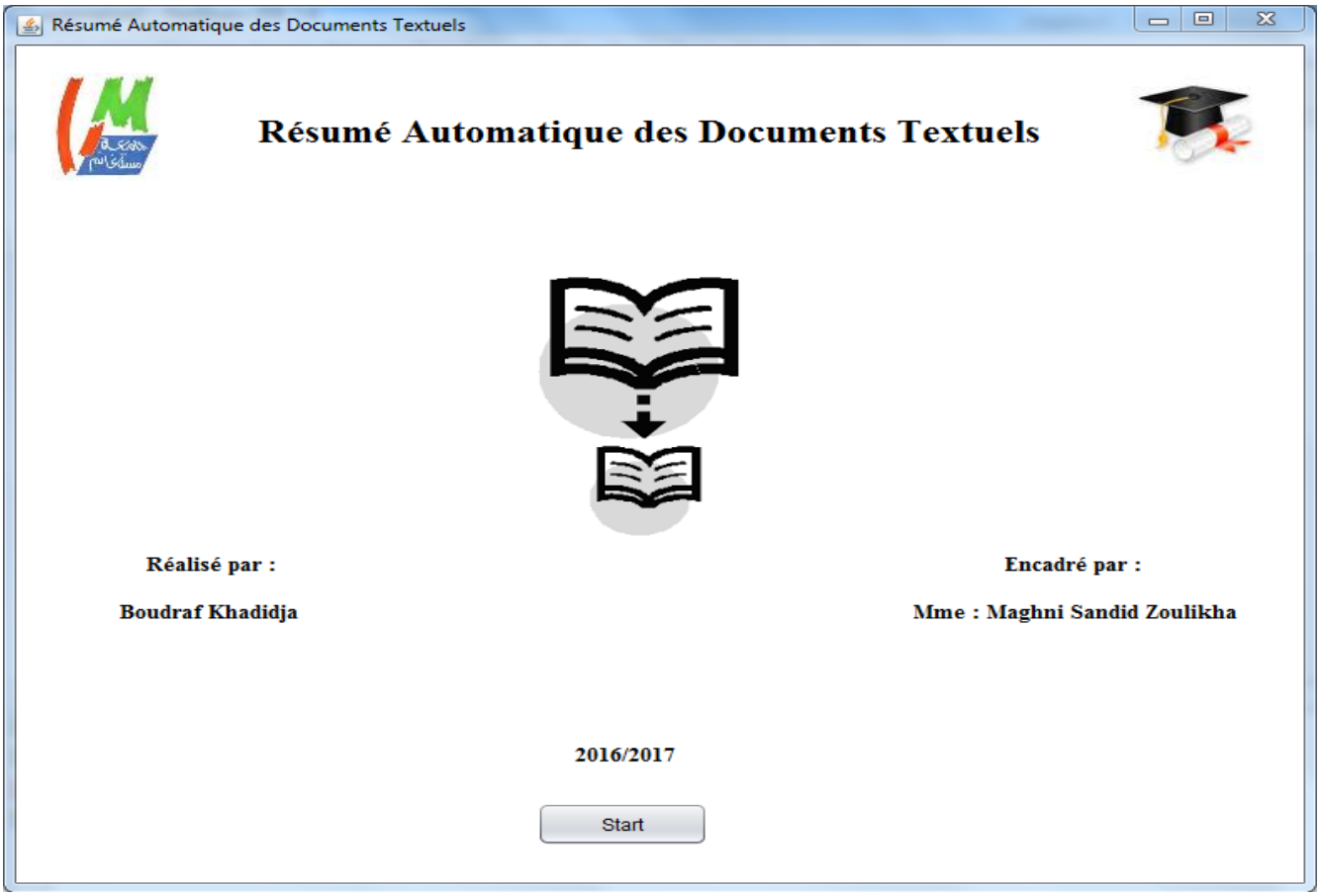

Figure 4.6 : L'interface principale de notre application

3.1. Menu Principal :

Notre application se compose d'un menu principal à partir de lequel l'utilisateur peut effectuer les traitements.

Le menu principal est montré dans la figure **4.7** :

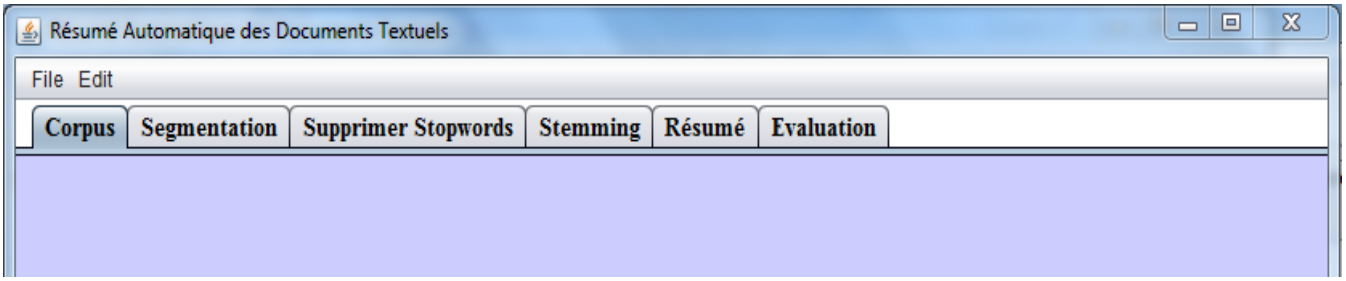

Figure 4.7 : Le Menu principal de notre application

3.1.1. Corpus :

L'interface dédiée pour les différentes tâches générales sur un corpus telles que l'ouverture et la consultation du contenu de chaque document.

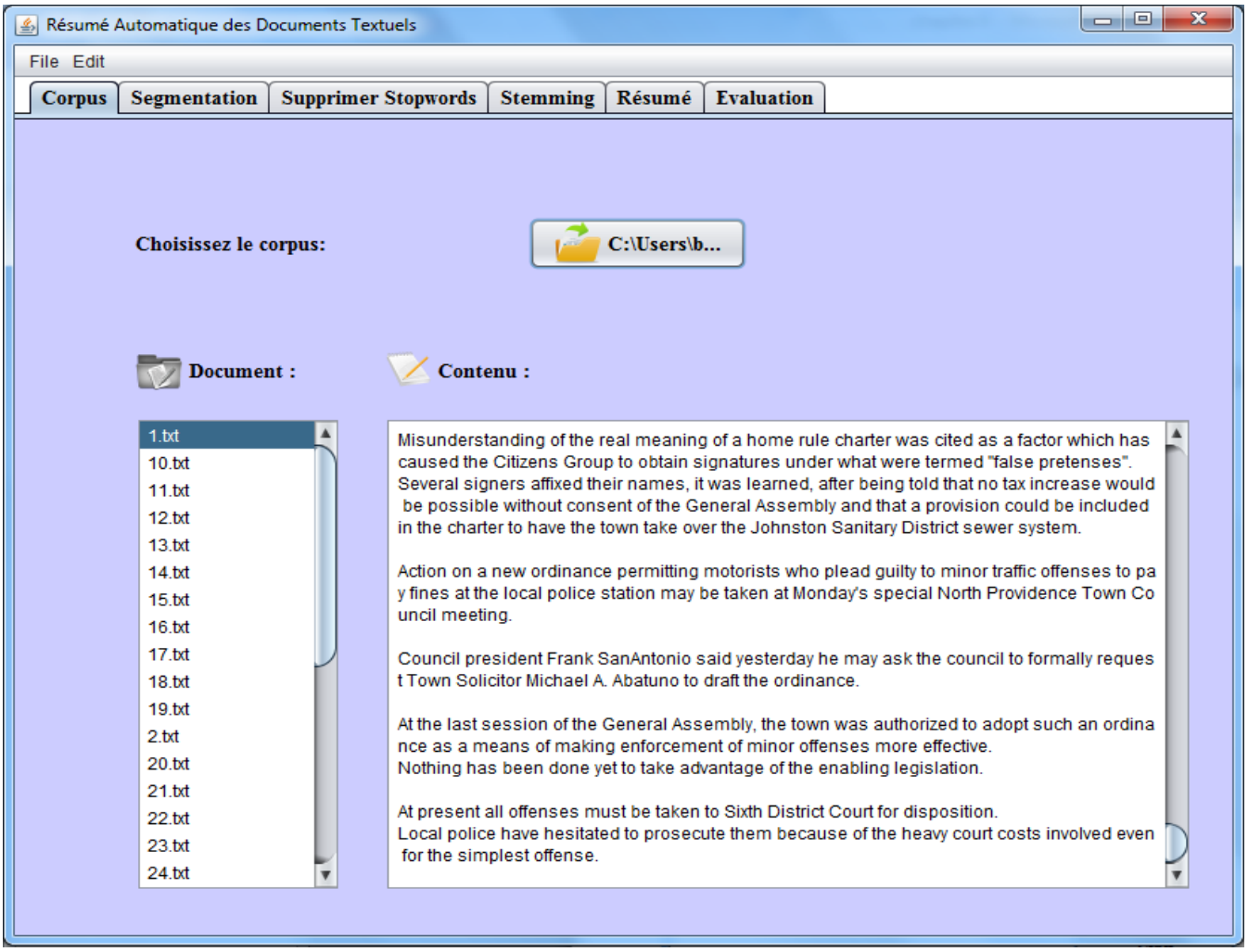

Figure 4.8 : Fenêtre de consultation d'un document

3.1.2. Segmentation :

L'interface dédiée pour consulter le document segmenté en des phrases à partir du document choisis dans la fenêtre de consultation en cliquant sur le bouton **Segmenter**.

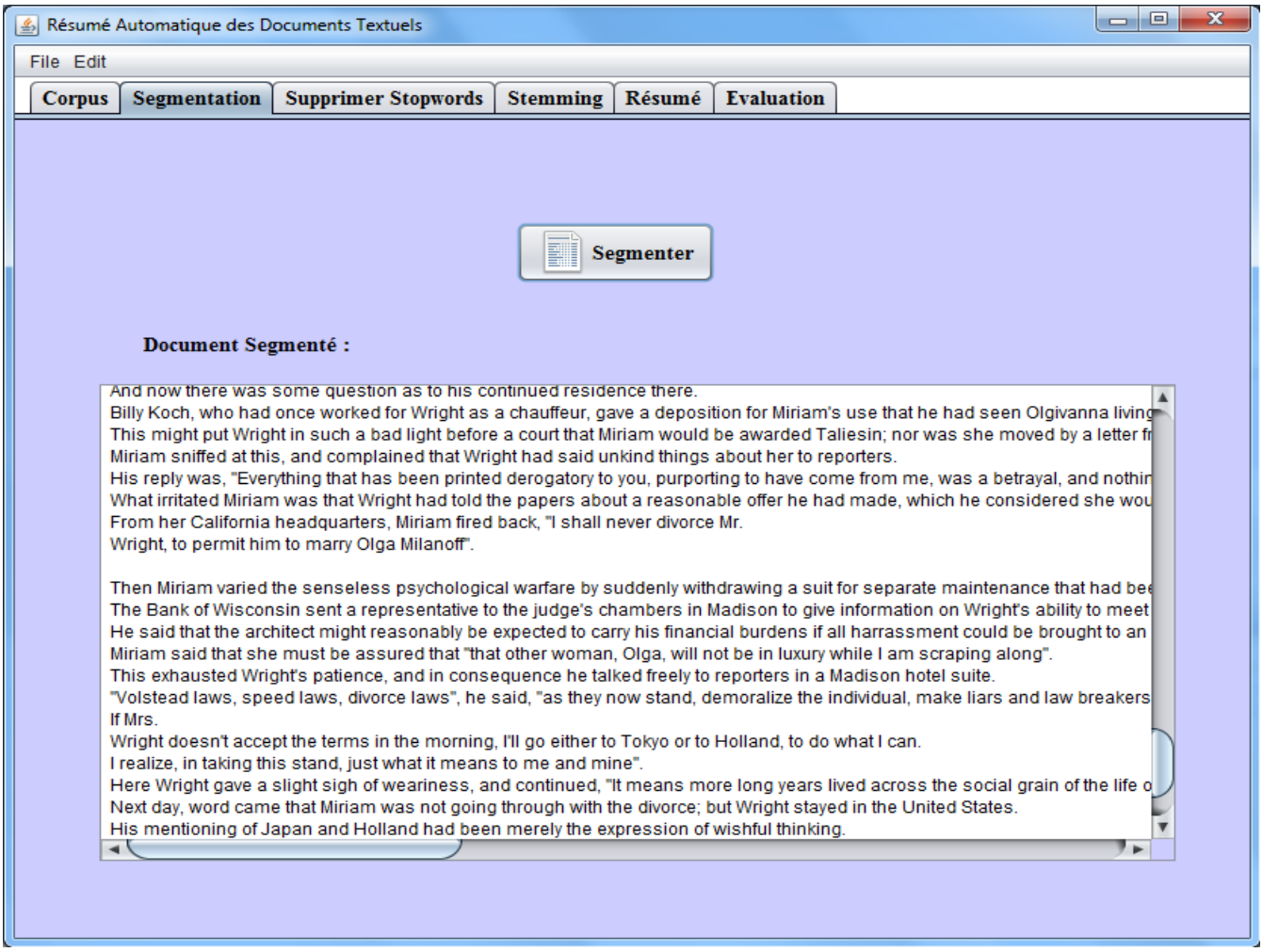

Figure 4.9 : Fenêtre de segmentation d'un document

Nous présentons dans les deux figures suivantes les résultats obtenues après la segmentation du document.

