Université Saad DAHLAB - Blida 1

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

Mlles. BOUDELLA Aicha Sirine et BENHALIMA Nour El Houda

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Traitement Automatique de la Langue

Sujet :

Conception et Réalisation d'un Système Multilingue de Restitution du Sommaire des Documents Numérisés

Soutenu le : 24-11-2020 devant le jury composé de

Dr M. Mezzi	Université blida 1	Présidente
Dr N.Lahiani	Université blida 1	Examinatrice
Dr M. Abbas	CRSTDLA	Encadrent
Prof N. Benblidia	Université blida 1	Promotrice

Résumé

Apres l'évolution du TLN, la détection des titres et l'extraction de la table des matières (TDM) sont devenus deux tâches indispensables pour cette première et l'analyse de documents, en particulier dans le domaine des finances ou les rapports sont généralement plus longs que dans d'autres domaines et qui ont un squelette complexe, dont la plus part de ces documents en format PDF ne contiennent pas la TDM à la création ce qui empêche l'obtention de l'information d'une manière rapide et claire ,c'est dans ce contexte que notre mémoire prend place afin d'étudier, concevoir et développer un système qui vise à extraire la TDM des documents PDF scannés et non scannés.

Pour bien gérer notre travail nous avons collecter plusieurs corpus ensuite nous avons utilisé des techniques d'extraction de l'information à partir des documents PDFs dans lesquelles nous avons appliqué de nombreuses méthodes telles que poppler et pyPDF2, pdfminer. Ces méthodes se concentrent sur le contenu textuel des documents numérisés, pour la comparaison entre ces techniques nous avons utilisé des algorithmes de la bibliothèque TextDistance, Les meilleurs résultats ont été obtenus en utilisant l'algorithme de l'entropie de Shannon. Les résultats que nous avons obtenu lors de nos expériences montre que poppler est le meilleur modèle utilisé pour notre étude avec un taux de 68.833% en utilisant des mémoires en master2 en anglais.

Mot clé : table des matières, documents financiers, extraction d'information, reconnaissance optique de caractères.

Abstract

After the evolution of the TLN, the detection of titles and the extraction of the table of contents (TOC) have become two indispensable tasks for the first one and the analysis of documents, especially in the field of finance where the reports are generally longer than in other fields and have a complex skeleton, most of these documents in PDF format do not contain the TOC at the time of creation, which prevents the information from being obtained quickly and clearly. It is at this stage that our project takes place in order to study, design and develop this system which aims to extract the TOC from scanned and non-scanned PDF documents.

In order to manage our work, we collected several corpora and then we used techniques to extract information from PDF documents in which we applied many methods such as poppler and pyPDF2, pdfminer. These methods focus on the textual content of the scanned documents, for the comparison between these techniques we used algorithms from the TextDistance library, the best results were obtained using the Shannon entropy algorithm. The best results were obtained using the Shannon entropy algorithm. The results we obtained in our experiments show that poppler is the best model used for our study with a rate of 68.833% using master2 theses in English.

Keywords: table of contents, financial documents, information extraction, Optical Character Recognition.

ملخص

بعد تطورNLP، أصبح اكتشاف العناوين واستخراج جدول المحتويات (TOC) مهمتين أساسيتين لهذا الأول وتحليل الوثائق، لا سيما في مجال التمويل حيث تكون التقارير بشكل عام أطول من الحقول الأخرى والتي تحتوي على هيكل معقد، ومعظم هذه المستندات بتنسيق PDF لا تحتوي على TOC عند الإنشاء مما يمنع الحصول على المعلومات بطريقة سريعة وواضحة، في هذه المرحلة تحدث ذاكرتنا من أجل دراسة وتصميم وتطوير هذا النظام الذي يهدف إلى استخراج TOC من مستندات PDF الممسوحة ضوئيًا وغير الممسوحة ضوئيًا.

من أجل إدارة عملنا بشكل جيد، قمنا بجمع العديد من الملفات ثم استخدمنا تقنيات لاستخراج المعلومات من مستندات PDF حيث قمنا بتطبيق العديد من الطرق مثل poppler وpyPDF2 وpdfminer. تركز هذه الطرق على المحتوى النصي للوثائق الممسوحة ضوئيًا، للمقارنة بين هذه التقنيات التي استخدمناها من مكتبة TextDistance، تم الحصول على أفضل النتائج باستخدام خوارزمية Shannon entropy. تظهر النتائج التي حصلنا عليها من تجاربنا أن poppler هو أفضل نموذج مستخدم لدراستنا بمعدل 68.833٪ باستخدام مذكرات تخرج ماستر2 باللغة الإنجليزية.

.

الكلمة الرئيسية: جدول المحتويات، المستندات المالية، استخراج المعلومات، التعرف البصري على الحروف.

Dédicaces

Je dédié ce modeste travail

A mes chers parents ma mère et mon père

Pour leur patience, leur amour, leur soutien et leurs

Encouragements.

A mes frères Rayane et mahmoud ainsi que Youcef,

A chaque membre de ma famille que j'aime

A mon meilleur ami qui a était toujours présent à mes côtés, à me soutenir, à me rendre heureuse je te remercie pour ta patience, pour tous ce que tu m'as apporté la joie, le bonheur, l'amour et que les belles choses

A Mes amis et mes camarades,

Sans oublier tous mes professeurs,

Que ce soit du primaire,

Du moyen,

Du secondaire et de

L'enseignement supérieur

Boudella Aicha Sirine

Dédicaces

Je tiens à dédier ce modeste travail

A mes parents

Pour votre amour, votre soutien et votre aide malgré les difficultés. Vous m'avez appris à persévérer, ne jamais baisser les bras et à me battre pour atteindre mes objectifs. Merci à vous pour accepter mes caprices, pour m'aimer et me choyer. Que Dieu vous préserve.

A la mémoire de ma grand-mère chérie

Ma-Khira *j'aurais tant aimé que vous soyez présente*

A mon grand père

Pour ton amour, ton authenticité et ta générosité.

A ma Sœur ***Maroua*** *et* ***mes deux frères*** *que j'aime beaucoup et que*

Dieu les garde pour moi. Merci d'être toujours à mes côtés

A ma famille

Cousins, cousines, oncles et tantes, amies de cœur.

A vous tous qui avez su me redonner le sourire quand ça n'allait pas. A vous qui m'avez écoutée et encouragée à continuer.

A mes camarades de promotion Master TAL

Je ne vous oublie pas, Vous figurez parmi mes plus belles rencontres.

Merci a toutes les personnes qui ont contribué de près ou de loin pour que ce projet soit possible.

Benhalima Nour El Houda

Remerciement

Nous remercions Allah le tout puissant

D'avoir nous donner le courage, la volonté et la patience

De mener à terme le présent travail.

Nous tenons à remercier notre encadreur

Mr abbas mourad et notre formateur

Mr mohamed lichouri ainsi que notre chère professeure

madame Mezzi melyara

Pour leur soutien,

Leurs conseils judicieux et leur grande bien vaillance durant l'élaboration de ce travail.

Enfin, nous remercions tous les enseignants qui ont contribué à notre Formation.

Table des matières

Liste des figures

Liste des tableaux

Liste d’acronymes

NLP Natural Language Processing

OCR Optical Character Recognition

TDM Table Des Matières

CRSTDLA Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe

COLING the International Conference on Computational Linguistics

FNP Financial Narrative Processing

FNS MultiLing Financial Summarisation

TA Traduction Automatique

TLN Traitement du langage Naturel

MT Machine Translation

TOC Table Of Contents

EDGAR Electronic Data Gathering, Analysis, and Retrieval

SEC Securities and Exchange Commission

AMF Autorité des marchés financiers

DICI Document d’information clé pour l’investisseur

OPC Les Organismes de placements collectifs

SICAV Société d’Investissement à Capital Variable

SPPICAV Sociétés à Prépondérance Immobilière à Capital Variable

UML Unified Modeling Language

Introduction générale

Une grande quantité de documents financiers sont créés et publiés en permanence dans des formats lisibles par machine (généralement des fichiers PDF), avec un minimum d'informations sur la structure. Les entreprises utilisent ces documents pour rendre compte de leurs activités, de leur situation financière ou de leurs projets d'investissement potentiels aux actionnaires, aux investisseurs et aux marchés financiers. Il s'agit essentiellement de rapports annuels d'entreprise contenant des informations financières et opérationnelles détaillées. [1]

Dans certains pays, comme aux États-Unis ou en France, les autorités de régulation comme EDGAR[1] SEC[2] ou l'AMF[3] exigent des entreprises qu'elles suivent un certain modèle lorsqu'elles communiquent leurs résultats financiers, afin d'assurer la normalisation et la cohérence des informations fournies par les entreprises. Dans d'autres pays européens, en revanche, les dirigeants ont généralement plus de latitude quant au contenu et à la manière des rapports, ce qui entraîne un manque de normalisation entre les documents financiers publiés sur un même marché. [1]

1. Problématique et objectifs

Dans cette tâche, nous nous concentrons sur l'analyse des rapports financiers, documents officiels au format PDF dans lesquels les fonds d'investissement décrivent précisément leurs caractéristiques et leurs modalités d'investissement. Bien que le contenu qu'ils doivent inclure soit souvent réglementé, leur format n'est pas standardisé et présente une grande variabilité allant du format texte simple à une présentation plus graphique et tabulaire des données et des informations. La majorité des prospectus sont publiés sans la table des matières, qui est généralement nécessaire pour aider les lecteurs à naviguer dans le document en suivant un simple schéma d'en-têtes et de numéros de page, et pour aider les équipes juridiques à vérifier si tout le contenu requis est bien inclus. Ainsi, l'analyse automatique des prospectus pour en extraire la structure devient de plus en plus vitale pour de nombreuses entreprises à travers le monde. [2]

[1] EDGAR est une base de données contenant les données relativement aux entreprises qui doivent être inscrites au registre des sociétés de la Securities

[2] SEC est l'organisme fédéral américain de réglementation et de contrôle des marchés financiers.

[3] AMF est chargée de veiller à la protection de l'épargne investie en produits financiers, à l'information des investisseurs et au bon fonctionnement des marchés.

Parmi les objectifs de notre application on trouve :

1. La conversion du PDF en texte.
2. L’extraction des titres.
3. L’extraction de la table des matières.

2. Organisation du mémoire

Pour résoudre notre problématique et atteindre les objectifs susmentionnés nous avons organisée notre mémoire comme suit :

Dans le premier chapitre intitulé « Étude de l’existant » nous introduisons une présentation du cadre de notre projet.

Le deuxième chapitre intitulé « les documents financiers » qui décrit une notion très importante dans le domaine des finances.

Le troisième chapitre intitulé « Conception et modélisation de la solution » est consacré à la modélisation de l’architecture de notre solution et la conception des diagrammes les plus importants pour la compréhension du fonctionnement de notre système.

Le quatrième chapitre intitulé « Implémentation et tests, évaluation » présente les langages et les outils de développement que nous avons utilisé pour implémenter le système, ce chapitre sera conclu par la présentation de l’application développée et ses différents écrans et enfin l’évaluation de la performance du système.

Enfin, nous clôturons ce mémoire par une conclusion dans laquelle nous résumons l’enchainement de notre travail et exposant quelques perspectives futures.

Chapitre I : Etude de l’existant

1.1 Introduction

Plusieurs documents financiers sont produits chaque jour pour différentes demandes financières. Ces documents sont généralement publiés dans un format lisible par machine (tel que Portable Document Format (fichiers PDF). Le besoin de ces fichiers à la table des matières joue un rôle assez important pour rendre leurs structures plus ordonner et sophistiquer.

D'où ce chapitre est une sorte de description générale du contexte de notre projet de fin d'études. Dans un premier temps nous présentons le cadre de notre projet (étude de l'existant, problématique…etc.). Dans un second temps nous définissions les applications et les principaux termes utilisées pour la réalisation de notre projet et on clôture par une conclusion de notre chapitre.

1.2 Présentation du sujet

1.2.1 Présentation de l'organisme d'accueil

Notre étude a été développer et améliorer dans Le Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (CRSTDLA) qui a pour mission de mettre en œuvre des recherches théoriques et appliquées sur le développement de la langue arabe et de la linguistique arabe, en coopération avec les institutions et établissements concernés par l'harmonisation et l'homologation de la terminologie. Le CRSTDLA a pour mission également de réaliser des projets de recherche dans les domaines des sciences et de la technologie du langage appliqués à la langue arabe et les langues à large diffusion en vue du développement de la langue arabe sur les plans didactiques et technologiques. [2]

1.2.2 Etude de l'existant

Notre étude est une compétition proposée par le (FNP-FNS 2020) qui s'agit du premier atelier conjoint entre les ateliers du traitement des documents financiers narratifs FNP et Synthèse financière Multilingue FNS, les deux ateliers se déroulant depuis plusieurs années avec beaucoup de succès.

L'atelier conjoint se concentrera sur l'utilisation des méthodes de traitement du langage naturel d'apprentissage automatique et de linguistique du corpus liées à tous les aspects de

l'exploration de textes financiers et du traitement des narrations financières, en plus de démontrer la valeur et les défis de l'application de la synthèse aux textes financiers multilingues, généralement appelés "informations financières narratives".

Il existe un intérêt croissant pour l'application d'approches automatiques et assistées par ordinateur pour l'extraction, la synthèse et l'analyse de données financières tant qualitatives que quantitatives. [3]

La 28éme Conférence internationale sur la linguistique informatique (COLING'2020) de la compétition ou on a eu l'opportunité de participer se tiendra à Barcelone, en Espagne, le 12 décembre 2020 au Centre de convention international de Barcelone.

1.2.3 Contexte général

Malgré l'essor de l'Extraction d'Information et le développement de nombreuses applications dédiées lors de ces vingt dernières années, cette tâche rencontre des problèmes lorsqu'elle est réalisée sur des documents d'une taille volumineuse.

Et ainsi que malgré l'importance de l'analyse documentaire longue, il y a peu de ressources disponibles et aucune dans un domaine à faibles ressources comme les finances. Dans cette tâche, nous nous concentrons sur l'extraction de la table des matières (TDM) des documents financiers qui sont des documents officiels en format PDF dans lesquels les fonds d'investissement décrivent précisément leurs caractéristiques et modalités d'investissement. La majorité des documents sont publiés sans la TDM, ce qui est d'une importance fondamentale pour les tâches complexes du NLP, comme l'extraction de l'information ou la réponse à des questions sur de longs documents. Bien que le contenu qu'ils doivent inclure soit souvent réglementé, leur format n'est pas normalisé et affiche une grande variabilité allant du format texte brut à une présentation plus graphique et tabulaire des données et de l'information, ce qui rend l'analyse de la structure du discours encore plus compliquée. [4]

Les travaux existants sur la reconnaissance de la table des matières des livres et des documents ont presque tous porté sur des ensembles de données de petite taille, dépendantes de l'application et spécifiques au domaine. Cependant, les tables des matières des documents de différents domaines considérablement dans leur présentation visuelle et leur style, ce qui fait de la reconnaissance des tables des matières un problème difficile pour une collection à grande

échelle de documents et de livres hétérogènes. Par rapport aux livres ordinaires (fournis pour la plupart en texte intégral avec des informations structurelles limitées telles que des pages et des paragraphes), les documents financiers, qui contiennent un contenu textuel et non textuel, ont une structure plus sophistiquée comprenant des parties, des sections, des sous-sections, des sous-sous-sections. [1]

1.2.4 Cadre du projet

Pour surpasser la problématique ci-dessus le CRSTDLA ou s'est déroulé notre stage a émis le désir de réaliser un système multilingue de restitution du sommaire permettant de gérer le contenu textuel des documents PDF dans les objectifs sont :

1. Recueil du corpus en anglais et en français et en arabe.
2. La conversion du PDF en texte.
3. La détection des titres.
4. L'extraction de la table des matières.

1.2.4.1 La différence entre un PDF scanné et non scanné

Le format de document portable, également connu sous le nom de PDF, est un format de document électronique d'Adobe, Lancé au début des années 1990, pour représenter des documents de façon à ce qu'ils soient séparés du système d'opération, de l'application ou du matériel où ils furent créés à l'origine. Un fichier PDF peut être de toute longueur, contenir n'importe quel nombre de polices de caractères et d'images, et est conçu pour permettre la création et le transfert de tout support prêt à l'impression [5]. Est devenu de nos jours le moyen de partage d'information le plus utilisé sur l'internet, ce format a deux types :

- **PDF non scanné :** il s'agit d'un fichier Word convertis en PDF.
- **PDF scanné :** il s'agit d'un livre scanné en image puis en reconnaissance optique de caractères[4].

Les document saisis avec des éditeurs textuels sont souvent plus clairs et mieux lisible que ceux qui sont scanné.

[4] Un système de reconnaissance optique des caractères analyse optiquement un texte et en produit une version informatique (textes modifiables).

1.3 Les travaux existants

Beaucoup d'efforts ont déjà été déployés pour étiqueter la structure des documents. Certains projets connus sont le projet Million Book (Linke, en 2003), l'Open Content Alliance (OCA) (Suber, en 2005), ou la numérisation de Google (Coyle, en 2006) (Doucet et al., en 2011). Ces projets visant à reconnaître automatiquement la structure d'un document prennent en entrée un document au format PDF ou son contenu obtenu par reconnaissance optique de caractères (OCR).

L'extraction de la structure d'un document est un problème bien étudié dans l'analyse des documents, et a été appliquée à des types de documents distincts et dans différents domaines. Les travaux sur ce sujet vont des articles scientifiques (Klampfl et al., 2014) (Bast et Korzen,2017) aux livres (Linke, 2003).Rangoni et al. (Rangoni et al., en 2012) Bitew (Bitew, 2018) comprend également trois catégories distinctes : les caractéristiques textuelles (similaires à la sémantique), les caractéristiques de balisage (similaires à la morphologie) et linguistiques (liées à la partie du discours). Comme décrit, certains auteurs regroupent les caractéristiques en catégories ; cependant, certaines études n'utilisent qu'une seule catégorie, notamment Kim et al. (Kim et al., 2017), qui n'utilisent les éléments morphologiques que pour l'extraction structurée logique. Les méthodes utilisées pour résoudre ce problème comprennent des approches basées sur des règles et l'apprentissage machine (Klampfl et Kern, 2013) (He, 2017) . [6]

1.4 L'extraction de la TDM

L'extraction d'Information ou EI (en anglais, Information Extraction ou IE) désigne une technologie récente qui vise à extraire et à structurer automatiquement un ensemble d'informations précises apparaissant dans un ou plusieurs documents textuels écrits en langue naturelle. [7]

Dans notre cas on s'intéresse à l'extraction de la TDM qui est une tâche très importante dans le domaine du TLN.

1.4.1 Définition de la TDM

La table des matières est un outil de repérage, qui représente en quelque sorte le plan du document, les titres des différentes divisons et subdivision du document apparaissent dans la TDM selon leur ordre d'apparition dans le texte ils sont suivis de leur numéro de page, auquel ils sont généralement liés par des points de conduite. [8]

Cette liste de titres apparait d'une façon structurée et hiérarchique où les titres renvoient directement au contenu par un hyperlien. La première page de la table des matières n'est toutefois pas numérotée, car elle commence par un titre important, celui de « Table des matières ». Ce titre est centré.

La TDM est parfois placée au début de document, après le remerciements et dédicaces, si elle est très longue, elle peut être remplacé par un sommaire et reportée en fin de document, contrairement au sommaire elle a l'avantage d'être plus détaillée.

1.4.2 Les exemples de la TDM

La TDM peut comporter au maximum 9 niveaux et au minimum un seul niveaux, vous trouverez ci-dessous différents exemples de table des matières :

1.4.2.1 TDM à un seul niveau

Pour les documents plus courts, un exemple de table des matières à un seul niveau peut être utilisé. Il s'agit d'un exemple de TDM courte et succincte qui n'utilise que des entrées à un niveau sur des sections ou des chapitres. L'exemple de table des matières suivant explore cette structure de base :

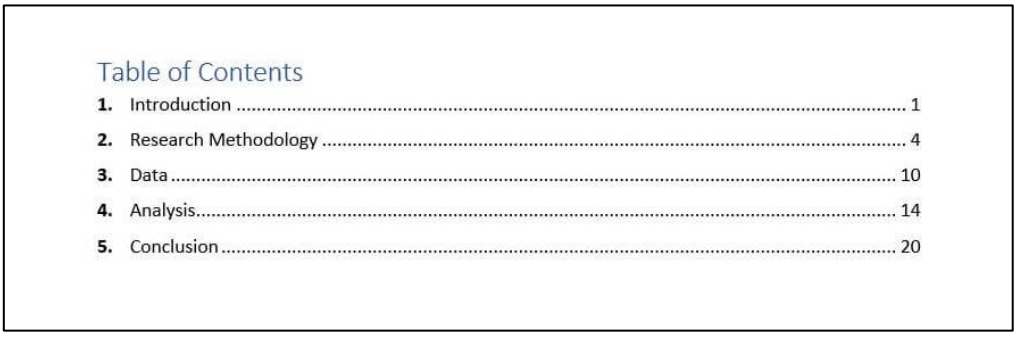


Figure 1 TDM à un seul niveau

1.4.2.2 TDM subdivisée

Un exemple de table des matières subdivisée est nécessaire pour les documents plus longs, offrant une subdivision des chapitres et des sections à l'intérieur des chapitres. Celles-ci sont plus détaillées et sont recommandées pour des documents plus riches en information comme les thèses de master ou de doctorat et sans oublier les livres. Il est courant (mais non nécessaire) de désigner chaque paragraphe par un chiffre (1.1, etc.)

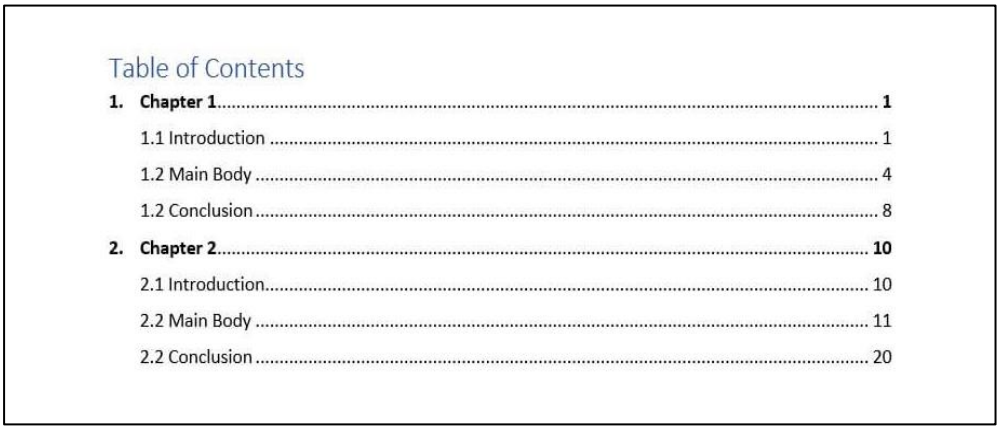


Figure 2 TDM subdivisée

1.4.2.3 TDM à plusieurs niveaux

L'ajout de niveaux supplémentaires à une table des matières est connu comme un exemple de table des matières à plusieurs niveaux. Celles-ci sont numérotés à partir de 1.1.1, etc.

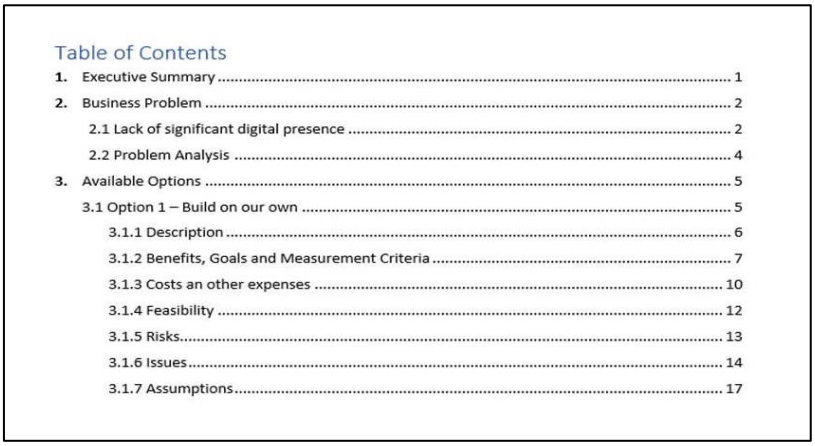


Figure 3 TDM à plusieurs niveaux

1.4.3 Type de numérotation de titres

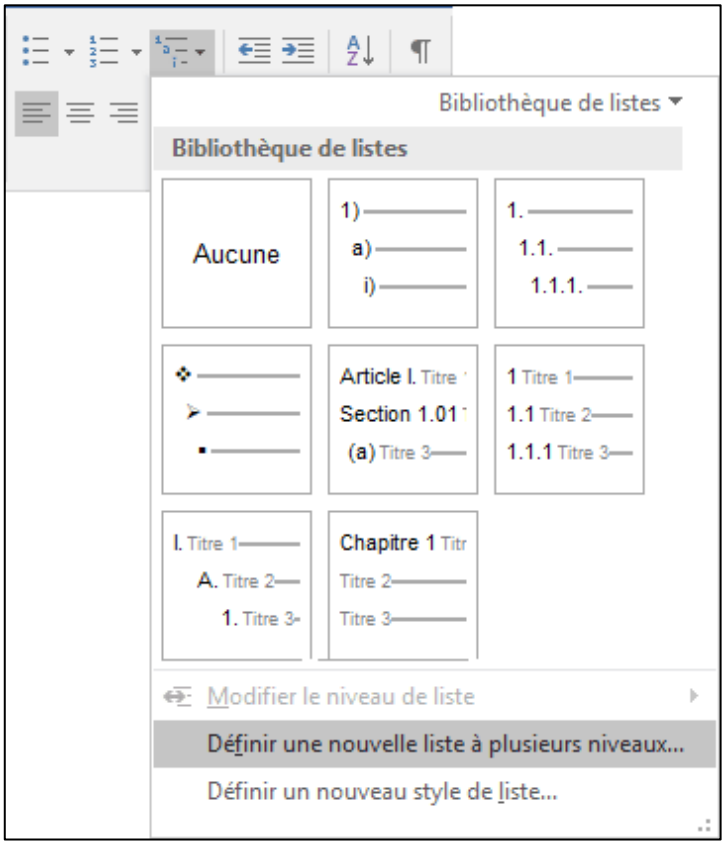


Figure 4 Différents type de numérotation de titres

1.4.4 Les avantages de la TDM

Parmi les avantages de la table des matières on trouve :

- Elle répond d'une manière précise suivant un ordre chronologique le plan de l'ouvrage, du mémoire ou de la thèse.
- Elle facilite l'accès aux informations des différents chapitres traités dans un document, mais d'une façon plus explicite que le sommaire.
- Elle permet d'avoir un aperçu de la structure et du contenu de l'ensemble du document.
- Un outil essentiel pour la consultation de textes ayant une certaine envergure.
- Elle aide à obtenir une vision synthétique du thème abordé dans un document.
- Permettre au lecteur de retourner facilement à la section dont il a besoin.

1.4.5 Les techniques de l'extraction de la TDM

L'extraction des données contenues dans des fichiers PDF est une tâche plutôt difficile et onéreuse en termes d'effort et de temps au vu du nombre de fichiers à extraire. Des milliers au minimum, parmi les techniques de l'extraction de la TDM on trouve :

La technique manuelle	La technique automatique
- Couteuse - Pénible - Lente	- Economique - Facile - Rapide

Tableau 1 les deux techniques de l'extraction de la TDM

C'est pour ces raisons, que le centre s'est intéressé à l'automatisation de cette tâche afin de pouvoir exploiter le grand nombre de documents numérisés qu'ils utilisent dans des cadres divers.

1.5 Conclusion

La TDM des documents numérisés de différents domaines diffère sensiblement dans leur mise en page visuelle et de style ce qui rend la reconnaissance de la TDM un problème difficile pour une collection à grande échelle, Nous avons observé que les TDMs peuvent être placés dans trois modèles de base, à savoir les TDMs à un seul niveau , les TDMs subdivisée , et à plusieurs niveaux , cette observation nous à aider à avoir des aperçus de comment réaliser l'analyse d'une TDM efficace En tant que tel, nous proposons une nouvelle approche de reconnaissance de la TDM qui détecte les titres et les extraire d'une manière fiable.

Le but de ce chapitre est de définir les principaux termes utilisés dans ce mémoire à travers les définitions et les techniques de l'extraction d'information de la TDM.

Dans le prochain chapitre nous introduirons en détails la notion des documents financiers.

Chapitre II : les documents financiers

2.1 Introduction

La compréhension des documents longs est encore un problème dans le traitement du langage naturel (NLP). La plupart des informations d'entreprise ou des connaissances académiques sont enfermées dans de longs documents (> 10 pages) avec une structure sémantique et de mise en page complexe. Les documents sont généralement convertis en texte brut et traités phrase par phrase, où la seule structure qui est facilement identifiable sont les paragraphes.

Dans ce chapitre on va décrire une notion très importante dans le domaine des finances qui s'agit des documents financiers.

2.2 Description des documents financiers

Dans le domaine financier, un grand nombre de documents sont publiés dans des formats lisibles par la machine pour les activités et la situation financière des entreprises déclarantes ou pour révéler des plants d'investissement partiels aux actionnaires, aux investisseurs et au marché financier.

Ces documents PDF sont les documents qui décrivent précisément les caractéristique et les modalités d'investissements, la plus par entre eux sont publié sans la TDM.

2.2.1 Définition des documents financiers

Ce sont des rapports utilisés dans le domaine financier et qui ont une grande variété de structure et de taille, donnent des informations sur la situation d'une entreprise, la performance, les flux des trésoriers, ces documents sont habituellement créés sur une base annuelle dans des formats lisible par la machine, et souvent avec une information minimale sur la structure.

2.2.2 Les objectifs des documents financiers

Les informations financières narratives représentent une grande partie des communications financières globales des entreprises avec les investisseurs. Les

commentaires textuels aident à clarifier les questions obscurcies par la complexité des méthodes comptables et des notes de bas de page.

En outre, les narratifs résument la stratégie de l'entreprise, mettent les résultats en contexte, expliquent les modalités de gouvernance, décrivent la politique de responsabilité sociale des entreprises et fournissent des informations prospectives aux investisseurs.

Les documents financiers servent à rendre compte des activités, de la situation financière, des plans d'investissement et de l'information opérationnelle aux actionnaires, aux investisseurs et aux marchés financiers.

2.2.3 Les besoins des documents financiers à la TDM

A cause de l'énorme volume d'information au niveau de documents financiers, l'existence de la table des matières est nécessaire car elle facilite la phase de recherche d'une information spécifique aux besoins ainsi que le prise de décisions dans les institutions financières par les investisseurs au niveau de marche.

2.2.4 Structure d'un document financier

Un document financier écrit en français se compose généralement de trois parties :

2.2.4.1 Le DICI

Le DICI (Document d'information clé pour l'investisseur) est, depuis le 1er juillet 2011, le nouveau document d'information remis aux investisseurs désireux d'investir leur épargne dans un OPC. C'est un document standardisé au niveau européen. Il doit donner, en 2 à 3 pages, une information claire, exacte et non trompeuse permettant à l'épargnant de prendre une décision d'investissement en connaissant les principales caractéristiques du produit. [9]

2.2.4.1.1 Les avantages du DICI

- Avec le DICI, vous avez accès aux informations nécessaires à votre prise de décision dans un document volontairement court.
- Le format standardisé du DICI vous permet d'obtenir une information claire et synthétique. Son but est de mettre à votre disposition tous les éléments

nécessaires à une meilleure compréhension et de vous aider à la comparaison des différents fonds français ou européens.

- L'affichage du niveau de risque et de rendement du produit sur une échelle allant de 1 à 7 vous permet de prendre facilement conscience du risque du produit.
- Le DICI affiche les frais courants réellement prélevés l'année précédente (les différents frais de gestion et de fonctionnement de l'OPC) ainsi que les frais d'entrée maximaux (négociables) et les éventuels frais de sortie. Cela vous offre une vision plus précise. [9]

2.2.4.1.2 Le guide de lecture du DICI

Les figures ci-dessous expliquent le mode de lecture des Document d'information clé pour l'investisseur.

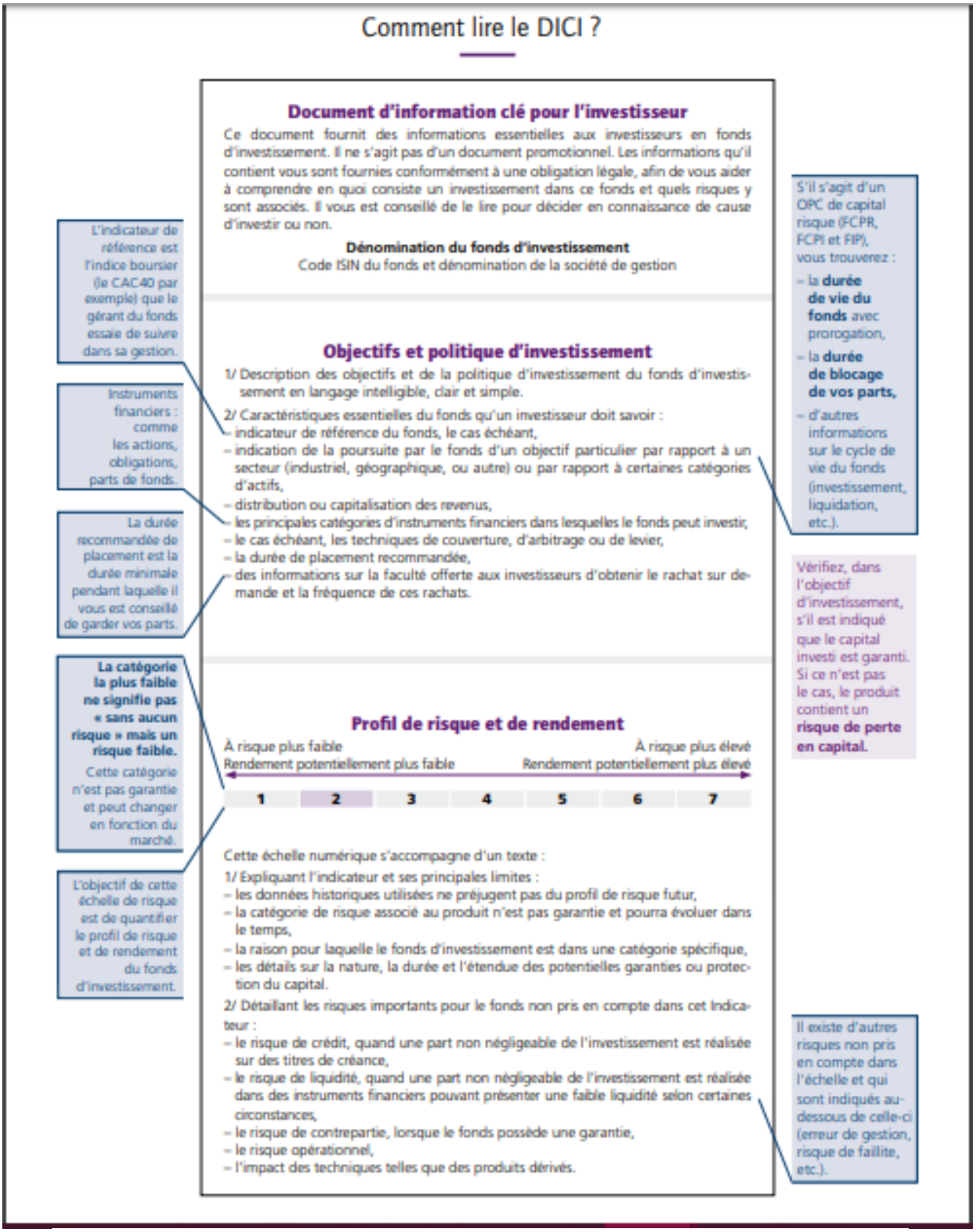
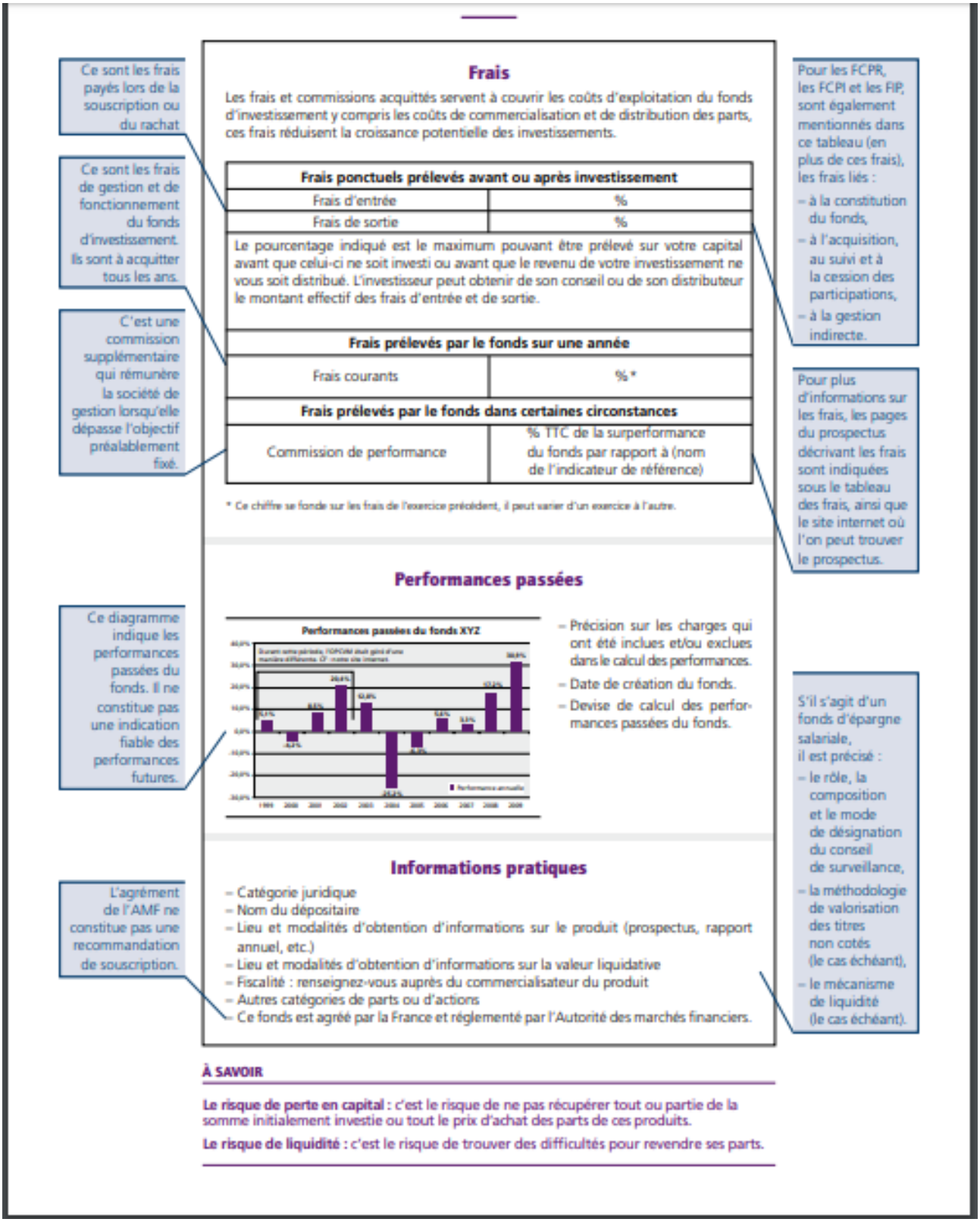


Figure 5 guide de lecture du DICI



Ce sont les frais payés lors de la souscription ou du rachat

Ce sont les frais de gestion et de fonctionnement du fonds d'investissement. Ils sont à acquitter tous les ans.