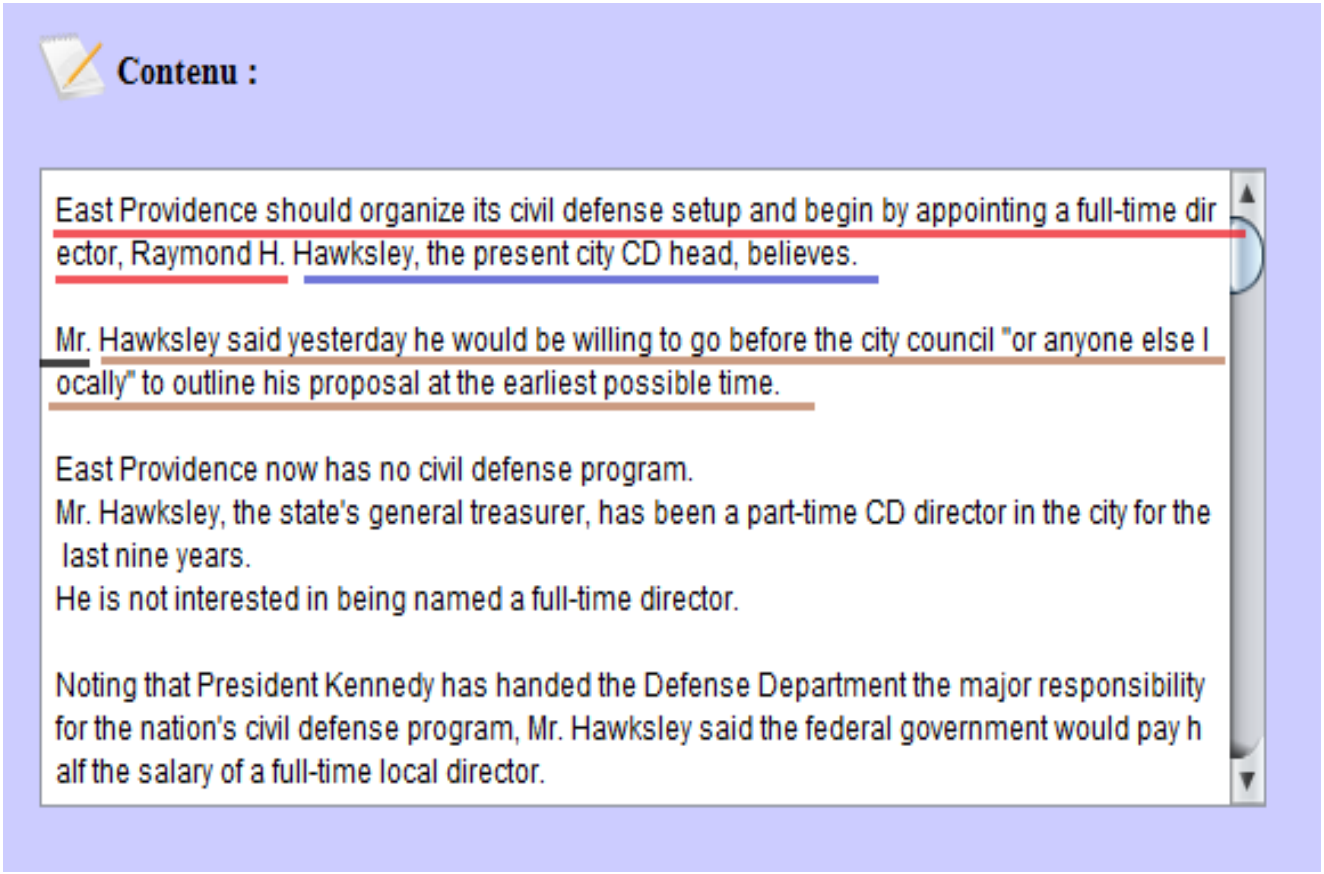

Figure 4.10 : Le document sans segmentation

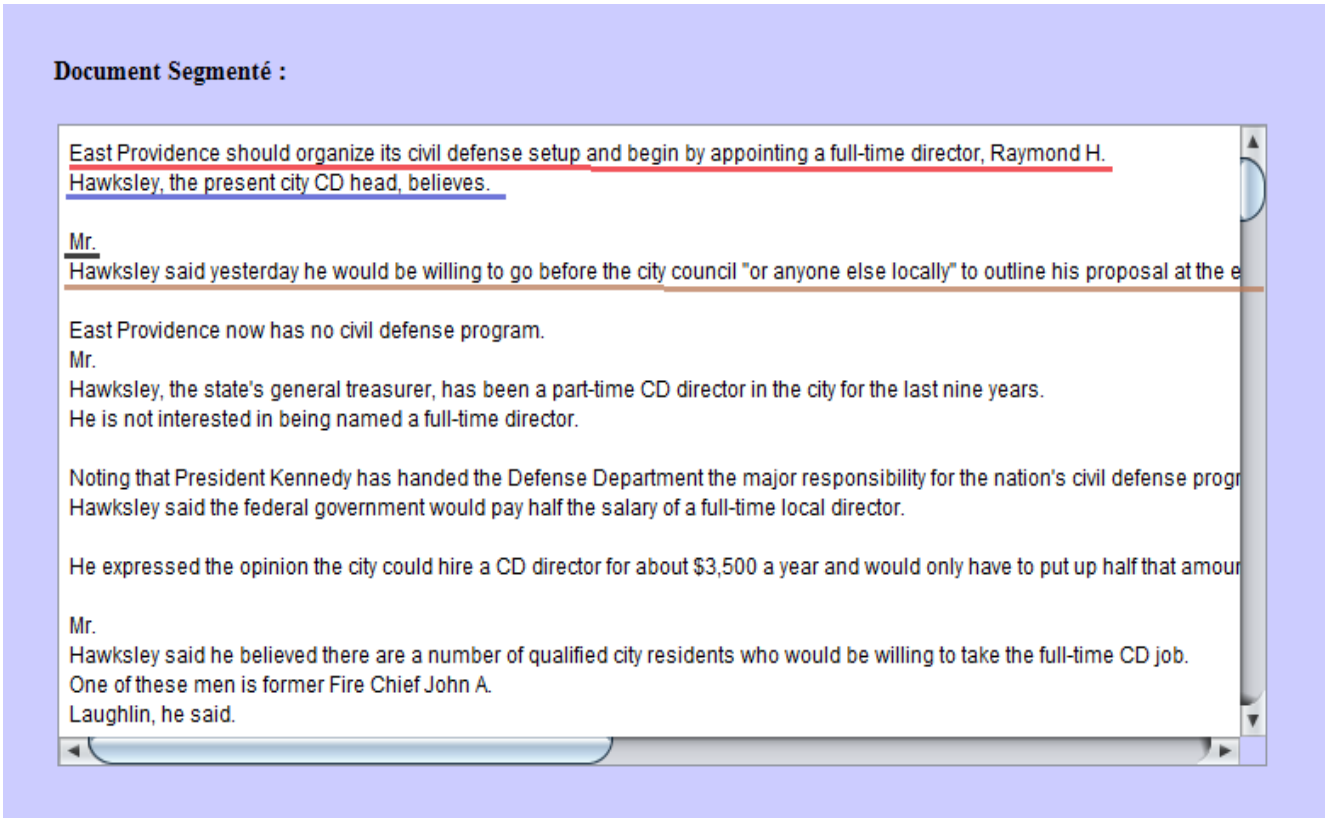

Figure 4.11 : Le résultat du document segmenté

3.1.3. Supprimer Mots-vides :

Nous présentons dans la figure suivante l'interface dédiée à la suppression des mots-vides d'un document.

Lorsqu'on choisit le document que nous voulons de le vider des mots inutiles, on clique sur le bouton **Supprimer**. Après ce traitement, un message d'information s'affiche sur l'interface indique que les mots-vides sont bien supprimés.

Figure 4.12: Fenêtre de la suppression des stop-words d'un document

Nous présentons dans les deux figures suivantes les résultats obtenues après l'élimination des mots non significatifs existants dans le document.

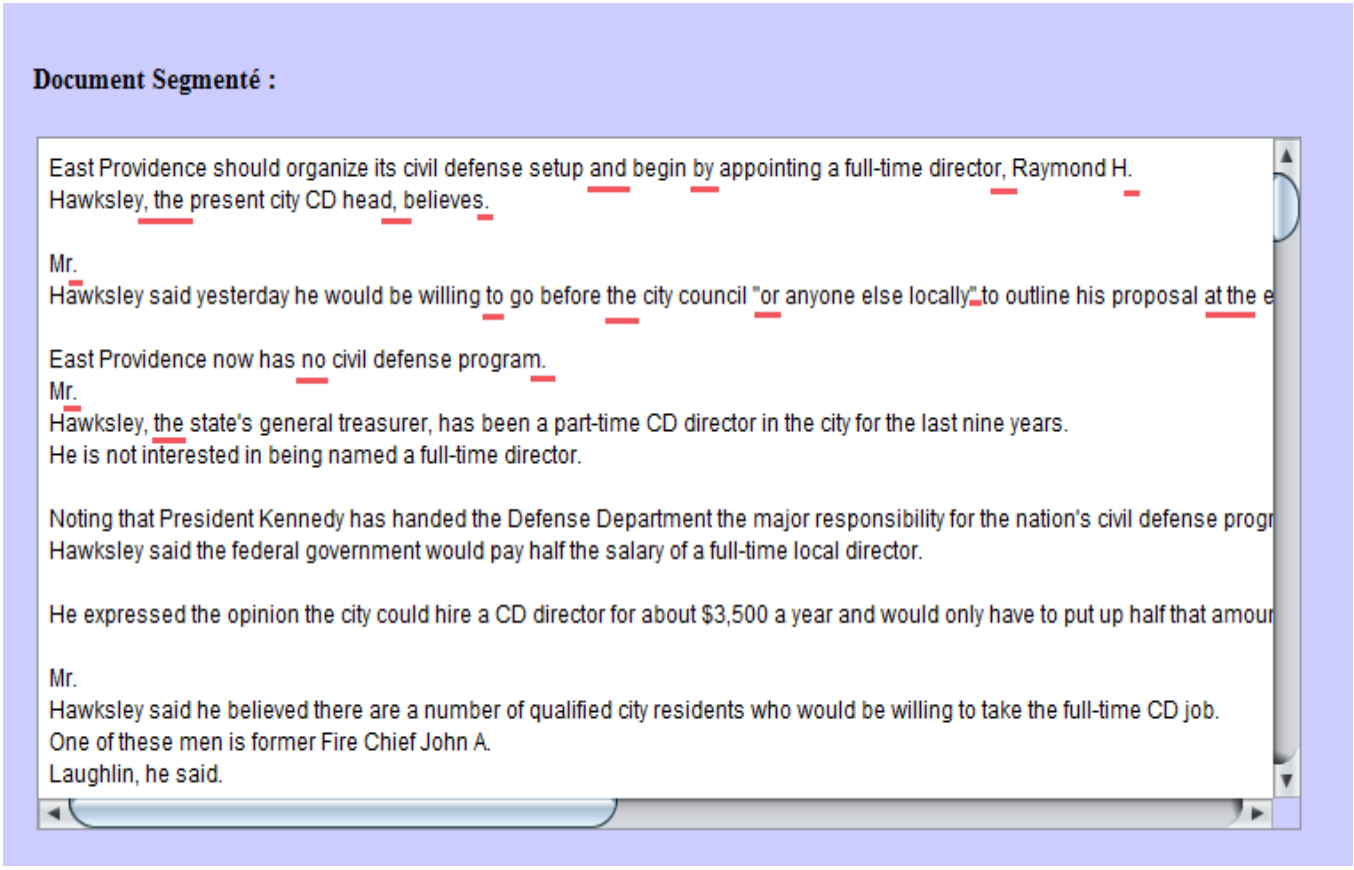

Figure 4.13 : Le document avec les stop-words

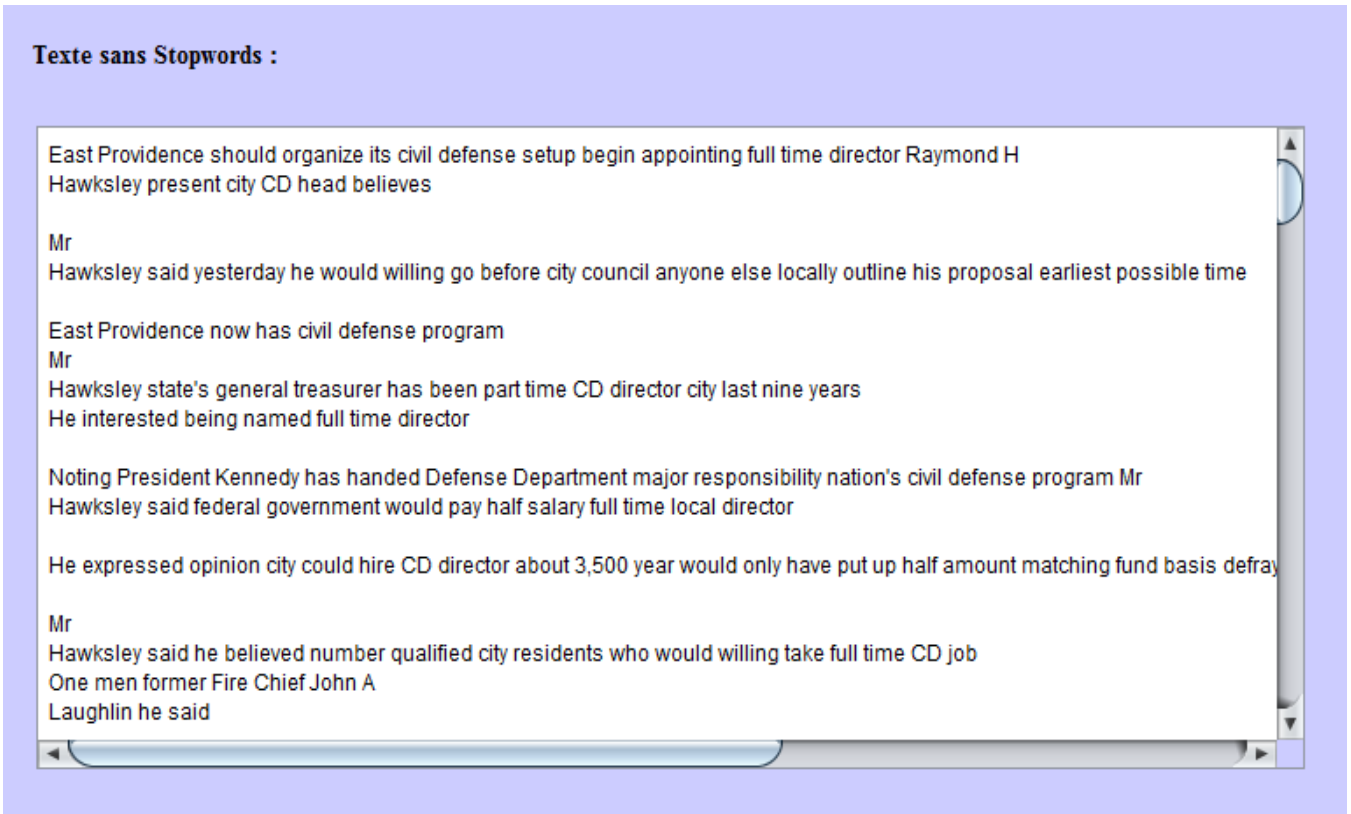

Figure 4.14 : Le document sans les stop-words

3.1.4. Stemming :

Nous présentons dans la figure suivante l'interface dédiée à le stemming d'un document :

Lorsqu'on choisit le document que nous voulons de transformer ces mots en leur forme canonique, on clique sur le bouton **Stemming**. Après ce traitement, un message d'information s'affiche sur l'interface indique que le document est normalisé.

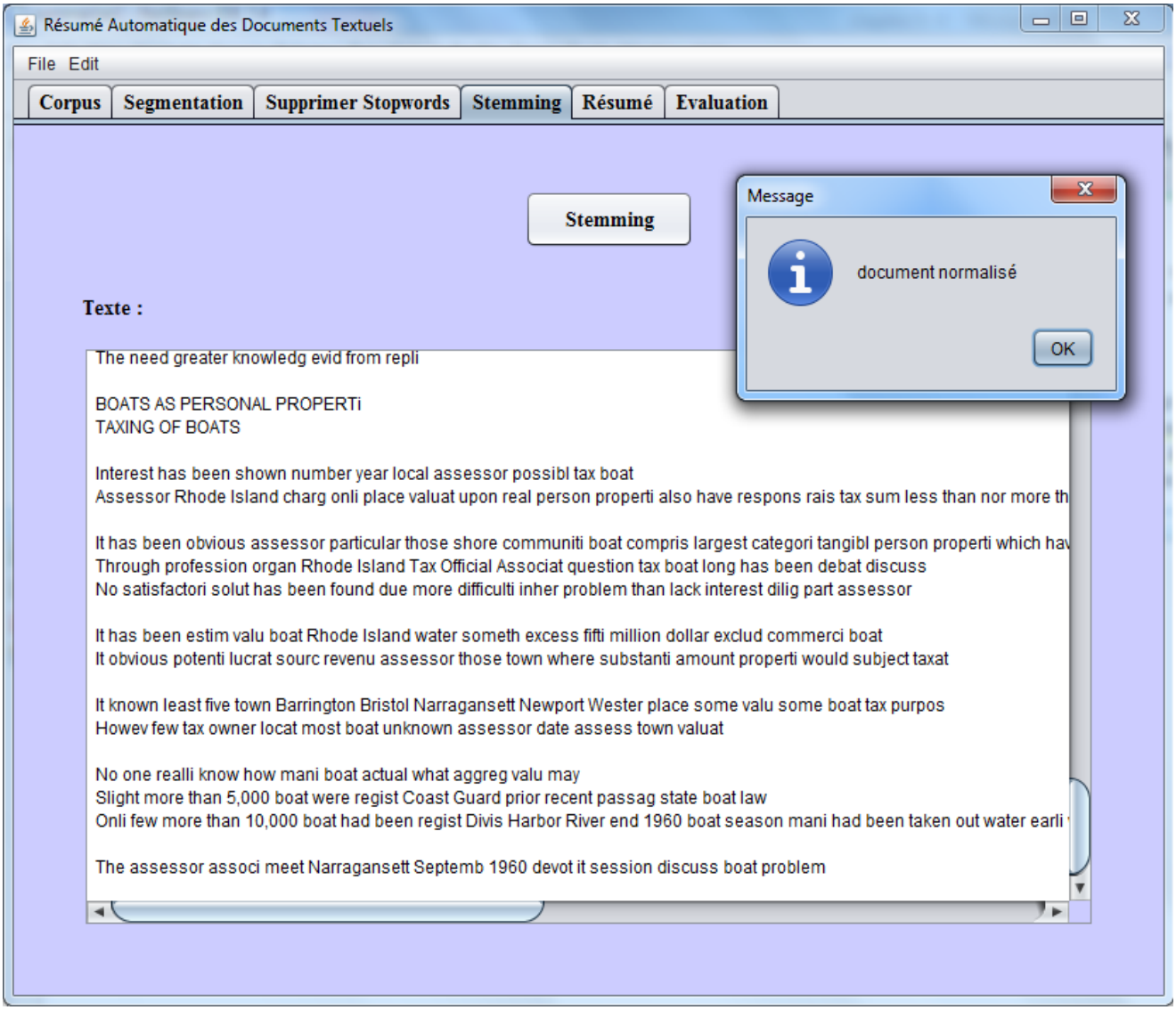

Figure 4.15 : Fenêtre du stemming

Nous présentons dans les deux figures suivantes les résultats obtenues après l'obtention de la forme canonique et originale de chaque mot existant dans le document.