C'est une commission supplémentaire qui rémunère la société de gestion lorsqu'elle dépasse l'objectif préalablement fixé.

Ce diagramme indique les performances passées du fonds. Il ne constitue pas une indication fiable des performances futures.

L'agrément de l'AMF ne constitue pas une recommandation de souscription.

Frais

Les frais et commissions acquittés servent à couvrir les coûts d'exploitation du fonds d'investissement y compris les coûts de commercialisation et de distribution des parts, ces frais réduisent la croissance potentielle des investissements.

Frais ponctuels prélevés avant ou après investissement	
Frais d'entrée	%
Frais de sortie	%

Le pourcentage indiqué est le maximum pouvant être prélevé sur votre capital avant que celui-ci ne soit investi ou avant que le revenu de votre investissement ne vous soit distribué. L'investisseur peut obtenir de son conseil ou de son distributeur le montant effectif des frais d'entrée et de sortie.

Frais prélevés par le fonds sur une année	
Frais courants	% *
Frais prélevés par le fonds dans certaines circonstances	
Commission de performance	% TTC de la surperformance du fonds par rapport à (nom de l'indicateur de référence)

* Ce chiffre se fonde sur les frais de l'exercice précédent, il peut varier d'un exercice à l'autre.

Pour les FCPR, les FCPI et les FIP, sont également mentionnés dans ce tableau (en plus de ces frais), les frais liés :
- à la constitution du fonds,
- à l'acquisition, au suivi et à la cession des participations,
- à la gestion indirecte.

Pour plus d'informations sur les frais, les pages du prospectus décrivant les frais sont indiquées sous le tableau des frais, ainsi que le site internet où l'on peut trouver le prospectus.

Performances passées

- Précision sur les charges qui ont été inclues et/ou exclues dans le calcul des performances.
- Date de création du fonds.
- Devise de calcul des performances passées du fonds.

S'il s'agit d'un fonds d'épargne salariale, il est précisé :
- le rôle, la composition et le mode de désignation du conseil de surveillance,
- la méthodologie de valorisation des titres non cotés (le cas échéant),
- le mécanisme de liquidité (le cas échéant).

Informations pratiques

- Catégorie juridique
- Nom du dépositaire
- Lieu et modalités d'obtention d'informations sur le produit (prospectus, rapport annuel, etc.)
- Lieu et modalités d'obtention d'informations sur la valeur liquidative
- Fiscalité : renseignez-vous auprès du commercialisateur du produit
- Autres catégories de parts ou d'actions
- Ce fonds est agréé par la France et réglementé par l'Autorité des marchés financiers.

À SAVOIR

Le risque de perte en capital : c'est le risque de ne pas récupérer tout ou partie de la somme initialement investie ou tout le prix d'achat des parts de ces produits.

Le risque de liquidité : c'est le risque de trouver des difficultés pour revendre ses parts.

Figure 6 suite du guide de la lecture du DICI

2.2.4.2 Les prospectus

Un prospectus est un document d'information détaillé qu'une société ou un fonds doit généralement produire pour pouvoir émettre des titres (par exemple, des actions) au grand public.

Le prospectus vise à fournir un exposé complet, véridique et clair de tous les faits importants des titres émis, afin que l'investisseur puisse prendre des décisions de placement éclairées. Le prospectus doit révéler tous les éléments susceptibles d'affecter la valeur ou le cours du titre faisant l'objet du placement.

L'information y est présentée de façon standard afin que l'investisseur puisse plus facilement comparer diverses possibilités de placement. [10]

2.2.4.2.1 Les avantages du prospectus

Le prospectus vous permet de vous informer en vous donnant des renseignements détaillés sur la société ou le fonds, ainsi que sur les actions ou les autres titres mis en vente. Vous pouvez y trouver réponse à bon nombre de questions avant d'investir pour vous aider à prendre une meilleure décision de placement.

2.2.4.2.2 La structure d'un prospectus

Le prospectus de l'entreprise ou du fonds est un document qui contient des renseignements détaillés sur :

- Les titres offerts ;
- Les activités ;
- La direction ;
- La situation financière. [11]

2.2.4.3 Le règlement du fonds ou les statuts de la SICAV (ou SPPICAV)

Les statuts définissent un ensemble de textes qui constitue la base d'une société ou d'une association. Ce sont ces statuts qui fixent les règles de fonctionnement de la société ou de l'association, en renseignant notamment les individus membres de leurs droits et de leurs obligations. Dans le monde économique, les statuts définissent plus spécifiquement la qualification juridique d'une entreprise. [12]

2.3 Synthèse :

Souvent l'investisseur trouve des difficultés pour accéder aux différentes informations au niveau des documents financiers qui sont très indispensables dans le marché financier. Et pour faciliter la tâche de restitution de l'information nous avons choisis de travailler avec ce type de documents.

2.4 Conclusion

Dans ce chapitre nous avons détaillé les documents du domaine de notre étude (les documents financiers).

Dans le chapitre suivant, nous verrons en détail l'architecture globale les diagrammes UML de la solution que nous proposons.

Chapitre III : Conception et modélisation de la solution

3.1 Introduction

La phase de la conception et la modélisation d'un système est l'une des phases les plus importante pour sa réalisation, cette dernière nous permet de mieux comprendre son fonctionnement, et de bien maitriser sa complexité, ce chapitre est consacré pour la modélisation de notre solution en utilisant certains outils de conception, d'abord on commence par la description de notre corpus. Ensuite nous allons présenter la solution logique qui représente l'architecture globale de notre système, enfin on va définir l'utilité de chaque sous tâche en utilisant les diagrammes UML pour la résolution de notre problématique.

3.2 Présentation du corpus

3.2.1 L'analyse et la description du corpus

Notre collection utilisée est un échantillon des documents financiers avec une grande variété de structure et de tailles. Ces documents sont disponibles en ligne, d'où le corpus est composé de deux dossiers, le premier est écrit en français (47 documents financiers) et l'autre en anglais (52 documents financiers de Luxemburg), Il s'agit de documents PDF ou la plupart ne contient pas de TDM, Chaque document d'entre eux fourni un ensemble d'informations nécessaires pour l'investisseur, sur la situation d'une entreprise, Ces informations peuvent concerner l'état de sa structure financière, la composition de son patrimoine, l'évaluation de ces performances .....etc., Ces documents s'adressent à tout intéressé : les associés, les dirigeants, les investisseurs, les équipes juridiques et à tout personnes trouvant un intérêt, Ils permettent au lecteur de prendre des décisions (les investisseurs les utilisent comme support, véritable outil d'aide à la décision), et de réaliser des comparaisons dans l'espace (les objectifs, les résultats d'une entreprise sont comparés à ceux d'une entreprise concurrente).

3.2.2 L’objectif d’analyse de corpus

Après la lecture attentive et l’analyse des documents financiers nous avons constatées les points en communs entre chaque document ainsi que leurs différences.

Une fois qu’on a terminées la comparaison des documents et que nous avons rassemblé tous les éléments on a pu organisées notre réponse sur les questions déjà posées, selon un classement on a essayé d’aller du générale au particulier et d’évident au moins évident. On a regroupé les points en communs dans un premier paragraphe et les différences dans le deuxième.

Les documents éventuellement ayant le même concept et partagent le même type d’information.

Et leur différence est au niveau de la langue sur lesquelles étaient écrits et aussi leurs structures qui est complètement différentes ainsi que leurs tailles.

3.3 Le besoin de construction d’un autre corpus

Plusieurs documents financiers sont produits, chaque jour, pour différentes demandes financières. Certains de ces documents sont obligatoires par la loi, mais ils ne sont pas créés selon la même norme et ont parfois une structure médiocre, ce qui fait il est difficile de retrouver les informations souhaitées. Ces documents sont généralement des fichiers PDF mais malheureusement, ils ne sont pas étiquetés ils n'ont pas d'étiquettes pour identifier les éléments de mise en page tels que les paragraphes, les colonnes, ou des tables. [6]

Afin d’avoir une possibilité de commencer notre projet nous avons choisis de construire un nouveau corpus qui contient un ensemble de mémoires de master 2 de différentes langues (en anglais, français, arabe) 50 mémoires par langue.

3.3.1 Description détaillée du nouveau corpus

Les tableaux ci-dessous englobent des informations supplémentaires concernant les mémoires collectés telles que : le nom du département, de l'université, l'année universitaire, et le nombre de mémoires pour chaque année universitaire :

	Département	**Université**	**Années universitaire**	**Nombre mémoire**	**Nombre totale**
Mémoire Master 2 en Anglais	**Anglais**	Abou bekr Belkaid Telemcen	2018-2019	25	50
		Mouhamed Boudiaf Msila	2016-2017	20	
			2018-2019	5	

Tableau 2 caractéristiques des mémoires en langue Anglaise

	Département	**Université**	**Années universitaire**	**Nombre mémoire**	**Nombre totale**
Mémoire Master 2 en français	**Génie électrique et électronique**	Abou bekr Belkaid Tlemcen	2018-2019	26	50
	Génie Mécanique	Abou bekr Belkaid Tlemcen	2018-2019	7	
	Architecture	Abou bekr Belkaid Tlemcen	2018-2019	1	
			2019-2020	2	
	Lettres et Langue Française	Abou bekr Belkaid Tlemcen	2018-2019	11	
			2019-2020	1	
		Kasdi Mebah Ouaragla	2018-2019	2	

Tableau 3caractéristiques des mémoires en langue Française

	Département	Université	Années universitaire	Nombre mémoire	Nombre totale
Mémoire Master 2 en Arabe	**Droit**	Abederahmane Mira Béjaia	2018-2019	8	50
		Larbi Tebessi Tbessa	2015-2016	1	
		Djilali Bounaama Khmis Miliana	2013-2014	1	
		Taha Moulay Saida	2014-2015	1	
	Philosophie	Abou bekr Belkaid Tlemcen	2019-2020	1	
		Abou bekr Belkaid Tlemcen	2017-2018	2	
			2018-2019	6	
			2019-2020	1	
	Arts	Abou bekr Belkaid Tlemcen	2018-2019	4	
	Langue et littérature Arabe	Abou bekr Belkaid Tlemcen	2016-2017	1	
			2018-2019	9	
		Abed el hamid Ben badis Mostaganem	2014-2015	1	
	Sciences Humaines	Hamma Lakhdar El ouad	2014-2015	1	
	Sciences de Gestion	Mouhamed Bougara Boumerdes	2016-2017	1	
	sociologies	8 mai 1945 Guelma	2017-2018	2	
		Abou bekr Belkaid Tlemcen	2011-2012	1	
	Sciences Commercials	Annex Universitaire Maghnia.	2015-2016	1	
	Histoire	Abou bakr Belkaid Tlemcen	2015-2016	2	
	Langue anglaise (traduction)	Abou bakr Belkaid Tlemcen	2018-2019	7	

Tableau 4 caractéristiques des mémoires en langue Arabe

3.3.2 Les caractéristiques communes entre les mémoires collectés

- Les mémoires sont construits d'une façon hiérarchique et structurer ou l'ordre des sections(chapitres) est bien respecté.
- La numérotation des titres.
- Chaque titre est suivi par son numéro de page dans la TDM.

3.3.3 Les caractéristiques en déférence entre les mémoires collectés

On vous présente ci-dessous les caractéristiques spécifiques pour chaque langue :

3.3.3.1 L'analyse du mémoire écrit en anglais

Parmi la caractéristique de tables des matières des mémoires écrites en anglais on trouve la numérotation romaine des pages du dédicace, le remerciement, le résumé, la table des matières, liste des tableaux et des figures ainsi que des acronymes comme le montre la figure 7

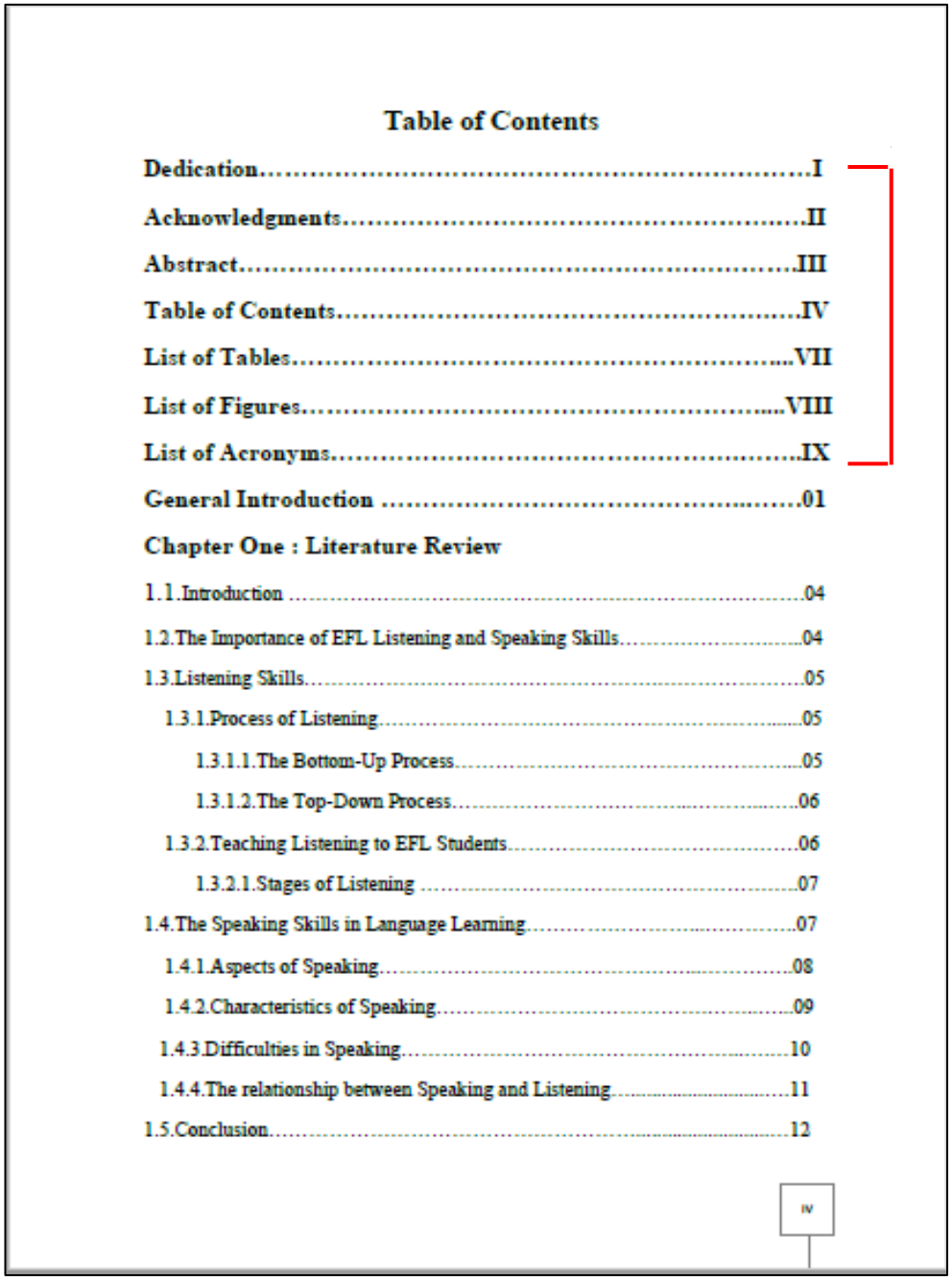
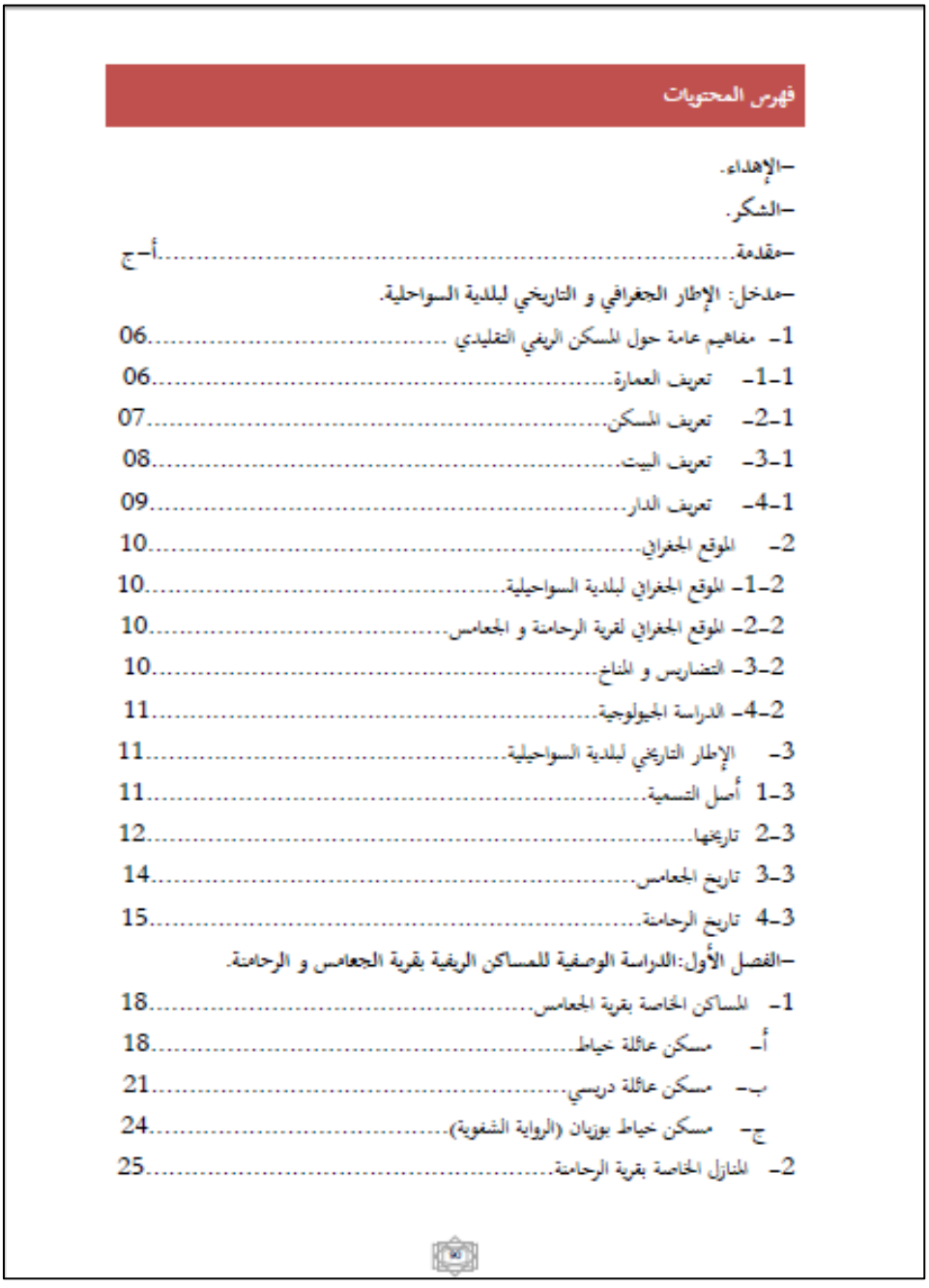


Table of Contents

Figure 7 TDM d'un mémoire en anglais

3.3.3.2 L'analyse du mémoire écrit en arabe

Parmi la caractéristique de tables des matières des mémoires écrites en arabe on a remarqué que cette dernière se situe généralement à la fin du mémoire.



فهرس المحتويات

Figure 8TDM d'un mémoire en arabe

3.3.3.3 L'analyse du mémoire écrit en français

La table de matières des mémoires en français (figure 9) est une table à plusieurs niveaux dont les titres des chapitres sont représentés dans le premier niveau et les sous-titres du chapitre dans le second niveau et ainsi de suite.

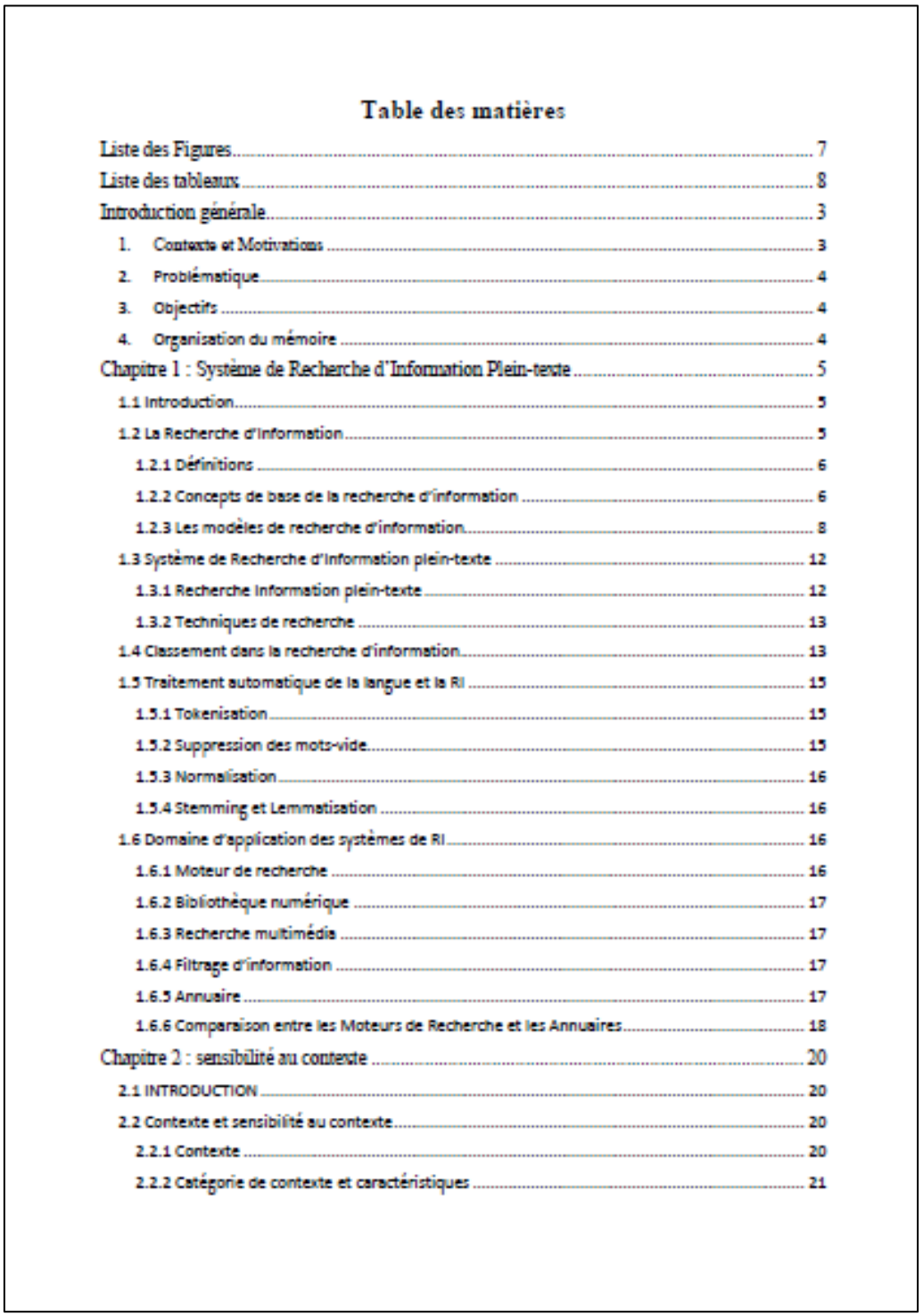


Table des matières

Figure 9TDM d'un mémoire en Français

3.4 L'architecture globale du système

Cette partie montre l'architecture détaillé de notre solution, Dans un premier lieu on a notre corpus sur lequel on va appliquer nos traitements, Nous présentons dans cette partie les trois étapes principales (conversion du PDF en texte, détection des titres, l'extraction de la TDM) et l'amélioration du contenu textuel comme étape supplémentaire, qui représentent la raison d'être de notre système pour avoir une vision globale de ces fonctionnalités. Cette architecture représente notre solution logique. Comme vous pouvez le voir dans la figure suivante.

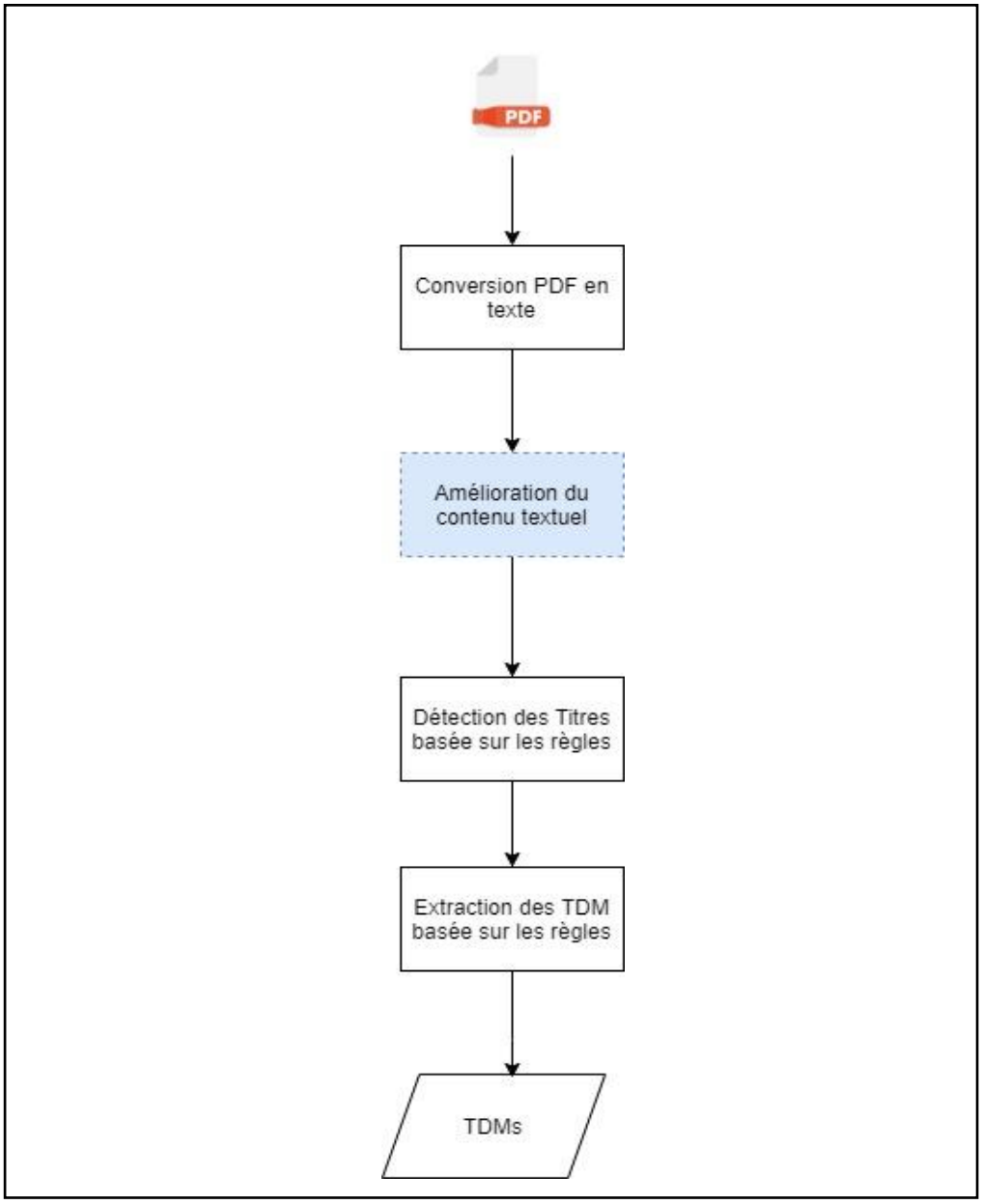


Figure 10 l'architecture globale du système multilingue de restitution de la TDM

3.4.1 Conversion du PDF en texte

La conversion du PDF en texte est jugée comme la première étape de notre solution proposée pour l'extraction du contenu textuel après la sélection du PDF.

3.4.2 L'amélioration du contenu textuel

L'Amélioration du contenu textuel inclus deux sous tâche importante pour la conversion parmi lesquels en trouve : l'Analyse de la structure de page on analysons la mise en page du fichier PDF et la Détection globale de la typographie.

La figure ci-dessous montre les deux sous tâche qui sont inclus dans la phase de l'amélioration du contenu textuel :

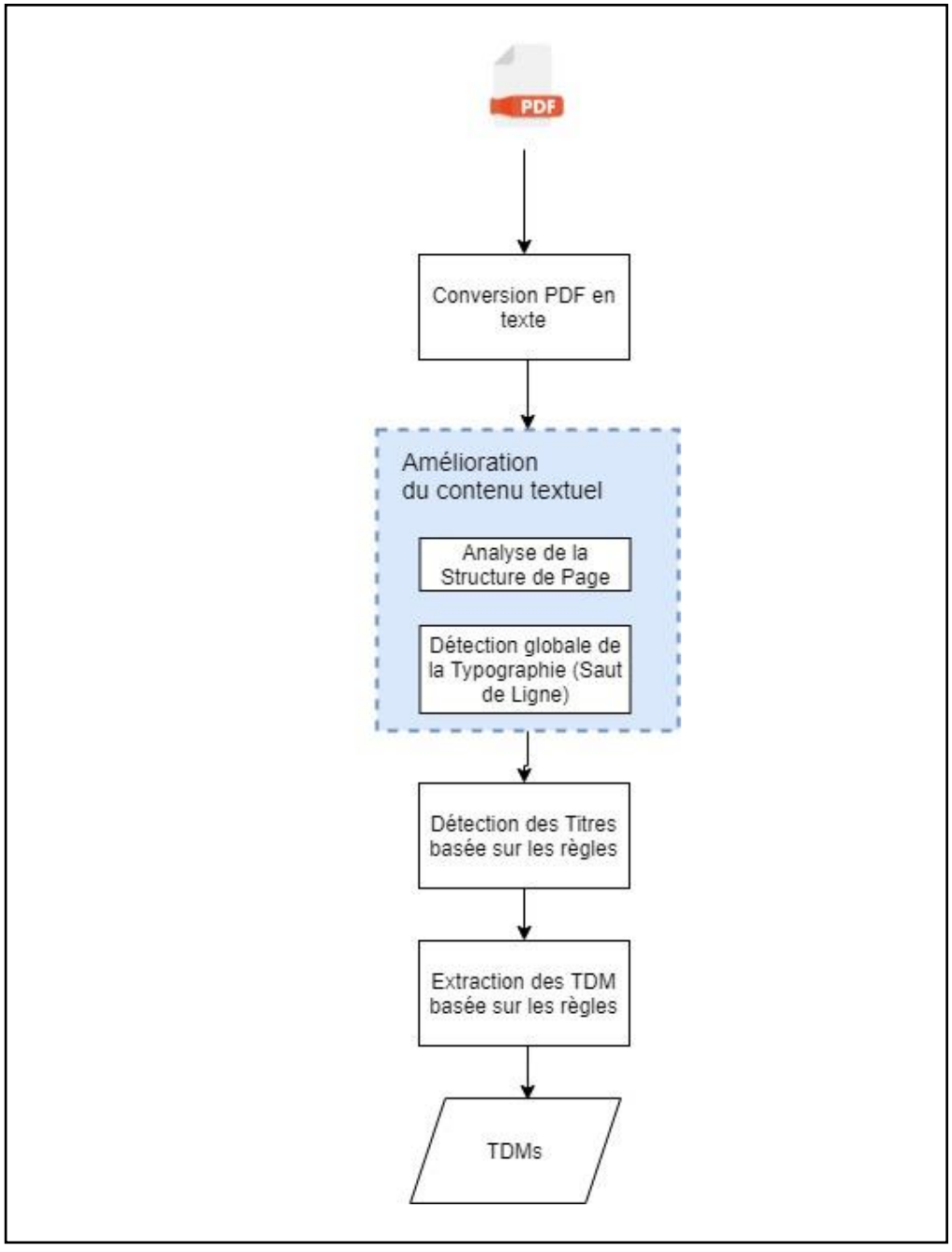


Figure 11 phase de l'amélioration du contenu textuel

3.4.2.1 L'analyse de la structure de page

Dans cette sous tache nous analysons la mise en page du fichier PDF on commence par l'ouverture de ce dernier ensuite par sa lecture et sans oublier l'extraction du nombre totale des pages du PDF sélectionner en utilisant ce script :

```
def pdfToTocInText(pdfFileName, tocFileName):
    pdfFileObj = open(pdfFileName, 'rb')

    # creating a pdf reader object
    pdfReader = PdfFileReader(pdfFileObj)

    # printing number of pages in pdf file
    numPage = pdfReader.numPages
    print(pdfReader.numPages)
```

Figure 12 Code de l'ouverture et la lecture du fichier PDF avec pyPDF2

3.4.2.2 La détection globale de la typographie

Cette sous tâche est consacrée à l'extraction du contenu textuel du PDF pour l'accomplir nous avons répartie cette sous tâche en deux sous tâches l'élimination des motifs indésirables[5] et la correction morphologique du texte afin d'avoir un format textuel facile a manipulé. Ces sous tâches sont spécifiques à une certaine langue ainsi que de techniques, l'élimination des motifs est généralement spécifique à la langue arabe et la correction morphologique du texte est spécifique pour la technique Pdfminer Comme vous pouvez les voir dans la figure suivante :

[5] Les motifs indésirables : les symboles Unicode : \u202a\ , \u202b\ , \u202c\ , \uf0fc\ , \u200c\ , \x0c

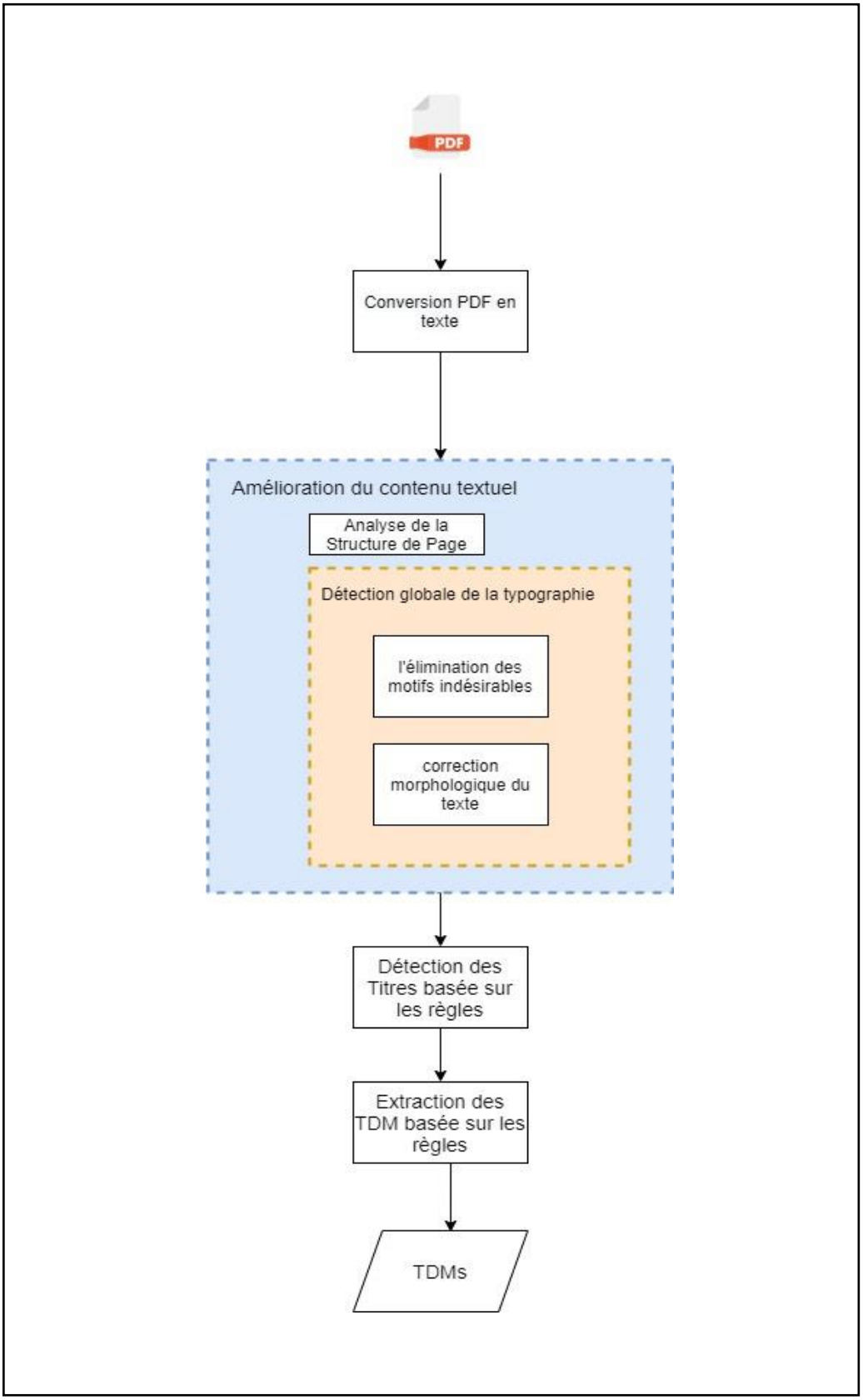


Figure 13 Détection globale de la typographie

3.4.2.2.1 L'extraction du contenu textuel

Dans cette étape nous avons convertis tous les fichiers PDF du dossiers pdfs en format textuel en sauvegardant le fichiers (avec une extension de pdf.txt) dans un dossier nommé texts .