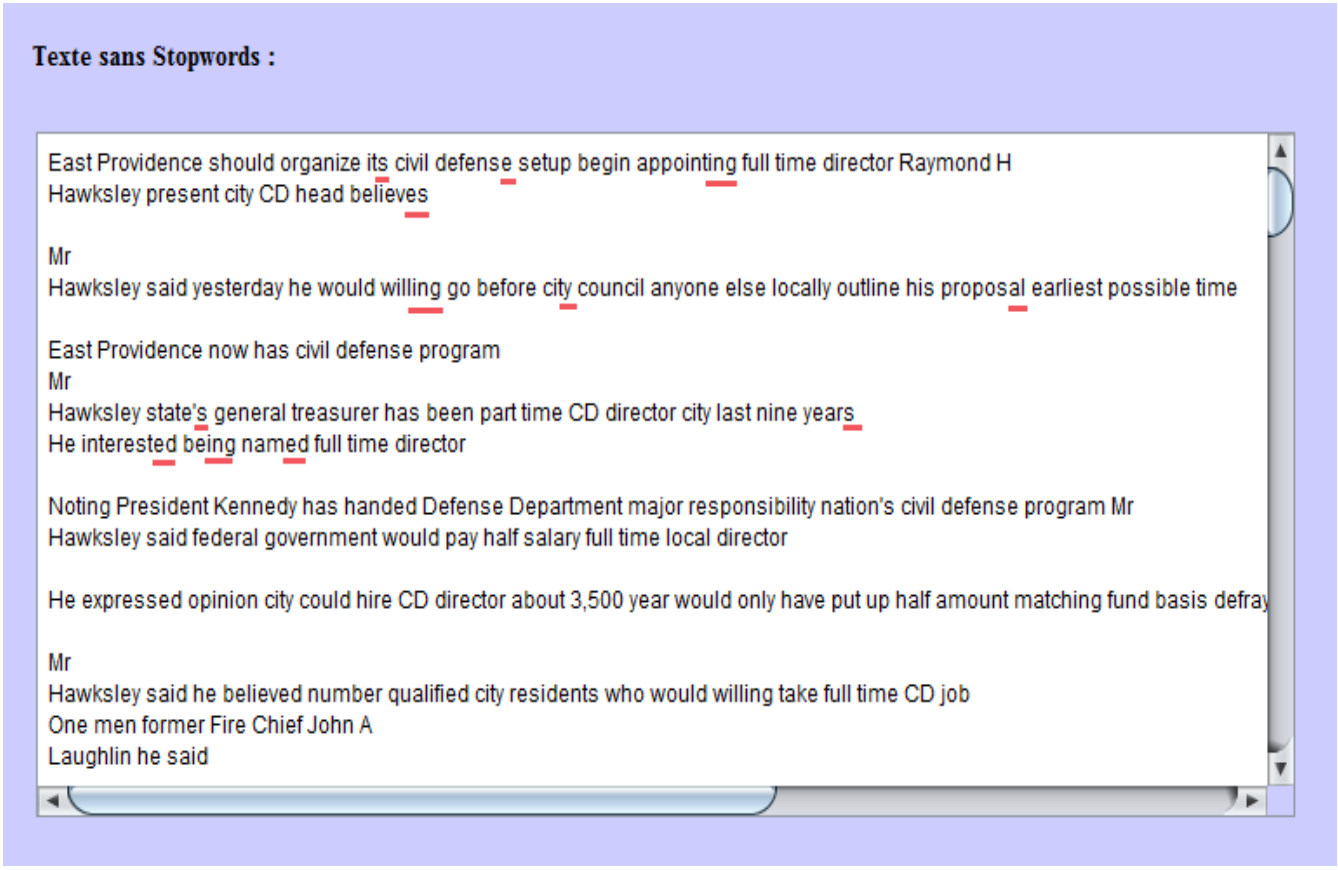

Figure 4.16 : Le document contenant les mots sans normalisation

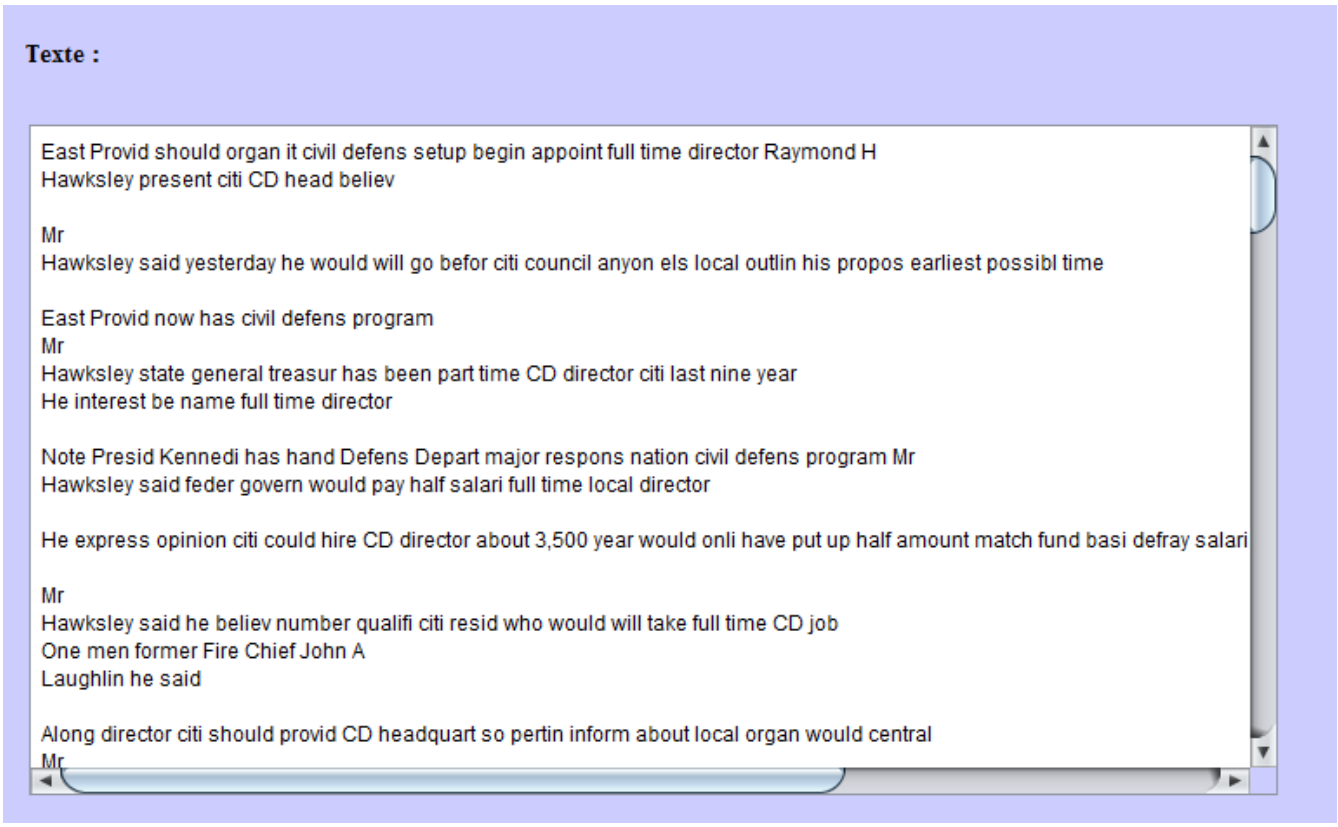

Figure 4.17 : Le document contenant les mots normalisés

3.1.5. Résumé :

Nous présentons dans les figures suivantes l'interface dédiée au résumé d'un document, chaque interface montre le résumé généré par les différentes méthodes utilisées et la combinaison entre eux:

La première interface présente un résumé généré par la méthode dépendant de la longueur de la phrase :

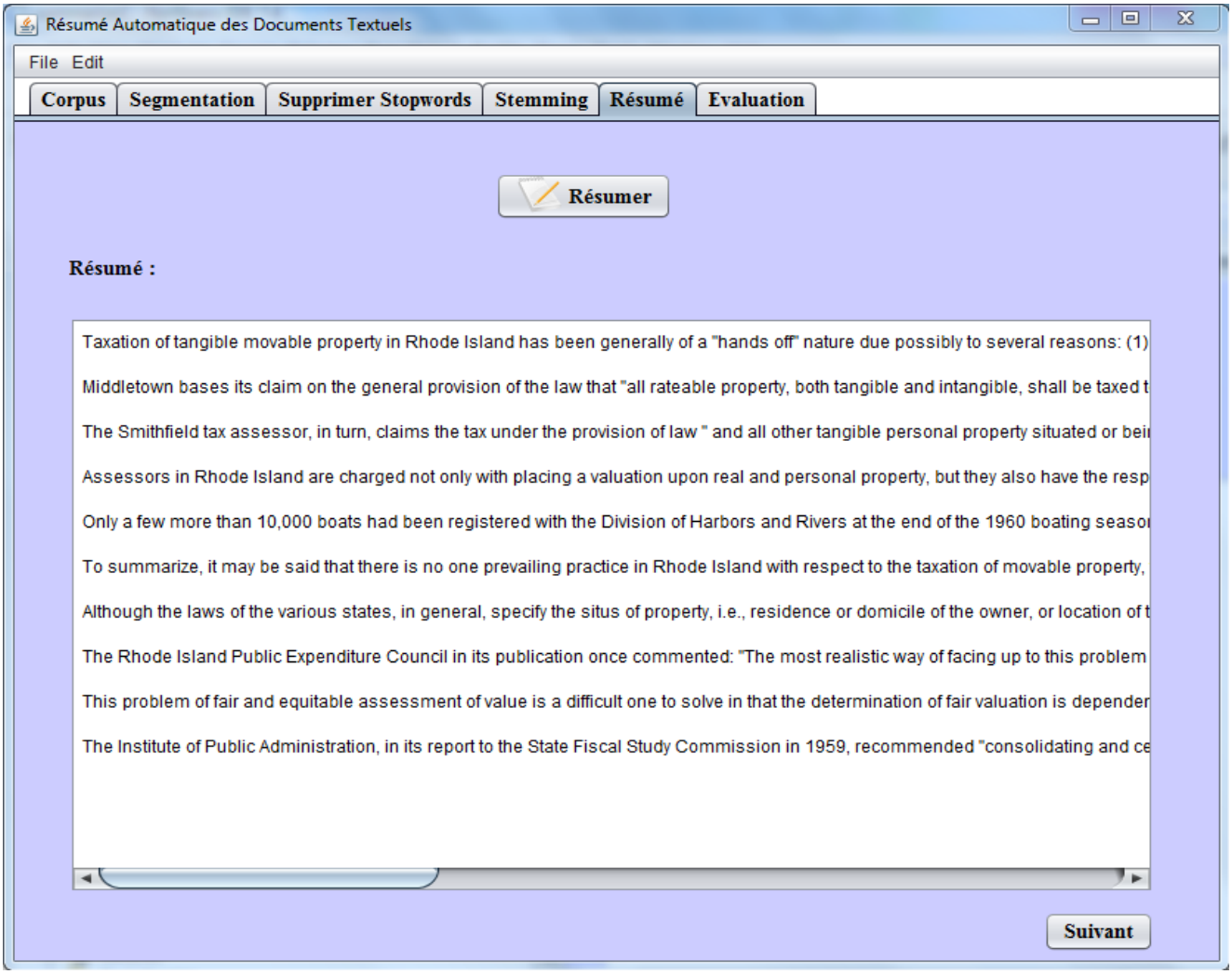

Figure 4.18 : Fenêtre de Résumé généré par la méthode dépendant de la longueur de la phrase

La deuxième interface présente un résumé généré par la méthode à base de la position de la phrase dans le document :

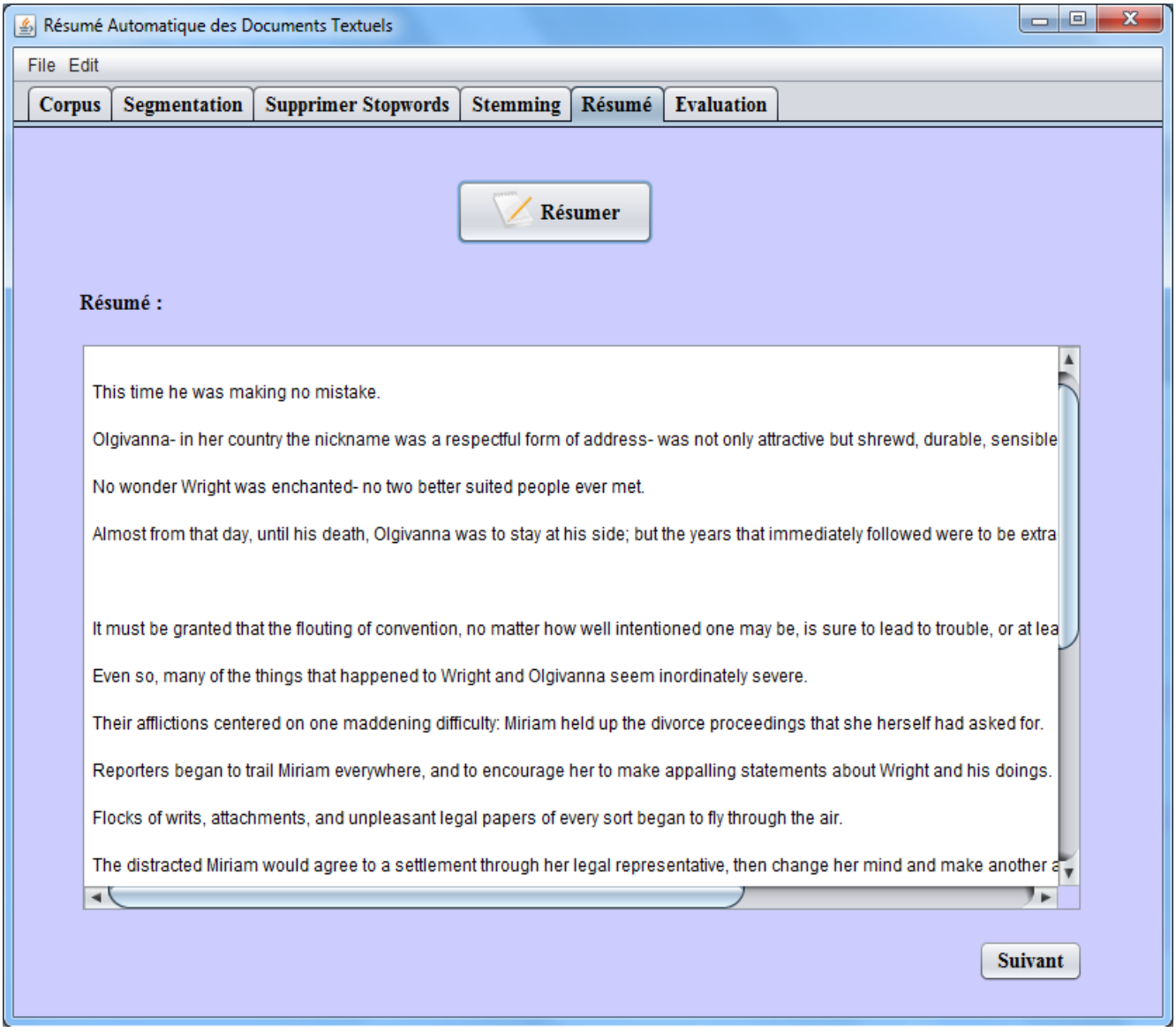

Figure 4.19 : Fenêtre de Résumé généré par la méthode à base de position de la phrase

La troisième interface présente un résumé généré par la méthode à base de mots-clés du document :