3.4.2.2.1.1 L'extraction du texte en utilisant poppler

```
@ECHO OFF
FOR /R %1 %%G IN (*.pdf) DO (
    ECHO Attempting to convert %%G
    "D:\M2TAL\pfe\poppler-0.68.0\bin\pdftotext.exe" "%%G" "%%G.txt"
    )
```

Figure 14 Code de la conversion du PDF en texte avec poppler

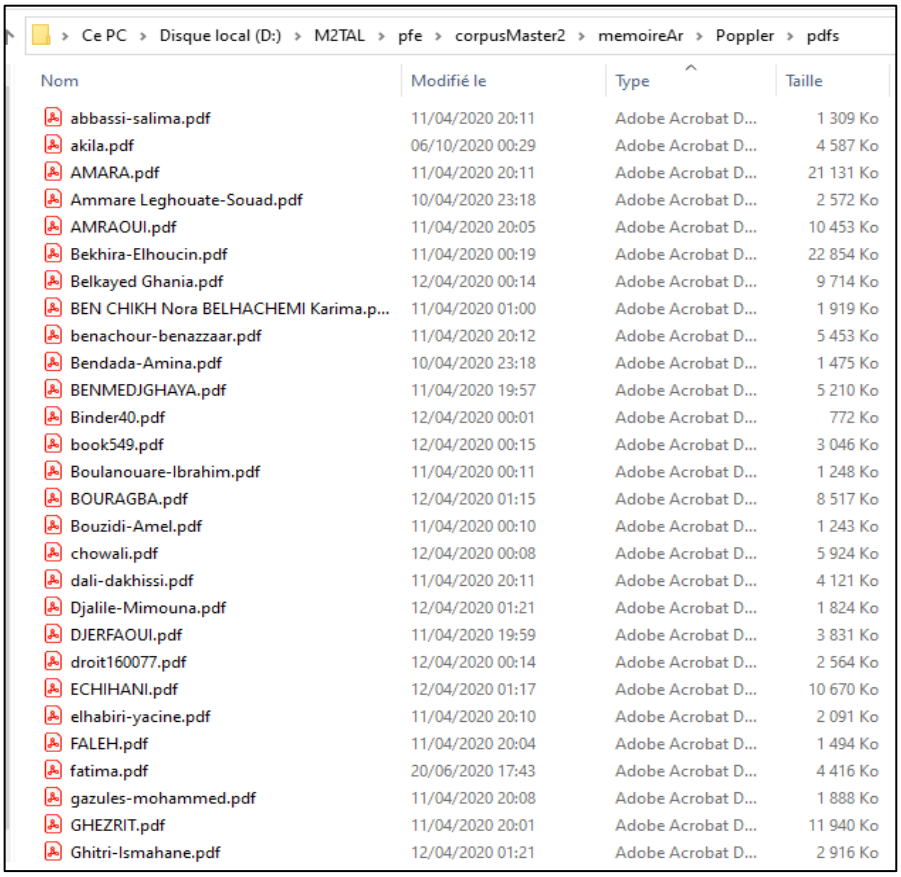


Figure 15 Dossier pdfs

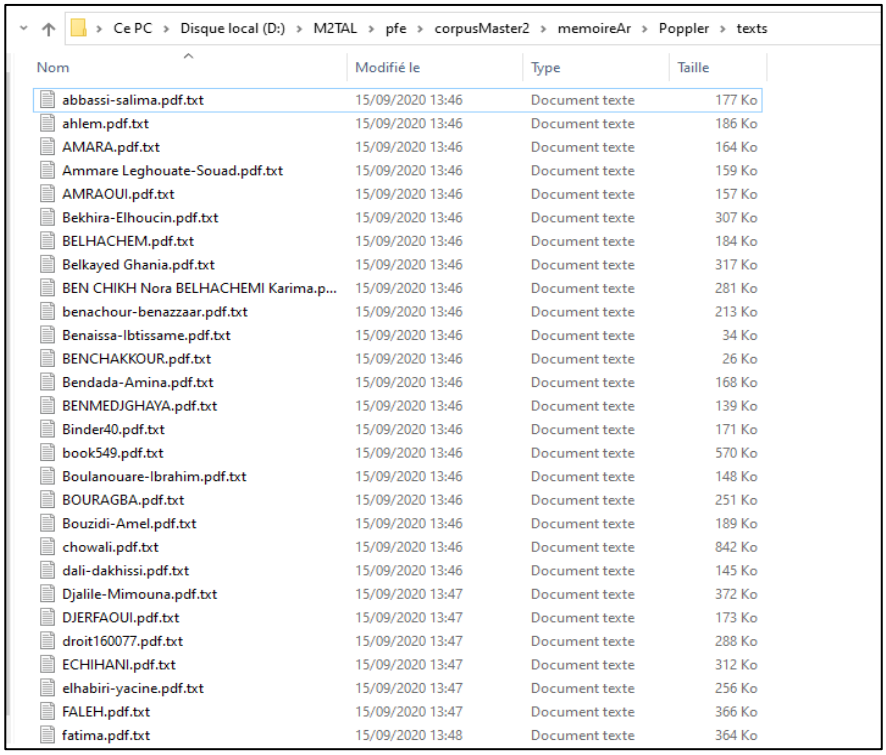


Figure 16 Dossier texts

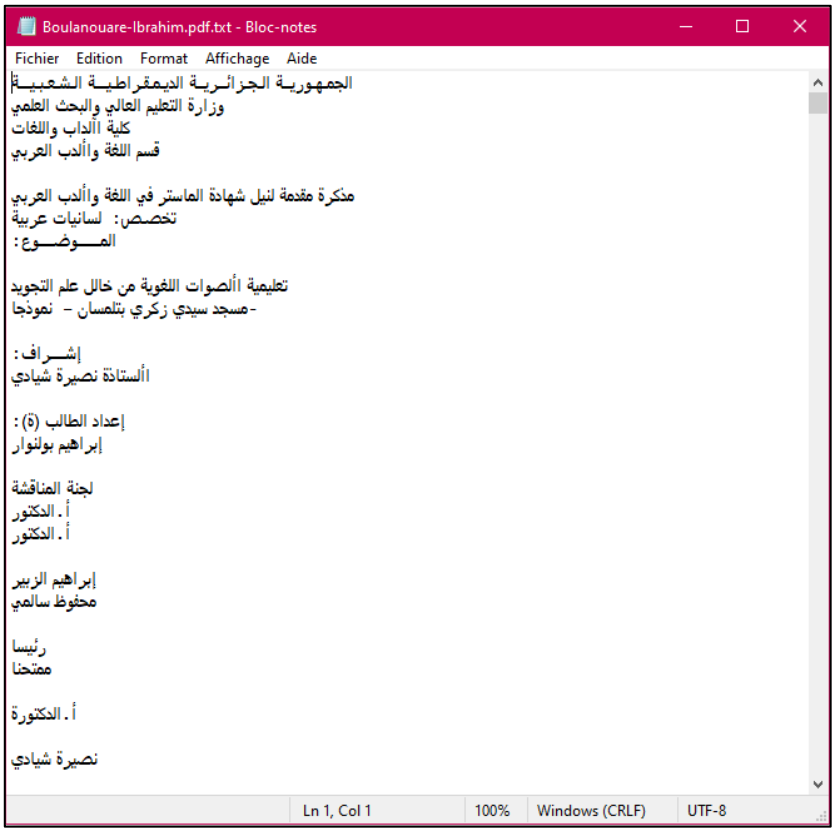


Figure 17 exemple d'un contenu textuel extrait avec poppler

3.4.2.2.1.2 L'extraction du texte en utilisant pyPDF2

```
# creating a page object
extractText = []
for n in range(numPage):
    pageObj = pdfReader.getPage(n)
    # extracting text from page
    #print(pageObj.extractText())
    extractText.append(pageObj.extractText())
# closing the pdf file object
pdfFileObj.close()
print(extractText)
```

Figure 18 Code de la conversion du PDF en texte avec pyPDF2

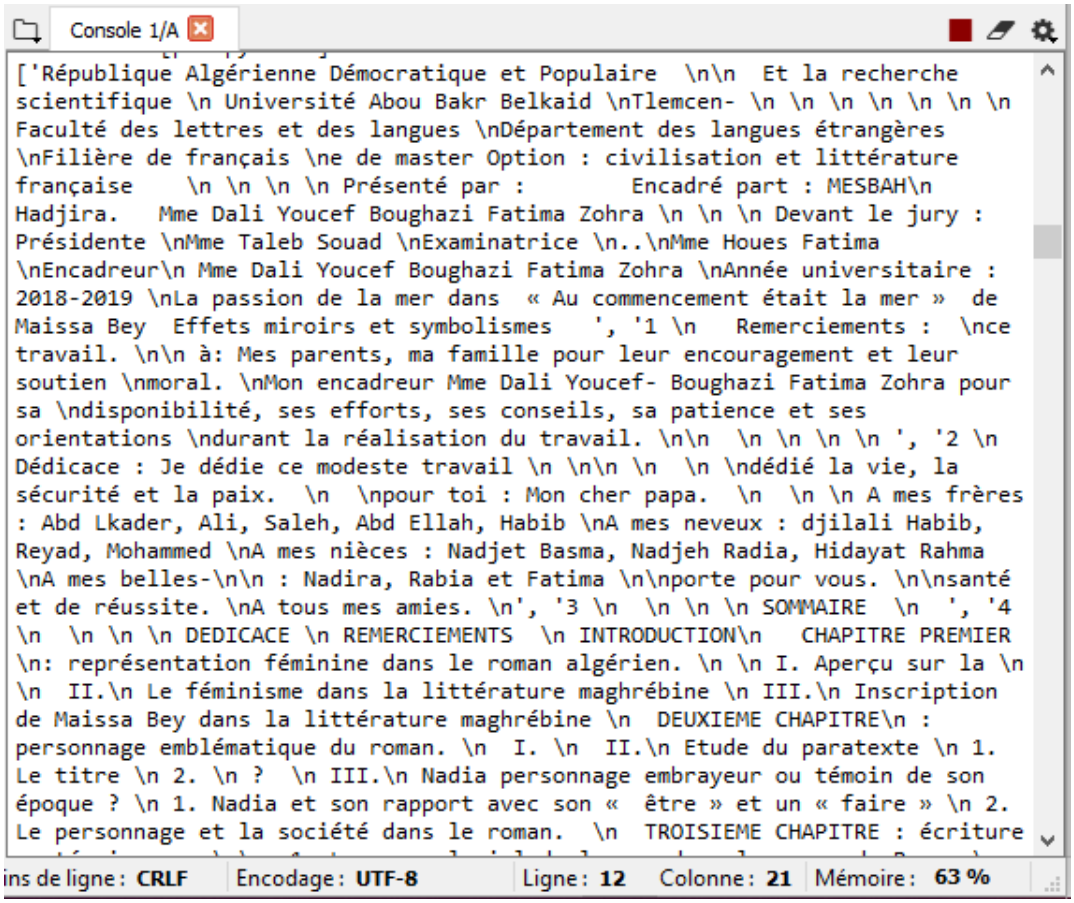


Figure 19 Contenu textuel en utilisant PyPDF2

3.4.2.2.1.3 L'extraction du texte en utilisant pdfminer

```
from pdfminer.high_level import extract_text
def extractTextPdfMiner(pdf_file):
    dobj = extract_text(pdf_file)
    out = repr(dobj)
    return out
```

Figure 20 Code de la conversion du PDF en texte avec pdfminer

3.4.2.2.2 L'élimination des motifs indésirables

Après avoir extraire le texte des mémoires du master2, et pour approfondir dans le contenu textuel extrait et voir la typographie du texte nous avons pensé à mettre ce dernier dans une liste (figure 22) :

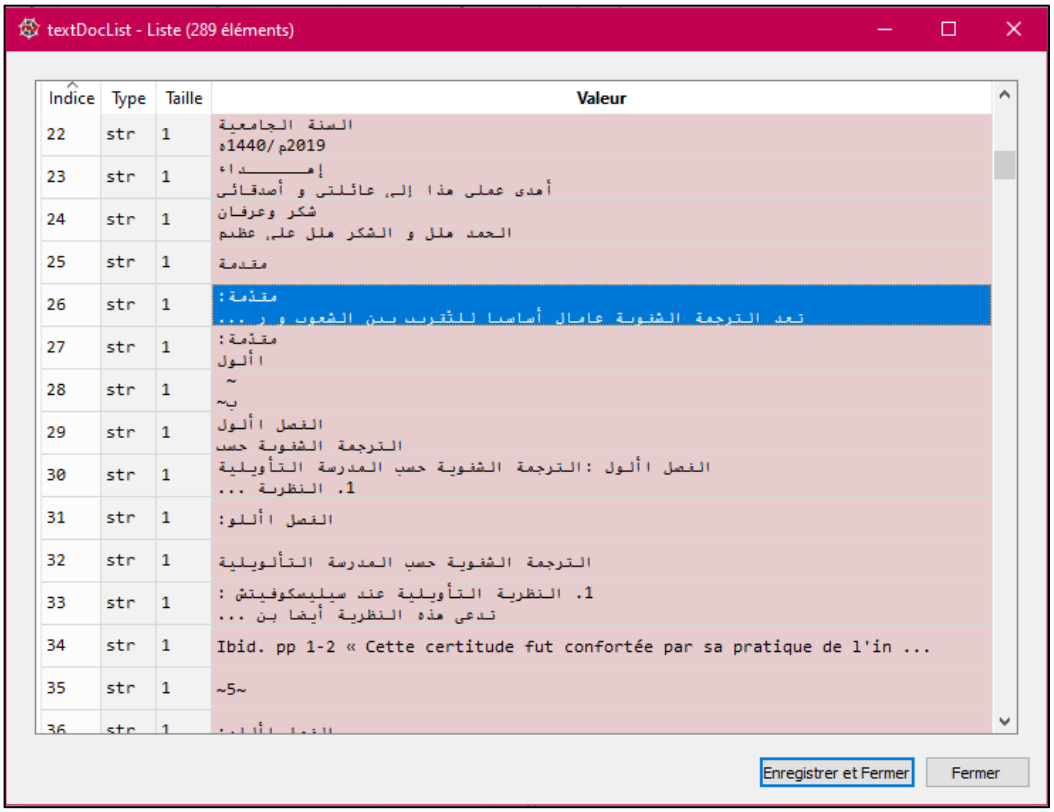


Figure 21 Résultat du contenu de la liste "textDocList" en utilisant poppler

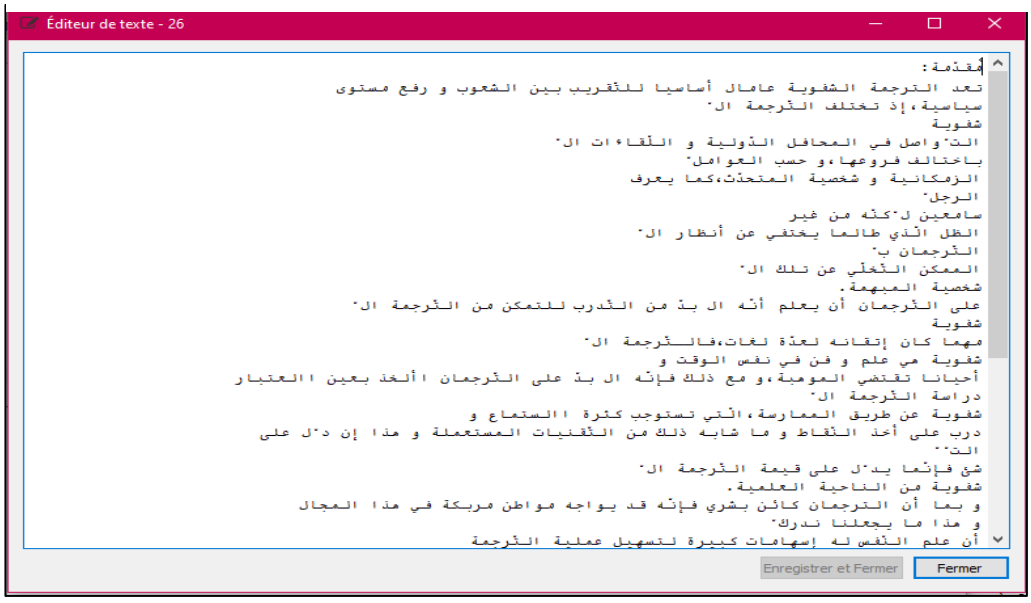


Figure 22 Résultat de l'objet 26

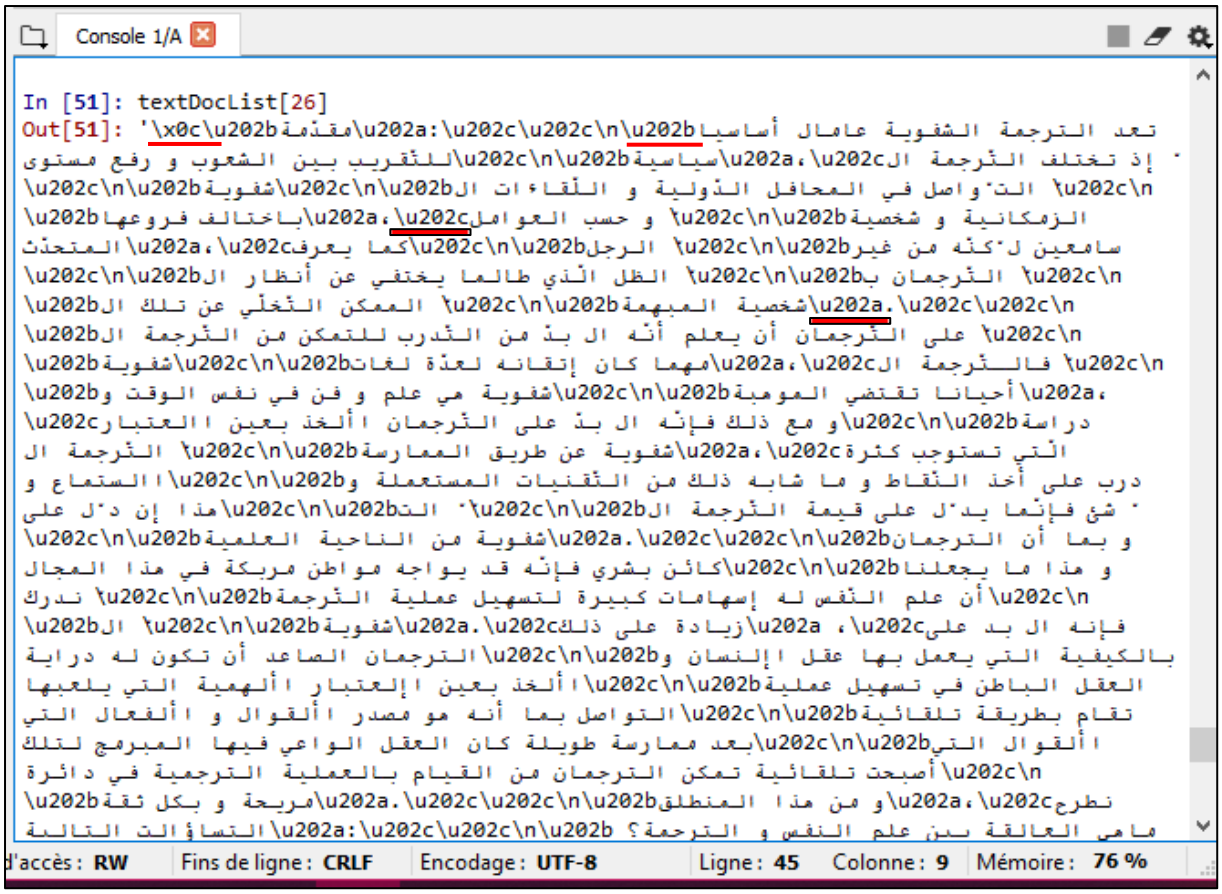


Figure 23 texte avec motifs indésirables

Afin d'éliminé les motifs indésirables nous avons créés une fonction pour résoudre ce problème (figure 24) :

```
def removeUnicodeChar(text):
    x = text.replace('\u202a', '')
    y = x.replace('\u202b', '')
    z = y.replace('\u202c', '')
    q = z.replace('\uf0fc', '')
    m = q.replace('\u200c', '')
    r = m.replace('\x0c', '')
    return r
```

Figure 24 Fonction de suppression des motifs indésirables

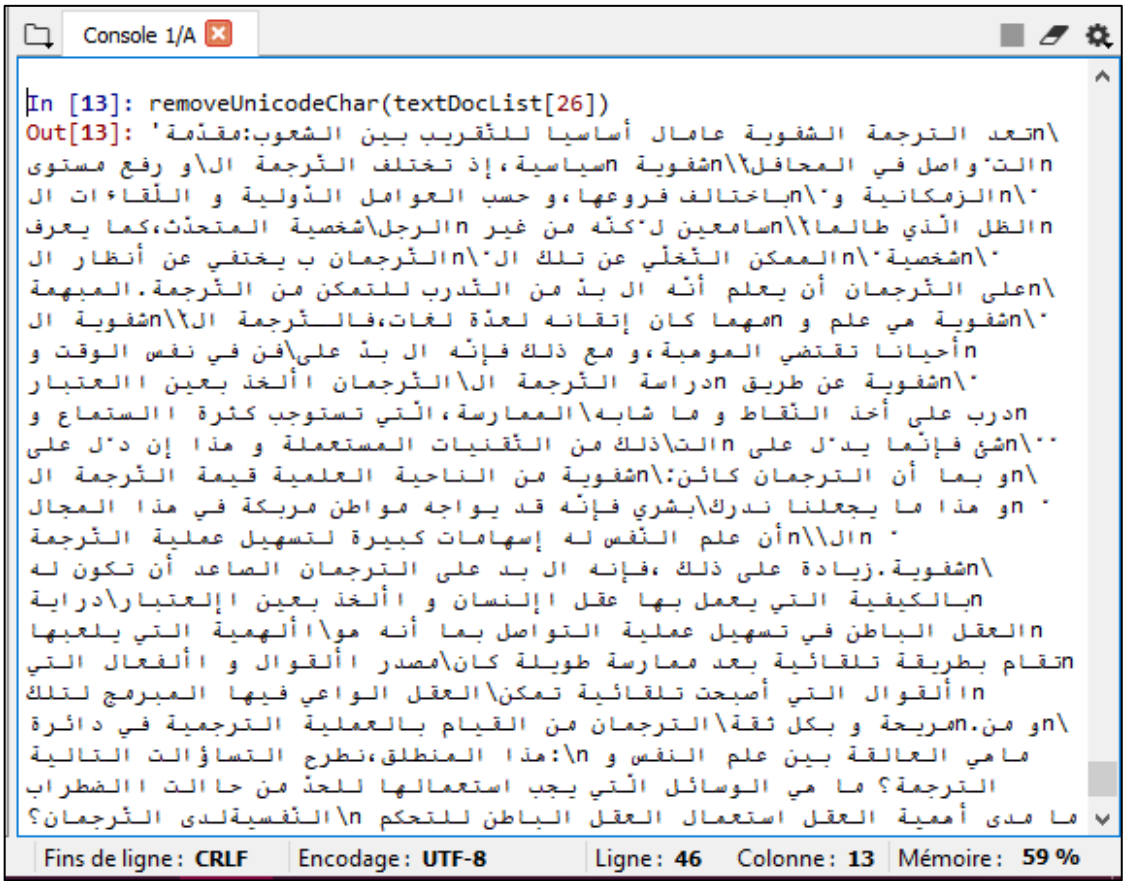


Figure 25 texte propre sans les motifs indésirables

Remarque :

- Nous avons laissé les (\n) pour les sauts de ligne.
- L'élimination des motifs est appliquée généralement pour les textes écrits en arabe.

3.4.2.2.3 La correction morphologique du texte

La correction morphologique du texte extrait est un processus très important pour avoir le contenu exact du fichier PDF

3.4.2.2.3.1 La correction morphologique du texte en utilisant pdfminer

Dans le cas par exemple de la conversion du PDF en texte en utilisant pdfminer pour la langue arabe nous avons observé que tous les caractères sont affichés d'une manière inverser car le pdfminer analyse le texte de gauche à droite (comme le montre la figure 26)ceci présente l'un des méfaits de cette technique.

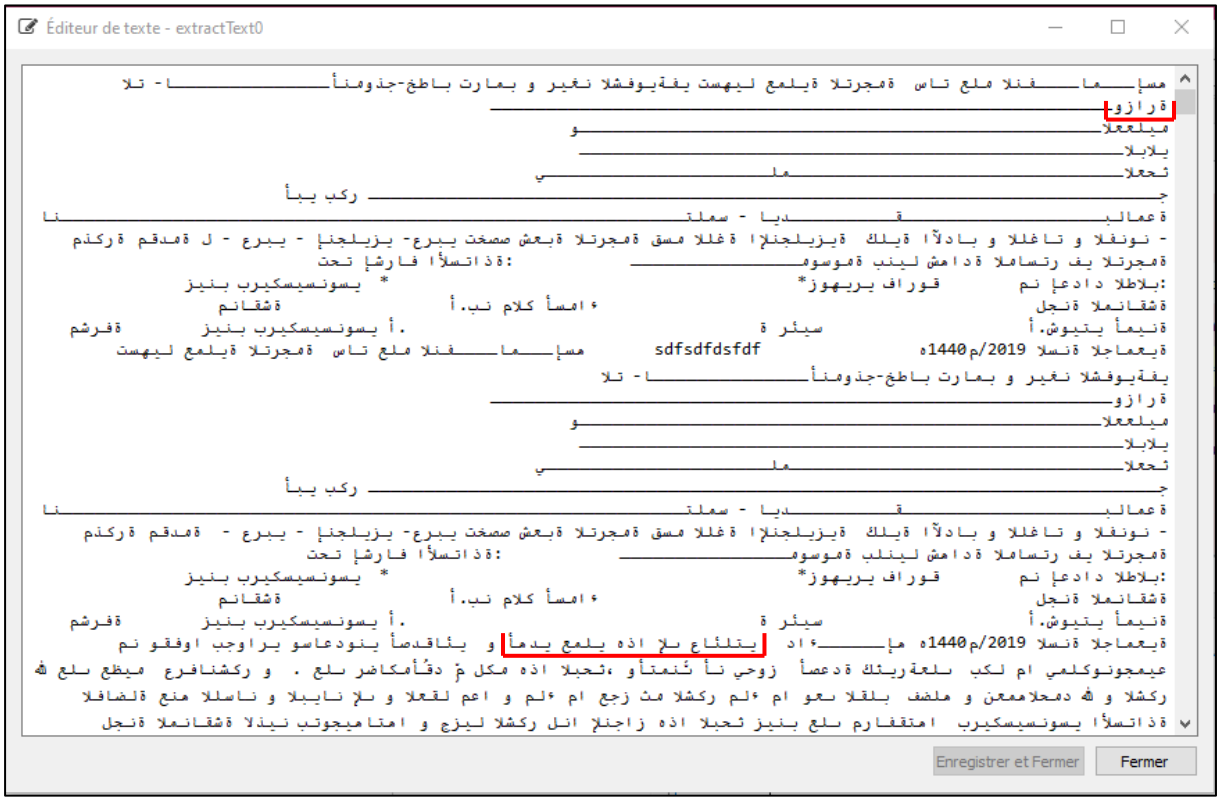


Figure 26 contenu textuel inversé par pdfminer

Pour corriger ce problème nous avons utilisé la méthode reverse_string3 qui inverse tous les caractères de droite à gauche comme le représente les figures (28-29) ci-dessous :

```
from pdfminer.high_level import extract_text
def reverse_string3(s):
    """Return a reversed copy of `s`"""
    chars = list(s)
    for i in range(len(s) // 2):
        tmp = chars[i]
        chars[i] = chars[len(s) - i - 1]
        chars[len(s) - i - 1] = tmp
    return ''.join(chars)
```

Figure 27 Code de la méthode reverse_string3()

Exemple :

```
In [6]: x='ةرازو'

In [7]: reverse_string3(x)
Out[7]: 'وزارة'
```

Figure 28 Résultat de teste d'un seul mot en utilisant reverse_string3()

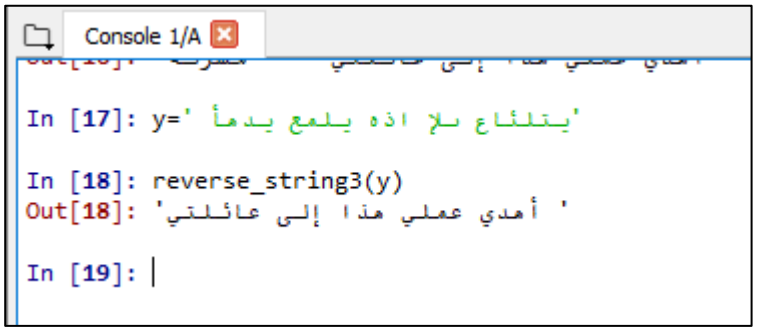


Figure 29 Résultat de teste d'une phrase en utilisant reverse_string3()

Après avoir obtenu la morphologie exacte des chaines de caractères nous avons rencontrés un autre problème l'affichage du texte a était inversé aussi dans cet ordre de la fin jusqu'au début (commençant par la références bibliographique) figure (30)

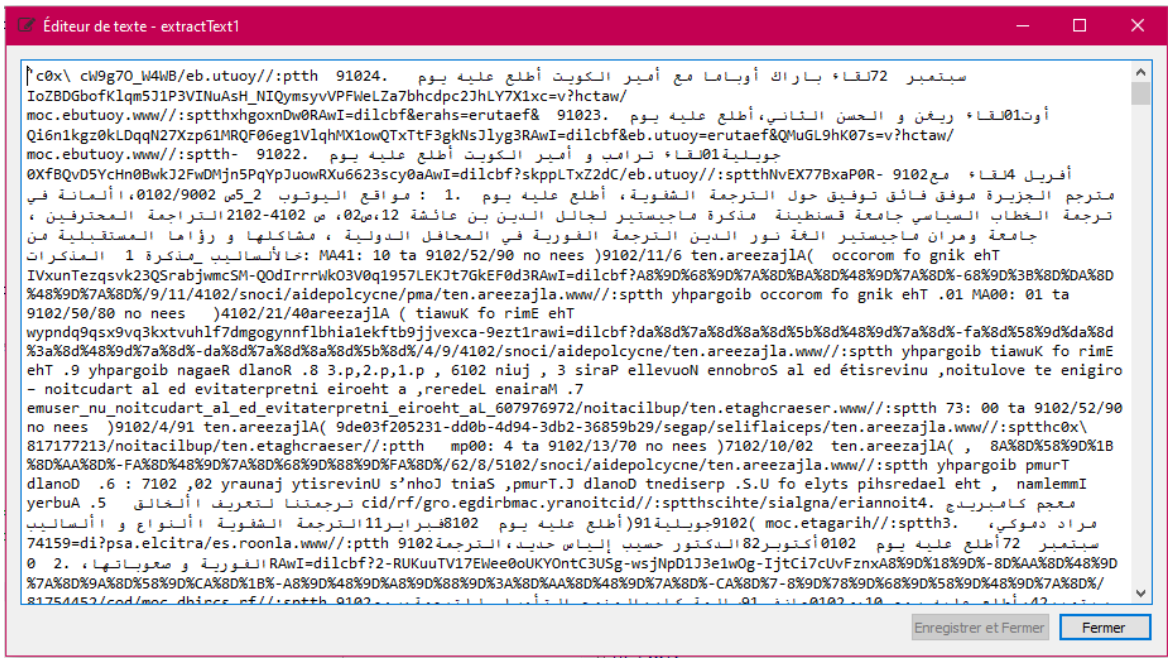


Figure 30 Résultat du texte après l'utilisation de la méthode reverse_string3()

Pour régler ce problème nous avons pensé à convertir notre chaine (extractText1) en liste(extractText2) on utilisant la méthode split() pour découper notre chaine initiale (figure32),ensuite inverser chaque chaine de caractères de cette liste dans le but d'ordonner l'affichage du début jusqu'à la fin avec la méthode reverse() (figure33),enfin nous avons fusionner cette liste en une seule chaine on utilisant la méthode de chaine ' '.join(liste inverser) comme le montre la figure (34).

```
extractText0=extract_text(pdfFileName)
extractText1=reverse_string3(repr(extractText0))#reverse list of caracter
extractText2=extractText1.split()
extractText2.reverse()
extractText3=' '.join(extractText2)# reverse list
```

Figure 31 Code de l'extraction exacte du texte

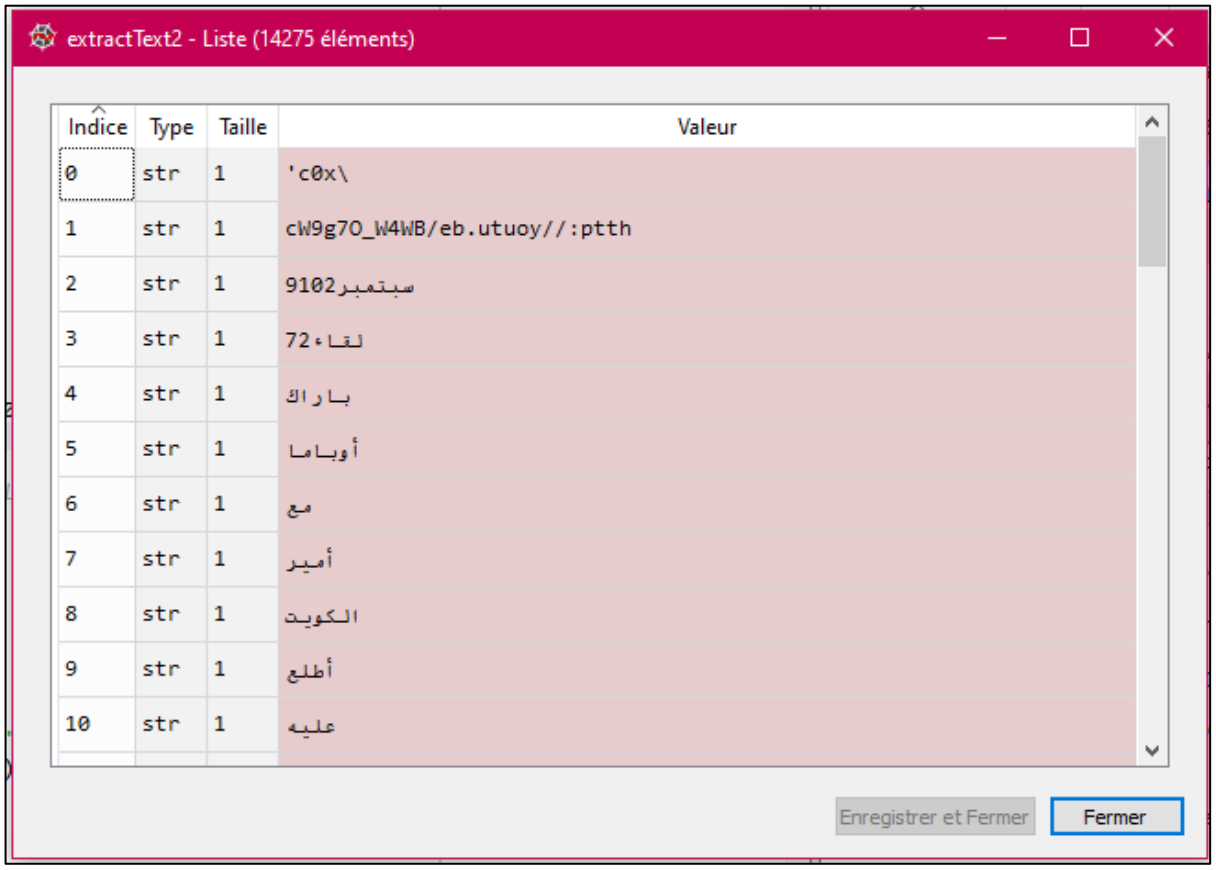


Figure 32 Résultat de la conversion de la chaine en liste

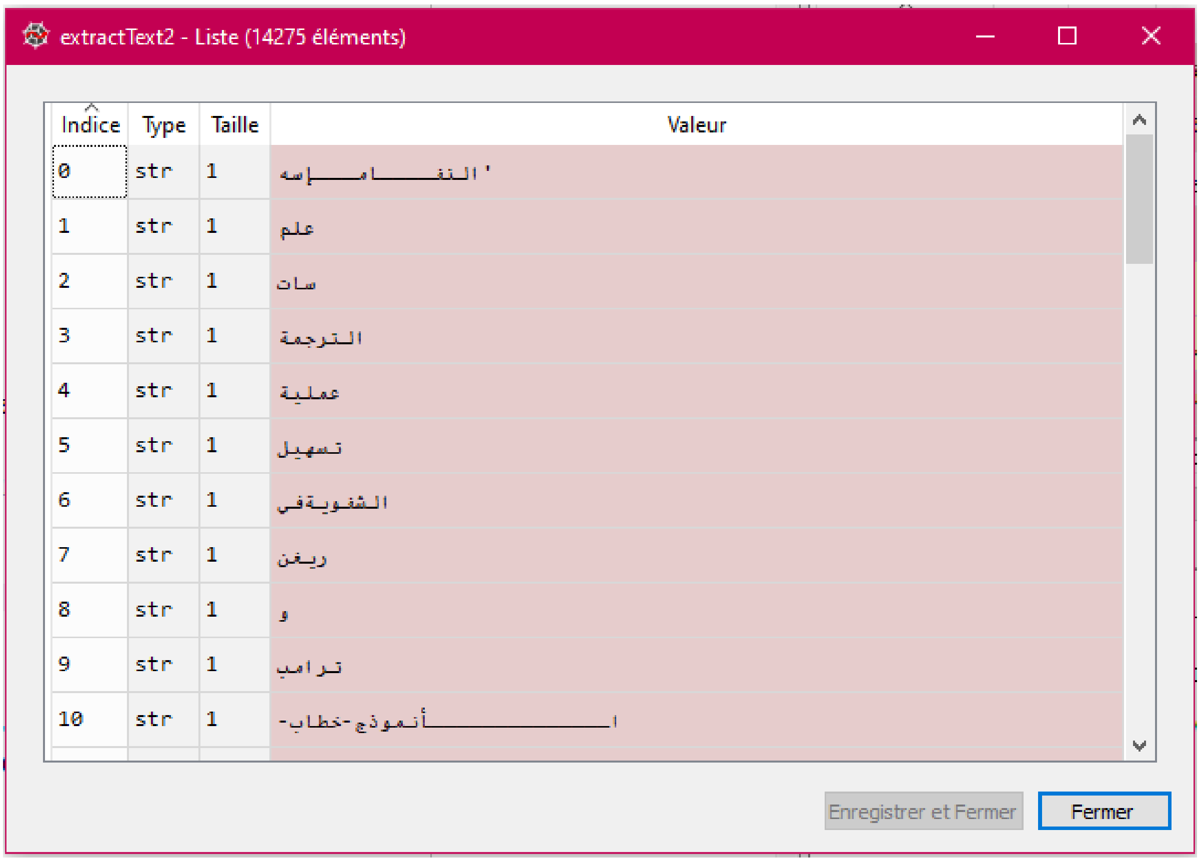


Figure 33 Résultat de l'inversion de la liste du début jusqu'à la fin

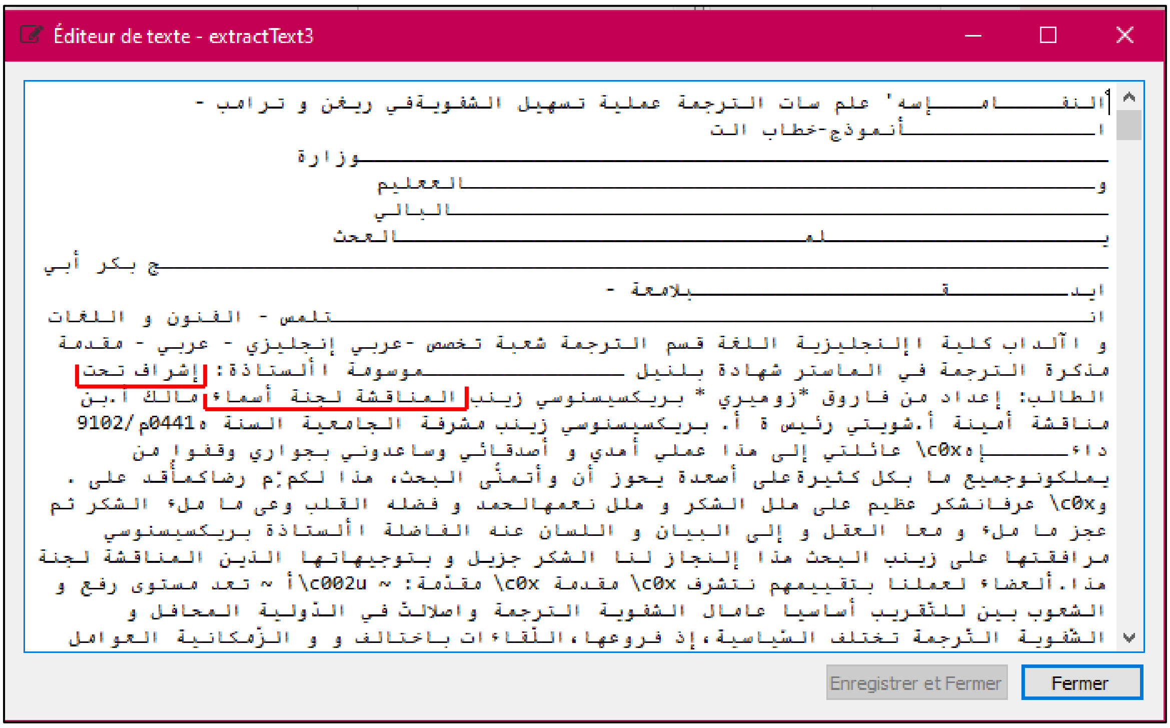


Figure 34 Résultat 1 de la conversion de la liste ordonnée en chaine

En analysant le résultat 1 de la conversion de liste ordonnée en chaine nous avons remarqué que les chaines de caractères sont affichées d'une manière inverser.

Exemple :

إشراف تحت

المناقشة لجنة أسماء

Afin d'ordonner les positions de chaque chaine de caractères nous avons améliorer notre script précèdent en décomposant dans un premier temps la dernière chaine obtenu (extractText3) par point comme le montre la figure36 ensuite pour chaque ligne de extractText4 on a l'divisé par espace et inversé les chaines de caractères pour avoir une phrase affichée de droite à gauche, après nous avons composé la liste (extractText5) qui contient les phrases ordonnées de droite à gauche(figure37) ,enfin nous avons fusionner cette liste en une seule chaine(extractText) on utilisant la méthode de chaine join comme le représente la figure38.

```
# split text by point '.'
extractText4 = extractText3.split('.')
extractText5 = []
for txt in extractText4:
    d = txt.split()
    d.reverse()
    extractText5.append(' '.join(d))
extractText = '\n'.join(extractText5)
textFileName = UPLOAD_FOLDER + '/inputPdf.txt'
with open(textFileName, 'w', encoding="utf-8") as fout:
    for d in extractText:
        fout.write(d)
```

Figure 35 la suite du code de l'extraction exacte du texte

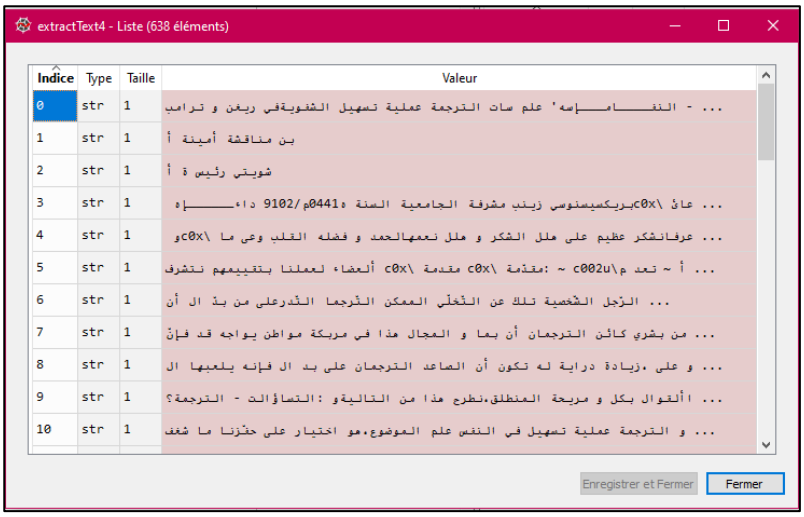


Figure 36 Résultat de la décomposition du contenu textuel par point

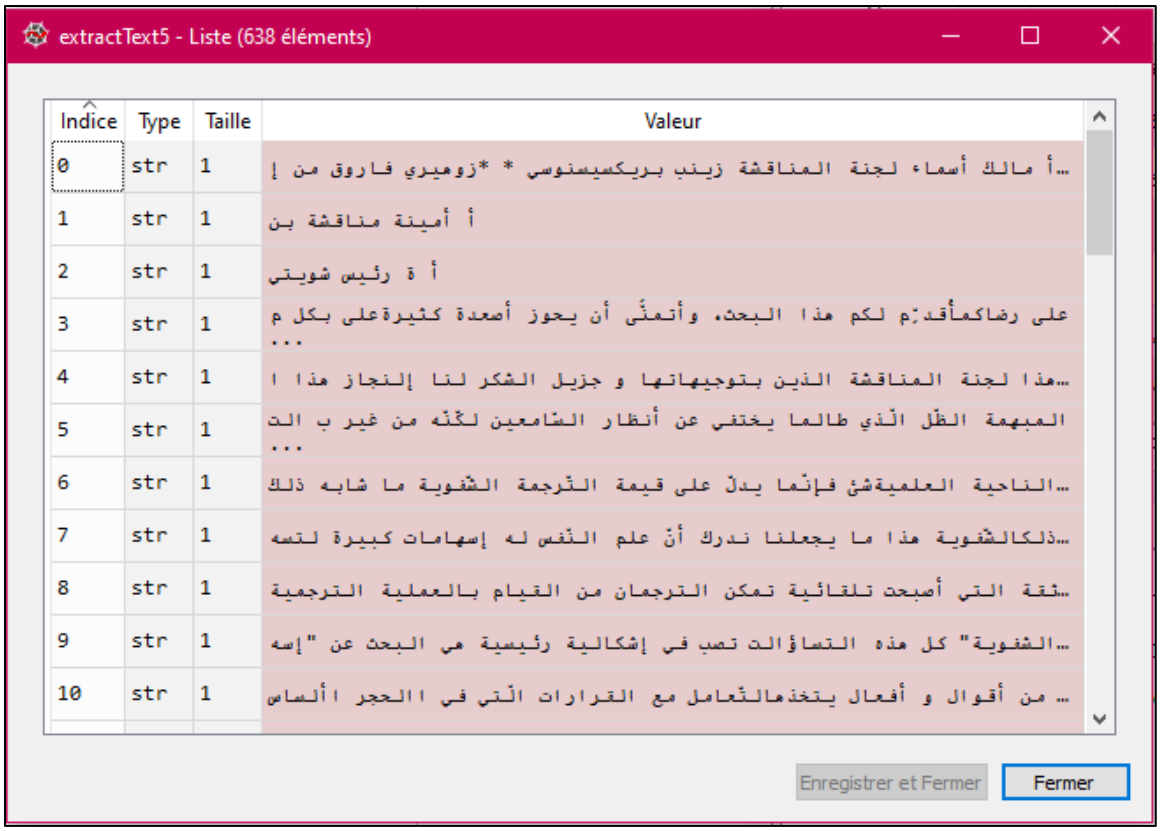


Figure 37 Résultat de l'inversion des positions des chaines de caractères

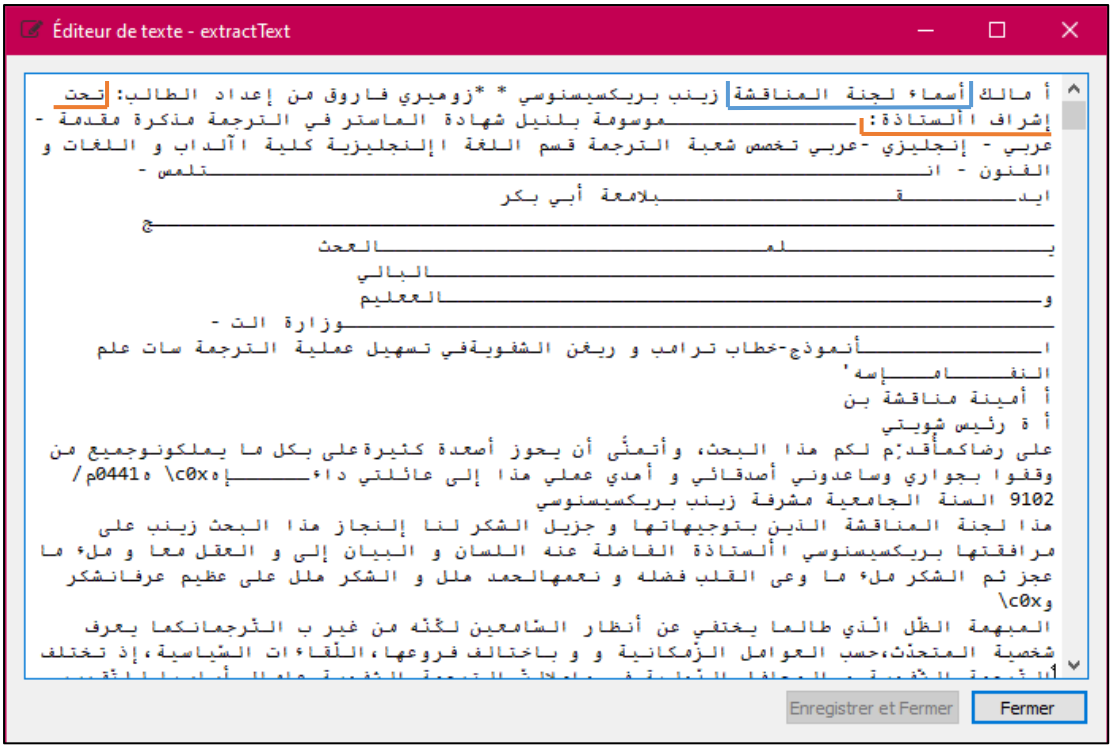


Figure 38 Résultat 2 de la conversion de la liste ordonnée en chaine

3.4.3 La détection des titres

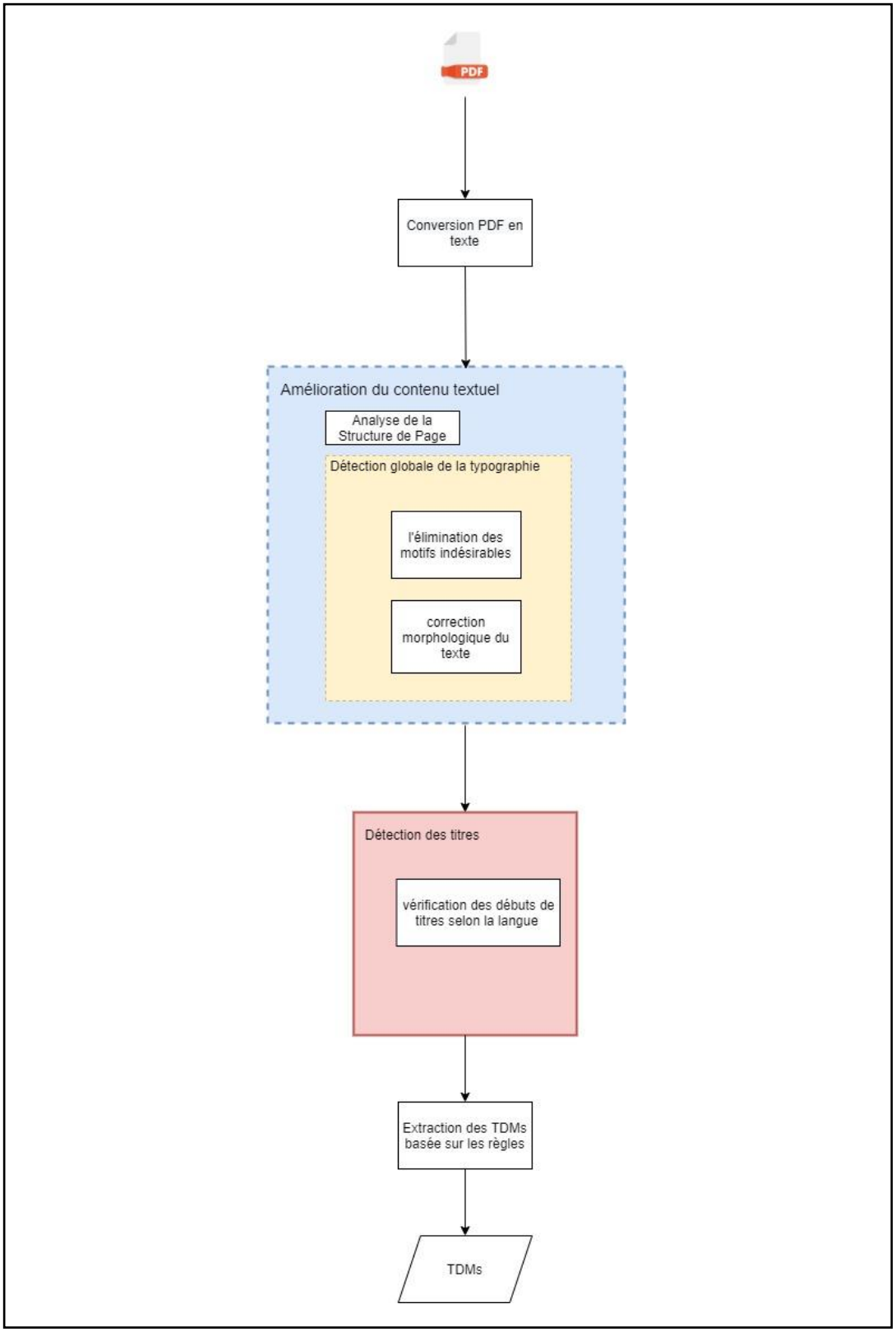


Figure 39 phase de la détection des titres

C’est l’étape la plus importante dans notre travail elle est considérée comme une tache préliminaire pour l’extraction de la TDM, elle aidera le système a une meilleure classification des titres ,Le document est divisé en blocs de texte (un bloc de texte regroupe les lignes qui ont la même disposition et qui sont spatialement proches les unes des autres) en détectant les renseignements dans les données textuelles déjà extraite dans la phase précédente, chaque bloc de texte doit être vérifier s’il s’agit d’un titre ou non.

Parmi les caractéristiques des titres traité dans notre programme on trouve :

Is_digit : True si le bloc de texte commence par une numérotation telle que : les nombres entiers : 1. ,1.1 ,2) ,3., 4.1.1, …, 9.etc, False sinon

Is-alpha : True si le bloc de texte commence par un alphabet telle que : a., b), c., d, ,…..,z ou l’alphabet romains : i) ,(ii), iii. , iv) ….etc. ou l’alphabet arabe : ا. , ب) , …..ي, False sinon

Is_upper : True si le bloc de texte commence par une majuscule telle que : A., B), C., D, , …..,Z ou l’alphabet romains : I) ,(II), III. , IV) ….etc. False sinon

Is_lower : True si le bloc de texte commence par une minuscule telle que : a., b), c., d, , …..,z ou l’alphabet romains : i) ,(ii), iii. , iv) ….etc. False sinon

```python
@author: DELL
"""

import glob
def hasNumbers(inputString):
    return any(char.isdigit() for char in inputString)

def startOfTitle(text):
    try:
        # (*)
        if(text[0]=='(' and text[-1]==')' and ':' not in text):
            return True
        # .1
        elif(text[0] in ',.-)/' and text[-1].isdigit() and ' ' not in text):
            return True
        # .ba
        elif(text[0] in ',.-)/' and text[-1].isalpha() and ' ' not in text):
            return True
        # 1.1.1.1
        elif('.' in text and text.replace('.','').isdigit() and len(text.replace('.',''))<10 and len(text.replace('.',''))>1 and ' ' not in text):
            return True
        else:
            return False
    except Exception:
        pass

def isTitle(text):
    start = text.split()[0]
    if startOfTitle(start):
        return True

    else:
        return False
```

Figure 40 fonction de la vérification des titres en arabe

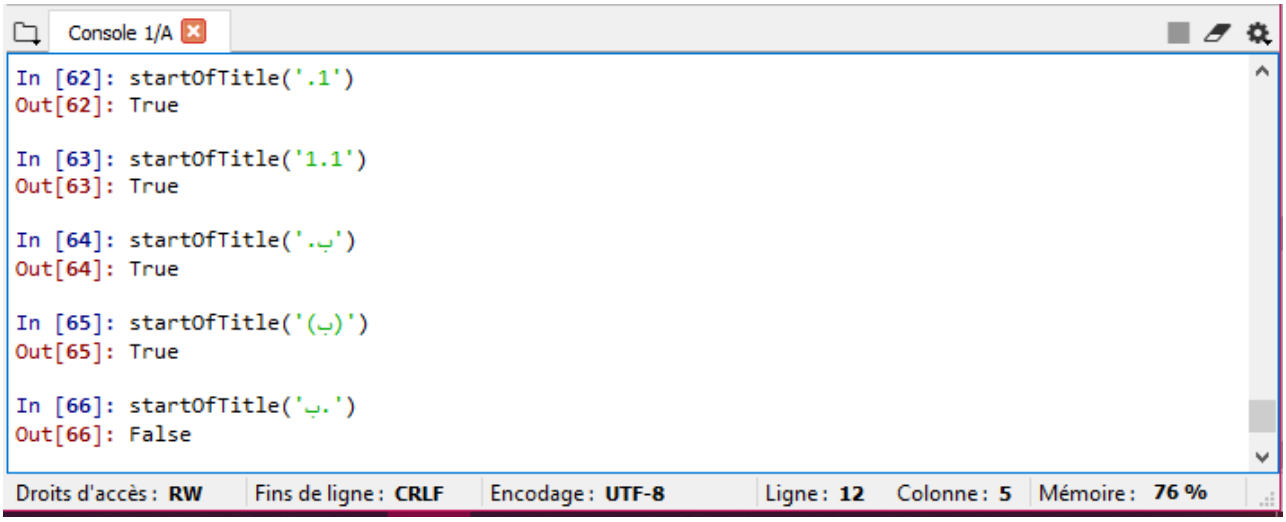


Figure 41 résultat du test de la fonction startOfTitle sur les débuts des titres écrites en Arabe

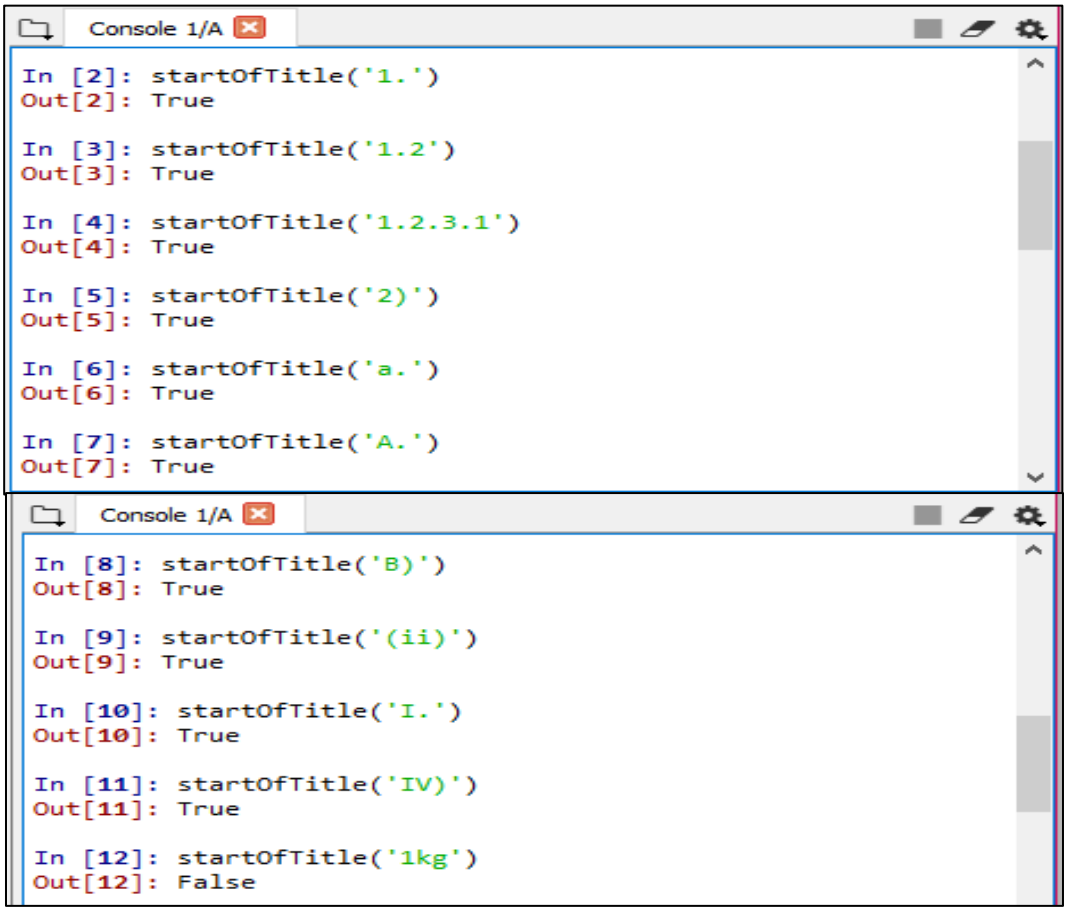


Figure 42 résultat du test de la fonction startOfTitle sur les débuts des titres écrites en anglais et en Français

3.4.4 L’extraction de la TDM

Dans cette étape l'objectif est d'identifier le niveau hiérarchique des titres et d'organiser les titres du document selon la structure hiérarchique pour produire la TDM finale après le filtrage des titres. La figure 43 montre les détails de cette phase :

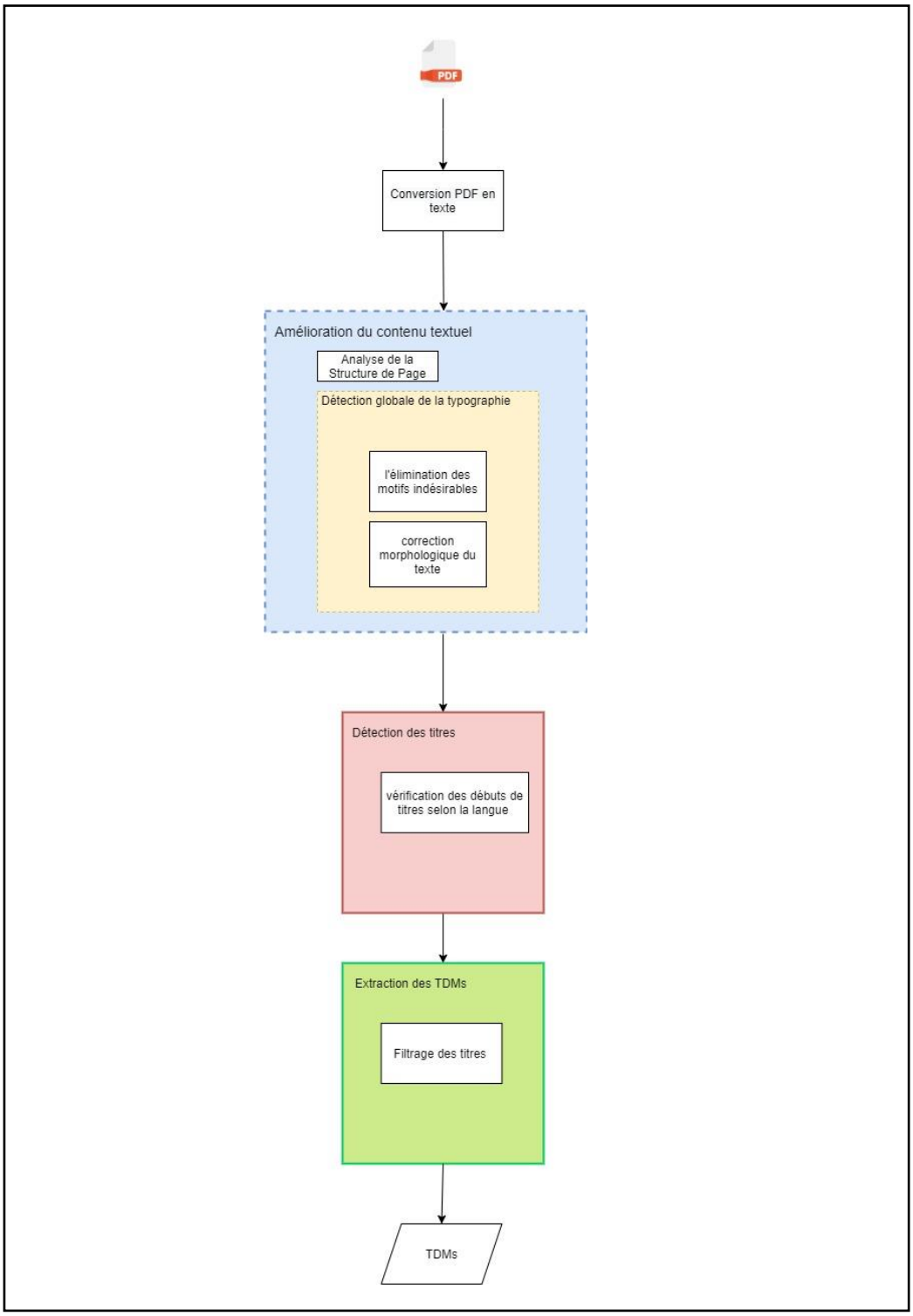


Figure 43 phase de l'extraction des TDMs

3.4.4.1 Filtrage des titres

Dans le but d'éliminer les phrases qui ne sont pas à la base des titres nous avons pensé à filtrer notre liste de titres extraite dans la phase précédente en limitant le nombre des chaines de caractères dans un titre, nous avons choisis de prendre pour chaque titre 5 chaines de caractères au maximum pour les documents écrits en anglais et en français.

D'abord nous avons crées des listes que nous l'avions nommés (pattern1-pattern2-pattern3) chacune d'elle contient des caractères qui apparaisent dans les chaines extraites dans la liste de titres et que nous nous désirons pas quelles soient afficher comme étant des titres dans notres TDM finale par exemple :

18.23%

1. introduction……12 (nous avons traité ce cas pour éliminer les titres de la TDM de base).

Education, art and culture and love (commence avec une majuscule et contient une virgule c'est pour sa que notre script l'as considérer comme titre avant le filtrage).

2010-01 (commence par un nombre et contient un tiret de six).

Afin de ne pas prendre les exemples précédents comme étant des titres nous avons développées ce script (figure45) qui vérifie pour chaque titre de la liste des titres extraite est ce qu'il s'agit bien d'un titre souhaité ou pas ,dans un premier temps la vérification du début de titre si il commence par un numéro ou pas ensuite ce dernier dois avoir au minimum 2 chaines de caractères et au maximum 5 chaines de caractères ,ce titre ne dois pas contenir un caractère indésirable comme vous pouvais les voir dans la figure 45 a l'interieur du titre. Si la condition est vérifée donc il l'ajoute dans la TDM finale (outToc) sinon il ne le prend pas en considération. Pour les titres qui comence par une majuscule la meme procédure sera effectuer pour le filtrage .

```
def tdmExtract(tocLst, lang):
    pattern1 = ['…','..','%',',',':']
    pattern2 = ['…','..','%',',','-','.']
    pattern3 = ['…','..','%',',','-']
    if lang == "Latin":
        outToc = []
        tocList = tocLst
        for ttl in tocList[:-1]:
            if ttl[0].isdigit() and len(ttl.split())<6 and len(ttl.split())>1 and not any([substring in ttl for substring in pattern1]):
                outToc.append(ttl.replace(':',''))
            elif ttl[0].isupper() and len(ttl.split())<6 and len(ttl.split())>1 and not any([substring in ttl for substring in pattern2]):
                outToc.append(ttl.replace(':','').rstrip())
```

Figure 44 Code de filtrage des titres en anglais et en français

Pour le filtrage des titres en arabe dois réspecter les memes condition que l'anglais et le français sauf que nous avons changée le nombre maximale de chaines de caractères du titres en 9 chaines de caractères car apres les expérimentation éffectuer nous avons remarqué que les titres en arabe sont plus long que d'autre langues.

```
elif lang == "Arabic":
    outToc = []
    tocList = tocLst
    for ttl in tocList[:-1]:
        if ttl[0].isdigit() and len(ttl.split())<10 and len(ttl.split())>1 and not any([substring in ttl for substring in pattern1]):
            outToc.append(ttl.replace(':',''))
        elif len(ttl.split())<10 and len(ttl.split())>1 and not any([substring in ttl for substring in pattern3]):
            outToc.append(ttl.replace(':',''))
```

Figure 45 Code de filtrage des titres en arabe

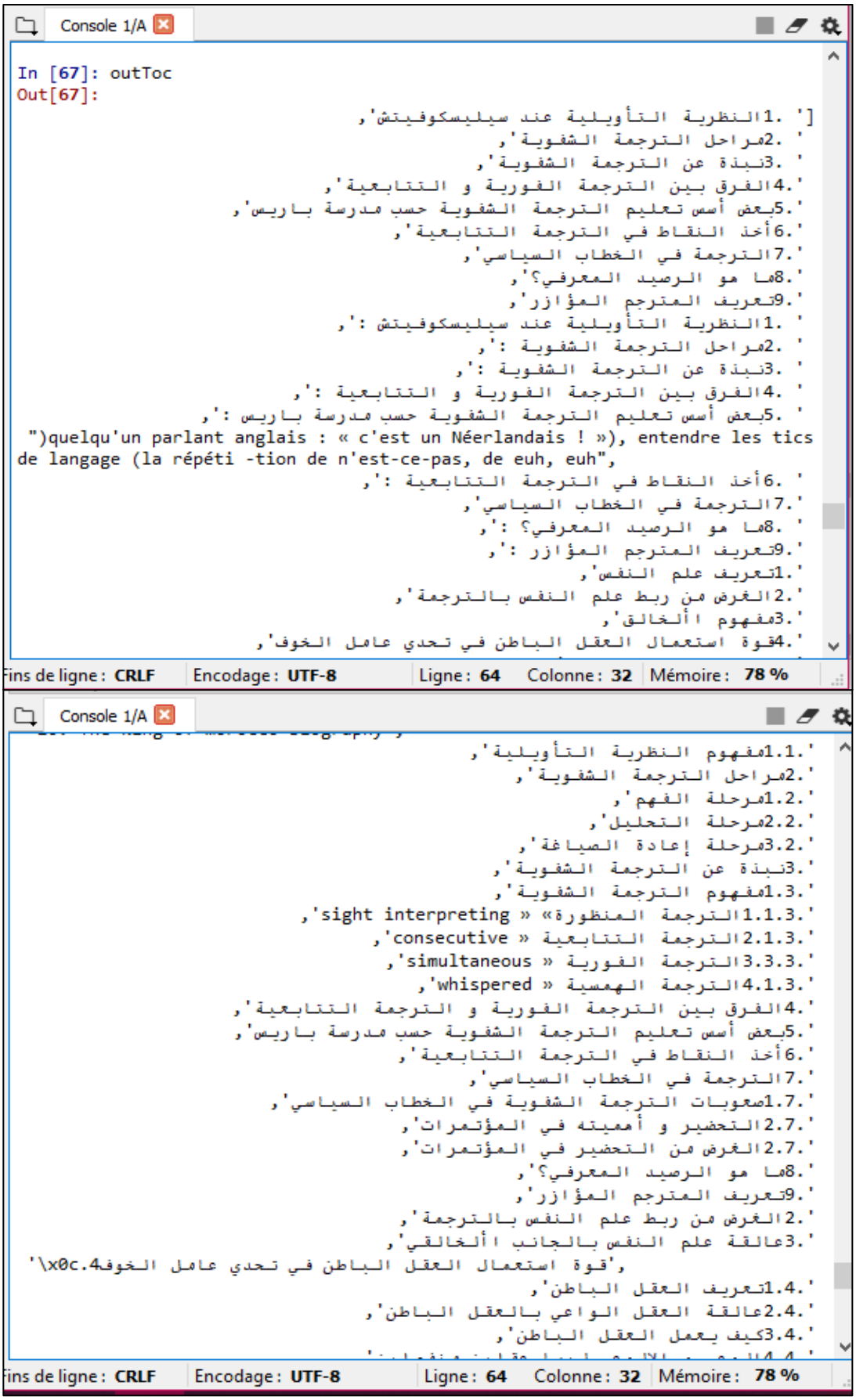


Figure 46 TDM extraite en utilisant poppler

3.5 Diagramme UML

Pour la conception du système on va vous présenter quelques diagrammes de modélisation, qu'on a jugé les plus importants pour la compréhension du fonctionnement de notre système.

3.5.1 Diagramme de cas d'utilisation

Nous avons choisi la conception de ce diagramme afin d'avoir une idée générale sur les fonctionnalités du futur système. Il est constitué d'un ensemble d'acteurs qui agit sur des cas d'utilisation.

3.5.1.1 Identification des acteurs

Au niveau de cette section, nous présentons les différents acteurs susceptibles d'interagir avec le système, pour notre cas La mise en marche de notre système nécessite essentiellement un seul acteur :

Acteur	**Rôles**
L'utilisateur	✓ Sélectionner un fichier PDF. ✓ Convertir un PDF en texte. ✓ Extraire les titres. ✓ Extraire la TDM.

Tableau 5 Les rôles de l'utilisateur

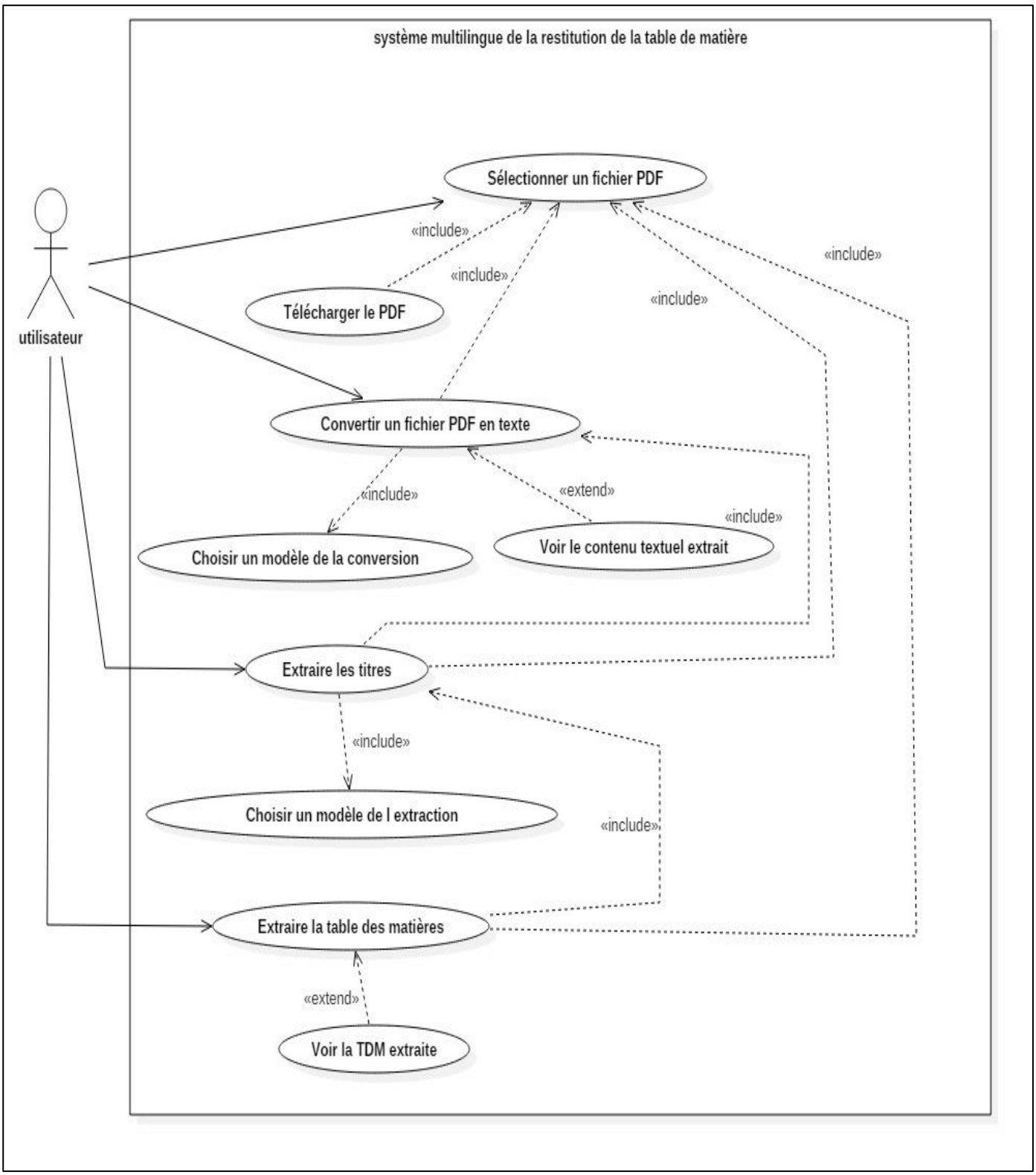


Figure 47 Diagramme de cas d'utilisation du système de restitution de la TDM

3.5.2 Diagramme de séquence et description textuelle

Afin d'obtenir une meilleure visualisation sur les interactions entre les éléments de notre système nous avions élaboré les diagrammes de séquence suivants qui représentent les principaux scénarios nominaux :

3.5.2.1 Diagramme de séquence du cas « Sélectionner un fichier PDF »

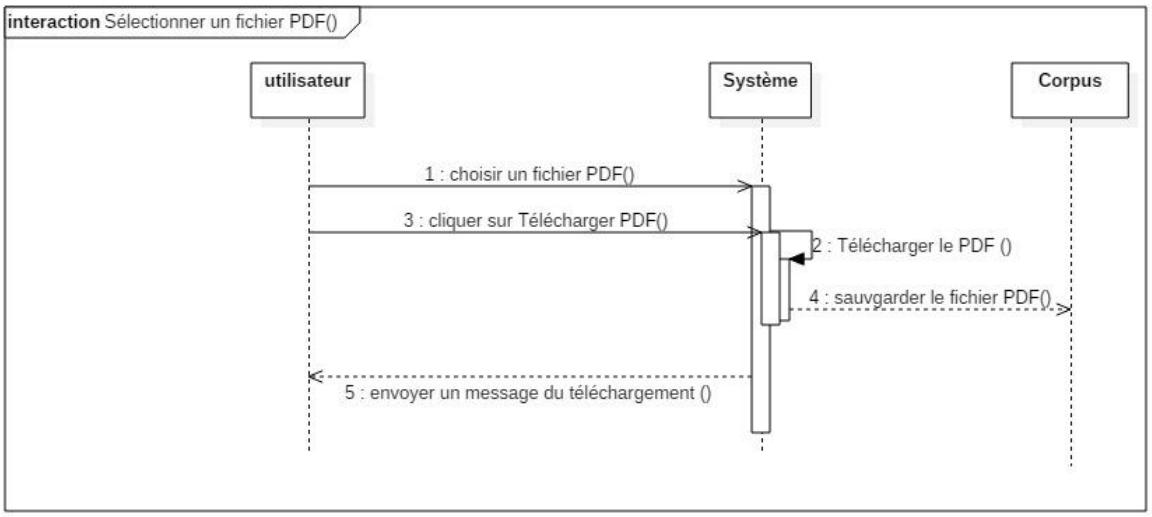


Figure 48 Diagramme de séquence de la sélection d'un fichier PDF

Cas d'utilisation	Sélectionner un fichier PDF.
Acteurs	Utilisateur
Résumé	L'utilisateur doit choisir un fichier PDF pour accéder à ses services.
Pré condition	L'utilisateur dois avoir un fichier PDF qui souhaite le traiter.
Post condition	Le PDF est téléchargé
Scénario	1. Le système affiche l'interface de la sélection.

	2. L'utilisateur choisis le fichier PDF souhaité. 3. L'utilisateur clique sur le bouton « téléchargerPDF » pour garder une trace sur le PDF choisi. 4. Le système télécharge le fichier PDF sélectionné et le sauvegarder dans le corpus. 5. Le système affiche un message de succès du téléchargement du PDF.

Tableau 6 table d'identification de cas d'utilisation "Sélectionner un fichier PDF"

3.5.2.2 Diagramme de séquence du cas « Convertir PDF en texte »

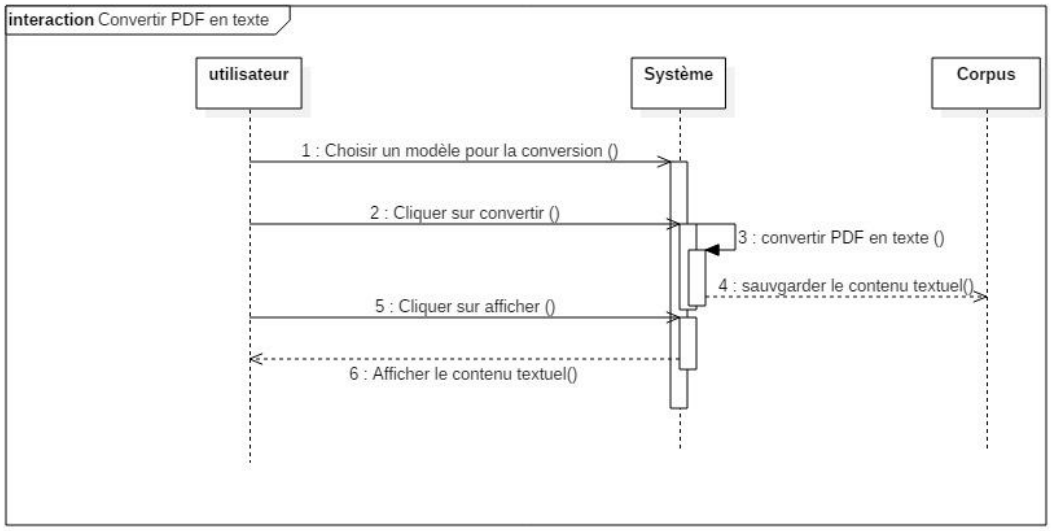


Figure 49 Diagramme de séquence du cas d'utilisation « Convertir PDF en texte »

Cas d'utilisation	Convertir PDF en texte
Acteurs	Utilisateur
Résumé	Permettre à l'utilisateur d'extraire le contenu textuel du fichier PDF choisi.
Pré condition	L'utilisateur dois choisir le modèle de la conversion qui lui convient.
Post condition	Fichiers PDF convertis en texte.
Scénario	1. Le système affiche la page de la conversion du PDF en texte. 2. L'utilisateur choisis un modèle de la conversion. 3. L'utilisateur clique sur le bouton convertir. 4. Le système convertis le PDF en texte et le sauvegarde dans le corpus. 5. L'utilisateur clique sur afficher pour voir le contenu textuel. 6. Le système affiche le contenu textuel.