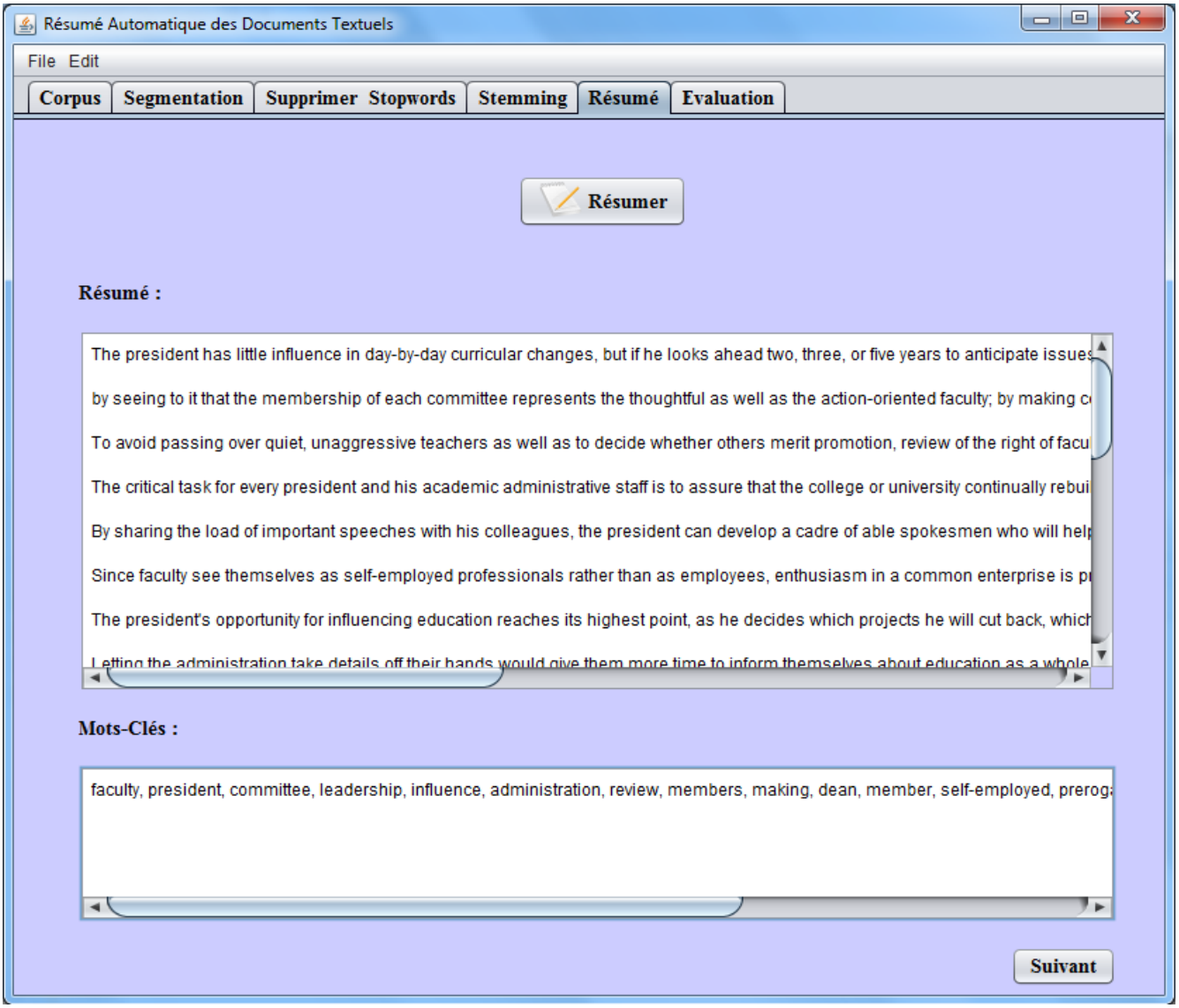

Figure 4.20 : Fenêtre de Résumé généré par la méthode à base de mots-clés du document

La quatrième interface présente un résumé généré par la méthode à base de mots-clés de la première phrase du document :

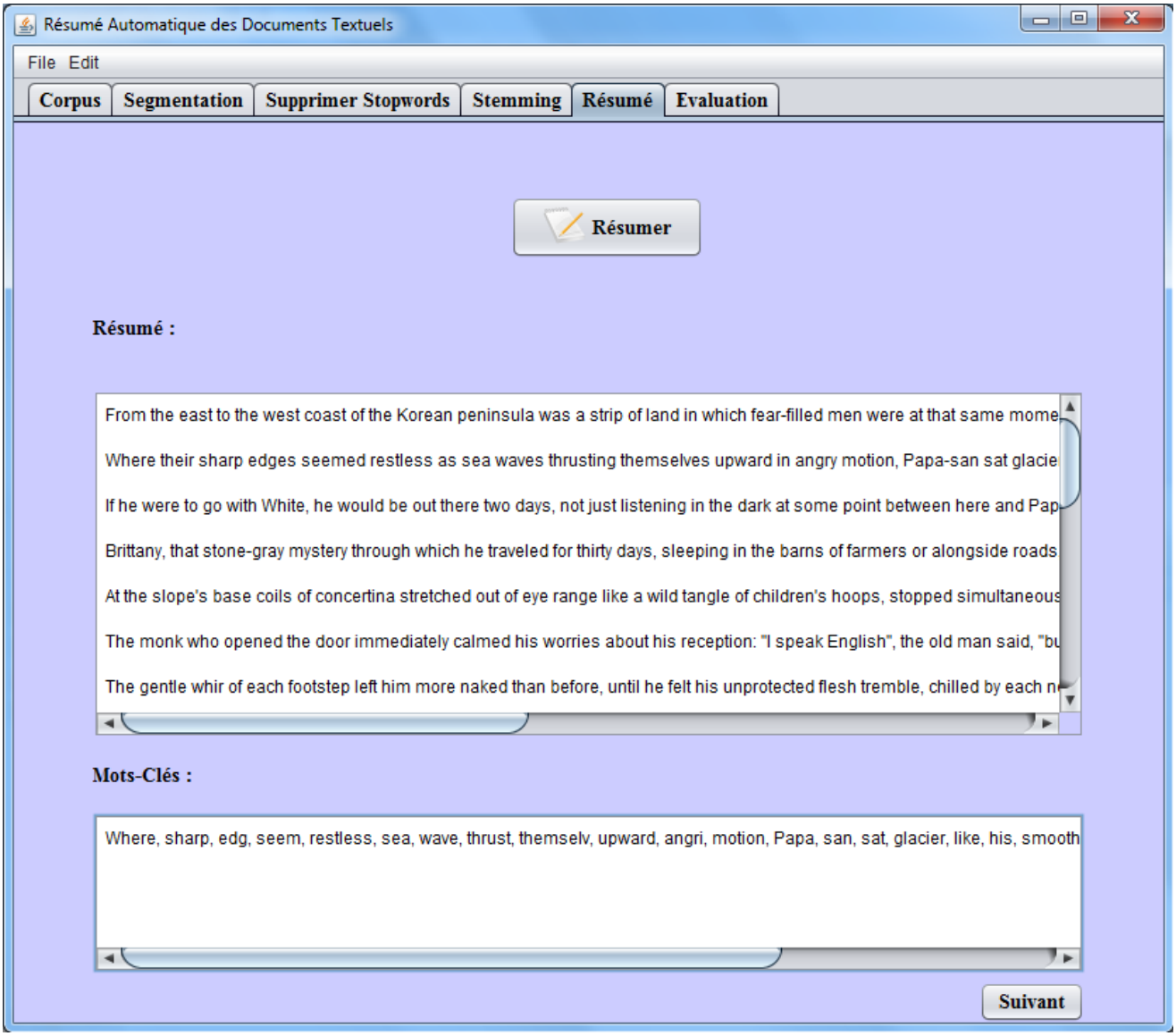

Figure 4.21 : Fenêtre de Résumé généré par la méthode à base de mots-clés de la première phrase du document

La dernière interface présente un résumé généré par la combinaison des trois premières méthodes :

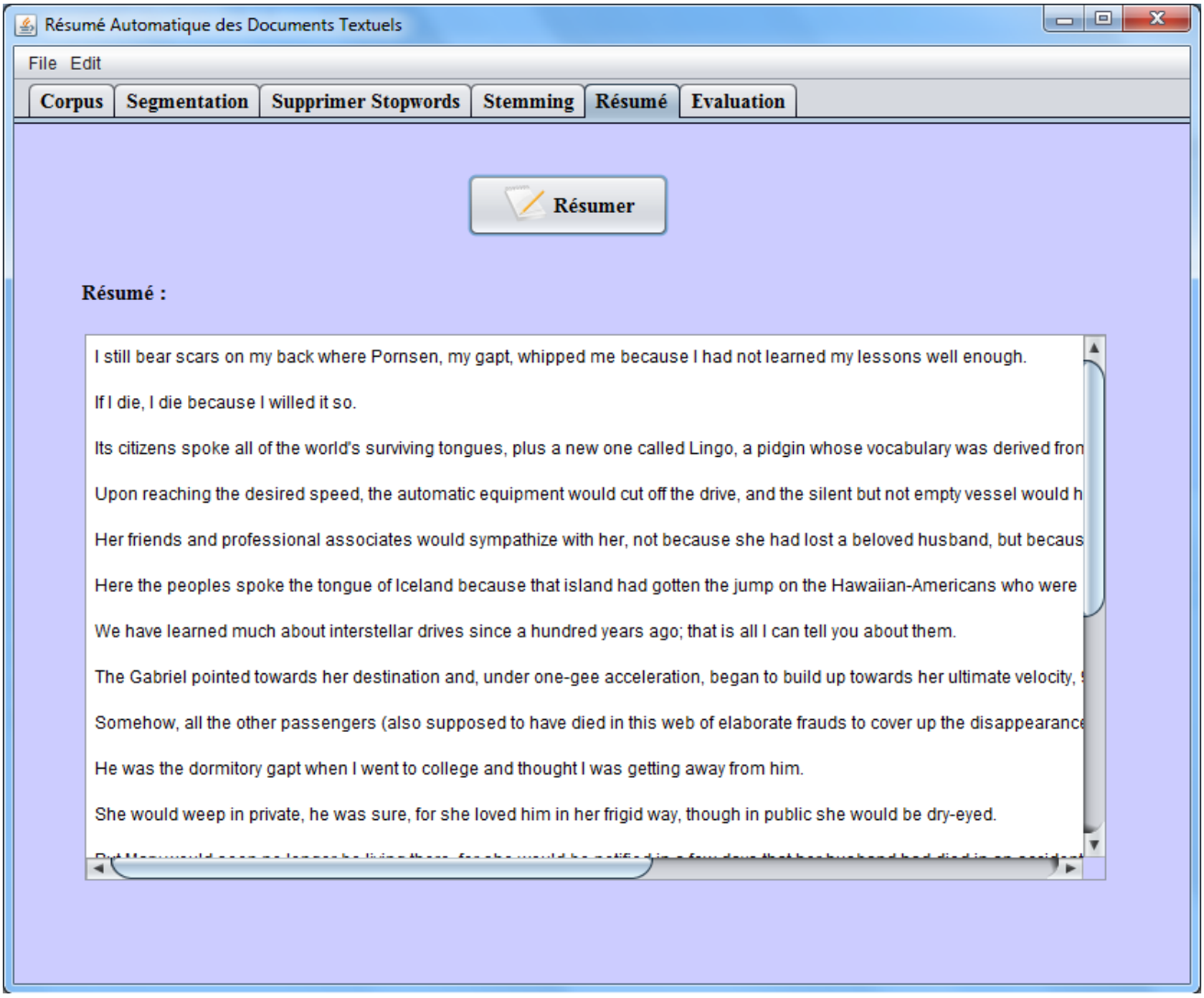

Figure 4.22 : Fenêtre de Résumé généré par la combinaison

Les figures suivantes présentent les résultats du chaque résumé généré par les quatre méthodes appliquées et la combinaison entre les trois premiers méthodes :