Tableau 7 table d'identification de cas d'utilisation "Convertir PDF en texte"

3.5.2.3 Diagramme de séquence du cas « Extraire les titres »

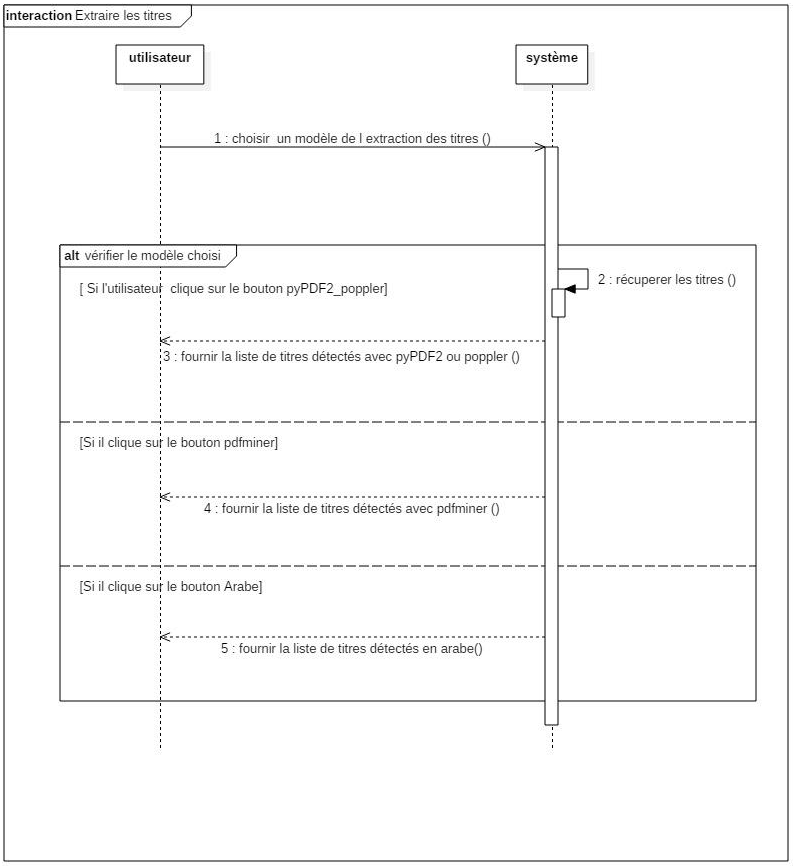


Figure 50 Diagramme de séquence du cas d'utilisation « Extraire les titres »

Cas d'utilisation	Extraire les Titres
Acteurs	Utilisateur
Résumé	Permettre à l'utilisateur d'extraire les titres
Pré condition	L'utilisateur dois choisir un modèle de l'extraction des titres.
Post condition	Une liste de titres.
Scénario	1. Le système affiche la page de l'extraction des titres.

	2. L'utilisateur choisis la méthode de l'extraction des titres selon le modèle choisi pour la conversion du PDF en texte. 3. Le système récupère tous les titres qui existent dans le contenu textuel. 4. L'utilisateur clique sur afficher pour voir les titres. 5. Le système renvoie une liste de titres détectés selon le choix de l'utilisateur.

Tableau 8 table d'identification de cas d'utilisation "Extraire les titres"

3.5.2.4 Diagramme du cas « Extraire la TDM »

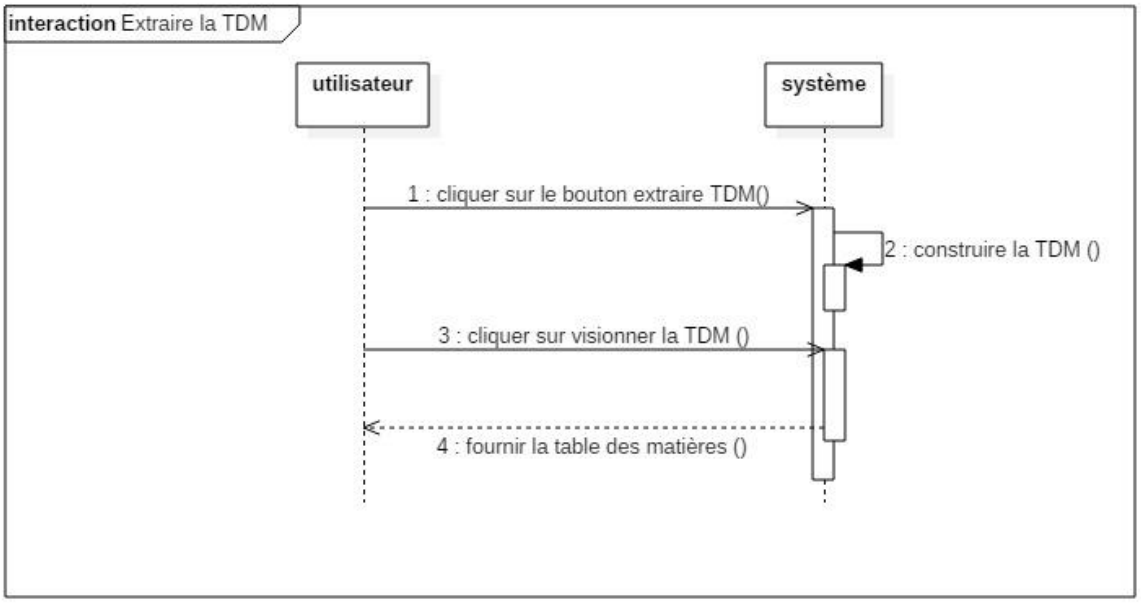


Figure 51 Diagramme de séquence du cas d'utilisation « Extraire la TDM »

Cas d'utilisation	Extraire la TDM
Acteurs	Utilisateur
Résumé	Permettre à l'utilisateur d'extraire la TDM à partir du fichier PDF sélectionné.
Pré condition	L'utilisateur dois cliquer sur extraire la TDM.
Post condition	Construction de la TDM
Scénario	1. Le système affiche la page de l'extraction la TDM. 2. L'utilisateur clique sur le bouton extraire la TDM. 3. Le système construit la TDM 4. L'utilisateur cliquer sur visionner la TDM pour l'apercevoir. 5. Le système affiche la TDM construite.

Tableau 9 table d'identification de cas d'utilisation "Extraire la TDM"

3.6 Conclusion

Nous avons présenté dans ce chapitre notre corpus ainsi que l'architecture général et la description de ces différentes phases. Et bien sûr les diagrammes UML ce concept nous a permis de structurer notre système et mieux comprendre son comportement.

Le prochain chapitre nous permettra de passer à la phase de l'implémentation et les tests de notre solution proposé. Et de décrire les outils utilisé pour réaliser ce projet en détail ainsi que l'évaluation de notre système.

Chapitre IV : Implémentation tests et évaluation de la solution

4.1 Introduction

A ce stade, le problème a été analysé en profondeur. Nous avons défini une conception mieux appropriée aux besoins de l'application. Nous pouvons alors entreprendre la prochaine étape du processus du développement, qui a comme objectif d'aboutir à un produit final, prêt a être en main des utilisateurs.

Dans cette phase nous allons présenter les outils de développement que nous avons utilisé pour implémenter les cas d'utilisation, les tester et enfin l'évaluation de notre système.

4.2 Environnement de développement

Nous présentons dans cette partie, le langage de programmation python utilisé et son environnement de développement spyder, ainsi que flask que nous avons utilisée pour la construction des interfaces (les composants) de notre système et sans oublier les utiles qui ont contribués à la réalisation du coté fonctionnel de notre application PyPDF2, Poppler.

4.2.1 Python

Python est un langage de programmation qui a été initialement conceptualisé par Guido van Rossum à la fin des années 1980 en tant que membre de l'Institut national de recherche en mathématiques et en informatique, Bien sûr, Python, comme d'autres langues, a connu plusieurs versions. Python 0.9.0 est sorti pour la première fois en 1991. En plus de la gestion des exceptions, Python comprenait des classes, des listes et des chaînes de caractères.

En 2000, Python 2.0 a été publié. Cette version était plutôt un projet à open-source émanant de membres de l'institut national de recherche en mathématique et en informatique. Cette version de python incluait des connaissances de listes, un collecteur d'ordures complet et la prise en charge de l'Unicode.

Python 3.0 était la version suivante et a été publié en décembre 2008 (la dernière version de Python est la 3.8.5). Bien que Python 2 et 3 soient similaires, il existe de subtiles différences. La différence la plus notable réside peut-être dans le

fonctionnement de la commande print, puisque dans Python 3.0, la commande print a été remplacée par une fonction print (). [13]

4.2.1.1 Caractéristiques de Python

Python offre de nombreuses fonctionnalités utiles qui le rendent populaire et précieux par rapport aux autres langages de programmation. Il supporte la programmation orientée objet, les approches de programmation procédurale et permet une allocation dynamique de la mémoire. Parmi les avantages de ce langage on trouve :

- Il est facile à apprendre et à utiliser.
- C'est un langage expressif car il peut effectuer des tâches complexes en utilisant quelques lignes de code.
- Python est un langage interprété, cela signifie que le programme Python est exécuté une ligne à la fois, cela rend le débogage facile et portable.
- Libre et open source. [14]

Figure 52 logo du langage python

4.2.2 Spyder

Spyder est un environnement scientifique puissant écrit en Python, pour Python, et conçu par et pour les scientifiques, les ingénieurs et les analystes de données. Il offre une combinaison unique de la fonctionnalité avancée d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec l'exploration des données, l'exécution interactive, l'inspection approfondie et de belles capacités de visualisation d'un paquet scientifique.

Au-delà de ses nombreuses fonctionnalités intégrées, ses capacités peuvent être étendues encore plus loin via son système de plugin et son API. En outre, Spyder peut également être utilisé comme une bibliothèque d'extension Pyqt5, permettant aux développeurs de construire sur sa fonctionnalité et intégrer ses composants, tels que la console interactive, dans leur propre logiciel Pyqt. [15]

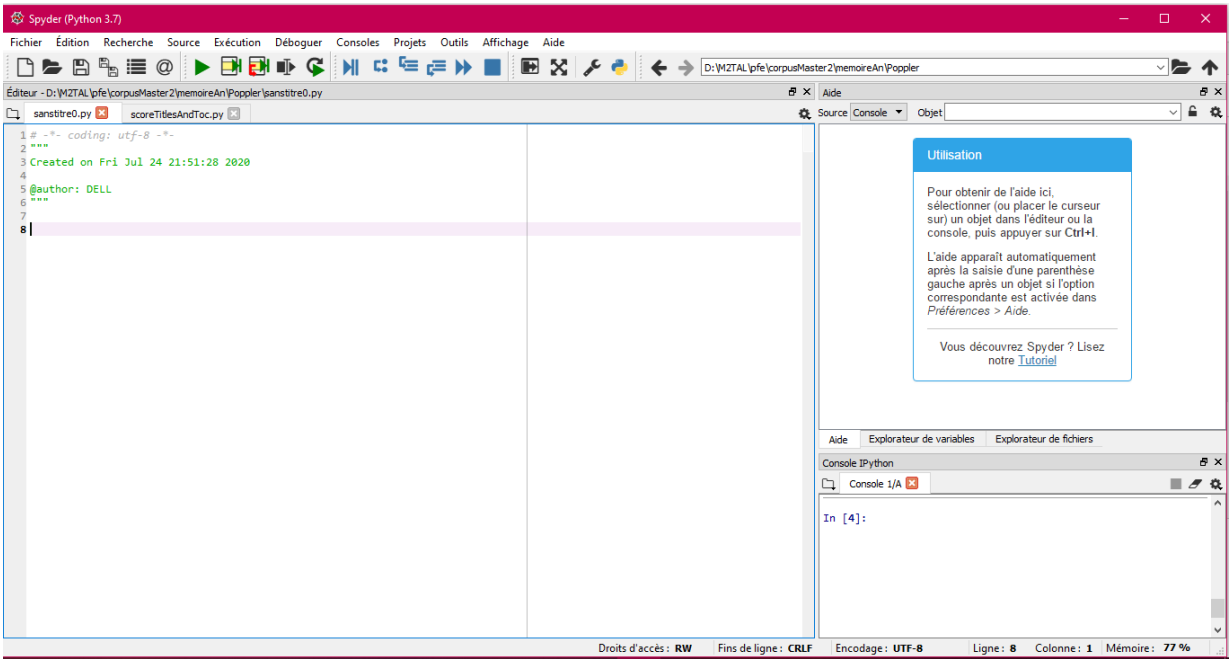


Figure 53 Environnement de développement Spyder

4.2.3 Flask

Flask est un micro Framework (une microstructure) Python sous licence BSD, basée sur Werkzeug et Jinja2. Le fait d'être un micro Framework ne le rend pas moins fonctionnel ; Flask est un cadre très simple mais très extensible. Cela donne aux développeurs la possibilité de choisir la configuration qu'ils souhaitent, ce qui facilite l'écriture d'applications ou de plugins. Flask a été initialement crée par Pocoo, une équipe de développeurs open source en 2010, et il est maintenant développé et maintenu par The Pallets Project qui alimente tous les composants derrière Flask. Flask est soutenu par une communauté de développeurs active et utile, qui comprend un canal IRC actif et une liste de diffusion. [16]

4.2.3.1 Caractéristiques de Flask

- Supports intégrés pour les essais unitaires
- Utilise le templating Jinja2
- Prise en charge des cookies sécurisés
- Vaste documentation
- Compatibilité des moteurs d'applications Google
- Demande de répartition
- Unicode based [17]

Figure 54 Logo de Flask

4.3 Les outils utilisés

4.3.1 PyPDF2

Pypdf2 est un paquet open-source en Python qui peut extraire des données de PDF au format textuel. Il ne reconnaît pas la disposition des tableaux et il extrait simplement les données au format texte. Ce paquet est utile pour extraire du texte de nombreux fichiers PDF qui peuvent être utilisés plus tard aux applications du NLP. [18]

Le package pyPdf d'origine a été publié en 2005. La dernière version officielle de pyPdf date de 2010. Après un laps de temps d'environ un an, une entreprise appelée Phasit à parrainé une fourchette de pyPdf appelée PyPDF2. Le code a été écrit pour être rétro compatible avec l'original et a très bien fonctionné pendant plusieurs années, sa dernière version étant en 2016. [19]

4.3.1.1 Caractéristique de PyPDF2

Une bibliothèque Pure-Python construite comme une boîte à outils PDF. Elle est capable de :

- D'extraire les informations d'un document (titre, auteur, ...)
- La division des documents page par page
- Fusion de documents page par page
- Recadrage des pages
- La fusion de plusieurs pages en une seule
- Le cryptage et le décryptage des fichiers PDF, et plus encore ![6]

4.3.2 Poppler

Poppler est un utilitaire gratuit qui permet de rendre des documents PDF (Portable Document Format). Parmi la liste des fonctionnalités très utiles, Poppler vous permet de convertir des fichiers .pdf en . txt vous permettant d'utiliser toute la formule de Foxtrot pour extraire des informations du document avec une grande précision, flexibilité et rapidité. [20]

4.3.2.1 Caractéristique de Poppler

- Lire une modification des métadonnées d'un document ;
- Lister et lire des documents intégrés ;
- Lister les polices utilisées par le document ;
- Rechercher ou extraire du texte sur une page donnée du document ;
- Rendre une page en une image brute ;
- Obtenir des informations sur les effets de transition entre les pages ;
- Lire la table des matières du document.[7]

[6] https://pypi.org/project/PyPDF2/
[7] https://pypi.org/project/python-poppler/

4.3.3 Pdfminer

PDFMiner est un outil pour extraire des informations à partir de documents PDF. Contrairement à d'autres outils liés au PDF, il se concentre entièrement sur l'obtention et l'analyse de données texte. PDFMiner permet d'obtenir l'emplacement exact du texte dans une page, ainsi que d'autres informations telles que les polices ou les lignes. Il comprend un convertisseur PDF qui peut transformer des fichiers PDF en d'autres formats de texte (tels que HTML). Il dispose d'un analyseur PDF extensible qui peut être utilisé à d'autres fins que l'analyse de texte.

4.3.3.1 Caractéristique de Pdfminer

- Entièrement écrit en Python.
- Analyser et convertir les documents PDF.
- Prise en charge de différents types de polices (Type1, TrueType, Type3 et CID).
- Conversion PDF en HTML (avec une application Web de convertisseur d'échantillons).
- Extraction des contours (TDM).
- Étiqueté l'extraction du contenu.
- Reconstruisez la mise en page originale en regroupant des morceaux de texte.[8]

4.4 Description de notre système

Pour le développement de notre système nous avons choisis de travailler avec trois méthodes qui sont : Poppler et PyPDF2 et pdfminer, afin de pouvoir détecter les titres et d'extraire la TDM, Nous avons procédé à plusieurs expérimentations afin de valider nos méthodes, Notre principale expérience a été réalisée sur le corpus (Master2), un corpus constitué de mémoires de niveau master2 de l'enseignement supérieure de différentes (langues (en anglais/français/arabe), université, taille). Une deuxième expérience a été effectuée sur un second corpus d'articles scientifiques. Afin d'évaluer la possibilité d'utiliser nos méthodes sur d'autres types de textes,

[8] https://pdfminer-docs.readthedocs.io/pdfminer_index.html

nous avons également réalisé une expérimentation sur un corpus du documents financiers.

4.5 Les Tests

4.5.1 Page d'accueil et de service

Avant de tester les fonctionnalités de notre application on vous présente tout d'abord la page d'accueil et de service de notre système, cette page est ergonomique, efficaces, faciles à utiliser d'où chaque composante est placée dans son endroit.

Figure 55 page d'accueil et de service

Dans cette page :

Nous pouvons accéder à n'importe quel service mais en respectant l'ordre des services suivantes car chaque étape dépend de l'étape précédente :

1. Lecture du PDF (c'est la première étape elle est obligatoire pour poursuivre les prochaines étapes).
2. Conversion d'un PDF en texte,
3. Extraction des titres,
4. Extraction de la TDM.

4.5.2 Lecture du PDF

Quand l'utilisateur clique sur le bouton lire PDF cette page apparait il doit choisir un fichier PDF qu'il désire ensuite il clique sur le bouton télécharger PDF afin d'avoir une trace sur le fichier PDF sélectionner

Figure 56 Page de lecture du PDF

Figure 57 la sélection d'un PDF

Figure 58 Téléchargement du PDF avec succès

Si l'utilisateur n'a pas sélectionner un fichier PDF et clique directement sur le bouton TéléchargerPDF ce message sera afficher pour l'obliger à introduire son fichier PDF.

Figure 59 Cas de la non sélection du PDF

4.5.3 Conversion d'un PDF en texte

Cette page permet aux utilisateurs de convertir leurs fichiers PDF, utiliseront l'une des trois différentes modèles de la conversion existantes au niveau de l'application (**PyPdf2**, **Poppler**, **Pfdfminer (**un choix pour l'arabe car comme déjà vu dans le chapitre précédent le pdfminer change la morphologie ainsi que le positionnement des chaines de caractères écrites en arabe et l'autre pour le français et l'anglais**))** après en cliquant sur le bouton « Convertir » le PDF sera convertis automatiquement en texte. L'utilisateur aura la possibilité d'afficher le contenue textuelle convertie en appuiera sur le bouton « Afficher ».

Figure 60 Page conversion d'un fichier PDF en texte

Figure 61 Modèles de conversion

Les figures que nous allons vous montrer prochainement représentes les différents résultats obtenus lors des tests effectués en utilisant de différentes techniques et de corpus pour la conversion du PDF en texte.

4.5.3.1 Conversion d'un mémoire en Français

Figure 62 Contenu textuel d'un mémoire en Français extrait avec la méthode poppler

Figure 63 Contenu textuel d'un mémoire en Français extrait avec la méthode pdfminer

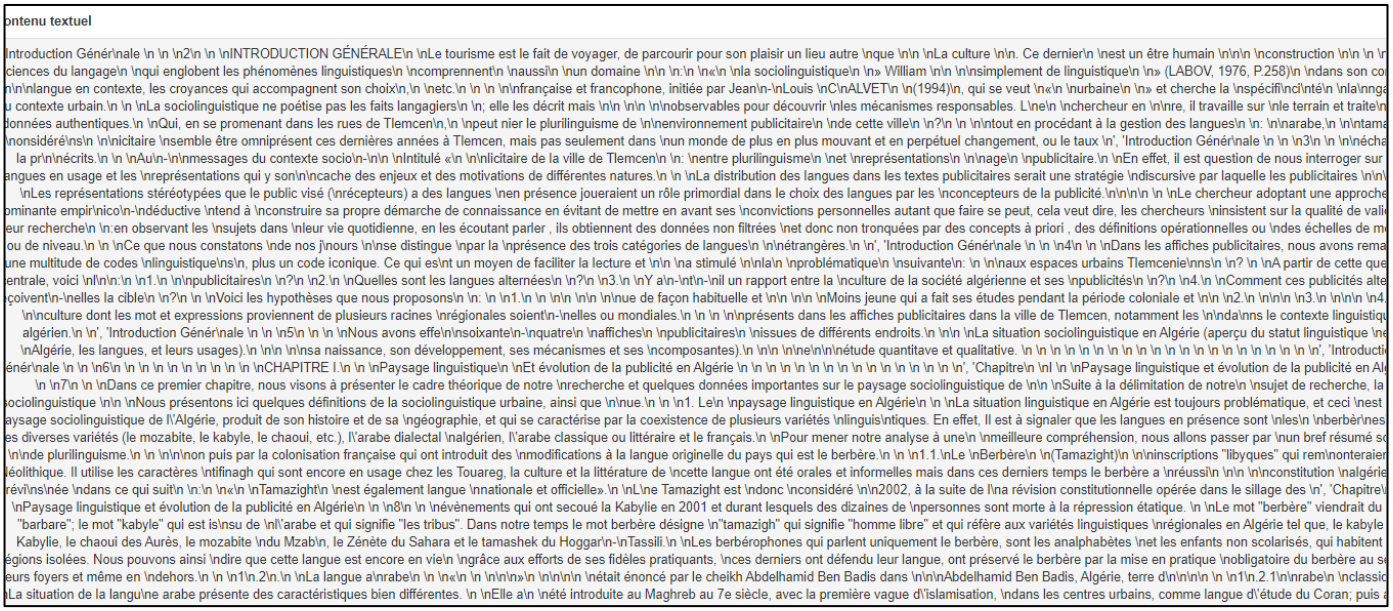


Figure 64 Contenu textuel d'un mémoire en Français extrait avec la méthode pyPDF2

4.5.3.2 Conversion d'un mémoire en Arabe

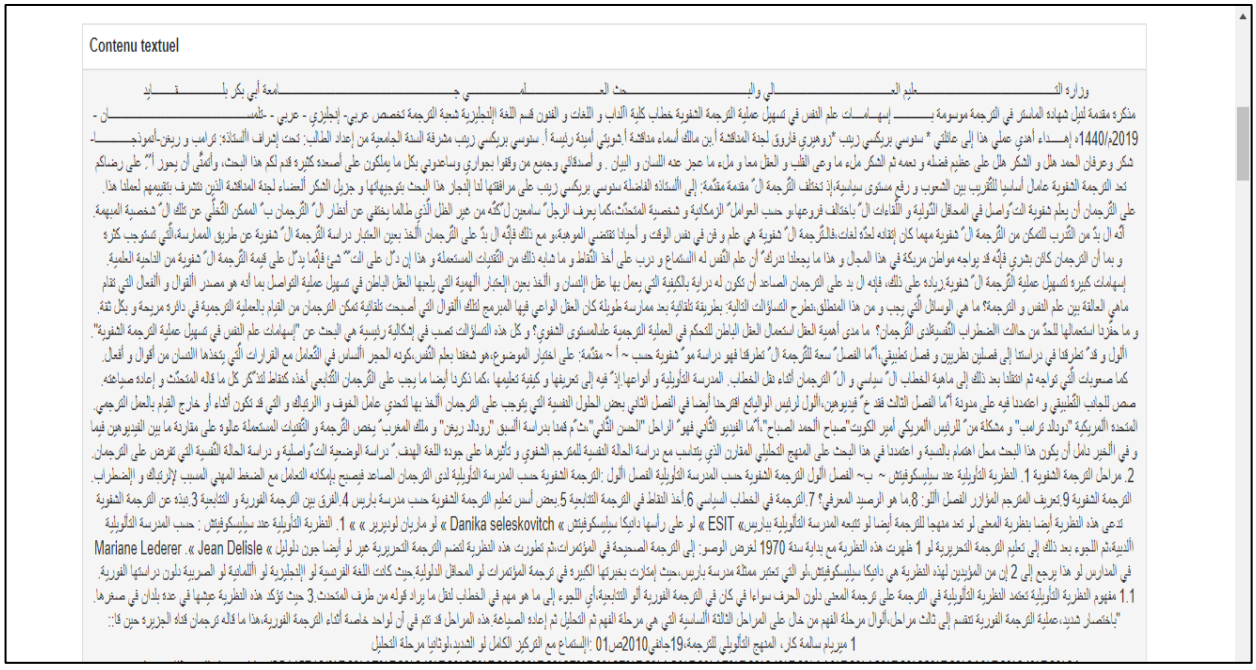

Contenu textuel

Figure 65 Contenu textuel d'un mémoire en Arabe extrait avec la méthode poppler


Contenu textuel

Figure 66 Contenu textuel d'un mémoire en Arabe extrait avec la méthode pdfminer

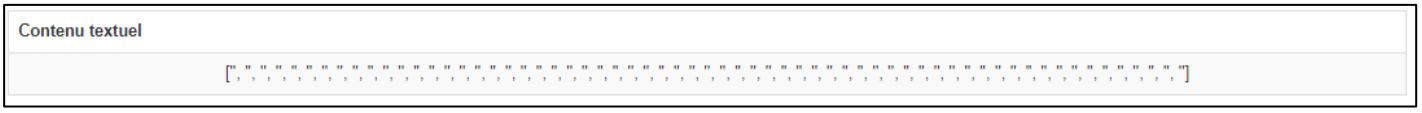
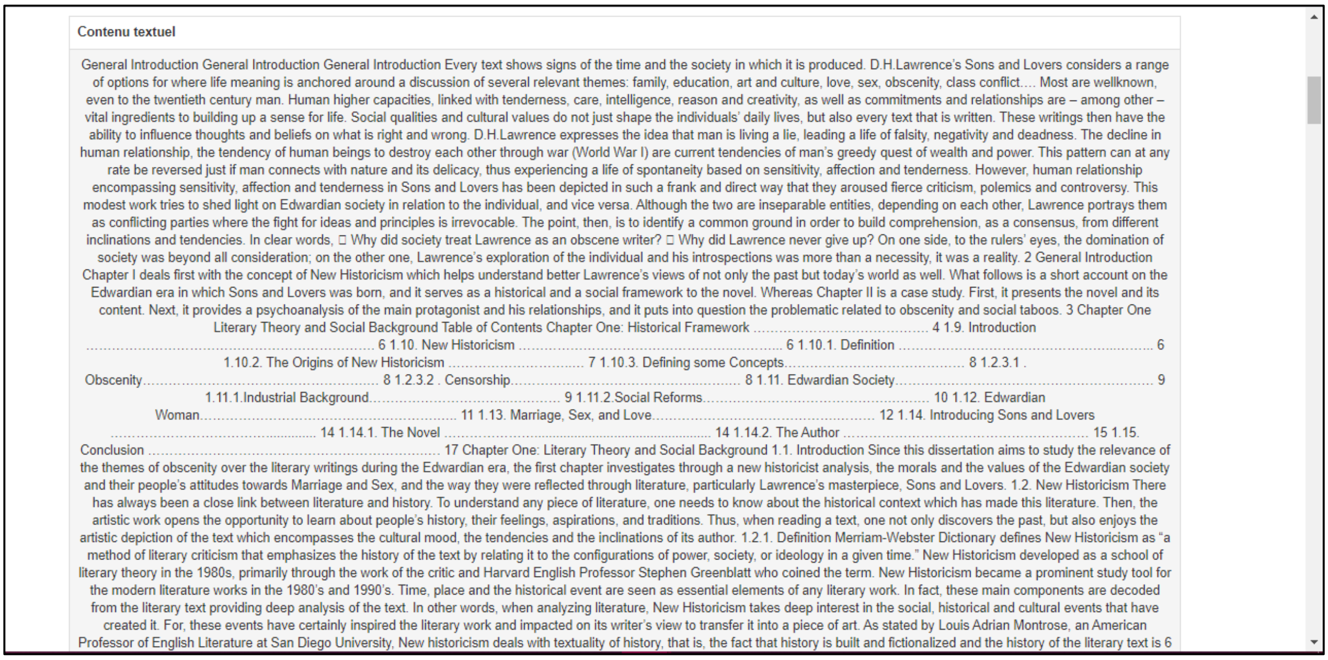

Contenu textuel

Figure 67Contenu textuel d'un mémoire en Arabe extrait avec la méthode pyPDF2

4.5.3.3 Conversion d’un mémoire en Anglais



Contenu textuel

General Introduction General Introduction General Introduction Every text shows signs of the time and the society in which it is produced. D.H.Lawrence’s Sons and Lovers considers a range of options for where life meaning is anchored around a discussion of several relevant themes: family, education, art and culture, love, sex, obscenity, class conflict.… Most are wellknown, even to the twentieth century man. Human higher capacities, linked with tenderness, care, intelligence, reason and creativity, as well as commitments and relationships are – among other – vital ingredients to building up a sense for life. Social qualities and cultural values do not just shape the individuals’ daily lives, but also every text that is written. These writings then have the ability to influence thoughts and beliefs on what is right and wrong. D.H.Lawrence expresses the idea that man is living a lie, leading a life of falsity, negativity and deadness. The decline in human relationship, the tendency of human beings to destroy each other through war (World War I) are current tendencies of man’s greedy quest of wealth and power. This pattern can at any rate be reversed just if man connects with nature and its delicacy, thus experiencing a life of spontaneity based on sensitivity, affection and tenderness. However, human relationship encompassing sensitivity, affection and tenderness in Sons and Lovers has been depicted in such a frank and direct way that they aroused fierce criticism, polemics and controversy. This modest work tries to shed light on Edwardian society in relation to the individual, and vice versa. Although the two are inseparable entities, depending on each other, Lawrence portrays them as conflicting parties where the fight for ideas and principles is irrevocable. The point, then, is to identify a common ground in order to build comprehension, as a consensus, from different inclinations and tendencies. In clear words, Why did society treat Lawrence as an obscene writer? Why did Lawrence never give up? On one side, to the rulers’ eyes, the domination of society was beyond all consideration; on the other one, Lawrence’s exploration of the individual and his introspections was more than a necessity, it was a reality. 2 General Introduction Chapter I deals first with the concept of New Historicism which helps understand better Lawrence’s views of not only the past but today’s world as well. What follows is a short account on the Edwardian era in which Sons and Lovers was born, and it serves as a historical and a social framework to the novel. Whereas Chapter II is a case study. First, it presents the novel and its content. Next, it provides a psychoanalysis of the main protagonist and his relationships, and it puts into question the problematic related to obscenity and social taboos. 3 Chapter One Literary Theory and Social Background Table of Contents Chapter One: Historical Framework …………………………………… 4 1.9. Introduction ……………………………………………………………… 6 1.10. New Historicism ……………………………………………………………… 6 1.10.1. Definition ……………………………………………………………… 6 1.10.2. The Origins of New Historicism ……………………………… 7 1.10.3. Defining some Concepts……………………………………… 8 1.2.3.1 . Obscenity………………………………………………… 8 1.2.3.2 . Censorship……………………………………………… 8 1.11. Edwardian Society……………………………………………………… 9 1.11.1.Industrial Background……………………………………… 9 1.11.2.Social Reforms…………………………………………………… 10 1.12. Edwardian Woman……………………………………………………… 11 1.13. Marriage, Sex, and Love…………………………………………… 12 1.14. Introducing Sons and Lovers ………………………………………… 14 1.14.1. The Novel ……………………………………………………… 14 1.14.2. The Author ……………………………………………………… 15 1.15. Conclusion ……………………………………………………… 17 Chapter One: Literary Theory and Social Background 1.1. Introduction Since this dissertation aims to study the relevance of the themes of obscenity over the literary writings during the Edwardian era, the first chapter investigates through a new historicist analysis, the morals and the values of the Edwardian society and their people’s attitudes towards Marriage and Sex, and the way they were reflected through literature, particularly Lawrence’s masterpiece, Sons and Lovers. 1.2. New Historicism There has always been a close link between literature and history. To understand any piece of literature, one needs to know about the historical context which has made this literature. Then, the artistic work opens the opportunity to learn about people’s history, their feelings, aspirations, and traditions. Thus, when reading a text, one not only discovers the past, but also enjoys the artistic depiction of the text which encompasses the cultural mood, the tendencies and the inclinations of its author. 1.2.1. Definition Merriam-Webster Dictionary defines New Historicism as “a method of literary criticism that emphasizes the history of the text by relating it to the configurations of power, society, or ideology in a given time.” New Historicism developed as a school of literary theory in the 1980s, primarily through the work of the critic and Harvard English Professor Stephen Greenblatt who coined the term. New Historicism became a prominent study tool for the modern literature works in the 1980’s and 1990’s. Time, place and the historical event are seen as essential elements of any literary work. In fact, these main components are decoded from the literary text providing deep analysis of the text. In other words, when analyzing literature, New Historicism takes deep interest in the social, historical and cultural events that have created it. For, these events have certainly inspired the literary work and impacted on its writer’s view to transfer it into a piece of art. As stated by Louis Adrian Montrose, an American Professor of English Literature at San Diego University, New historicism deals with textuality of history, that is, the fact that history is built and fictionalized and the history of the literary text is 6

Figure 68 Contenu textuel d'un mémoire en Anglais extrait avec la méthode poppler

Contenu textuel

[' \nGeneral Introduction\n \n', 'General Introduction\n \n \n2\n \n \nGeneral Introduction\n \n \nEvery text shows signs of the time and the society\n \nin which it is produced\n. \n\nSons and Lovers \nconsiders a range of options for where life \nmeaning is anchored around a discussion of several relevant themes: family, \n\nl\n-\nknown, even to the twentieth century man. H\numan \nhigher capacities, \nlinked\n \nwith \ntenderness, care, intelligence, reason and creativity, as well as commitments and \nrelationships are \n\n \namong other \n\n \nvital ingredients to building up a sense \nfor life\n. \nSocial qualities and cultural values do not just shape \n\n \ndaily lives, \nbut also every text that is written. These writings then have the ability to influence \nthoughts and beliefs on what is\n \nright and\n \nwrong.\n \nD.H.Lawrence\n \nexpresses\n \nthe\n \nidea\n \ntha\nt\n \nman\n \nis\n \nliving\n \na\n \nlie,\n \nleading\n \na\n \nlife\n \nof\n \nfalsity,\n \nnegativity\n \nand deadness. The decline in human relationship, the tendency of human \nbeings to destroy each other through war (World War I) are current tendencies of \n\ny quest of\n \nwealth and power. Thi\ns pattern can at any rate be reversed \njust if man connects with nature and its delicacy, thus experiencing a life of \nspontaneity based on sensitivity, affection and\n \ntenderness.\n \n \nHowever,\n \nhuman relationship encompassing\n \nsensitivity, affection and tenderness \nin \nSons and Lovers \nha\ns\n \nbeen\n \ndepicted in such a\n \nfrank and direct way that they \naroused fierce criticism, polemics and controversy. \nThis modest work tries to shed \nlight \non Edwardian \nsociety in relation to the individual\n, and \nvice versa\n. Although \nthe two are inseparable entities, depending on each other, \nLawrence portrays them \nas conflicting parties where the fight for ideas and principles is irrevocable.\n \nThe \npoint, then, is to identify a common ground\n \nin order to build c\nomprehension, as a \nconsensus, from different inclinations and tendencies.\n \nIn clear words, \n \n\n \nWhy\n \ndid society treat Lawrence\n \nas \nan \nobscene\n \nwriter\n?\n \n \n\n \nW\nhy did Lawrence never give up?\n \n \neyes,\n \nthe domination of society was beyond all \nconsideration\n\n \nand his \nintrospections\n \nwas more than a necessity, it was a reality. \n \n', 'General Introduction\n \n \n3\n \n \nChapter I\n \ndeals\n \nfirst with t\nhe concept of New Historicism which helps \nunderstand \n\n. What \nfollows is \na short account on the Edwardian era\n \nin which \nSons and Lovers\n \nwas born, \nand\n \nit\n \nserves as a historical\n \nand a social\n \nframework to the novel\n.\n \nWhereas C\nhapter \nII\n \nis a case study. First, it \npresents the novel\n \nand its content. Next, it provides a \npsychoanalysis of the \nmain protagonist\n \nand his relationships\n, and it puts into \nquestion the problematic related to obscenity and social taboos. \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n', ' \n \n \nC\nhapter One\n \n \nLiterary Theory \n \nand \n \nSocial Background\n', ' \n \nTable of Contents\n \n \nChapter One: Historical Framework\n \n\n.\n\n \n4\n \n1.9.\n \n \nIntroduction\n \n\n \n6\n \n1.10.\n \n\n \n6\n \n1.10.1.\n \n\n \n6\n \n1.10.2.\n \n\n \n7\n \n1.10.3.\n \n\n \n8\n \n1.2.3.1\n \n\n \n8\n \n1.2.3.2\n \n. \n\n \n8\n \n1.11.\n \nEdwardian \n\n \n9\n \n1.11.1.\n \n\n \n9\n \n1.11.2.\n \n\n \n10\n \n1.12.\n \nEdwardian \nWoman\n\n\n\n \n11\n \n1.13.\n \n\n \n12\n \n1.14.\n \nIntroducing \nSons and Lovers\n \n\n..\n...\n.........\n \n14\n \n1.14.1.\n \n\n \n14\n \n1.14.2.\n \n\n \n15\n \n1.15.\n \n\n \n17\n \n \n \n \n \n \n \n \n', 'Chapter One: Literary Theory and Social Background\n \n6\n \n \n1.1.\n \nIntroduction\n \n \nSince this dissertation aims to study the relevance of the themes of obscenity over \nthe literary writings during the Edwardian era, the first chapter investigates through a \nnew historicist analysis, the morals and the values of the Edwardian society and th\neir \n\n\nSons and Lovers.\n \n \n1.2.\n \nNew Historicism\n \n \nThere has always been a close link between literature and history. To understand \nany piece of literature, one needs to know about the historical context which has made \n\nhistory, their feelings, aspirations, and traditions. Thus, when reading a text, o\nne not \nonly discovers the past, but also enjoys the artistic depiction of the text which \nencompasses the cultural mood, the tendencies and the inclinations of its author.\n \n1.2.1.\n \nDefinition\n \nMerriam\n-\n\ny \ncriticism that emphasizes the history of the text by relating it to the configurations of \n\nof literary theory in the 1980s, primarily through the work of the critic and Ha\nrvard \nEnglish Professor Stephen Greenblatt who coined the term. New Historicism became \n\nplace and the historical event are seen as essential elements of any literary work\n. In \nfact, these main components are decoded from the literary text providing deep analysis \nof the text. \n \nIn other words,\n \nwhen analyzing literature,\n \nNew Historicism \ntakes deep interest in \nthe social, historical and cultural events that have created \nit. For\n, these events have \n\na piece of art.\n \nAs stated by Louis Adrian Montrose, an American Professor of English Literature \nat San Diego University,\n \nNew historicism deals with textuality of history, that is, the fact that \nhistory is built and fictionalized and the history of the literary text is \n', "Chapter One: Literary Theory and Social Background\n \n7\n \n \nwithout a doubt found within the socio\n-\ncultural and political conditions \nsurrounding its conception and inte\nrpretation. (\nMontrose, 2006)\n \nThe main concern of New Historicism is the exploration of history and the \nmechanisms which are inherent in the way institutions and people \n\n \nrulers, monarchs, \nchurch, men and women of different classes \n\n \ninteract for the buildi\nng and the \ndevelopment of this history. \nThe historical components, encompassing social events, \ncultural factors and political contexts help understand the literary text and the degrees \nof impact and influence they have on the artist.\n \nNew Historicism recogn\nizes and bears \nthe idea that, as times goes on and changes, the perception of literary works develops.\n \n1.2.2.\n \nThe Origins of New Historicism\n \nTo understand New Historicism it is of importance to also understand where it \ncame