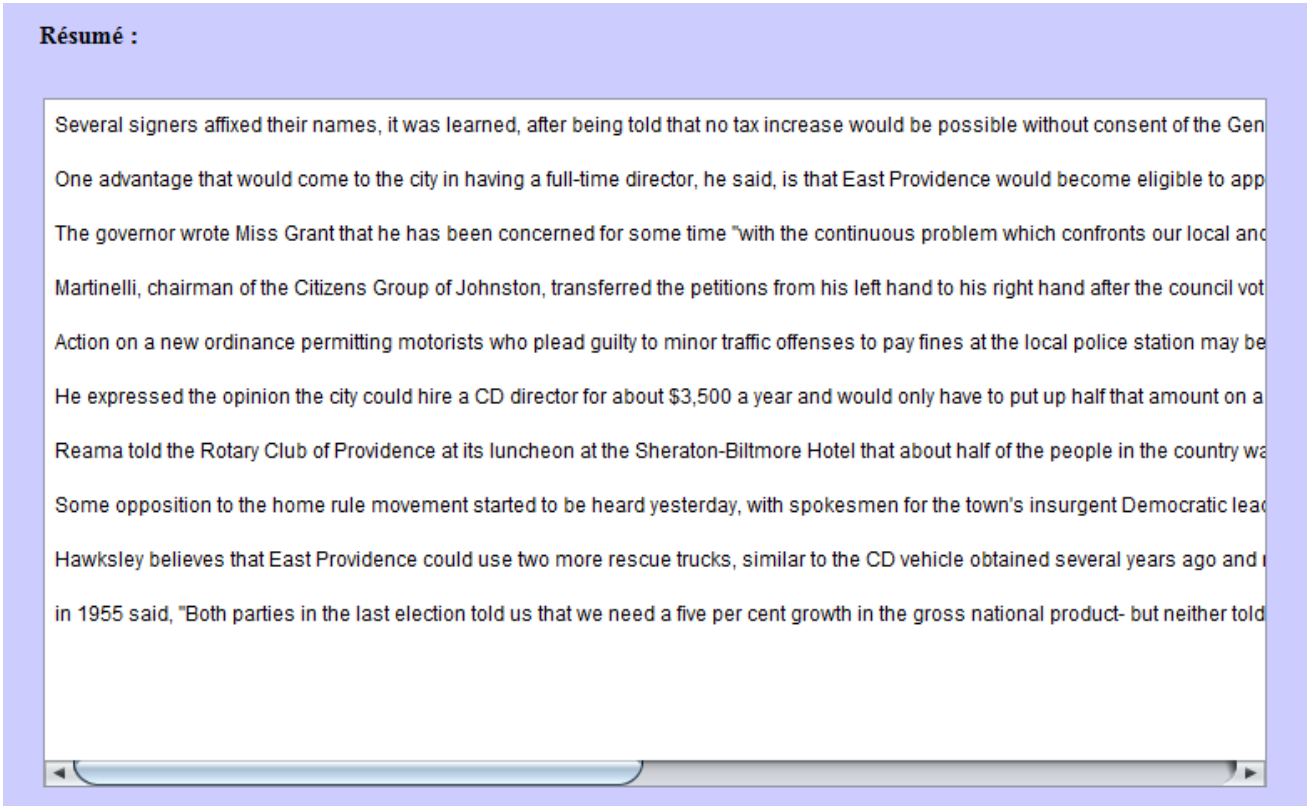

Figure 4.23 : Résultat d'un résumé d'un document généré par notre système (méthode1)

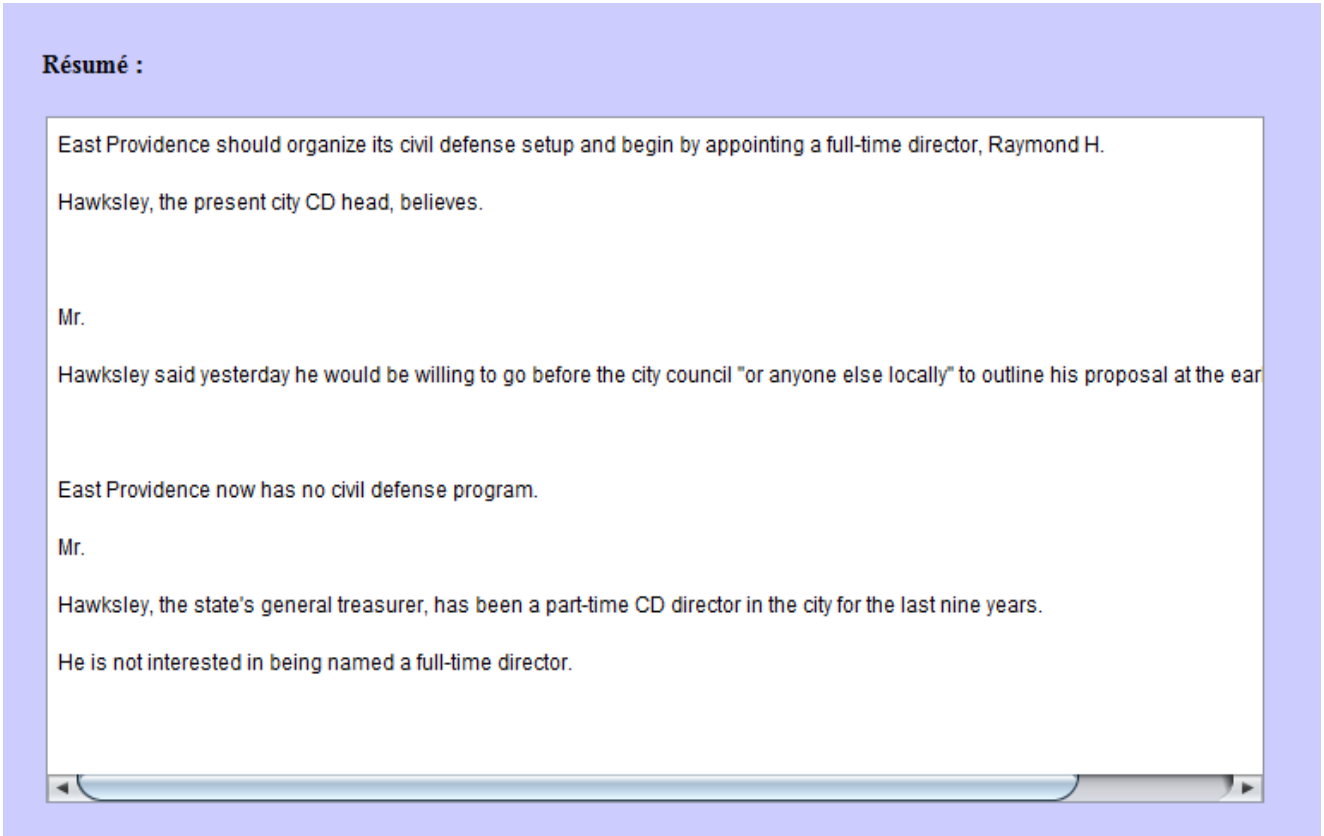

Figure 4.24 : Résultat d'un résumé du même document généré par notre système (méthode2)

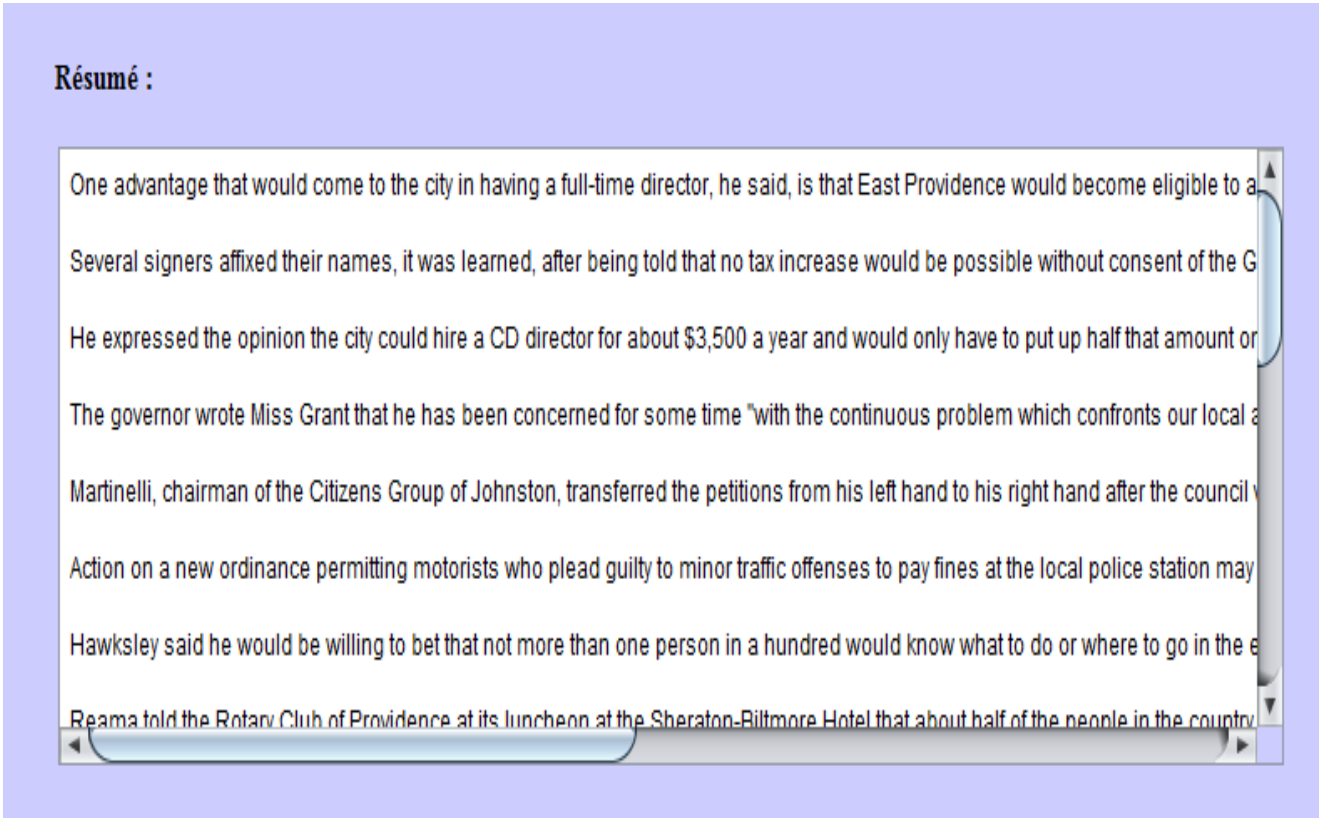

Figure 4.25 : Résultat d'un résumé du même document généré par notre système (méthode3)

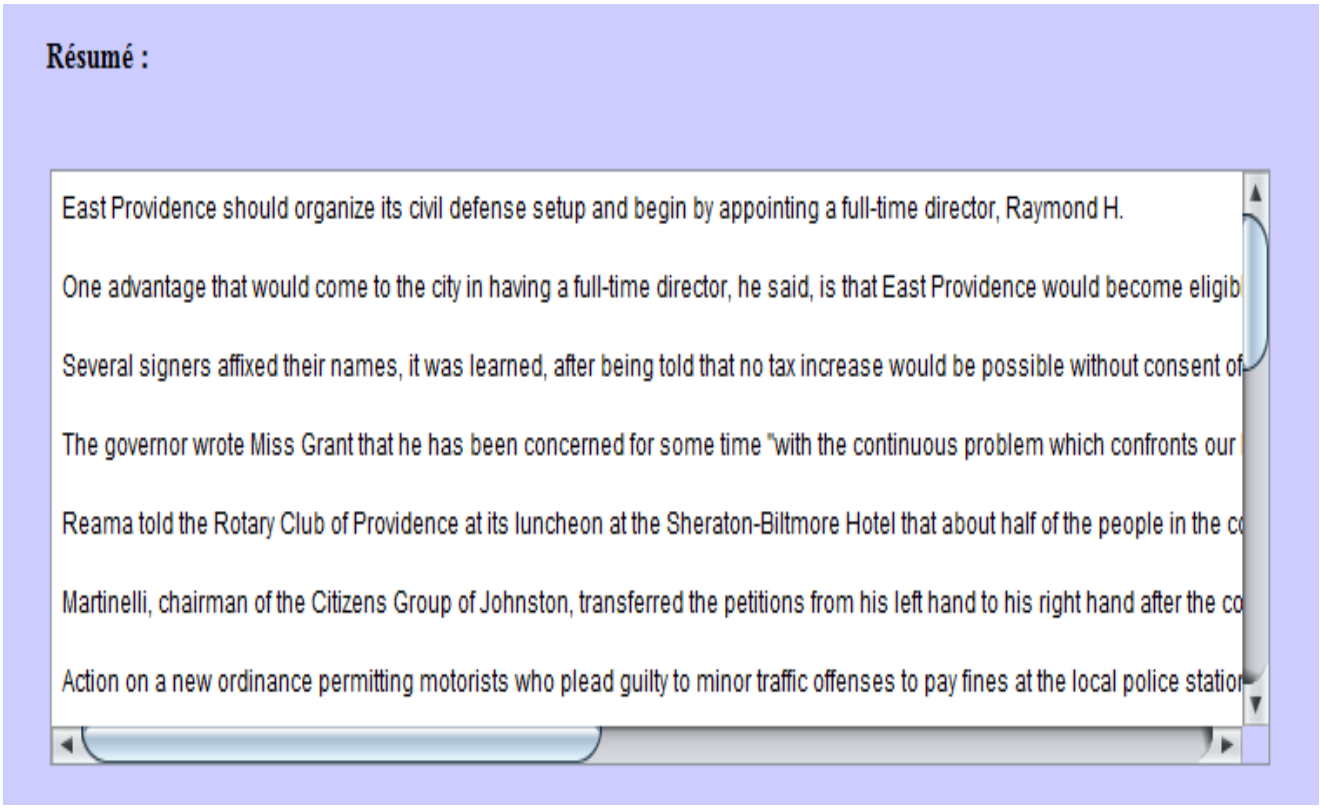

Figure 4.26 : Résultat d'un résumé du même document généré par notre système (méthode4)