Figure 69 Contenu textuel d'un mémoire en Anglais extrait avec la méthode pyPDF2

Contenu textuel

'General Introduction \n\n \n\n\x0cGeneral Introduction \n\nGeneral Introduction \n\n \n\nEvery text shows signs of the time and the society in which it is produced. \n\nD.H.Lawrence's Sons and Lovers considers a range of options for where life \n\nmeaning is anchored around a discussion of several relevant themes: family, \n\neducation, art and culture, love, sex, obscenity, class conflict…. Most are well-\n\nknown, even to the twentieth century man. Human higher capacities, linked with \n\ntenderness, care, intelligence, reason and creativity, as well as commitments and \n\nrelationships are – among other – vital ingredients to building up a sense for life. \n\nSocial qualities and cultural values do not just shape the individuals' daily lives, \n\nbut also every text that is written. These writings then have the ability to influence \n\nthoughts and beliefs on what is right and wrong. \n\nD.H.Lawrence expresses the idea that man is living a lie, leading a life of falsity, \n\nnegativity and deadness. The decline in human relationship, the tendency of human \n\nbeings to destroy each other through war (World War I) are current tendencies of \n\nman's greedy quest of wealth and power. This pattern can at any rate be reversed \n\njust if man connects with nature and its delicacy, thus experiencing a life of \n\nspontaneity based on sensitivity, affection and tenderness. \n\nHowever, human relationship encompassing sensitivity, affection and tenderness \n\nin Sons and Lovers has been depicted in such a frank and direct way that they \n\naroused fierce criticism, polemics and controversy. This modest work tries to shed \n\nlight on Edwardian society in relation to the individual, and vice versa. Although \n\nthe two are inseparable entities, depending on each other, Lawrence portrays them \n\nas conflicting parties where the fight for ideas and principles is irrevocable. The \n\npoint, then, is to identify a common ground in order to build comprehension, as a \n\nconsensus, from different inclinations and tendencies. In clear words, \n\n\uf0b7 Why did society treat Lawrence as an obscene writer? \n\n\uf0b7 Why did Lawrence never give up? \n\nOn one side, to the rulers' eyes, the domination of society was beyond all \n\nconsideration; on the other one, Lawrence's exploration of the individual and his \n\nintrospections was more than a necessity, it was a reality. \n\n2 \n\n \n\n \n\n\x0cGeneral Introduction \n\nChapter I deals first with the concept of New Historicism which helps understand \n\nbetter Lawrence's views of not only the past but today's world as well. What \n\nfollows is a short account on the Edwardian era in which Sons and Lovers was born, \n\nand it serves as a historical and a social framework to the novel. Whereas Chapter \n\nII is a case study. First, it presents the novel and its content. Next, it provides a \n\npsychoanalysis of the main protagonist and his relationships, and it puts into \n\nquestion the problematic related to obscenity and social taboos. \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n3 \n\x0c \n\n \n\n \n\n \n\nChapter One \n\nLiterary Theory \n\nand \n\nSocial Background\n\n \n\n \n\n\x0cTable of Contents \n\nChapter One: Historical Framework .... \n\n1.9. Introduction .... \n\n1.10. New Historicism .... \n\n1.10.1. Definition .... \n\n1.10.2. The Origins of New Historicism .... \n\n1.10.3. Defining some Concepts.. \n\n1.2.3.1 . Obscenity.. \n\n1.2.3.2 . Censorship.. \n\n1.11. Edwardian \n\nSociety.. \n\n1.11.1. Industrial Background... \n\n1.11.2. Social Reforms.. \n\n10 \n\n1.12. Edwardian \n\nWoman.. \n\n1.13. Marriage, Sex, and Love.. \n\n1.14. Introducing Sons and Lovers .... \n\n1.14.1. The Novel .... \n\n1.14.2. The Author .... \n\n1.15. Conclusion .... \n\n \n\n4 \n\n6 \n\n6 \n\n6 \n\n7 \n\n8 \n\n8 \n\n8 \n\n9 \n\n9 \n\n11 \n\n12 \n\n14 \n\n14 \n\n15 \n\n17 \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n \n\n\x0cChapter One: Literary Theory and Social Background \n\n1.1. Introduction \n\nSince this dissertation aims to study the relevance of the themes of obscenity over \n\nthe literary writings during the Edwardian era, the first chapter investigates through a \n\nnew historicist analysis, the morals and the values of the Edwardian society and their \n\npeople's attitudes towards Marriage and Sex, and the way they were reflected through \n\nliterature, particularly Lawrence's masterpiece, Sons and Lovers. \n\n1.2. New Historicism \n\n There has always been a close link between literature and history. To understand \n\nany piece of literature, one needs to know about the historical context which has made \n\nthis literature. Then, the artistic work opens the opportunity to learn about people's \n\nhistory, their feelings, aspirations, and traditions. Thus, when reading a text, one not \n\nonly discovers the past, but also enjoys the artistic depiction of the text which \n\nencompasses the cultural mood, the tendencies and the inclinations of its author. \n\n1.2.1. Definition \n\nMerriam-Webster Dictionary defines New Historicism as "a method of literary \n\ncriticism that emphasizes the history of the text by relating it to the configurations of \n\npower, society, or ideology in a given time." New Historicism developed as a school \n\nof literary theory in the 1980s, primarily through the work of the critic and Harvard \n\nEnglish Professor Stephen Greenblatt who coined the term. New Historicism

Figure 70 Contenu textuel d'un mémoire en Anglais extrait avec la méthode pdfminer

4.5.3.4 Conversion d'un article scientifique

Contenu textuel

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/323515420 Logical Structure Extraction from Digitized Books: Benchmarking State-ofthe-Art Systems Chapter · April 2018 DOI: 10.1142/9789813229273_0001 CITATIONS READS 0 208 1 author: Antoine Doucet La Rochelle Université 113 PUBLICATIONS 812 CITATIONS SEE PROFILE Some of the authors of this publication are also working on these related projects: Daniel: Multilingual Epidemic Surveillance View project AméliOCR View project All content following this page was uploaded by Antoine Doucet on 09 August 2018. The user has requested enhancement of the downloaded file. Document Analysis and Text Recognition Downloaded from www.worldscientific.com by 212.68.24.82 on 08/09/18. Re-use and distribution is strictly not permitted, except for Open Access articles. "9x6" b3022 Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems Chapter 1 Logical Structure Extraction from Digitized Books Antoine Doucet 1.1 Introduction Mass digitization projects, such as the Million Book Project, efforts of the Open Content Alliance, and the digitization work of Google, are converting whole libraries by digitizing books on an industrial scale [5]. The process involves the efficient photographing of books, page-by-page, and the conversion of the image of each page into searchable text through the use of optical character recognition (OCR) software. Current digitization and OCR technologies typically produce the full text of digitized books with only minimal structure information. Pages and paragraphs are usually identified and marked up in the OCR, but more sophisticated structures, such as chapters, sections, etc., are not recognized. In order to enable systems to provide users with richer browsing experiences, it is necessary to make such additional structures available, for example, in the form of XML markup embedded in the full text of the digitized books. The Book Structure Extraction competition aims to address this need by promoting research into automatic structure recognition and extraction techniques that could complement or enhance current OCR methods and 3 b3022_Ch-01.indd 3 25-01-2018 12:17:29 b3022 Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems Document Analysis and Text Recognition Downloaded from www.worldscientific.com by 212.68.24.82 on 08/09/18. Re-use and distribution is strictly not permitted, except for Open Access articles. 4 "9x6" Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems lead to the availability of rich structure information for digitized books. Such structure information can then be used to aid user navigation inside books as well as to improve search performance [35]. The chapter is structured as follows. We start by placing the competition in the context of the work conducted at the Initiative for the Evaluation of XML Retrieval (INEX) Evaluation Forum [22]. We then describe the setup of the competition, including its goals and the task that has been set for its participants. The book collection used in the task is also detailed. The ground-truth-creation process and its outcome are next described, together with the corresponding evaluation metrics used and the final results, alongside brief descriptions of the participants' approaches. We conclude with a summary of the competition and how it could be built upon. 1.1.1 Background Motivated by the need to foster research in areas relating to large digital book repositories (see e.g., [21]), the Book Track was launched in 2007 [22] as part of the INEX. Founded in 2002, INEX is an evaluation forum that investigates focused retrieval approaches [14] where structure information is used to aid the retrieval of parts of documents, relevant to a search query. Focused retrieval over books presents a clear benefit to users, enabling them to gain direct access to those parts of books (of potentially hundreds of pages in length) that are relevant to their information needs. One major limitation of digitized books is the fact that their structure is physical, rather than logical. Following this, the evaluation and relevance judgments based on the book corpus have essentially been based on whole books and selections of pages. This is unfortunate considering that books seem to be the key application field for structured information retrieval (IR). The fact that, for instance, chapters, sections, and paragraphs are not readily available has been a frustration for the structured IR community gathered at INEX, because it does not allow us to test the techniques -created for collections of scientific articles and for Wikipedia. Unlike digitally born content, the logical structure of digitized books is not readily available. A digitized book is often only split into pages with b3022_Ch-01.indd 4 25-01-2018 12:17:29 "9x6" b3022 Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems Document Analysis and Text Recognition Downloaded from www.worldscientific.com by 212.68.24.82 on 08/09/18. Re-use and distribution is strictly not permitted, except for Open Access articles. Logical Structure Extraction from Digitized Books 5 possible paragraphs, lines, and word markup. This was also the case for the 50,000 digitized book collection of the INEX Book Search track [22]. The use of more meaningful structure, e.g., chapters, table of contents (ToC), bibliography, or back-of-book index, to support focused retrieval has been explored for many years at INEX and has been shown to increase retrieval performance [35]. To encourage research aiming to provide the logical structure of digitized books, we created the Book Structure Extraction competition, which we later brought to the community of document analysis. Starting from 2008, within the second round of the INEX Book Track, we entirely created the methodology to evaluate the Structure Extraction process from digitized books: problem description, submission procedure, annotation procedure (and corresponding software), metrics, and evaluation. 1.1.2 Context and Motivation The overall goal of the INEX Book Track is to promote interdisciplinary research investigating techniques for supporting users in reading, searching, and navigating the full texts of digitized books

Figure 71 Contenu textuel d'un article extrait avec la méthode poppler

Figure 72 Contenu textuel d'un article extrait avec la méthode pdfminer

Figure 73 Contenu textuel d'un article extrait avec la méthode pyPDF2

4.5.3.5 Conversion d'un documents financiers en Anglais

Contenu textuel

Ba nt leon D y na mic B a n t l e o n Tr e n d Ba nt leon Strateg ie Ba nt leon Yield Ba nt leon Return Bantleon AnleihenFonds Sales Prospectus with Management Regulations Sales Prospectus with Management Regulations »Bantleon AnleihenFonds« with the sub-funds Bantleon Return Institutional Class »IA« – LU0109659770 Institutional Class »IT« – LU0524467833 Retail Class »PA« – LU0430091412 Retail Class »PT« – LU0524467676 Bantleon Yield Institutional Class »IA« – LU0261192784 Institutional Class »IT« – LU0532347472 Retail Class »PA« – LU0261193329 Retail Class »PT« – LU0524467916 Bantleon Strategie Institutional Class »IA« – LU0104810238 Institutional Class »IT« – LU0524468054 Retail Class »PA« – LU0430091503 Retail Class »PT« – LU0532346581 Bantleon Trend Institutional Class »IA« – LU0150854106 Institutional Class »IT« – LU0524468211 Retail Class »PA« – LU0532346748 Retail Class »PT« – LU0532347043 Bantleon Dynamic Institutional Class »IA« – LU0117465517 Institutional Class »IT« – LU0532345930 Retail Class »PA« – LU0532346151 Retail Class »PT« – LU0532346318 March 2012 Remark: This is a translation of the German prospectus. The German version shall be binding for the interpretation of the sales prospectus and the management regulations. Table of Contents 04 A. Sales Prospectus The Fund | Investment Policy | Risk Warning | Investor Profile | The Management C ompany | The Depositary Bank | Management and Transfer Agent | Prevention of Money Laundering | Shares | Share Prices | Acquisition, Redemption and Exchange of Shares, Purchase Price Payment | Charges and Fees | Further Notes | Publications and Information | The Management Regulations | Combating Market Timing and Late Trading Activities 10 Fund overview 11 Management and Administration 12 B. Management Regulations 12 I. Management/Organisation The Fund | The Management Company | The Investment Manager | The Depositary Bank | Management and Transfer Agent 13 II. General Investment Policy Guidelines Investment Objective | Listed Securities | Unlisted Securities and Other Securitised Rights | New Issuers | Investment Limits | Interest Rate Futures | Liquid Assets | Further Investment Restrictions | Loans and Encumbrance Prohibitions 14 III. Issues and Redemptions, Further Conditions Shares in the Sub-Funds | Issue of Shares | Calculation of the Net Asset Value per Share | Suspension of Calculation of the NAV per Share | Redemption of Shares | Exchange of Shares | Dividend Distributions | Duration and Closing of the Fund and the Sub-Funds and the Merger of Funds or Sub-Funds | Fees, Charges | Fiscal Year and Audit of Annual Accounts | Statute of Limitations | Amendments Publications | Applicable Law, Jurisdiction and Language of Contract | Entry into Force 17 C. Special REGULATIONS Bantleon Return | Bantleon Yield | Bantleon Strategie | Bantleon Trend | Bantleon Dynamic 23 D. ANNEX 23 I. Issuers Sales Prospectus 1. The Fund 2. Investment Policy Bantleon Return, Bantleon Yield, Bantleon Strategie, Bantleon Trend and Bantleon Dynamic are sub-funds of the »BANTLEON ANLEIHENFONDS« (hereinafter referred to as »the sub-funds«). Bantleon Invest S. A., a subsidiary of Bantleon Bank AG is responsible for managing the sub-funds. Bantleon Return Bantleon Bank AG is specialised in the investment management of high-quality bonds. The different management styles of the individual sub-funds offer investors the opportunity to participate in this core competency of Bantleon Bank AG. All sub-funds invest solely in bonds in accordance with the investment restrictions detailed in the relevant Special Section. The liquidity of the relevant sub-funds is held by the depositary bank, the state banks of Bavaria and Baden-Wuerttemberg and Deutsche Bank AG. The investment of liquid assets is restricted to 20 % of the fund's assets per counterparty. Investments must be in euro. For Eurozone investors there is thus no exchange rate risk. The »BANTLEON ANLEIHENFONDS« was established in accordance with Part 1 of the Luxembourg Law on Undertakings for Collective Investment in Transferable Securities of 30 March 1988 (UCITS) as a public fund (fonds commun de placement) on 1 March 2000 under the name BANTLEON STRATEGIE NO. 1 for an indefinite period. On 1 June 2003 it was transformed into an umbrella fund and on 1 December 2005 into an investment fund in accordance with Part 1 of the Luxembourg Law on Undertakings for Collective Investment of 20 December 2002. The »BANTLEON STAATSANLEIHENFONDS« was renamed the »BANTLEON ANLEIHENFONDS« on 21 January 2010. On 1 July 2011, it was transformed into an investment fund in accordance with Part I of the Luxembourg Law on Undertakings for Collective Investment of 17 December 2010. The »BANTLEON ANLEIHENFONDS« and its subfunds comply with Directive 2009/65/EC of the European Parliament and the Council. The sub-funds are legally and financially independent of one another. Each sub-fund is liable only for its own obligations with respect to third parties and particularly creditors. Shares of the fund may not be offered, sold or delivered within the United States. The fund may neither be offered, sold or delivered to U.S. citizens or persons residing in the U.S. and/or other natural or legal persons whose income and/or earnings, regardless of origin, are subject to the U.S. income tax, and persons that are subject to Regulation S under the U.S. Securities Act of 1933 and/or the U.S. Commodity Exchange Act as amended. 4 Sales Prospectus with Management Regulations Bantleon Return is based on the immunisation strategy of Bantleon Bank AG and thus optimises the earnings of highquality bonds across the entire yield curve. The objective is to fully capitalise the income potential through the close dovetailing of duration adjustment, yield curve management, spread management and inflation indexing. Modified duration of the sub-fund's assets: 2.0 to 6.0. Bantleon Yield Working within the immunisation strategy of Bantleon Bank AG, Bantleon Yield focuses more specifically on maximising the interest income and on spread management. The sub-fund invests in bonds across the entire yield curve. Modified

Figure 74 Conversion d'un documents financiers en Anglais avec poppler

es Prospectus with Management Regulations\n\nn\nr\nu\nt\ne\nR\n\n \n\nn\no\ne\nl\nt\nn\na\nB\n\nd\n\nl\ne\ni\n\nY\n\n \n\nn\no\ne\nl\nt\nn\na\nB\n\ne\ni\ng\ne\nt\na\nr\nt\nS\n\n \n\nn\no\ne\nl\nt\nn\na\nB\n\nd\nn\ne\nr\nT\n\n \n\nn\no\ne\nl\nt\
rospectus with Management Regulations \n \n»BAntleon AnlelhenFonDS« \nwith the sub-funds\n\nBantleon Return\nInstitutional Class »IA« – LU0109659770\nInstitutional Class »IT« – LU0524467833\nRetail Class »PA« – LU0430091412\nR
!784\nInstitutional Class »IT« – LU0532347472\nRetail Class »PA« – LU0261193329\nRetail Class »PT« – LU0524467916\n\nBantleon Strategie\nInstitutional Class »IA« – LU0104810238\nInstitutional Class »IT« – LU0524468054\nRetail Cla
itutional Class »IA« – LU0150854106\nInstitutional Class »IT« – LU0524468211\nRetail Class »PA« – LU0532346748\nRetail Class »PT« – LU0532347043\n\nBantleon Dynamic\nInstitutional Class »IA« – LU0117465517\nInstitutional Class :
32346318\n\nMarch 2012\n\nRemark: This is a translation of the German prospectus. The German version shall be binding for the interpretation of the sales prospectus and the management regulations.\n\n\x0ctABle oF ContentS\n\n04 \n\nA.
The Management Company | The\xa0 Depositary \nBank | Management and Transfer Agent | Prevention of Money Laundering | Shares | Share Prices | Acquisition, \nRedemption and Exchange of Shares, Purchase Price Payment | Charges a
| Combating Market Timing and Late Trading Activities\n\n10 \n\nFunD oveRvIew \n\n11 \n\nMAnAgeMent AnD ADMInIStRAtIon\n\n12 \n\nB. MAnAgeMent RegulAtIonS \n\n12 \n\nI. MAnAgeMent/oRgAnISAtIon\n\n \n\n \n\n \n\n \n\n \n\n23 \n\
ment Manager | The Depositary Bank | \nManagement and Transfer Agent\n\n13 \n\nII. geneRAl InveStMent PolICy gulDelIneS\n\n Investment Objective | Listed Securities | Unlisted Securities and Other Securitised Rights | New Issuers | Inves
| Loans and Encumbrance Prohibitions\n\n14 \n\nIII. ISSueS AnD ReDeMPtIonS, FuRtheR ConDItIonS\n\n Shares in the Sub-Funds | Issue of Shares | Calculation of the Net Asset Value per Share | Suspension of Calculation of \nthe NAV per
and Closing of \nthe Fund and the Sub-Funds and the Merger of Funds or Sub-Funds | Fees, Charges | Fiscal Year and Audit of Annual \n Accounts | Statute of Limitations | Amendments Publications | Applicable Law, Jurisdiction and Languag
rn | Bantleon Yield | Bantleon Strategie | Bantleon Trend | Bantleon Dynamic\n\n\x0cSAleS PRoSPeCtuS\n\n1. the Fund\n\nBantleon Return, Bantleon Yield, Bantleon Strategie, Bantleon \nTrend and Bantleon Dynamic are sub-funds of the »B
tleon Invest S. A., a subsidiary of Bantleon Bank AG is \nresponsible for managing the sub-funds. \n\nBantleon Bank AG is specialised in the investment manage-\nment of high-quality bonds. The different management styles \nof the individual
Bantleon Bank AG.\n\n2. Investment Policy\n\nBantleon Return\n\nBantleon Return is based on the immunisation strategy of \n Bantleon Bank AG and thus optimises the earnings of high-\nquality bonds across the entire yield curve. The object
ition adjustment, yield curve management, spread \nmanagement and inflation indexing.\n\nModified duration of the sub-fund's assets: 2.0 to 6.0.\n\nAll sub-funds invest solely in bonds in accordance with the \ninvestment restrictions detailed i
nds is held by the depositary \nbank, the state banks of Bavaria and Baden-Wuerttemberg and \nDeutsche Bank AG. The investment of liquid assets is restricted \nto 20 % of the fund's assets per counterparty. Investments must \nbe in euro. F
NFONDS« was established in ac-\ncordance with Part 1 of the Luxembourg Law on Undertakings \nfor Collective Investment in Transferable Securities of 30 March\n1988 (UCITS) as a public fund (fonds commun de placement) on \n1 March
e 2003 it was transformed into \nan umbrella fund and on 1 December 2005 into an investment \nfund in accordance with Part 1 of the Luxembourg Law on Un-\ndertakings for Collective Investment of 20 December 2002. The \n»BANTLEON S
on 21 January 2010. On 1 July \n2011, it was transformed into an investment fund in accordance \nwith Part I of the Luxembourg Law on Undertakings for Collec-\ntive Investment of 17 December 2010.\n\nThe »BANTLEON ANLEIHENFONDS
d the \nCouncil.\n\nThe sub-funds are legally and financially independent of one \nanother. Each sub-fund is liable only for its own obligations with \nrespect to third parties and particularly creditors.\n\nShares of the fund may not be offered, sol
elivered to U.S. citizens \nor persons residing in the U.S. and/or other natural or legal \npersons whose income and/or earnings, regardless of origin, are \nsubject to the U.S. income tax, and persons that are subject to \nRegulation S under the
\n\nWorking within the immunisation strategy of Bantleon Bank \nAG, Bantleon Yield focuses more specifically on maximising \nthe interest income and on spread management. The sub-fund \ninvests in bonds across the entire yield curve.\n\nN
duration management, Bantleon Strategie strives to \nachieve a high and consistent return in euro with simultaneously \nlow volatility. Investment management decides between bonds \nwith long maturities or money market characteristics (abs
d\n\nThrough an active duration management Bantleon Trend pursues \nto achieve high interest income and capital gains from high-quality \nbonds with maturities from 1 to 10 years using. In times of markedly \nfalling returns, interest income
of completely \nrebalancing the portfolio in favour of near-money-market bonds \nis to avoid larger capital losses in times of rapidly rising returns. \nEvery investment decision is based on the »absolute return« model, \nnex·i·t |concept of Bantle
Dynamic\n\nUsing active duration management, Bantleon Dynamic strives \nto achieve attractive interest income and significant bond-price \ngains in times of falling returns by investing in bonds with long \nmaturities. The limited sale of intere
nS\n\n\x0cthe fund to make capital gains in times of falling bond prices \n(absolute return plus). \n\nSubfunds\n\nModified duration of the sub-fund's assets: –3.0 to +8.0.\n\nGlobal risk\ncalculation\nmethod\n\nExpected\nleverage value\n(as a
res involves risks. These risks can include interest \nrate, credit, liquidity and counter party risks as well as volatility \nrisks or political risks, or be associated with it. Each of these risks\ncan also occur together with other risks. Some of these ris
ruments that are used in the specified investment \npolicy. Also investors should fully understand the risks associ-\nated with the investment in shares and only make investment \ndecisions, if they match the advise given by their own legal, tax
n investment in shares, taking into account \ntheir personal financial or tax situation and other circumstances \nand the advice in this prospectus as well as the investment policy \nof this fund.\n\nIt should be noted that investments in funds offe
determined by the price fluctuations of the assets contained \nin it. The value of shares can rise or fall accordingly below the \npurchase price.\n\nNo warranty can therefore be given, that the objectives of an \ninvestment policy will be achieved
oach \nand the value-at-risk approach is applied pursuant to the applica-\nble laws and regulatory provisions.\n\nLeverage\nLeverage is defined pursuant to the applicable ESMA directives \nand the CSSF circular 11/512 as the total of the nom
s definition, leverage may result in artificially increased lever-\nage amounts as some derivatives that can be used for hedging \npurposes may be included in the calculation. Consequently, this \ninformation does not necessarily reflect the prec
leverage is expressed in the table below \nas a ratio between the total of the nominal value of the used de-\nrivatives and the net asset value of the respective sub-fund and is \nbased on historical data. For sub-funds which have not yet been \

Figure 75 Conversion d'un documents financiers en Anglais avec pdfminer

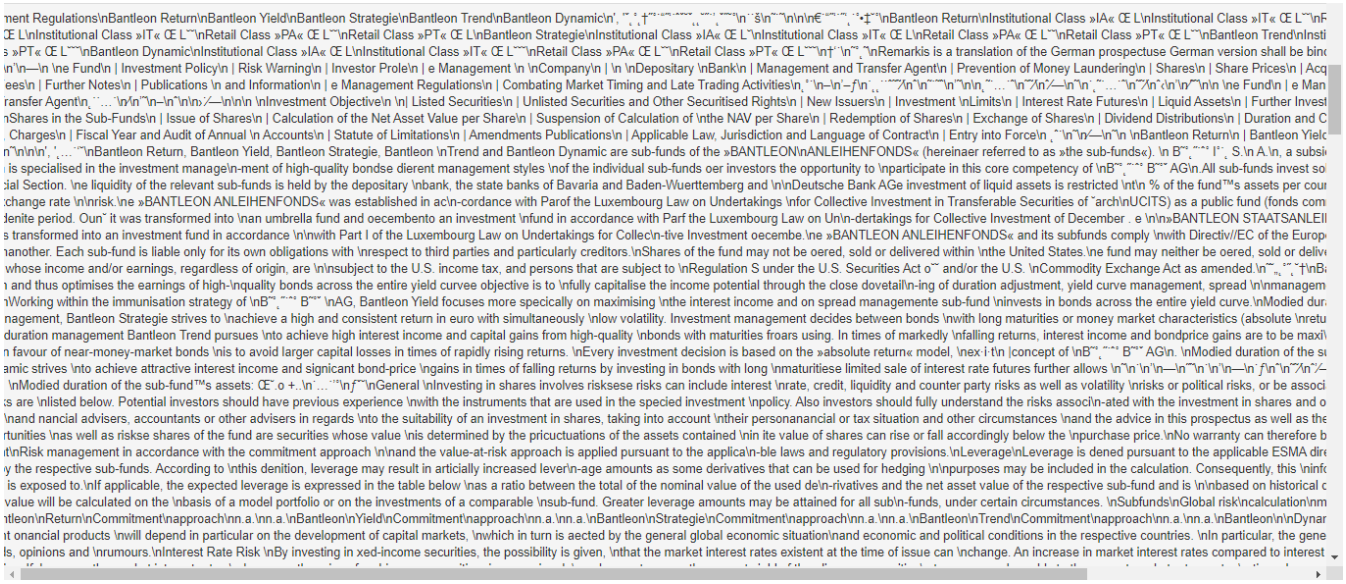


Figure 76 Conversion d'un documents financiers en Anglais avec pyPDF2

4.5.4 Détection des titres

Passant maintenant à l'étape de l'extraction des titres, d'abord l'utilisateur clique sur le même choix du modèle choisis dans l'étape précédente afin de pouvoir détecter les titres susceptibles à partir du contenu textuel, pour les fichiers en arabe il suffit juste cliquer sur le bouton arabe. L'utilisateur aura la possibilité d'afficher les titres détectés en appuiera sur le bouton « Afficher ».

Figure 77 Page d'extraction des titres

Les figures que nous allons vous montrer prochainement représentes les différents résultats obtenus lors des tests effectués en utilisant de différentes techniques et de corpus pour l'extraction des titres.

4.5.4.1 Détection des titres pour un mémoire en Fançais

Les titres
grammaticale, lexicale, en orthographe et sémantique, etc.
1- Les étudiants de la troisième année français font-ils la différence entre les verbes
intransitifs, les verbes transitifs directs et les verbes transitifs indirects.
1
p. 49
11
2- Les étudiants de la troisième année français qui sont en contact permanent avec le
3- Ces étudiants peuvent utiliser les verbes dans des constructions transitives et
conclusion.
12
1 .1. Définition du verbe
Etymologiquement, le mot verbe est issu du latin vebrum, qui signifie «parole», mot par
mots, aux groupes de mots qui se construisent autour de lui. Il a pour rôle admettre à
correspondant, sur le plan de la signification, au nombre, à la personne, au
temps, et au mode (qui peuvent également déterminer des variations du
radical). La voix, le temps et l'aspect.3
arrive) ou il est suivi d'un complément du verbe essentiel ou facultatif (exemple :
2
Michel, P., Dictionnaire grammaire et difficultés grammaticales, Armand Colin, paris, 1998, p.396
Martin, R. et al. , Grammaire méthodique du français, P.U.F. Collection Quadrige, Paris, 1994, p. 435
4
Ibid., P.435.
3
14
1.2. Le groupe verbal
complément.
1.3. Locution verbale
verbe.
1. Locution nominal : pomme de terre.
2. Locution verbale : rendre visite.
3. Locution prépositionnelle : à l'aide de.
4. Locution adjectivale : avoir beau.
1.4. Espèces de verbe
1. Auxiliaires : il se construit avec le participe passé d'un autre verbe, ses formes
5

Figure 78 Titres détectés dans un mémoire en français avec le modèle poppler

Les titres
grammaticale, lexicale, en orthographe et sémantique, etc.
1- Les étudiants de la troisième année français font-ils la différence entre les verbes
intransitifs, les verbes transitifs directs et les verbes transitifs indirects.
11
2- Les étudiants de la troisième année français qui sont en contact permanent avec le
3- Ces étudiants peuvent utiliser les verbes dans des constructions transitives et
conclusion.
12
1 .1. Définition du verbe
Etymologiquement, le mot verbe est issu du latin vebrum, qui signifie «parole», mot par
mots, aux groupes de mots qui se construisent autour de lui. Il a pour rôle admettre à
arrive) ou il est suivi d'un complément du verbe essentiel ou facultatif (exemple :
14
1.2. Le groupe verbal
complément.
1.3. Locution verbale
verbe.
1. Locution nominal : pomme de terre.
2. Locution verbale : rendre visite.
3. Locution prépositionnelle : à l'aide de.
4. Locution adjectivale : avoir beau.
1.4. Espèces de verbe
1. Auxiliaires : il se construit avec le participe passé d'un autre verbe, ses formes
15
3. Les verbes supports qui se combine avec d'autres parties du discours qu'ils
1.5. Les constructions du verbe
1.5.1. Verbes transitifs (verbes objectifs)
chose, c\'est-à-dire l'objet de l'action. Ils se construisent avec un complément d'objet
1.5.1.1. Transitif direct
1.5.1.2. Transitif indirect
16
1.5.2. Verbes intransitifs (verbes subjectifs)
phrase. Mais, un verbe
circonstanciel. Exemple : Il a neigé pendant cinq jours. Le verbe « neigé » est un verbe
intransitif. (Pendant cinq jours : complément circonstanciel de temps).
1.5.3. Verbes doublement transitifs (ditransitifs)
objet. Exemple : Tu offre des fleurs à sa mère. Le verbe offrir admet un COD (des

Figure 79 Titres détectés dans un mémoire en français avec le modèle pdfminer

Les titres
Aujourd™hui, la recherche des nouveaux designs permettan
t, soit d™assurer une
inte, soit d™alléger une structure est de la plus haute im
n, nous ne
ées, le premier objectif de la
action, cisaillement par
traction. Le deuxième objectif est d™identifier les par
I.1. Principe du soudage par résistance
-à dire qu™ils exploitent le phénomène de l™échauffement d™un
l™élément, I, l™intensité efficace du courant
etc.
épaisseurs, sont rigoureusement temporisées et se déroulent automatiquement.
soudage. Ils assurent une bonne régularité des conditions de contact des pièces et favorisent
supermarché, est également assimilé au soudage par bossage, car il présente de nombreuses
recouvrement, continues et étanches.
puissance, et porte les parties en contact à la température de soudage ;
similaire.
-t-on recours
alisé. Il permet de souder des sections de
10 000 mm
2
a. Accostage
achine.
b. Effort
fusion.
c. Soudage
d. Temps de soudage
e. Intensité
f. Maintien
g. Forgeage
recherchée. Par analogie à d'autres procédés d'assemblage discontinus, tels que le vissage ou
soudé.
soudage. Un cycle de soudage par point complet est typiquement effectué en moins d'une
seconde.
chauffage, une partie du carbone peut être remis en solution. Au refroidissement, on peut
refroidissement.
thermique. Au chauffage, la nouvelle structure austénitique, non homogène et globalement

Figure 80 Titres détectés dans un mémoire en français avec le modèle pyPDF2

4.5.4.2 Détection des titres pour un mémoire en Arabe

Les titres
النظرية التأويلية عند سيليسكوفيتش1.
مراحل الترجمة الشفوية2.
نبذه عن الترجمة الشفوية3.
الفرق بين الترجمة الفورية و التتابعية4.
بعض أسس تعليم الترجمة الشفوية حسب مدرسة باريس5.
أخذ النقاط في الترجمة التتابعية6.
الترجمة في الخطاب السياسي7.
ما هو الرصيد المعرفي؟8.
تعريف المترجم المؤازر9.
النظرية التأويلية عند سيليسكوفيتش1. :
مراحل الترجمة الشفوية2. :
نبذه عن الترجمة الشفوية3. :
الفرق بين الترجمة الفورية و التتابعية4. :
بعض أسس تعليم الترجمة الشفوية حسب مدرسة باريس5. :
)quelqu'un parlant anglais : « c'est un Néerlandais ! »), entendre les tics de langage (la répéti -tion de n'est-ce-pas, de euh, euh
أخذ النقاط في الترجمة التتابعية6. :
الترجمة في الخطاب السياسي7.
ما هو الرصيد المعرفي؟8. :
تعريف المترجم المؤازر9. :
تعريف علم النفس1.
الغرض من ربط علم النفس بالترجمة2.
مفهوم الخالق3.
قوه استعمال العقل الباطن في تحدي عامل الخوف4.
مصدر القلق و الخوف5.
ضروره التواصل عن طريق العقل الحوفي6.
التحرر من التفكير السلبي يجعل التواصل أكثر إيجابية7.
تعريف علم النفس1. :
الغرض من ربط علم النفس بالترجمة2. :
علاقة علم النفس بالجانب الخالقي3. :
قوه استعمال العقل الباطن في تحدي عامل الخوف4. :
تعريف العقل الباطن1.4. :
علاقة العقل الواعي بالعقل الباطن2.4. :
كيف يعمل العقل الباطن3.4. :
الوعي و الالوعي ليسا عقلين مختلفين منفصلين4.4. :
كيف يتحكم الالوعي في جميع وظائف الجسم5.4. :
مصدر القلق و الخوف5. :