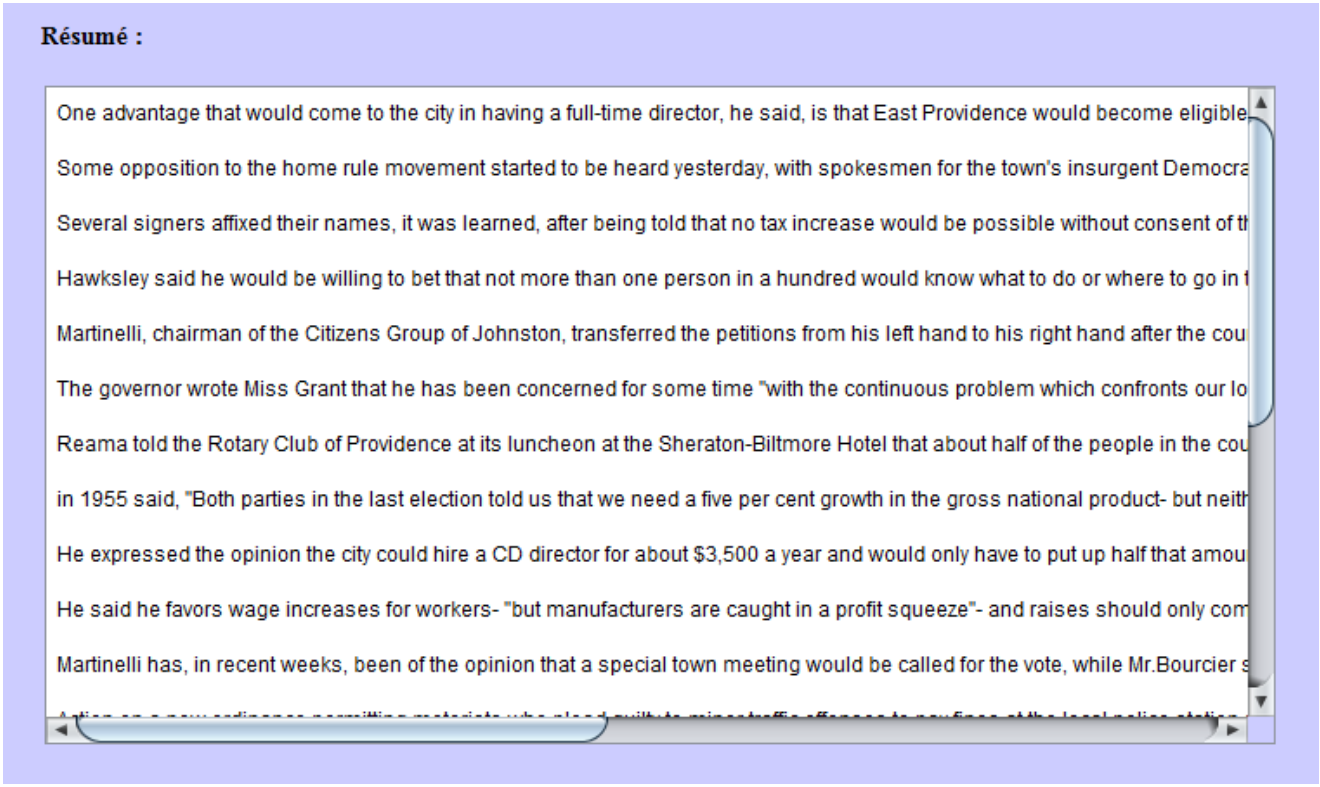

Figure 4.27 : Résultat d'un résumé du même document généré par la combinaison

La figure suivante présente un résultat d'un résumé d'un document par le système Intellexer Summarizer :

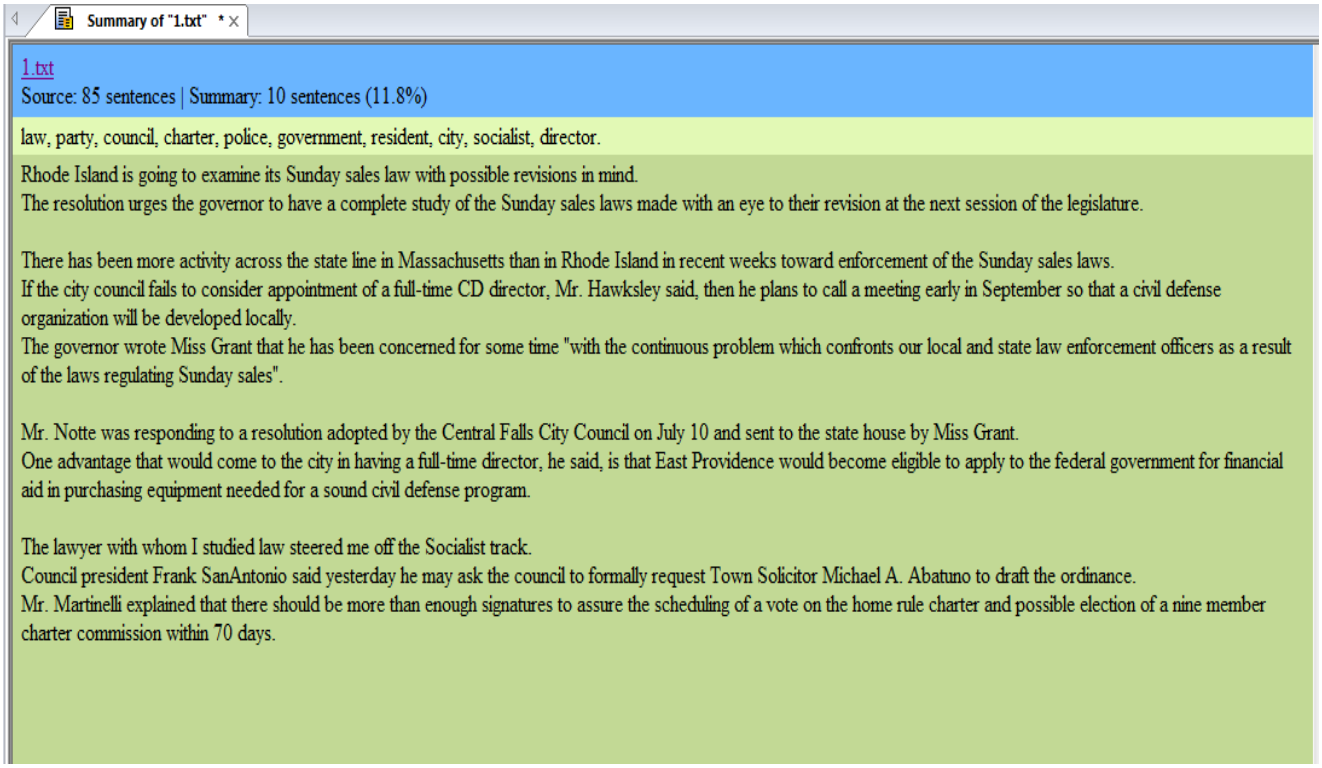

Figure 4.28 : Résultat d'un résumé du même document généré par Intellexer Summarizer

3.1.6. Evaluation :

Nous allons utiliser le corpus de jugement mentionné dans le chapitre précédent afin de tester notre système et les résultats des résumés générés par les différentes méthodes utilisées en calculant la mesure de similarité : rouge qui est basée sur le calcul des mesures suivantes : rappel, précision et F-score.

Nous présentons dans la figure suivante notre interface d'évaluation : l'utilisateur a la possibilité de sélectionner un corpus jugement d'une côté et de l'autre côté de sélectionner un corpus des documents résumés :

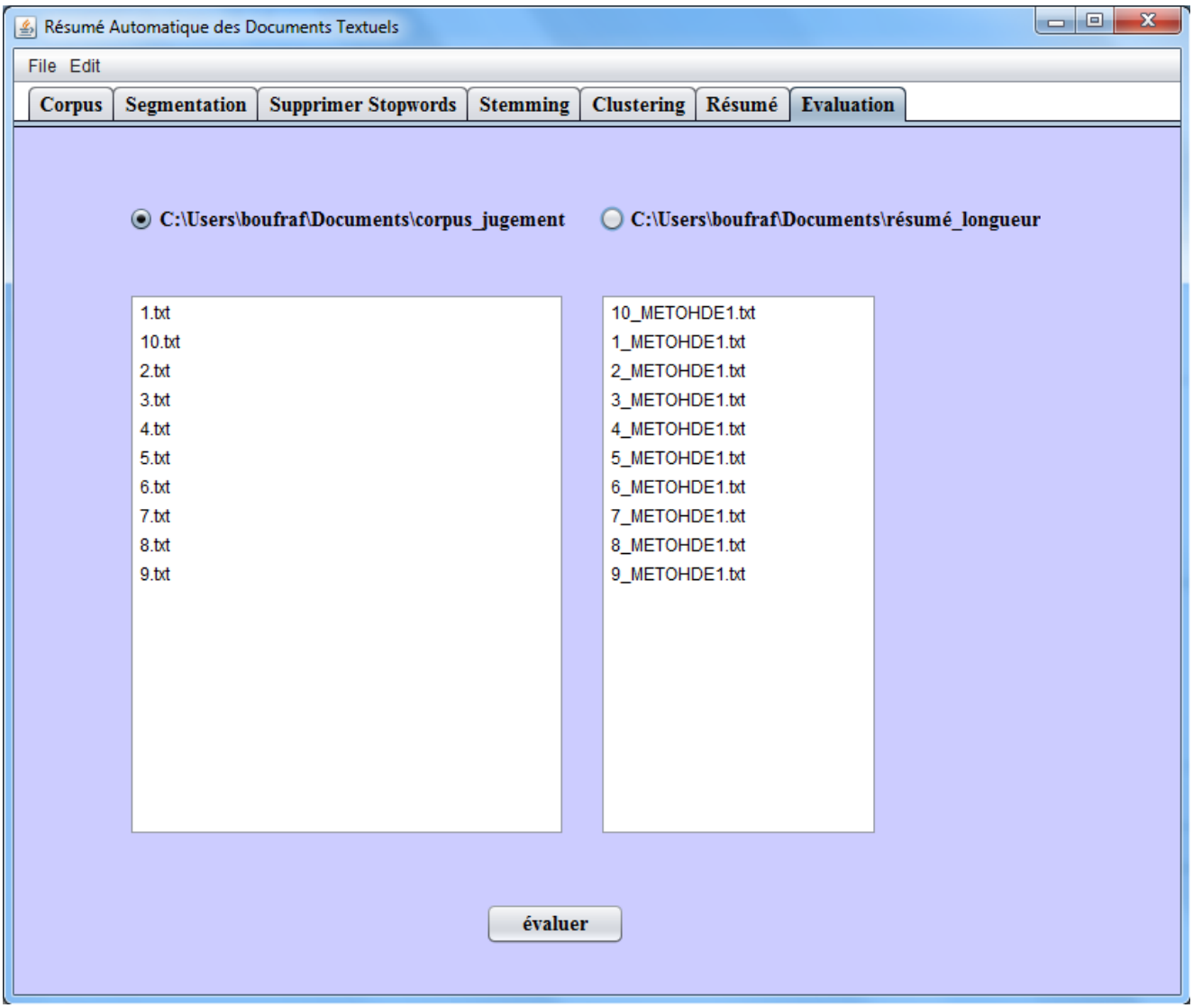

Figure 4.29 : Interface d'évaluation

- **Mesures d'évaluation :**

Dans nos expérimentations, nous nous intéressons en particulier aux mesures suivantes :

Rappel, Précision, calculés selon les équations définies dans le **Chapitre 2**, et F-score qui est calculée de la manière suivante :

F-score = 2 * Rappel * Précision / Rappel + Précision

ROUGE-N est un rappel de n-gramme entre un résumé de candidat et un ensemble de résumés de référence.

ROUGE-N est calculé comme suit:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

Où : **n** représente la longueur du n-gramme, **Count(gram$_n$)** et **Count$_{match}$ (gram$_n$)** est le nombre maximal de n-grammes co-occurrent dans un résumé candidat et un ensemble de résumés de référence. Il est clair que **ROUGE-N** est une mesure de rappel parce que le dénominateur de l'équation (1) est la somme totale du nombre de n-grammes qui se produisent à la côté du système de référence.

Les figures suivantes représentent les résultats obtenues après l'évaluation des résumés générés par notre système pour chaque méthode utilisée avec les résumés obtenus par le système « Intellexer Summarizer » :

	A	B	C	D	E	F
1	ROUGE-Type	System Name	Rappel	Precision	F-Score	Num Reference Summaries
2	ROUGE-1	METOHDE1.TXT	0,41772	0,34555	0,37822	1
3	ROUGE-1	METOHDE1.TXT	0,55385	0,29752	0,38710	1
4	ROUGE-1	METOHDE1.TXT	0,38235	0,23214	0,28889	1
5	ROUGE-1	METOHDE1.TXT	0,24806	0,18079	0,20915	1
6	ROUGE-1	METOHDE1.TXT	0,33951	0,23913	0,28061	1
7	ROUGE-1	METOHDE1.TXT	0,36364	0,21667	0,27154	1
8	ROUGE-1	METOHDE1.TXT	0,15686	0,09959	0,12183	1
9	ROUGE-1	METOHDE1.TXT	0,30952	0,16596	0,21607	1
10	ROUGE-1	METOHDE1.TXT	0,48031	0,29048	0,36202	1
11	ROUGE-1	METOHDE1.TXT	0,32192	0,19106	0,23980	1

Figure 4.30 : Résultats de l'évaluation des résumés générés par la méthode1

	A	B	C	D	E	F
1	ROUGE-Type	System Name	Rappel	Precision	F-Score	Num Reference Summaries
2	ROUGE-1	METHODE2.TXT	0,25316	0,35088	0,29412	1
3	ROUGE-1	METHODE2.TXT	0,40769	0,27041	0,32515	1
4	ROUGE-1	METHODE2.TXT	0,30147	0,36937	0,33198	1
5	ROUGE-1	METHODE2.TXT	0,23256	0,19231	0,21053	1
6	ROUGE-1	METHODE2.TXT	0,27160	0,25287	0,26190	1
7	ROUGE-1	METHODE2.TXT	0,43357	0,30392	0,35735	1
8	ROUGE-1	METHODE2.TXT	0,54902	0,32061	0,40482	1
9	ROUGE-1	METHODE2.TXT	0,53968	0,32692	0,40719	1
10	ROUGE-1	METHODE2.TXT	0,51969	0,36872	0,43137	1
11	ROUGE-1	METHODE2.TXT	0,41781	0,40940	0,41356	1

Figure 4.31 : Résultats de l'évaluation des résumés générés par la méthode2

	A	B	C	D	E	F
1	ROUGE-Type	System Name	Rappel	Precision	F-Score	Num Reference Summaries
2	ROUGE-1	METHODE3.TXT	0,39873	0,38182	0,39009	1
3	ROUGE-1	METHODE3.TXT	0,46923	0,26872	0,34174	1
4	ROUGE-1	METHODE3.TXT	0,51471	0,33333	0,40462	1
5	ROUGE-1	METHODE3.TXT	0,17829	0,12432	0,14650	1
6	ROUGE-1	METHODE3.TXT	0,32716	0,24651	0,28117	1
7	ROUGE-1	METHODE3.TXT	0,25874	0,18408	0,21512	1
8	ROUGE-1	METHODE3.TXT	0,19608	0,12448	0,15228	1
9	ROUGE-1	METHODE3.TXT	0,34127	0,19907	0,25146	1
10	ROUGE-1	METHODE3.TXT	0,46457	0,28502	0,35329	1
11	ROUGE-1	METHODE3.TXT	0,32192	0,19583	0,24352	1

Figure 4.32 : Résultats de l'évaluation des résumés générés par la méthode3

	A	B	C	D	E	F
1	ROUGE-Type	System Name	Rappel	Precision	F-Score	Num Reference Summaries
2	ROUGE-1	METHODE4.TXT	0,41772	0,35676	0,38484	1
3	ROUGE-1	METHODE4.TXT	0,55385	0,29752	0,38710	1
4	ROUGE-1	METHODE4.TXT	0,38235	0,23214	0,28889	1
5	ROUGE-1	METHODE4.TXT	0,24806	0,18079	0,20915	1
6	ROUGE-1	METHODE4.TXT	0,33951	0,23913	0,28061	1
7	ROUGE-1	METHODE4.TXT	0,36364	0,21667	0,27154	1
8	ROUGE-1	METHODE4.TXT	0,15686	0,09959	0,12183	1
9	ROUGE-1	METHODE4.TXT	0,30952	0,16596	0,21607	1
10	ROUGE-1	METHODE4.TXT	0,48031	0,29048	0,36202	1
11	ROUGE-1	METHODE4.TXT	0,32192	0,19106	0,23980	1

Figure 4.33 : Résultats de l'évaluation des résumés générés par la méthode4

	A	B	C	D	E	F
1						
2	ROUGE-Type	System Name	Rappel	Precision	F-Score	Num Reference Summaries
3	ROUGE-1	COMBINAISON.TXT	0,57595	0,27164	0,36917	1
4	ROUGE-1	COMBINAISON.TXT	0,63077	0,20707	0,31179	1
5	ROUGE-1	COMBINAISON.TXT	0,66176	0,24793	0,36072	1
6	ROUGE-1	COMBINAISON.TXT	0,41860	0,16265	0,23427	1
7	ROUGE-1	COMBINAISON.TXT	0,47531	0,20479	0,28625	1
8	ROUGE-1	COMBINAISON.TXT	0,53846	0,20588	0,29787	1
9	ROUGE-1	COMBINAISON.TXT	0,34641	0,12156	0,17997	1
10	ROUGE-1	COMBINAISON.TXT	0,61111	0,20588	0,30800	1
11	ROUGE-1	COMBINAISON.TXT	0,69291	0,23978	0,35628	1
12	ROUGE-1	COMBINAISON.TXT	0,56164	0,19385	0,28822	1
13						

Figure 4.34 : Résultats de l'évaluation des résumés générés par la combinaison des méthodes

Discussion :

D'après les résultats obtenus nous constatons que la valeur de la colonne « Précision » qui montre la valeur de notre système, d'une part, change d'une méthode à une autre et d'autre part change d'un résumé à un autre.

D'après ses valeurs, nous remarquons que la méthode à base de la position a eu les meilleures valeurs par rapport aux méthodes pour chaque résumé généré, et la combinaison a eu les valeurs les plus basses.

On déduit que cette méthode est mieux pour générer des résumés des textes automatiquement par ce qu'elle permet de déterminer les phrases importantes en extrayant celles qui sont en tête dont le résumé généré est plus proche au document original.

IV.4. Conclusion

Dans ce chapitre on a présenté l'implémentation de notre application, l'objectif principal de cette implémentation est de produire des résumés à partir de la combinaison des différentes méthodes utilisées, ainsi l'évaluation de notre système avec le système de référence intégré en commentant les résultats obtenus.

Conclusion Générale

La notion de résumé automatique devient un des grands thèmes du Traitement Automatique des Langues. Plutôt que de diffuser les documents entiers, n'est-il pas préférable de diffuser seulement les résumés qui contiendraient les informations vraiment pertinentes ? En effet, il est plus facile de lire quelques lignes ou quelques pages pour s'apercevoir qu'aucune information nouvelle ne s'y trouve. Un document textuel devra donc être maintenant géré en même temps que son résumé qui sera, par ailleurs, un des moyens d'accès au contenu du document. Notre travail s'inscrit dans le cadre l'amélioration du résumé automatique des documents textuels. Ainsi un système de résumé automatique doit prendre en considération ses caractéristiques et proposer des outils et des techniques afin de permettre son fonctionnement.

Dans le premier chapitre nous avons défini certaines caractéristiques de résumé classique, le résumé d'une façon générale est la réduction d'un texte en un nombre limité de mots sachant qu'on garde les informations et les idées principales du texte original. Nous avons présenté aussi les concepts clés de la recherche d'information et du traitement automatique des langues.

Dans le second chapitre nous avons étudié l'état de l'art du résumé automatique en citant quelques méthodes et quelques étapes pour développer un système de résumé automatique ainsi qu'une étude sur l'évaluation de ce système au moyen de quelques évaluations pour le comparer avec d'autres techniques de production de résumé automatique.

Dans le troisième chapitre nous avons abordé les aspects de développement de notre solution. Nous avons décrit l'architecture et l'approche de développement du notre système de résumé automatique des documents textuels.

Le dernier chapitre présente l'implémentation de notre application,nous avons expliqué les différentes étapes de l'implémentation et d'expérimentation de notre système ainsi que les résultats obtenus.

Nous avons décrit notre système, dont l'objectif est d'élaborer un système capable de résumer automatiquement des textes.

Les méthodes d'extraction se sont avérées adaptables et nous avons pu faire nos expérimentations sur le corpus de jugement qui regroupe un ensemble des résumés de notre corpus de test générés par le système « Intellexer Summarizer ».

La méthode à base de position est mieux pour générer des résumés des textes automatiquement par ce qu'elle permet de déterminer les phrases importantes en extrayant celles qui sont en tête dont le résumé généré est plus proche au document original.

Bibliographie :

[1] Juan, Manuel Torres, Moreno, Livre : Résumé automatique de documents, une approche statique.

[2] MR. ABDELKRIM BOURAMOUL, Thèse pour obtenir le grade de Docteur en Sciences : RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE SUR LE WEB, Année Universitaire : 2010/2011

[4] Fouad Soufiane Douzidia, Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique, Thème : Résumé automatique de texte arabe, Septembre, 2004.

[5] Ricco Rakotomalala, Cours : Introduction au Text Mining, Principes et Application.

[6] Saidi Imene, Thèse en vue de l'obtention du diplôme du Doctorat 3ème cycle (LMD), Thème : Contributions aux techniques de recherche d'information, Année Universitaire : 2014/2015.

[7] Ould Hadri Imene Mansouria, MEMOIRE DE FIN D'ETUDES Pour l'Obtention du Diplôme de Master en Informatique, Thème : Conception et intégration d'un analyseur morphologique arabe dans un moteur de recherche, Année Universitaire 2015/2016.

[12] ARIES Abdelkrime, Mémoire pour l'obtention du Magister de l' Ecole Nationale Supérieure d'Informatique (ESI), Thème : Résumé automatique des textes, le 26/06/2013.

[13] Maâli Mnasri, (1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, F-91191 France. (2) Univ. Paris Sud, Orsay, France.

Webographie :

[3] http://www.technolangue.net/imprimer.php3?id_article=274 visité le 01/12/2016.

[7] http://www.quentinfily.fr/tf-idf-pertinence-lexicale/ visité le 01/12/2016.

[8] http://ldelafosse.pagesperso-orange.fr/Glossaire/Tal.htm visité le 01/12/2016.

[9] http://fis.ucalgary.ca/Brian/ecrire/e-resume.htm visité le 19/11/2016.

[10] https://www.mpl.ird.fr/documentation/indexation/resume.htm visité le 19/11/2016.

[11] https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2004-3-page-200.htm visité le 16/11/2016.

[14] http://rali.iro.umontreal.ca/rali/?q=fr/Resume%20automatique visité le 18/11/2016.

[15] https://www.java.com/fr/download/faq/whatis_java.xml visité le 02/02/2017.

[16] https://netbeans.org/index_fr.html visité le 02/02/2017.

[17] http://summarizer.intellexer.com/ visité le 18/04/2017.