Figure 81 Titres détectés dans un mémoire en Arabe avec le modèle poppler

Les titres
تعليم الترجمة الشفوية حسب مدرسة باريس بعض 5 6
صغرها دالمدارس لو هذا يرجع إلى عيشها في عده بلدافي حيث 2
في 1
قوله من طرف المتحدثل ما يرالقى ما هو مهم في الخطاب في كان في الترجمة الفورية ألو التتابعية،أي اللجوء إلتعتمد النظرية التأويلية في الترجمة على ترجمة المعنى دلون الحرف سواءا مفهوم النظرية التأويلية 1 3
,syap srueisulp snad tnemevisseccus essenuej as snad ucév tnaya ,ic ,siaçnarf el tialrap te tiassiannoc noitcudart al ed evitaterpretni– « 1 egap , 6102 niuj , 3 siraP ellevuoN ennobroS al ed étisrevinu ,noitulove te enigiro www//:sptthemuser_nu_noitcudart_
hctivokseleS-
3 elocé'l à seugnal sec tnemelbatirév erdnerppa à ue riova siamaj snas sialgna'l te ebres el ,dnamella'l »elleC
بطريقة تلقائية ألل لو هي المحترف يستعملها الترجماذإ يذكر جل المترجمين الشفويين المراحل الثالثة التي الموضوع المتداول:لومات التي تخص ملما أيضا بالمع يكونكما البد عليه أن يكون دائما محايدابكافر" فالترجمان ينقل فقط إيديولوجية المتكلم لو ال : 2 يجب أن يستعمل مشاعره،أي يقا: "ناقل الكفر ليس يكون ملما بالفكار لو الخلفيات الإيديولوجة للمتحدث لو كماخالتها يكون الترجمان قد تقمص شخصية المتحدث التي بها يعرف نواياه،أي فمن هالوراأي الإصغاء على رأسمتواتره مراحل إلى الشفوية الترجمة نقسمت : مراحل الترجمة الشفوية 2
كوفي ستفدإلتباه له في عملية الترجمة الشفويةيقو: من الناحية الإصغاء» yevoC nehpetS « : "أن تفهم لو بعدها ستفهم 2
:حالو" 2
أما إذا كان الخطاب علمياً لوفنياً سوف يكون التركيز على المضمون الترجيب لوالشكر لوالتقدير سوف يكون التركيز هنا على فإذا تضمن الخطابلوطبيعة اللغة لومستواها لونوع الخطاب المعالجتحليل الخطاب باختالف الموضوع ترجمتها لويختلف ثم ، 3 بالفرنسية بـ لو هذه المرحلة تسمان يحلّل خطاباً دلون فهمه على الفهم إذ ال يمكن يبدى الذي لو الكالميحالو: تحليل الترجمان ان يعني أن : مرحلة التحليل noitatsilabrevedالصاللغة متلقى الرسالة بين الإصغاء لو لو المحصوره
,)denoitnem ton raey(noitide traehylf ,elpoep evitceffe ylhgih fo stibah neves ehT,yevoC nehpetS 2 utuoy//:sptth
الترجمة الشفوية حسب المدرسة التأويلية لو: الفصل ال ~ 7 ~ 2 c0x\ www//:ptth
لوال يصاب بالتعب لوالإعياء الراحة لكي يستعيدمحصوراً على الخطاب لولذلك يتطلب من المترجم الفوري أن يتمتع بقسط من تركيزه فكرياً لوعصبياً كبيراً لوالبد من ان يكون في حالة استفار قصوى لويكونلوكما هو معرلوف لدى أصحاب الخبره ان 2 الترجمة الفورية تتطلب جهداً كبيره
تبذه عن الترجمة الشفوية 3 : 3
إذ تنقسم الترجمة الشفوية إلى أربعة أنواع لو هي المواضيع : 3
1
المرجع السابقالدكتور حسيب إلياس حديد، 1 المرجع السابقموفق فائق توفيق ، بد منه لو إال قال يستطيع ألجنبي التعامل مع قوانين ضرلوري2 c0x\الترجمة الشفوية حسب المدرسة التأويلية لو: الفصل ال ~ 8 ~ 3
1
من الترجمة الترجمة 3
1
للوقت لو تركيز عب 3
1
بالنسبة للمترجم لوالمستمع على حد سواءلوال يستخدم هذا النوع من الترجمة في جلسات المناقشات الطويلة ألنها تسبب المصدر، لوكما بد: عليها إسمها فإن الترجمان يهمس ترجمته في أذن المستمع، النوع من الترجمة حين يتعذر على شخص ألو شخصين 1 الترجمة الهمسية 4)gniterpretni derepsihw(: فهم اللغة هذا يصلح
« c0x\لو زمن إعاده سيالهالي طفيف في الزمن ال يتجالوز بضع ثوان بين زمن تل عملية التلفظ مع فارق ~ 9 ~ الترجمة الشفوية حسب المدرسة التأويلية لو: الفصل ال 1
)etunim/stom 022 te 021 ertne(elorap al ed lamron emhtyr ua nifne revirra te tnemevissergorp rerélécca ruop tnemetnel sélucitra sruocsid sed erdnetne eriaf rap recnemmoc ed noitseuq sap tnemmedivé tse'\n li ,essetiv ettec erdnietta ruoP laro
عن المشكالت نفسهالو الداطقين باسمهم، لو بين اللغة العادية التي يستعملها رجل الشارع البسيط حين الواقع سواء كان داخليا ألو خارجيا تحديدا كما يظهر في كالم السابة لو مستشاريهم ~ 21 ~ الترجمة الشفوية حسب المدرسة التأويلية لو: الفصل 1 ص 0102/9002جامعة قسنطينة المادة في ترجمة الخطاب السياسي ، مذكرة ماجيستير، جالل الدين بن عائشة الف بين نوع الخطاب السياسي الذي يتدالو: مشكالت ألخير ألوجه التشابه لو الختيوصف كل استعما: للغة على أنه سلوك لفظي15 \c0xال يؤدي إلى التواصل، لو لهذا 1ترجمة المعنىعلى ألسلوب الذي قد يضع الترجمان في حيره من أمره بين ترجمة ألسلوب ألو لو يقرأ من لوراء السطور لو أيضا فيما أحيانا ألسلوبفي صعابالهذه ظهرلوتالترجمات، من أصعب كما تشير ألبحاث تعد ترجمة الخطاب السياسي بالتاليلو السياسة
يتحدث 7
المترجم الفوري لوخصوصية1 7
2
الفصل الثاني من هذا لتوجيه c0x\الترجماندور علم النفس في توجيه الفصل الثاني c0x\ دور علم النفس في توجيه الترجمان : الفصل الثاني 1
لسلبي يجعل التواصل أكثر إيجابيةالتحرر من التفكير 7 c0x\ دور علم النفس في توجيه الترجمادالثاني الفصل ~ 12 ~ 1
كما جاء عند أنطوني بيم 3، أي مجموعة من المعتقدات» myP ynohtnA «و ذلك يكون عن طريق إلبتعاد عن الذاتية و الترجمة بكل موضوعيةمادة الترجمية لنقل الثقافتين ثقافة اللغة ألصل و ثقافة اللغة ألا ما يبدى عليهي ن ألخالق في الترجمة فإ 4

Figure 82 Titres détectés dans un mémoire en Arabe avec le modèle pdfminer

4.5.4.3 Détection des titres d' un mémoire en Anglais

Les titres
education, art and culture, love, sex, obscenity, class conflict.... Most are wellknown, even to the twentieth century man. Human higher capacities, linked with
tenderness, care, intelligence, reason and creativity, as well as commitments and
However, human relationship encompassing sensitivity, affection and tenderness
point, then, is to identify a common ground in order to build comprehension, as a
consensus, from different inclinations and tendencies. In clear words,
2
3
4
1.9. Introduction .....
6
1.10. New Historicism .....
6
1.10.1. Definition .....
6
1.10.2. The Origins of New Historicism .....
7
1.10.3. Defining some Concepts.....
8
1.2.3.1 . Obscenity.....
8
1.2.3.2 . Censorship.....
8
1.11. Edwardian
9
9
10
1.12. Edwardian
Woman.....
11
1.13. Marriage, Sex, and Love.....
12
1.14. Introducing Sons and Lovers .....
14
1.14.1. The Novel .....
14
1.14.2. The Author .....
15
1.15. Conclusion .....
17
1.1. Introduction
literature, particularly Lawrence's masterpiece, Sons and Lovers.
1.2. New Historicism
history, their feelings, aspirations, and traditions. Thus, when reading a text, one not
1.2.1. Definition
power, society, or ideology in a given time." New Historicism developed as a school
fact, these main components are decoded from the literary text providing deep analysis
6
church, men and women of different classes – interact for the building and the
1.2.2. The Origins of New Historicism
Thus, the author is a subject to his context and therefore to his historical
background, living circumstances, class, family and world view. Another central aspect
7
possible.
1.3. Edwardian Society
1.3.1. Industrial Background

Figure 83 Titres détectés dans un mémoire en Anglais avec le modèle poppler

Les titres
education, art and culture, love, sex, obscenity, class conflict.... Most are well-\n\nknown, even to the twentieth century man. Human higher capacities, linked with
tenderness, care, intelligence, reason and creativity, as well as commitments and
However, human relationship encompassing sensitivity, affection and tenderness
point, then, is to identify a common ground in order to build comprehension, as a
consensus, from different inclinations and tendencies. In clear words,
2
3
1.9. Introduction .....
1.10. New Historicism .....
1.10.1. Definition .....
1.10.2. The Origins of New Historicism .....
1.10.3. Defining some Concepts.....
1.2.3.1 . Obscenity.....
1.2.3.2 . Censorship.....
1.11. Edwardian
1.11.1. Industrial Background.....
1.11.2. Social Reforms.....
10
1.12. Edwardian
Woman.....
1.13. Marriage, Sex, and Love.....
1.14. Introducing Sons and Lovers .....
1.14.1. The Novel .....
1.14.2. The Author .....
1.15. Conclusion .....
4
6
6
6
7
8
8
8
9
9
11
12

Figure 84 Titres détectés dans un mémoire en Anglais avec le modèle pdfminer

Les titres
2
known, even to the twentieth century man. H
tenderness, care, intelligence, reason and creativity, as well as commitments and
wrong.
lie,
falsity,
tenderness.
However,
sensitivity, affection and tenderness
point, then, is to identify a common ground
omprehension, as a
consensus, from different inclinations and tendencies.
eyes,
3
4
1.9.
6
1.10.
6
1.10.1.
6
1.10.2.
7
1.10.3.
8
1.2.3.1
8
1.2.3.2
8
1.11.
9
1.11.1.
9
1.11.2.
10
1.12.
11

Figure 85 Titres détectés dans un mémoire en Anglais avec le modèle pyPDF2

4.5.4.4 Détection des titres d'un article scientifique

Les titres
0
1 author:
1.1 Introduction
recognized. In order to enable systems to provide users with richer
available, for example, in the form of XML markup embedded in the full
3
4
described, together with the corresponding evaluation metrics used and
1.1.1 Background
users, enabling them to gain direct access to those parts of books
Track, we entirely created the methodology to evaluate the Structure
procedure, annotation procedure (and corresponding software), metrics,
1.1.2 Context and Motivation
However, since the collection was made of digitized books, the only
file, as a result of the scanning process. In addition, a few other elements
pages, words (detected as regions of text separated by horizontal space),
Hence, there is a clear gap to be filled between research in structured
IR, which relies on a logical structure (chapters, sections, etc.), and the
6
1.1.2.1 Structured Information Retrieval Requires Structure
moment, it seems, however, that users are still attached to the concept of
3 millennia until the Roman codex brought up the concept of a page.
usable. This motivated the design of the Book Structure Extraction
competition, to bridge the gap between the digitized books and the
1.1.2.2 Context
8
measures, and a first test set of 100 books built by the organizers.
(ICDAR) [9] where it was accepted as an official competition. This
year, whereas ICDAR runs every second year).
systems, as well as the challenges and contributions that this work involved.
1.2 Book Collection
books, biographies, literary studies, religious texts and teachings, reference
works, encyclopedias, essays, proceedings, novels, and poetry.
book, headers may include chapter/section titles and logical page numbers
separately. Coordinates that correspond to the four points of a rectangle
10
1.3 Setting Up the Competition

Figure 86 Titres détectés dans un Article scientifique avec le modèle poppler

Les titres
1.1 Introduction
-ing whole libraries by digitizing books on an industrial scale [5]. The
recognized. In order to enable systems to provide users with richer
experiences, it is necessary to make such additional structures
available, for example, in the form of
4 Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems
-tion in the context of the work conducted at the Initiative for the
described, together with the corresponding evaluation metrics used and
1.1.1 Background
users, enabling them to gain direct access to those parts of books
structure, e.g., chapters, table of contents
-tized books, we created the Book Structure Extraction competition, which
Track, we entirely created the methodology to evaluate the Structure
procedure, annotation procedure (and corresponding software), metrics,
1.1.2 Context and Motivation
However, since the collection was made of digitized books, the only
-ily identified from the fact that it corresponds to one and only one image
file, as a result of the scanning process. In addition, a few other elements
OCR, as we can see with the DjVu file
pages, words (detected as regions of text separated by horizontal space),
Hence, there is a clear gap to be filled between research in structured
IR, which relies on a logical structure (chapters, sections, etc.), and the
6 Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems
books, but it remains clearly insufficient.
1.1.2.1 Structured Information Retrieval Requires Structure
moment, it seems, however, that users are still attached to the concept of
-pear in the long run. All physical
-tured IR, while on the other, the collection™s logical structure was hardly
usable. This motivated the design of the Book Structure Extraction
competition, to bridge the gap between the digitized books and the
(structured) IR research community.
1.1.2.2 Context
8 Document Analysis and Text Recognition: Benchmarking State-of-the-Art Systems
measures, and a first test set of 100
(ICDAR) [9] where it was accepted as an official competition. This
year, whereas ICDAR runs every second year).
systems, as well as the challenges and contributions that this work involved.

Figure 87 Titres détectés dans un Article scientifique avec le modèle pyPDF2

Les titres
1.1.1 Background\n\nMotivated by the need to foster research in areas relating to large digital \nbook repositories (see e.g., [21]), the Book Track was launched in \n2007\xa0[22] as part of the INEX. Founded in 2002, INEX is an evaluation \nforum that investigates focused retrieval approaches [14] where structure \ninformation is used to aid the retrieval of parts of documents, relevant to \na search query. Focused retrieval over books presents a clear benefit to \nusers, enabling them to gain direct access to those parts of books \n(of\xa0 potentially hundreds of pages in length) that are relevant to their \n information needs.\n\nOne major limitation of digitized books is the fact that their structure \nis physical, rather than logical. Following this, the evaluation and relevance \njudgments based on the book corpus have essentially been based on whole \nbooks and selections of pages. This is unfortunate considering that books \nseem to be the key application field for structured information retrieval \n(IR). The fact that, for instance, chapters, sections, and paragraphs are not \nreadily available has been a frustration for the structured IR community \ngathered at INEX, because it does not allow us to test the techniques \n created for collections of scientific articles and for Wikipedia.
1.1.2 Context and Motivation\n\nThe overall goal of the INEX Book Track is to promote interdisciplinary \nresearch investigating techniques for supporting users in reading, search-\ning, and navigating the full texts of digitized books and to provide a forum \nfor the exchange of research ideas and contributions. In 2007, the Track \nfocused on IR tasks [24].\n\nHowever, since the collection was made of digitized books, the only \nstructure that was readily available was that of pages, each page being eas-\nily identified from the fact that it corresponds to one and only one image \nfile, as a result of the scanning process. In addition, a few other elements \ncan easily be detected through OCR, as we can see with the DjVu file \nformat (an example of which is given in Figure 1.1). This markup denotes \npages, words (detected as regions of text separated by horizontal space), \nlines (regions of text separated by vertical space), and “paragraphs” (regions \nof text separated by a significantly wider vertical space than other lines). \nThose paragraphs, however, are only defined as internal regions of a page \n(by definition, they cannot span over different pages).\n\nHence, there is a clear gap to be filled between research in structured \nIR, which relies on a logical structure (chapters, sections, etc.), and the \ndigitized book collection, which contains only the physical structure. From
1.1.2.2 Context\n\nIn 2008, during the second year of the INEX Book Track, the Book \nStructure Extraction task was introduced [25] and set up with the aim to
1.2 Book Collection\n\nThe INEX Book Search corpus contains 50,239 digitized, out-of- \ncopyright books, provided by Microsoft Live Search and the Internet \nArchive [22]. It consists of books of different genres, including history \nbooks, biographies, literary studies, religious texts and teachings, reference \nworks, encyclopedias, essays, proceedings, novels, and poetry.\n\nEach book is available in three different formats: image files as port-\nable document format (PDF), DjVu XML containing the OCR text and \nbasic structure markup as illustrated in Figure 4.1, and BookML, contain-\ning a more elaborate structure constructed from the OCR and illustrated \nin Figure 1.2.\n\nDjVu format. An <OBJECT> element corresponds to a page in a digi-\ntized book. A page counter, corresponding to the physical page number,
1.3.1 Def ining the Evaluation Corpus\n\nIn 2009, 2011, and 2013, the Book Structure Extraction evaluation corpus \nconsisted of 1,000 distinct book subsets of the Book Search Track’s 50,239 \nbook corpus. Therefore, it consisted of a representative set of books of \n different genres, including history books, biographies, literary studies, \nreligious texts and teachings, reference works, encyclopedias, essays, \n proceedings, novels, and poetry.
semi-automatic) #REQUIRED \ntoc-source (book-toc | no-book-toc |
xml\npdf\n\n(yes|no) #REQUIRED \n(yes|no) #REQUIRED>
(#PCDATA) #REQUIRED \ntitle\npage (#PCDATA) #REQUIRED>
Finally, the annotation effort was shared among all participants. \nTeams that submitted runs were required to contribute a minimum of \n50\xa0 books, while others were required to contribute a minimum of \n100\xa0books (20% of those books did not contain a printed ToC). The cre-\nated ground-truth was made available to all contributing participants for \nuse in future evaluations.
1.3.4.2 Collected Ground-truth Data\n\nIn 2009, seven teams participated in the ground-truth annotation process, \nfour of which did not submit runs.
1.3.6 Metrics\n\nThe automatically generated ToCs submitted by participants were evalu-\nated by comparing them to a manually built ground-truth. The evaluation \nrequired the definition of a number of basic concepts.\n\nTable 1.1. The score sheet measuring annotator agreement for \nthe 61 books that were assessed independently by two distinct \ninstitutions.
Complete, except depth
Precision\n(%)
74.32
82.45
82.45
73.57
83.91
75.00
82.87
82.87
74.25
82.86
74.04

Figure 88 Titres détectés dans un Article scientifique avec le modèle pdfminer

4.5.4.5 Détection des titres d'un document financier en Anglais

Les titres
04
A. Sales Prospectus
10
11
12
B. Management Regulations
12
I. Management/Organisation
13
II. General Investment Policy Guidelines
14
III. Issues and Redemptions, Further Conditions
17
C. Special REGULATIONS
D. ANNEX
I. Issuers
1. The Fund
2. Investment Policy
bank, the state banks of Bavaria and Baden-Wuerttemberg and
risk.
1 March 2000 under the name BANTLEON STRATEGIE NO. 1
2011, it was transformed into an investment fund in accordance
Council.
another. Each sub-fund is liable only for its own obligations with
4
AG, Bantleon Yield focuses more specifically on maximising
return).
maturities. The limited sale of interest rate futures further allows
3. Risk Warning
rate, credit, liquidity and counter party risks as well as volatility
policy. Also investors should fully understand the risks associated with the investment in shares and only make investment
decisions, if they match the advise given by their own legal, tax
launched, the expected leverage value will be calculated on the
sub-fund. Greater leverage amounts may be attained for all subfunds, under certain circumstances.
value)
n.a.
n.a.

Figure 89 Détection des titres d'un document financier en Anglais avec poppler

Les titres
A. SAleS PRoSPeCtuS\n\n The Fund | Investment Policy | Risk Warning | Investor Profile | The Management Company | The\xa0 Depositary \nBank | Management and Transfer Agent | Prevention of Money Laundering | Shares | Share Prices | Acquisition, \nRedemption and Exchange of Shares, Purchase Price Payment | Charges and Fees | Further Notes | Publications \nand Information | The Management Regulations | Combating Market Timing and Late Trading Activities\n\n10
11
B. MAnAgeMent RegulAtIonS
12
I. MAnAgeMent/oRgAnISAtIon\n\n
D. AnneX
I. ISSueRS\n\n The Fund | The Management Company | The Investment Manager | The Depositary Bank | \nManagement and Transfer Agent\n\n13
II. geneRAl InveStMent PolICy guIDelIneS\n\n Investment Objective | Listed Securities | Unlisted Securities and Other Securitised Rights | New Issuers | Investment \nLimits | Interest Rate Futures | Liquid Assets | Further Investment Restrictions | Loans and Encumbrance Prohibitions\n\n14
III. ISSueS AnD ReDeMPtIonS, FuRtheR ConDItIonS\n\n Shares in the Sub-Funds | Issue of Shares | Calculation of the Net Asset Value per Share | Suspension of Calculation of \nthe NAV per Share | Redemption of Shares | Exchange of Shares | Dividend Distributions | Duration and Closing of \nthe Fund and the Sub-Funds and the Merger of Funds or Sub-Funds | Fees, Charges | Fiscal Year and Audit of Annual \n Accounts | Statute of Limitations | Amendments Publications | Applicable Law, Jurisdiction and Language of Contract | \nEntry into Force
17
C. SPeCIAl RegulAtIonS\n\n Bantleon Return | Bantleon Yield | Bantleon Strategie | Bantleon Trend | Bantleon Dynamic\n\n\x0cSAleS PRoSPeCtuS\n\n1. the Fund\n\nBantleon Return, Bantleon Yield, Bantleon Strategie, Bantleon \nTrend and Bantleon Dynamic are sub-funds of the »BANTLEON\nANLEIHENFONDS« (hereinafter referred to as »the sub-funds«). \n Bantleon Invest S. A., a subsidiary of Bantleon Bank AG is \nresponsible for managing the sub-funds.
7. Management and transfer Agent\n\nThe management and transfer agent is UBS Fund Services \n(Luxembourg) S. A., with its registered office at 33A, avenue \nJ.F. Kennedy, 1855 Luxembourg, a 100 % subsidiary of UBS AG, \nSwitzerland.\n\nUBS Fund Services (Luxembourg) Bantleon Invest S. A. \nis, as the management and transfer agent, responsible for the \ngeneral administrative tasks required within the framework of \nfund management and prescribed under Luxembourg law. These \nservices comprise primarily domiciliation, the calculation of the \nnet asset value (NAV) of the shares, accounting for the fund and \nreporting.\n\n8. Prevention of Money laundering\n\nThe sales offices of the management company are obliged to \nobserve the rules of the Luxembourg Law as of 19 February 1973. \nthe law on the financial sector as of 5 April 1993 and the law as \nof 12
10.000.001 to 25.000.000
25.000.001 to 50.000.000
12. Charges and Fees\n\n0.10 %\n\n0.08 %\n\n0.05 %\n\n0.03 %\n\nA lump sum management fee is applied to the sub-funds, the \nlevel of which is regulated in the Special Section of the Manage-\nment Regulations.\n\nThe overall management fee includes, in particular, the costs of \nthe\n\n■ depositary bank\n■ management and transfer agent\n■ management company\n■ distributor\n■ registration with supervisory authorities\n■ auditing company\n■ issue and distribution of annual and semiannual reports\n■ issue and distribution of all other reports and documentation\n■ sales documentation and prospectuses\n■ accounting\n■ determination on each business day of the share price and its
15. the Management Regulations\n\nThe Management Regulations printed below are subdivided into \na General Section and the relevant Special Sections. The General \nSection details the legal principles and the general investment \nguidelines. The Special Sections of the Management Regulations \ncontain the sub-fund specific details and the investment policy of \nthe relevant sub-fund.\n\n16. Combating Market timing and late trading Activities\n\nMarket timing is the arbitrage method whereby the investor \nsystematically subscribes to and redeems or exchanges shares of \na fund within a short time frame, utilising the time lags and/or \nimperfections or weaknesses in the evaluation system of the NAV \nof the fund. Market timing is however excluded in the case of \nall sub-funds of the »BANTLEON ANLEIHENFONDS«, as the \ndetermination of the relevant NAV per share is based on prices, \nwhich are set without exception after the latest point in time for \nsubscription, exchange or redemption applications.\n\nLate trading is the acceptance and execution of a subscription, \nexchange or redemption application, which is received after \nthe expiry of the period allowed for acceptance of applications \n (cut-off time) on the relevant day. The periods specified in the \nManagement Regulations for acceptance of subscription, ex-\nchange or redemption applications must be adhered to without \nexception and expire before the point in time when the NAV \nper share is determined. The NAV per share is therefore always \nunknown at the time the order for purchases, sales and exchange \ntransactions is issued.
Shareholders, who are not resident in Luxembourg or do not \nmaintain any permanent establishments there, are not required \nto pay either income tax, inheritance tax or wealth tax in Lux-\nembourg on their shares or income from shares. The relevant \nnational tax laws apply to them.\n\nThe preceding paragraphs are only a summary of tax implications \nand do not claim any completeness. It shall be for the interested \ninvestors and shareholders to catch up with the laws and all regu-\nlations governing the purchase, the holding and possible sale of \nshares in connection with their place of residence or nationality.\n\n14. Publications and Information\n\nThis sales prospectus, the Management Regulations including \nthe Special Regulations and the annual and semi-annual reports \nas well as the sales documents can be obtained from Bantleon \nInvest S. A. and all paying agents and distributors.\n\n9\n\n\x0cSAleS PRoSPeCtuS\n\nFunD oveRvIew\n\nBantleon \nReturn\n\nBantleon \nYield\n\nBantleon \nStrategie\n\nBantleon \nTrend
0.10 %\n0.08 %\n0.05 %\n0.03 %\n\n0.40 %\n\n0.50 %\n\n0.50 %\n\n0.50 %\n\n1.00 %\n\n1.00 %\n\n1.25 %\n\n1.25 %\n\n1.25 %\n\n2.50 %\n\nEnd of fiscal year\n\n30 November\n\n30 November\n\n30 November\n\n30 November\n\n30 November\n\nUse of income (»IA«/»PA«)\n\n(min. semi-annual)\n\nDividend \ndistribution\n31 May and\n30 November\n\nDividend \ndistribution\n31 May and\n30 November\n\nDividend \ndistribution\n31 May and\n30 November\n\nDividend \ndistribution\n31 May and\n30 November\n\nDividend \ndistribution\n31 May and\n30 November\n\nUse of income (»IT«/»PT«)\n\naccumulating\n\nDate of the \nannual report\n\nDate of the \nsemi-annual\xa0report\n\n30 November, released mid-February\n\n31 May, released mid-July\n\nExchange quotation\n\nLuxembourg\n\nLuxembourg\n\nLuxembourg\n\nLuxembourg\n\nLuxembourg\n\n10\n\nSAleS PRoSPeCtuS wIth MAnAgeMent RegulAtIonS\n\n\x0cMAnAgeMent AnD ADMInIStRAtIon\n\nAuditor of the Management Company \nand the Fund\n\nManagement Company\n\nBantleon Invest S. A.\n33A, avenue J. F. Kennedy\n1855 Luxembourg, Luxembourg\n\nBoard of Directors of the Management Company\nMarcel Rösch (Chairman)\nMember of the Management of\nBantleon Bank AG, Zug\n\nWerner A. Schubiger (Member)\nVice President of the Board of Directors \nof Bantleon Bank AG, Zug;\nLawyer in Zürich\n\nGünter Lutgen (Member)\nExecutive Director,\nUBS Fund Services (Luxembourg) S. A.,\nLuxembourg\n\nAloyse Hemmen (Member)\nExecutive Director\nUBS Fund Services (Luxembourg) S. A.,\nLuxembourg\n\nResponsible for the day-to-day

Figure 90 Détection des titres d'un document financier en Anglais avec pdfminer

Les titres
A.
-ment of high-quality bondse dierent management styles
.All sub-funds invest solely in bonds in accordance with the
bank, the state banks of Bavaria and Baden-Wuerttemberg and
risk.
-cordance with Parof the Luxembourg Law on Undertakings
UCITS) as a public fund (fonds commun de placement) on
-dertakings for Collective Investment of December . e
-tive Investment oecembe.
Council.
another. Each sub-fund is liable only for its own obligations with
-ing of duration adjustment, yield curve management, spread
AG, Bantleon Yield focuses more specically on maximising
return).
-mised by overweighting longer maturitiese aim of completely
rate, credit, liquidity and counter party risks as well as volatility
policy. Also investors should fully understand the risks associ
-ated with the investment in shares and only make investment
decisions, if they match the advise given by their own legal, tax
-ble laws and regulatory provisions.
-age amounts as some derivatives that can be used for hedging
-age risk that the investor is exposed to.
-rivatives and the net asset value of the respective sub-fund and is
launched, the expected leverage value will be calculated on the
sub-fund. Greater leverage amounts may be attained for all sub
value)
n.a.
n.a.
n.a.
n.a.
n.a.
n.a.
n.a.
n.a.
n.a.
n.a.

Figure 91 Détection des titres d'un document financier en Anglais avec pyPDF2

4.5.5 L'extraction de la TDM

Par ailleurs, notre application web offre la possibilité d'extraire la TDM qui représente la tâche principale de notre système, En cliquant sur le bouton « ExtraireTDM » la restitution de la TDM sera effectuée, pour l'extraction de la TDM des fichiers écrits en arabe il suffit de cliquer sur le bouton « ExtraireTDM_Arabic », L'utilisateur aura la possibilité d'afficher la TDM en appuiera sur le bouton « Visionner la TDM ». (Voir la figure 92)

Figure 92 Phase de la restitution de la TDM

Les figures que nous allons vous montrer prochainement représentes les différents résultats obtenus lors des tests effectués en utilisant de différentes techniques et de corpus pour l'extraction de la TDM.

4.5.5.1 L'extraction de la TDM d'un mémoire en Français

Table des matières
1 .1. Définition du verbe
1.2. Le groupe verbal
1.3. Locution verbale
1.4. Espèces de verbe
1.5. Les constructions du verbe
1.5.1. Verbes transitifs (verbes objectifs)
1.5.1.1. Transitif direct
1.5.1.2. Transitif indirect
1.5.2. Verbes intransitifs (verbes subjectifs)
1.5.3. Verbes doublement transitifs (ditransitifs)
1.5.4. Exceptions
1.5.5. Verbes attributifs
1.5.5.1. Espèce d'attribut
1.5.5.1.1. Attribut de sujet
1.5.6.1. Types des verbes pronominaux
1.5.6.1.1. Sens réfléchi
1.5.6.1.2. Sens réciproque
1.5.6.1.3. Sens passif
1.5.6.1.4. Sens essentiellement pronominaux
1.5.7. Verbes impersonnels
1.6. Complément du verbe
1.6.1. Définition du complément
1.6.2. Types du complément
1.6.2.1. Complément d'objet direct
1.6.2.2. Complément d'objet indirect
1.6.2.3. Complément d'objet second(COS)
1.6.2.4. Complément circonstanciel (CC)
1.6.2.5. Complément d'agent
1.6.3. Pronoms personnels compléments
1.6.3.1. Les pronoms directs
1.6.3.2. Les pronoms indirects
1.7. Formes du verbe
1.7.1. La personne
1.7.2. Le nombre
1.7.3. Les modes du verbe
1.7.4. Temps du verbe
1.7.5. Aspect verbal
1.7.5.1. 4. Inchoatif/Terminatif
1.7.5.1. 5. Semelfactif/Itératif
1.7.5.1. Principaux aspects du verbe
1.7.5.1.1. Accomplis/Inaccompli
1.7.5.1.2. Perfectif/Imperfectif
1.7.5.1.3. Séant/Non-séant
1.7.5.1.6. Aspect progressif
1.7.6. Voix du verbe
1.7.6.1. La voix active
1.7.6.2. La voix passive
2. 2. Analyse des données
2. Pronom « y »
2.2.1. Analyse de l'exercice n°1
2.2.2.1. Analyse du deuxième exercice
2.2.3. Analyse de troisième exercice

Figure 93 L'extraction de la TDM d'un mémoire en Français avec poppler

Table des matières
1 .1. Définition du verbe
1.2. Le groupe verbal
1.3. Locution verbale
1.4. Espèces de verbe
1.5. Les constructions du verbe
1.5.1. Verbes transitifs (verbes objectifs)
1.5.1.1. Transitif direct
1.5.1.2. Transitif indirect
1.5.3. Verbes doublement transitifs (ditransitifs)
1.5.4. Exceptions
1.5.5. Verbes attributifs
1.5.5.1. Espèce d'attribut
1.5.5.1.1. Attribut de sujet
1.5.6.1. Types des verbes pronominaux
1.5.6.1.1. Sens réfléchi
1.5.6.1.2. Sens réciproque
1.5.6.1.3. Sens passif
1.5.6.1.4. Sens essentiellement pronominaux
1.5.7. Verbes impersonnels
1.6. Complément du verbe
1.6.1. Définition du complément
1.6.2. Types du complément
1.6.2.1. Complément d'objet direct
1.6.2.2. Complément d'objet indirect
1.6.2.3. Complément d'objet second(COS)
1.6.2.4. Complément circonstanciel (CC)
1.6.2.5. Complément d'agent
1.6.3. Pronoms personnels compléments
1.6.3.1. Les pronoms directs
1.6.3.2. Les pronoms indirects
1.7. Formes du verbe
1.7.1. La personne
1.7.2. Le nombre
1.7.3. Les modes du verbe
1.7.4. Temps du verbe
1.7.5. Aspect verbal
1.7.5.1. 4. Inchoatif/Terminatif
1.7.5.1. 5. Semelfactif/Itératif
1.7.5.1. Principaux aspects du verbe
1.7.5.1.1. Accomplis/Inaccompli
1.7.5.1.2. Perfectif/Imperfectif
1.7.5.1.3. Séant/Non-séant
1.7.5.1.6. Aspect progressif
1.7.6. Voix du verbe
1.7.6.1. La voix active
1.7.6.2. La voix passive
2. Pronom « y »
2. 2. Analyse des données
2.2.1. Analyse de l'exercice n°1
2.2.2.1. Analyse du deuxième exercice
2.2.3. Analyse de troisième exercice

Figure 94 L'extraction de la TDM d'un mémoire en Français avec pdfminer

Table des matières
00608595.pp. 31.37
02 affiches
03 affiches
05 affiches
07 affiches
09 affiches
1. Arabe standard
1. Arabe standard
1. La pré
1. Le
1. Publicité
1. Une langue sacrée
1.1 Les affich
1.2 Les affiches plurilingues
17 affiches
18 ans et au
18 ans à 30 ans
19 siècle
2. 1.2.Notre échantillon
2. Arabe Algérien
2. Arabe algérien
2. Une langue pratique
2. Une langue pratique
2.1. Notre
3. Berbère
3. Berbère
4. Français
4. Une langue
4. Une langue utile
4.3. La publicité nous attire
4.5. Les langues p
4.6.2. La langue pratiquée
4.6.4. Y a
4/ Préf
4/ Préférez
5. Anglais
5. Anglais

Figure 95 L'extraction de la TDM d'un mémoire en Français avec pyPDF2

4.5.5.2 L’extraction de la TDM d’un mémoire en Anglais

Table des matières
1.1. Introduction
1.11. Edwardian
1.12. Edwardian
1.2. New Historicism
1.2.1. Definition
1.3. Edwardian Society
1.3.1. Industrial Background
1.3.2. Social Reforms
1.4. Edwardian Woman
1.6. Introducing Sons and Lovers
1.6.1. The Novel
1.6.2. The Author
1.7. Defining Some Concepts
1.7.1. Obscenity
1.7.2. Censorship
1.8. Conclusion
2.1. Introduction
2.4.1. Plot and Setting
2.4.2. Characters
2.4.2.1. Gertrude Morel
2.4.2.2. Walter Morel
2.4.2.3. William Morel
2.4.2.4. Paul Morel
2.4.2.5. Miriam Leivers
2.4.2.6. Clara Dawes
2.6. Lawrence’s Portrayal of Sexuality
2.6.1. Lawrence’s Perception of Sex
2.6.2. Lawrence’s Depiction of Sexuality
2.7. Reasons behind Obscenity
2.8.1. Freud’s Theory
2.8.2. Oedipus Complex
2.9. Conclusion

Figure 96 L’extraction de la TDM d’un mémoire en Anglais avec poppler

Table des matières
2.1. Introduction

Figure 97 L'extraction de la TDM d'un mémoire en Anglais avec pyPDF2

Table des matières
1.1. Introduction
1.11. Edwardian
1.12. Edwardian
1.2. New Historicism
1.2.1. Definition
1.3. Edwardian Society
1.3.1. Industrial Background
1.3.2. Social Reforms
1.4. Edwardian Woman
1.6. Introducing Sons and Lovers
1.6.1. The Novel
1.6.2. The Author
1.7. Defining Some Concepts
1.7.1. Obscenity
1.7.2. Censorship
1.8. Conclusion
2.1. Introduction
2.4.1. Plot and Setting
2.4.2. Characters
2.4.2.1. Gertrude Morel
2.4.2.2. Walter Morel
2.4.2.3. William Morel
2.4.2.4. Paul Morel
2.4.2.5. Miriam Leivers
2.4.2.6. Clara Dawes
2.6. Lawrence's Portrayal of Sexuality
2.6.1. Lawrence's Perception of Sex
2.6.2. Lawrence's Depiction of Sexuality
2.7. Reasons behind Obscenity
2.8.1. Freud's Theory
2.8.2. Oedipus Complex
2.9. Conclusion

Figure 98 L'extraction de la TDM d'un mémoire en Anglais avec pdfminer

4.5.5.3 L’extraction de la TDM d’un mémoire en Arabe

Figure 99 L’extraction de la TDM d’un mémoire en Arabe avec poppler

Figure 100 L'extraction de la TDM d'un mémoire en Arabe avec pdfminer

4.5.5.4 L'extraction de la TDM d'un article scientifique

Figure 101 L'extraction de la TDM d'un article scientifique avec poppler

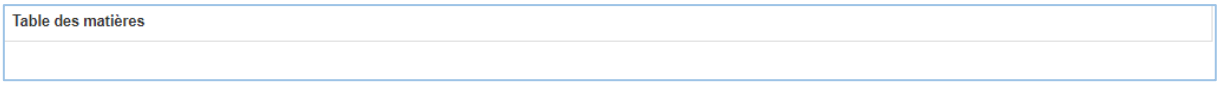


Figure 102 L'extraction de la TDM d'un article scientifique avec pdfminer

Table des matières
1.1 Introduction
1.1.1 Background
1.1.2 Context and Motivation
1.1.2.2 Context
1.2 Book Collection
1.3 Setting Up the Competition
1.3.1 Defining the Evaluation Corpus
1.3.3 Task Description
1.3.4.1 Annotation Process
1.3.4.2 Collected Ground-truth Data
1.3.6 Metrics
1.3.6.1 Alternative Measure and Discussion
1.3.7 Approaches Presented
1.3.8 Summary
1.4 Related Publications
1.5 Conclusions and Perspectives
82.45 82.87 81.83 Complete entries
83.51 83.91 82.86 Levels

Figure 103 L'extraction de la TDM d'un article scientifique avec pyPDF2

4.5.5.5 L'extraction de la TDM d'un document financier en Anglais

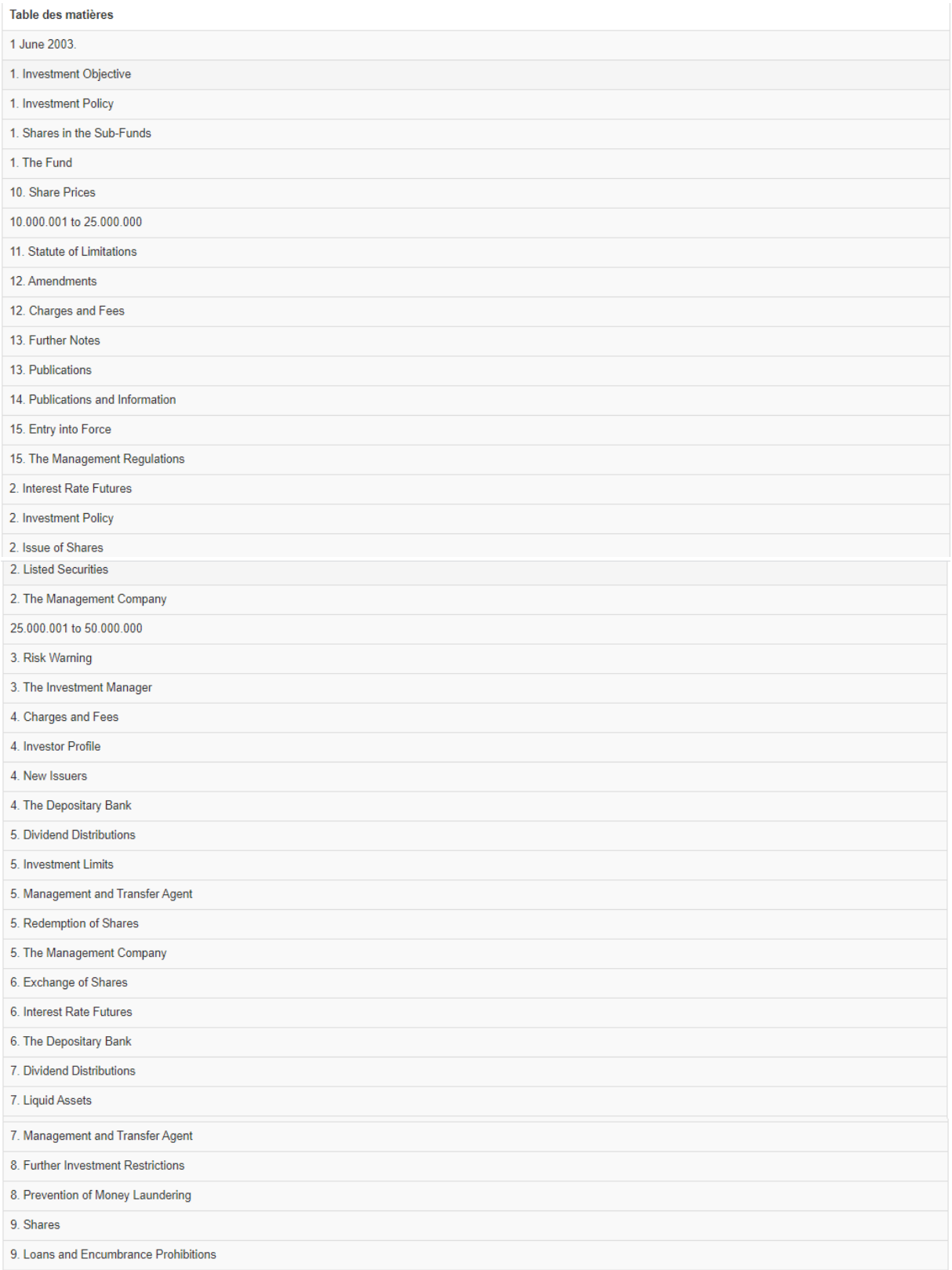


Table des matières

Figure 104 l'extraction de la TDM d'un document financier avec poppler

Figure 105 l'extraction de la TDM d'un document financier avec pdfminer

Figure 106 l'extraction de la TDM d'un document financier avec pyPDF2

4.5.6 Page « à propos de nous »

Comme vous pouvez le remarquer dans la figure 107 , cette page nous représente ainsi que les avantages et le mode d'emploi de notre application

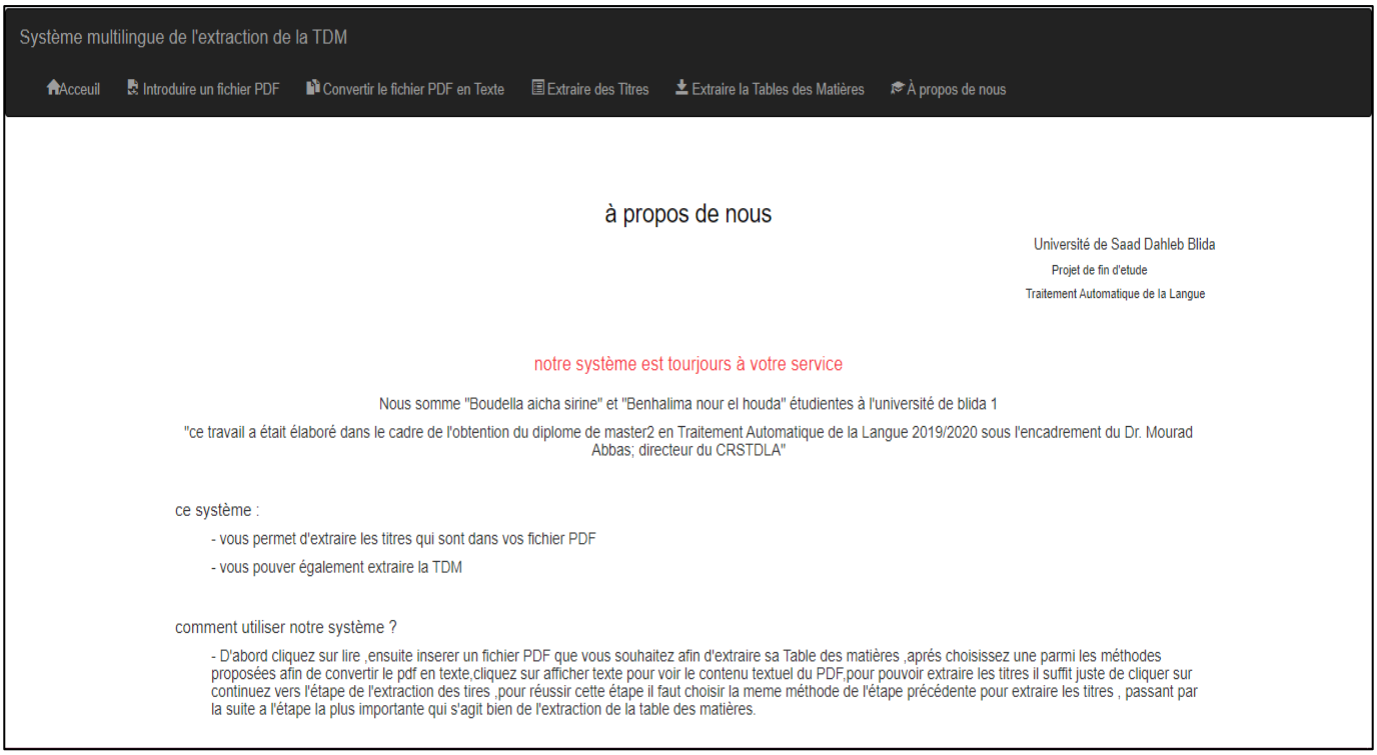


Figure 107 Page "à propos de nous"

4.6 L’évaluation du système

Dans le but d’évaluer la qualité et la performance de notre système, une étape d’évaluation est nécessaire pour prouver nos résultats obtenus, Nos expérimentations sont évaluées en comparant les résultats de l’extraction manuelle de la TDM avec les résultats de l’extraction automatique en utilisant les algorithmes fournis par la bibliothèque « textdistance ».

4.6.1 Similarité entre textes

Évaluer la similarité entre documents textuels est une des problématiques importantes de plusieurs disciplines comme l'analyse de données textuelles, la recherche d'information ou l'extraction de connaissances à partir de données textuelles (Text Mining). Dans chacun de ces domaines, les similarités sont utilisées pour différentes traitements :

- En analyse de données textuelles, les similarités sont utilisées pour la description et l'exploration de données.
- En recherche d'information, l'évaluation des similarités entre documents et requêtes est utilisée pour identifier les documents pertinents par rapport à des besoins d'information exprimés par les utilisateurs.
- En Text Mining, les similarités sont utilisées pour produire des représentations synthétiques de vastes collections de documents.

Les techniques mises en œuvre pour calculer les similarités varient bien évidemment selon les disciplines, mais elles s'intègrent cependant le plus souvent dans une même approche générale en deux temps :

1. Les documents textuels sont d'abord associés à des représentations spécifiques qui vont servir de base au calcul des similarités. Bien que la nature précise des représentations utilisées dépende fortement du domaine d'application, il faut noter que, presque dans tous les cas, les documents sont représentés sous la forme d'éléments d'un espace vectoriel de grande dimension.
2. Un modèle mathématique est choisi pour mesurer les similarités. [21]

4.6.1.1 Similarité syntaxique

En mathématiques et en informatique, une mesure permettant de comparer des documents textuels, consiste à comparer des chaines de caractères. C'est une métrique qui mesure la similarité ou la dissimilarité entre deux chaînes de caractères. Par exemple, les chaînes de caractères "Sam" et "Samuel" peuvent être considérées comme similaires (très proches). Une telle mesure sur les chaines de caractères fournit une valeur obtenue algorithmiquement. Parmi de telles mesures de similarité, citons par exemple, la distance de Levenshtein (ou distance d'édition), le coefficient de Dice, l'indice de Jaccard, la distance euclidienne, le cosinus, ... [21]

Nous présentons dans ce qui suit les mesures de similarité syntaxique les plus utilisées

4.6.1.1.1 Distance de Levenshtein

En théorie de l'information et en informatique, la distance de Levenshtein est une métrique permettant de mesurer la différence entre deux séquences (c'est-à-dire la distance de montage). La distance de Levenshtein entre deux chaînes de caractères est donnée par le nombre minimum d'opérations nécessaires pour transformer une chaîne en une autre, où une opération est une insertion, une suppression ou une substitution d'un seul caractère. Une généralisation de la distance de Levenshtein (distance Damerau-Levenshtein) permet la transposition de deux caractères comme une opération. [22]

4.6.1.1.2 Similarité cosinus

La similarité cosinus est fréquemment utilisée [Baeza-Yates and Ribeiro-Neto, 1999] en tant que mesure de ressemblance entre deux documents d1 et d2. Il s'agit de calculer le cosinus de l'angle entre les représentations vectorielles des documents à comparer. La similarité obtenue simcosinus(d1, d2) ∈ [0, 1]. [21]

$$sim_{cosinus}(d_1, d_2) = \frac{\overrightarrow{d_1} \cdot \overrightarrow{d_2}}{\|\overrightarrow{d_1}\| \|\overrightarrow{d_2}\|}$$

4.6.1.1.3 La distance euclidienne

La distance euclidienne calcule la similarité entre deux documents d1 et d2 comme la distance entre leurs représentations vectorielles ramenées à un seul point.

$$sim_{euclidienne}(d_1, d_2) = \|\overrightarrow{d_1} - \overrightarrow{d_2}\| = \sqrt{\sum_{i=1}^{n}(d_{1i} - d_{2i})^2}$$

Où n est le nombre total de termes représentés, i.e. la taille des vecteurs. [21]

4.6.1.1.4 Coefficient de Jaccard

L'indice de Jaccard ou coefficient de Jaccard [Jaccard, 1901] est le rapport entre la cardinalité (la taille) de l'intersection des ensembles considérés et la cardinalité de l'union des ensembles. Il permet d'évaluer la similarité entre les ensembles. Les documents d1 et d2 sont donc représentés, non pas comme des vecteurs, mais comme des ensembles de termes. La similarité obtenue simjaccard(d1, d2) ∈ [0, 1]. [21]

$$sim_{jaccard}(d_1, d_2) = \frac{\|d_1 \cap d_2\|}{\|d_1 \cup d_2\|}$$

4.6.1.1.5 Indice de Dice

L'indice de Dice mesure la similarité entre deux documents d1 et d2 en se basant sur le nombre de termes communs à d1 et d2.

$$sim_{dice}(d_1, d_2) = \frac{2N_c}{N_1 + N_2}$$

Où Nc est le nombre de termes communs à d1 et d2, et N1 (resp. N2) est le nombre de termes de d1 (resp. d2). [21]

4.6.1.1.6 Entropie de Shannon

L'entropie de Shannon, due à Claude Shannon, est une fonction mathématique qui, intuitivement, correspond à la quantité d'information contenue ou délivrée par une source d'information. Cette source peut être un texte écrit dans une langue

donnée, un signal électrique ou encore un fichier informatique quelconque (collection d'octets).[9]

4.6.2 Textdistance

Est une bibliothèque python pour comparer la distance entre deux séquences ou plus par de nombreux algorithmes.[10]

Il existe de nombreuses approches différentes pour comparer deux textes (chaînes de caractères). Chacun a ses propres avantages et inconvénients et n'est bonne que pour une gamme de cas d'utilisation spécifiques.

4.6.3 Calcule de similarité avant le filtrage des titres

Parmi les algorithmes utilisé pour calculer la similarité entre le contenu de chaque fichier du documents refs (TDMs extraites manuellement) et le contenu de chaque fichier du documents tocs (TDMs extraites automatiquement). On trouve :

[9] https://fr.wikipedia.org/wiki/Entropie_de_Shannon
[10] https://pypi.org/project/textdistance/

	Algorithme Corpus	Distance de Levenshtein	Similarité cosinus	Entropie de Shannon
Poppler	Mémoires M2 Anglais	4.384%	15.473%	69.927%
	Mémoires M2 Arabe	0.059%	0.088%	77.185%
	Mémoires M2 Français	4.500%	13.919%	78.778%
	Articles scientifiques	4.541%	13.221%	63.036%
	Documents financiers en anglais	1.736%	6.025%	60.561%
	Documents financiers en français	-	-	-

Tableau 10Résultats du score de la méthode poppler avant le filtrage des titres

	Algorithme Corpus	Distance de Levenshtein	Similarité cosinus	Entropie de Shannon
Pdfminer	Mémoires M2 Anglais	0.494%	2.147%	68.202%
	Mémoires M2 Arabe	0.022%	0.050%	73.410%
	Mémoires M2 Français	0.112%	0.500%	77.118%
	Articles scientifiques	17.579%	28.485%	70.572%
	Documents financiers en anglais	1.736%	6.025%	60.561%
	Documents financiers en français	-	-	-

Tableau 11 Résultats du score de la méthode Pdfminer avant le filtrage des titres

	Algorithme Corpus	Distance de Levenshtein	Similarité cosinus	Entropie de Shannon
PyPDF2	Mémoires M2 Anglais	0.092%	0.390%	62.605%
	Mémoires M2 Arabe	0.198%	1.067%	14.956%
	Mémoires M2 Français	0.005%	0.015%	61.252%
	Articles scientifiques	0.417%	2.041%	18.291%
	Documents financiers en anglais	10.809%	11.861%	30.418%
	Documents financiers en français	-	-	-

Tableau 12 Résultats du score de la méthode PyPDF2 avant le filtrage des titres

4.6.4 Calcule de similarité après le filtrage des titres

	Algorithme Corpus	Distance de Levenshtein	Similarité cosinus	Entropie de Shannon
Poppler	Mémoires M2 Anglais	0.229%	0.468%	68.833%
	Mémoires M2 Arabe	0.125%	0.611%	17.672%
	Mémoires M2 Français	0.032%	0.084%	55.809%
	Articles scientifiques	0.417%	0.589%	52.942%
	Documents financiers en anglais	4.423%	7.191%	64.321%
	Documents financiers en français	-	-	-

Tableau 13 Résultats du score de la méthode Poppler aprés le filtrage des titres

	Algorithme Corpus	Distance de Levenshtein	Similarité cosinus	Entropie de Shannon
Pdfminer	Mémoires M2 Anglais	0.229%	0.640%	66.853%
	Mémoires M2 Arabe	0.125%	0.155%	72.681%
	Mémoires M2 Français	0.032%	0.063%	53.714%
	Articles scientifiques	0.417%	2.041%	52.384%
	Documents financiers en anglais	5.698%	7.906%	57.962%
	Documents financiers en français	-	-	-

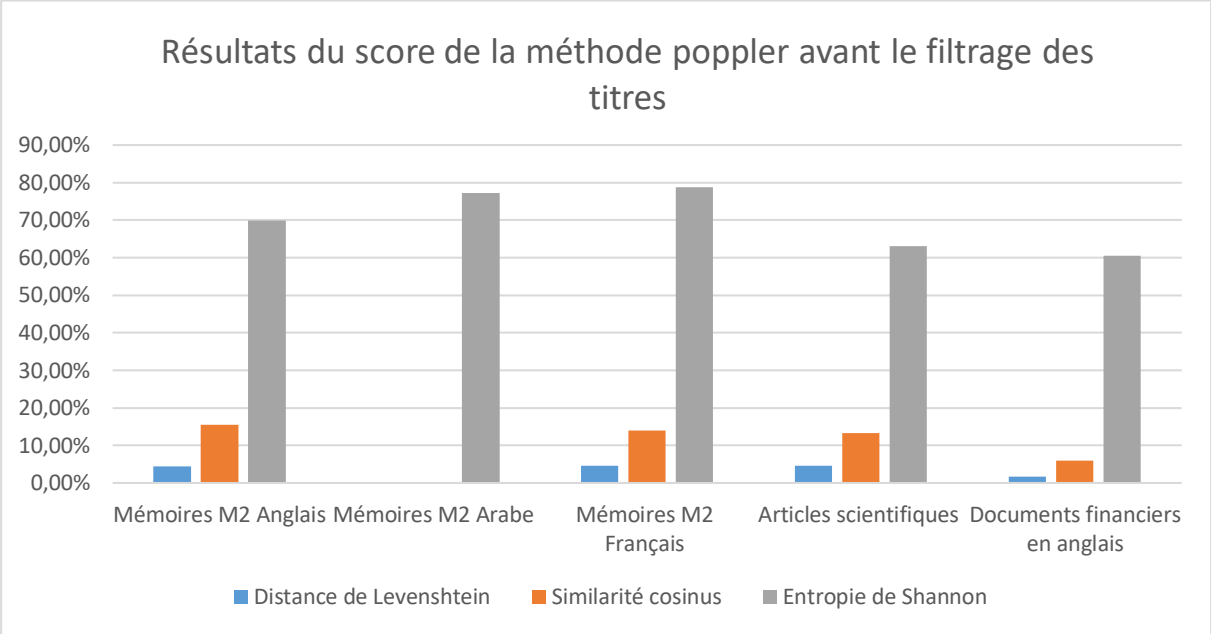
Tableau 14 Résultats du score de la méthode Pdfminer aprés le filtrage des titres

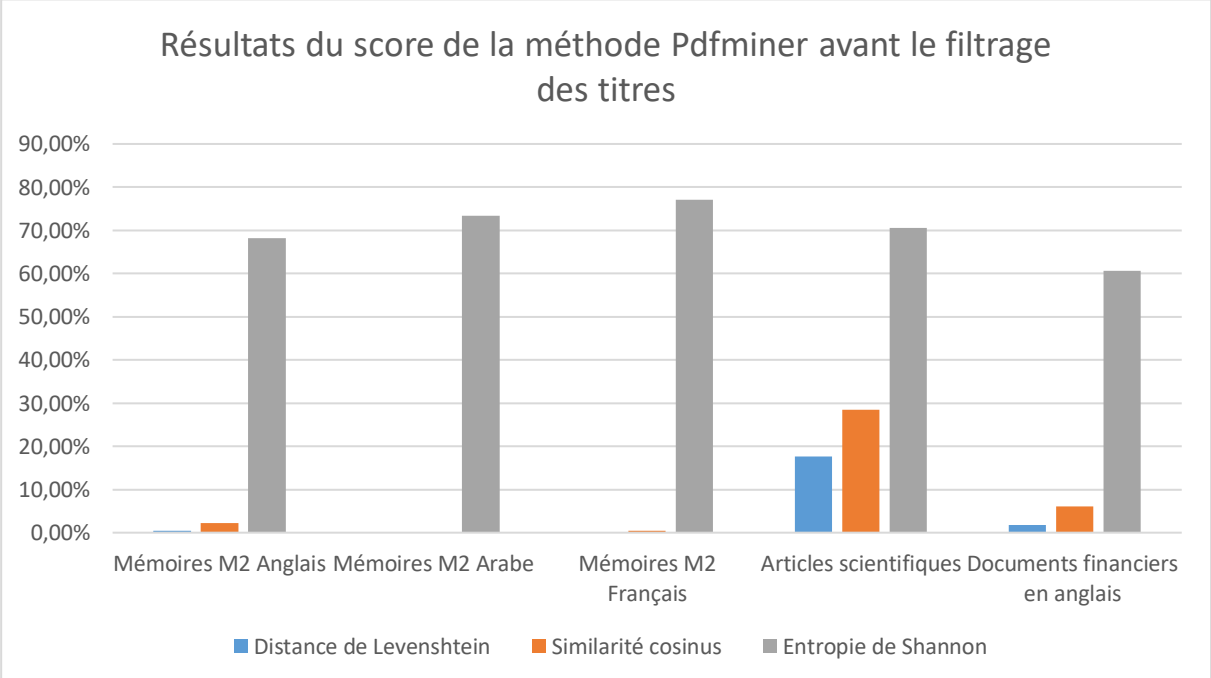
	Algorithme Corpus	Distance de Levenshtein	Similarité cosinus	Entropie de Shannon
PyPDF2	Mémoires M2 Anglais	0.229%	0.767%	64.290%
	Mémoires M2 Arabe	0.198%	1.067%	23.694%
	Mémoires M2 Français	0.042%	0.125%	51.050%
	Articles scientifiques	0.417%	2.041%	30.870%
	Documents financiers en anglais	5.750%	8.172%	50.947%
	Documents financiers en français	-	-	-

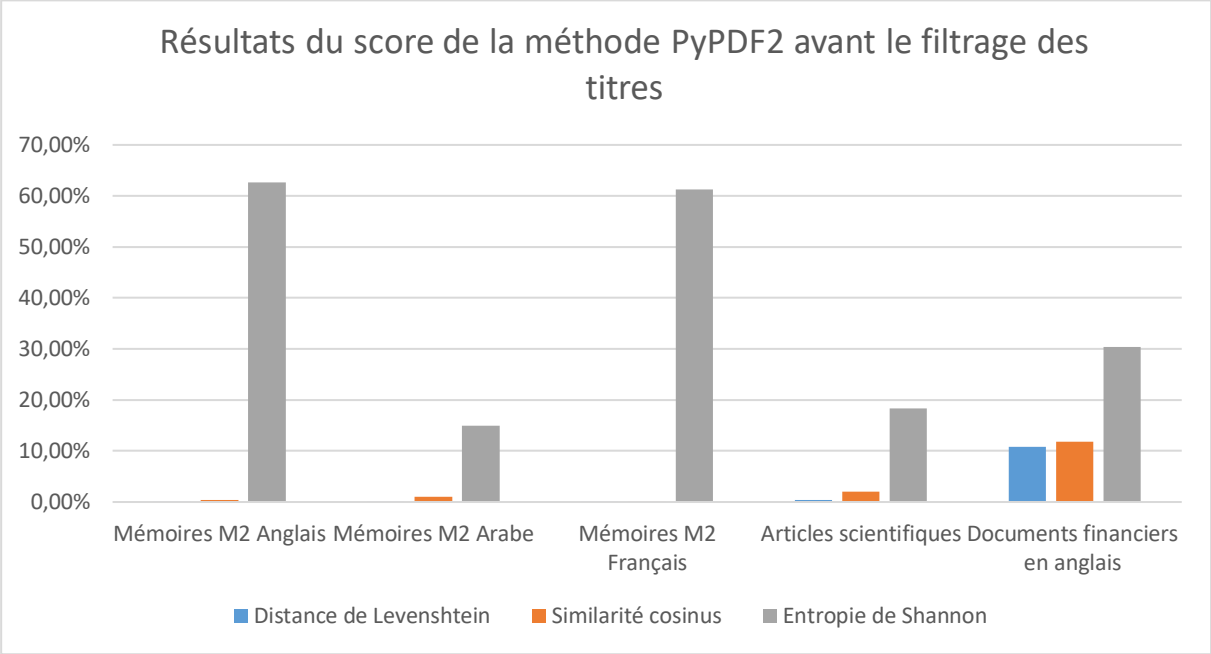
Tableau 15 Résultats du score de la méthode PyPDF2 aprés le filtrage des titres

4.7 Les histogrammes

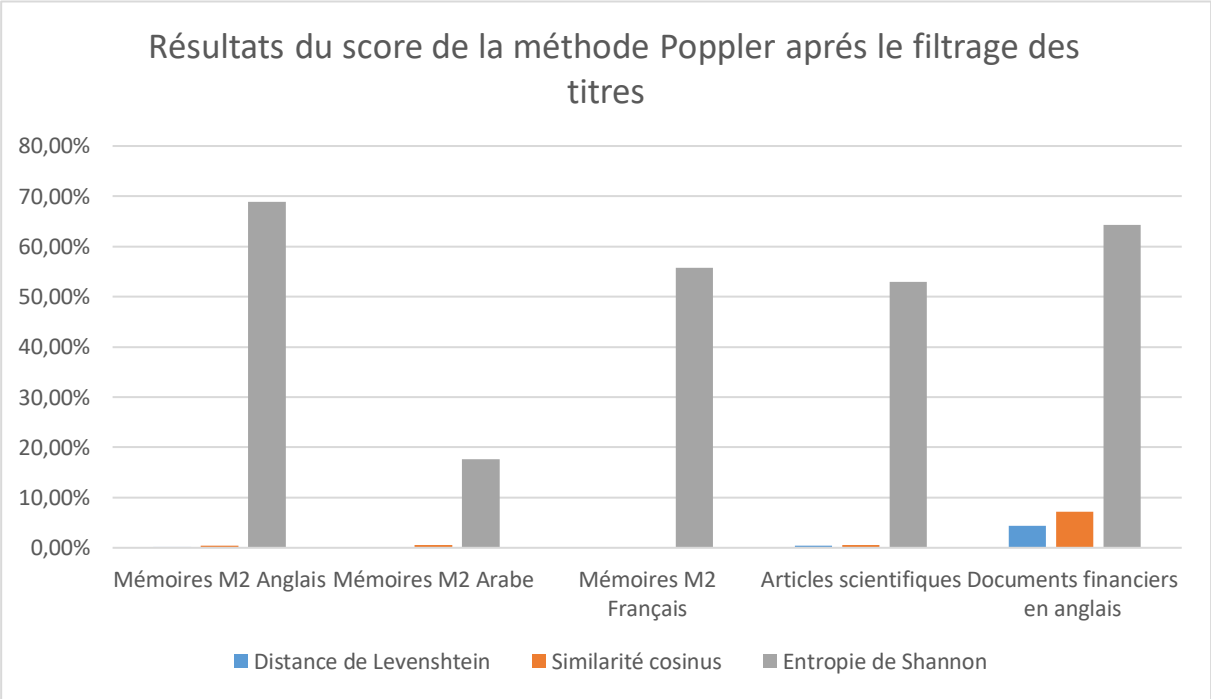
4.7.1 Avant le filtrage des titres

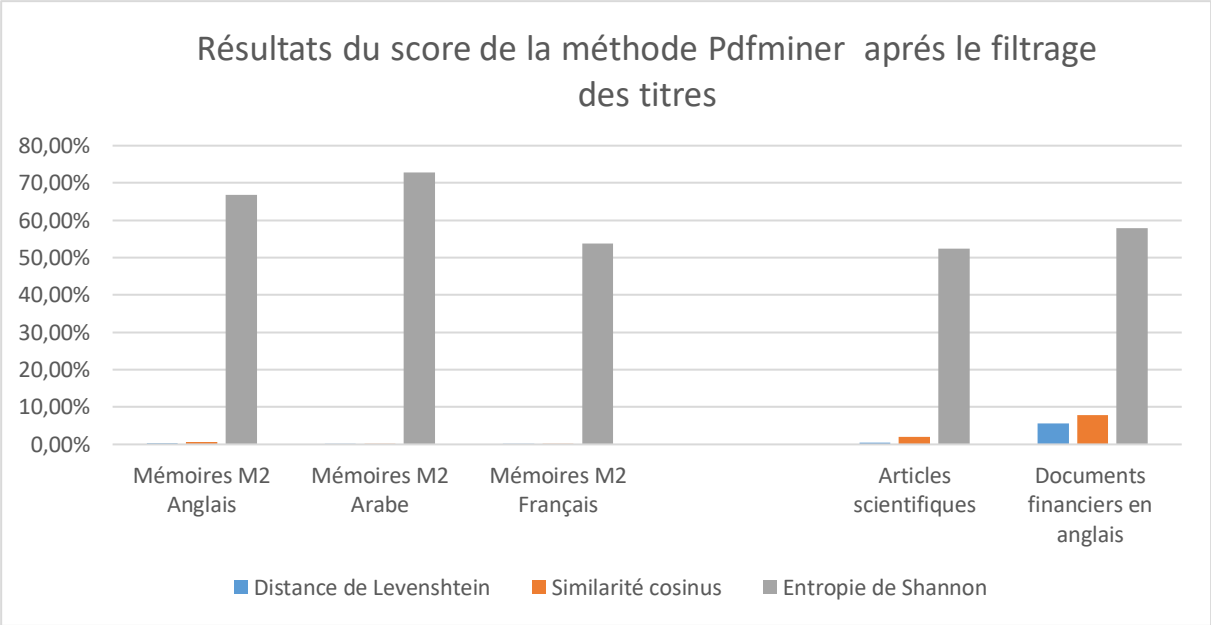


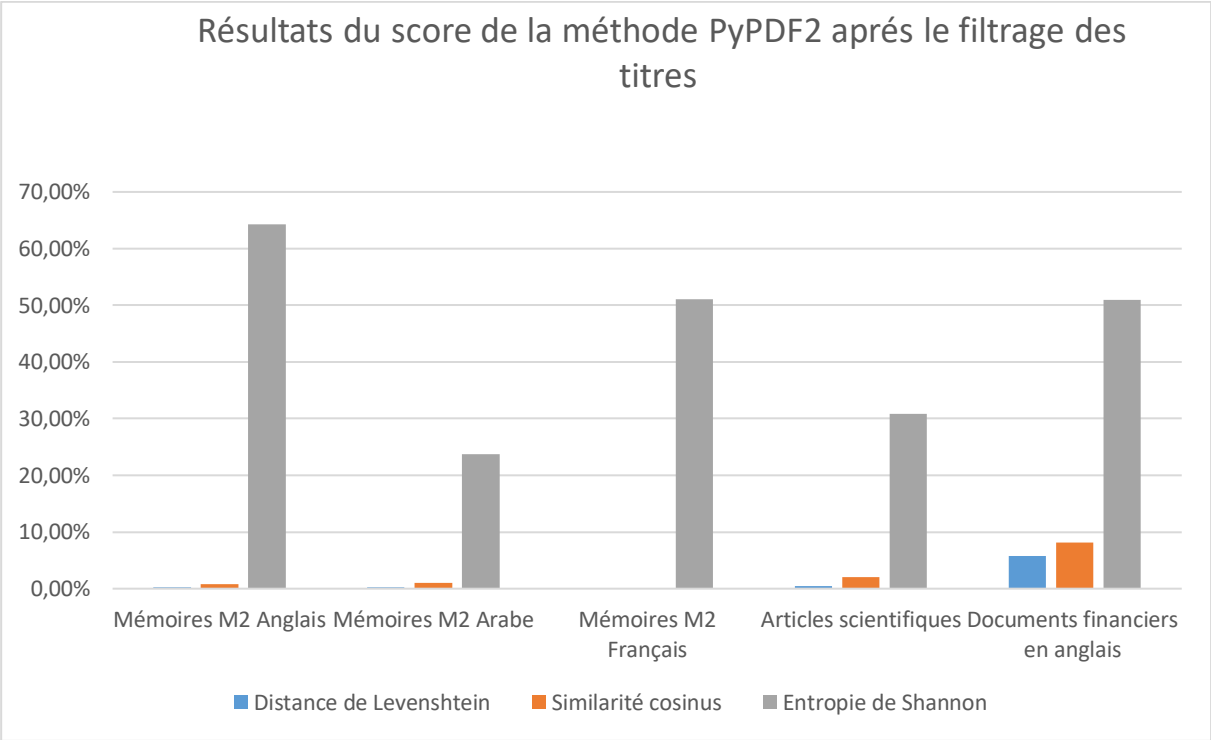




4.7.2 Après le filtrage des titres




Résultats du score de la méthode Pdfminer aprés le filtrage des titres
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Mémoires M2 Anglais
Mémoires M2 Arabe
Mémoires M2 Français
Articles scientifiques
Documents financiers en anglais
Distance de Levenshtein
Similarité cosinus
Entropie de Shannon


Résultats du score de la méthode PyPDF2 aprés le filtrage des titres
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Mémoires M2 Anglais
Mémoires M2 Arabe
Mémoires M2 Français
Articles scientifiques
Documents financiers en anglais
Distance de Levenshtein
Similarité cosinus
Entropie de Shannon

4.8 Interprétation des résultats

Après avoir appliqué les trois différents algorithmes de calcule de similarité sur nos corpus avant et après le filtrage des titres il est temps de discuter a propos des résultats obtenu afin de connaitre la meilleure technique utilisée pour l'extraction des titres et la restitution de la TDM.

4.8.1 Avant le filtrage des titres

À travers les résultats nous avons remarqué qu'ils étaient différents d'une méthode à une autre et d'un algorithme à un autre.

L'algorithme de l'Entropie de Shannon as marqué des résultats satisfaisants pour les trois méthodes. (Poppler, Pdfminer, Pypdf2), en commençant par Poppler là ou l'intervalle des résultats étaient entre 60 % et 80 %, Passons à Pdfminer d'où la valeur minimale était 60 % et la valeur maximale était 77%, Ensuite un recule au niveau des résultats concernant le modèle Pypdf2 surtout pour le corpus de la langue arabe (14,95%).

4.8.2 Après le filtrage des titres

Après avoir évalué les trois différentes méthodes nous avons observé quelques problèmes au niveau des titres afficher qui ne sont pas souvent pertinent avec les titres de la table des matières. Sur la base de ces observations nous avons cherché à résoudre les problèmes en appliquant des procédures de vérification qui nous a aidés à améliorer nos résultats, ensuite nous avons effectué l'évaluation à nouveau à l'aide de mêmes algorithmes d'évaluation.

Après la deuxième analyse effectuée nous avons observé que l'algorithme Entropie de shannon a marqué des résultats élevés par rapport aux deux autres algorithmes mais concernant le corpus de mémoire en arabe les résultats d'évaluation après le filtrage des titres étaient moins de 50 % pour les deux techniques Poppler et Pypdf2.

4.9 Evaluation des techniques par rapport aux fonctionnalités de notre système

	Conversion des PDF en texte	**Extraction des titres**	**Extraction de la TDM**
Poppler	Excellent	Bien	Très bien
Pdfminer	Bien	A bien	Bien
PyPDF2	A bien	Bien	A bien

Tableau 16 Evaluation des techniques

4.10Les avantages et les inconvénients des techniques utilisés dans notre étude

Tableau 17 Les avantages et les inconvénients des techniques

	Avantages	**Inconvénients**
Poppler	- La rapidité du temps d'exécution. - L'extraction des titres et de la TDM est généralement très bonne pour toutes les langues traitées sauf l'arabe.	- L'extraction des titres et de la TDM est généralement médiocre pour la langue arabe.
Pdfminer	- L'extraction des titres et de la TDM est généralement bonne pour	- Temps d'exécution très long. - L'analyse du contenu textuel du gauche à droite (donc pdfminer change la morphologie de la

	toutes les langues traitées sauf l'arabe.	langue arabe pendant la conversion du pdf). - La détection des titres et l'extraction de la tdm avec Pdfminer est médiocre pour les articles scientifiques.
PyPDF2	- La rapidité du temps d'exécution.	- PyPDF2 ne supporte pas la langue arabe.

4.11Conclusion

Dans ce chapitre, nous avons présentés les différents outils, langages et Framework que nous avons utilisés au cours de développement de cette application, nous avons également présenté quelque testes de notre système et sans oublier Les expérimentations qui nous ont permis de parcourir les résultats et l'évaluation de l'extraction des titres et de la TDM.

Conclusion et perspectives

L'augmentation quotidienne de l'informations disponibles dans l'internet crée le besoin d'outils capables de l'extraire et la traiter. Des sources d'information importantes sont à l'origine créées sous forme de documents textuels. Bien que stockés dans des ordinateurs, ces documents ne contiennent pas d'indication formelle sur les types de données qu'ils contiennent ou sur leur propre structure. Ce manque d'indication formelle empêche la manipulation de l'information pour répondre aux besoins spécifiques de l'utilisateur lors de l'accès, la demande, la recherche. [23]

Pour rendre ces connaissances traitables par l'ordinateur, il est nécessaire de comprendre la structure des documents, d'encoder leurs connaissances et de développer des algorithmes pour combler le fossé entre les documents texte et les représentations traitables par l'ordinateur.

La principale contribution de notre travail est le développement d'un système guidé par l'utilisateur pour la détection des titres et la restitution de la TDM automatiquement en utilisant des méthodes basées sur nos recherches, il est destiné spécifiquement aux trios de la langues anglaise, française et arabe, il nécessite un faible espace de stockage et un temps de traitement réduit. Cependant, au cours de cette tâche, nous avons rencontré quelques difficultés affectant négativement la qualité de l'extraction, mais que nous sommes parvenus à résoudre grâce à des observations, des analyses et des expérimentations.

Parmi la difficulté que nous avons rencontrée dans notre projet est la confusion entre des listes numérotées et les titres des sections et des blocs ainsi que le défi de la langue arabe car le sens de la rédaction en arabe est de droite à gauche donc la détection des titres en arabe diffère à celles écrites en anglais ou en français. Et sans oublier l'extraction des phrases qui ne sont pas des titres à la base, Dans ce cas la qualité de notre système diminue considérablement.

Afin d'obtenir de meilleurs résultats et améliorer la qualité de l'extraction de la TDM on a essayé plusieurs expériences avec différents corpus en termes de type (mémoires de master 2, documents financiers, articles scientifiques) en langues (en arabe, français et en anglais) et de taille ainsi que nous avons amélioré nos résultats

de l’extraction de la TDM en filtrant les titres jusqu'à ce que nous sommes arrivés à un système acceptable qui traite différents types de documents numérisés.

Nous considérons que nous avons pu montrer que cet objectif a été atteint avec un certain succès. Bien que certaines améliorations doivent être apportées, Néanmoins, nous estimons que le système nécessite encore une période d'expérimentation afin d'évoluer dans l’avenir (le rajout de la pagination et l’amélioration du contenu textuel extrait en arabe avec pdfminer, l’élimination des titres en arabe en double, l’utilisation d’autre techniques open source afin d’améliorer la qualité de texte obtenu, l’amélioration de l’étude pour les PDFs scannés).

L’élaboration de ce travail nous a permis, d’une part, d’approfondir les connaissances et le savoir-faire acquis durant nos cinq ans d’études, et d’autre part, de préparer notre intégration à la vie professionnelle.

Références bibliographiques

[1] «FinTOC 2020,» 1 Décembre 2019. [En ligne]. Available: http://wp.lancs.ac.uk/cfie/fintoc2020/. [Accès le 12 avril 2020].

[2] «Présentation du centre,» CRSTDLA, [En ligne]. Available: http://www.crstdla.dz/fr/?Introduction. [Accès le 08 09 2020].

[3] «FNP Workshop Series,» 2020. [En ligne]. Available: http://wp.lancs.ac.uk/cfie/fnp2020/. [Accès le 12 avril 2020].

[4] R. Juge, N.-I. Bentabet et S. Ferradans, «The FinTOC-2019 Shared Task:Financial Document Structure Extraction,» p. 1.

[5] «zamzar,» [En ligne]. Available: https://www.zamzar.com/fr/convert/pdf-to-txt/. [Accès le 03 09 2020].

[6] H. L. C. E. O. Carla Abreu1, «FinDSE@FinTOC-2019 Shared Task,» p. 69–73, 2019.

[7] F. Even, «Extraction d'Information et modélisation de connaissances à partir de Notes de Com- munication Orale.,» Université de Nantes, 2005.

[8] «La Banque de dépannage linguistique,» l'Office québécois de la langue française, [En ligne]. Available: http://bdl.oqlf.gouv.qc.ca/bdl/gabarit_bdl.asp?Th=2&t1=&id=3744. [Accès le 13 06 2020].

[9] A. d. m. financiers, «Le Document d'information clé pour l'inverstisseur,» [En ligne]. [Accès le 14 06 2020].

[10] «Autorité des Marchés financiers,» [En ligne]. Available: https://lautorite.qc.ca/grand-public/investissements/fonds/prospectus/. [Accès le 11 05 2020].

[11] «Autorité des Marchés financiers,» [En ligne]. Available: https://lautorite.qc.ca/grand-public/investissements/fonds/prospectus/dispenses-de-prospectus. [Accès le 07 09 2020].

[12] «journal du net,» [En ligne]. Available: https://www.journaldunet.fr/business/dictionnaire-economique-et-financier/1199263-statuts-definition-traduction/. [Accès le 08 09 2020].

[13] J. Wolfe, «A Brief History of Python,» 5 mars 2018. [En ligne]. Available: https://medium.com/@johnwolfe820/a-brief-history-of-python-ca2fa1f2e99e. [Accès le 11 juillet 2020].

[14] «Python Features,» [En ligne]. Available: https://www.javatpoint.com/python-features. [Accès le 12 07 2020].

[15] «Spyder,» [En ligne]. Available: https://www.spyder-ide.org/. [Accès le 25 juillet 2020].

[16] K. Relan, Building REST APIs with Flask, B. C. Apress, Éd., New Delhi, Delhi, 2019.

[17] F. A. Aslam et H. N. Mohammed, «Efficient Way Of Web Development Using Python And Flask,» *International Journal of Advanced Research in Computer Science,* vol. 6, n° %12, p. 2, 2015.

[18] A. J. A. G. P. P. S. H. Aditya Kekare, «Techniques for Detecting and Extracting Tabular Data from PDFs and Scanned Documents: A Survey,» *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, 2020.

[19] «Comment travailler avec un PDF en Python,» [En ligne]. Available: https://www.codeflow.site/fr/article/pdf-python?fbclid=IwAR3U6LgtvEpLgTJudhLa58ixa2MQsuv573fF_k8rDe30dIoIofOf2FiuVhc. [Accès le 10 10 2020].

[20] M. Balsløw, «How-To Work With Poppler Utility Library (PDF Tool),» [En ligne]. Available: https://support.foxtrotalliance.com. [Accès le 26 juillet 2020].

[21] E. Negre, «Comparaison de textes: quelques approches,» 17 10 2013. [En ligne]. Available: https://hal.archives-ouvertes.fr/hal-00874280/document. [Accès le 11 10 2020].

[22] A. F. V. J. M. Frederic P. Miller, Levenshtein Distance, V. Publishing, Éd., 2009.

[23] M. N. a. R. F., «Extracting Structure, Text and Entities from PDF Documents of the Portuguese Legislation,» pp. 123-131, 2012.

[24] P. M. e. a. Nadkarni, «Natural language processing: an introduction,» *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, pp. 544-51, september 2011.