



Mémoire de Projet de Fin d'Études

Pour l'Obtention du

Master Sciences Et Techniques en

*Informatique Et Modélisation Des Systèmes
Complexes*

Rapport sur

Développement d'une solution
intelligente pour la
reconnaissance et la détection
du texte depuis des documents

Réalisé par :

OUAYACH FATIMA ZAHRA

Encadré par :

MR. BALOUKI YOUSSEF
MR. OIHI JAMAL

ANNÉE UNIVERSITAIRE 2019-2020

Dédicaces

Je dédie ce modeste travail comme un témoignage d'affection, de respect et d'admiration :

À mes très chers parents,

Pour tout ce que vous avez fait pour moi, pour les efforts que vous avez consentis pour mon éducation et ma formation. Je ferai de mon mieux pour rester un sujet de fierté à vos yeux avec l'espoir de ne jamais vous décevoir.

A mes chères soeur-frère,

Pour votre présence dans ma vie, pour votre précieux soutien moral et matériel, pour vos encouragements continus, votre affection et vos soutiens m'ont été d'un grand secours au long de ma vie.

A mes chers Amis (es),

Pour votre amitié, pour les meilleurs souvenirs, pour les bons moments, pour l'encouragement et le soutien.

A mes enseignants,

Pour votre aide tout au long de mon cursus universitaire, je serai toujours reconnaissante.

A tous ceux que j'aime et à tous ceux qui m'aiment

Remerciement

Au nom d'Allah le tout miséricordieux, ce travail, ainsi accompli, n'aurait point pu arriver à terme, sans le guidage d'Allah.

*J'adresse mes plus vifs remerciements à Monsieur **Youssef BALOUKI**, pour avoir accepté de m'encadrer, pour l'intérêt qu'il a porté à mon sujet, pour son encadrement, pour ses remarques pertinentes ainsi que pour sa patience. Son suivi, son encadrement et ses conseils m'ont été d'un appui considérable.*

*Mes remerciements s'adressent aussi à mon encadrant de stage, Monsieur **OIHI Jamal**, pour m'avoir accordé sa confiance dans ce projet et pour toute l'aide qu'il m'a apporté.*

Je profite également pour remercier l'agence PYXICOM et leur témoigner toute ma reconnaissance pour leur accueil sympathique et chaleureux durant tous les quatre mois.

Mes vifs remerciements s'adressent à tous ceux qui m'ont aidés, de près ou de loin, à accomplir ce travail trouvent ici l'expression de mes remerciements les plus distingués.

Résumé

Ce document a pour but de décrire le déroulement de mon projet « Reconnaissance et Extraction de texte » effectué dans l'agence Pyxicom, dont l'objectif est de développer une solution intelligente pour la reconnaissance, la détection et l'extraction de texte depuis des documents.

La première partie du projet consiste à développer un système qui peut reconnaître, détecter et extraire du texte depuis les fichiers pdf, word, excel et même des images documentaires ou naturelles et enregistrer le résultat dans un fichier texte modifiable.

La deuxième partie consiste à développer un deuxième système qui a pour but l'extraction des informations personnelles des candidats ayant déposé leurs CVs dans l'agence afin de répondre à une offre d'emploi.

La dernière partie consiste à assembler les deux premières parties en une application intelligente développée par python et flask.

Abstract

This document aims to describe the progress of my project « Recognition and text Extraction » made in Pyxicom agency, whose objective is the development of an smart solution which recognises and detects and extracts text from documents.

The first part of this project consists of developing a system that can recognise, detect and extract text from pdf, word, excel files, also from documentary images and natural ones. And eventually save the result within an editable text file.

The second part of this project consists of developing a system that can extract personal informations of candidates who have submitted their resumes at the agency in order to respond to a job offer.

The last part of this project consists of combining the two first parts into an intelligent web application developed with python and flask.

Table des matières

1	CADRE GENERAL DU PROJET	14
1.1	<i>Introduction</i>	14
1.2	<i>Présentation de l'organisme d'accueil</i>	14
1.2.1	<i>Environnement de stage</i>	14
1.2.2	<i>Historique de Pyxicom</i>	15
1.2.3	<i>Activités de Pyxicom</i>	16
1.3	<i>Cahier des Charges</i>	16
1.3.1	<i>Étude préalable</i>	16
1.3.2	<i>Présentation de la problématique</i>	17
1.3.3	<i>Analyse de l'existant</i>	17
1.3.4	<i>Solution Proposée</i>	18
1.3.5	<i>Outils de développement</i>	19
1.3.6	<i>Phase de développement</i>	19
1.4	<i>Méthodologie de travail</i>	20
1.4.1	<i>Principe des méthodes agiles</i>	20
1.4.2	<i>L'agilité avec SCRUM</i>	21
1.4.3	<i>Les rôles SCRUM</i>	22
1.5	<i>Pilotage du projet</i>	22
1.5.1	<i>Planification du projet</i>	22
1.5.2	<i>Diagramme de Gantt</i>	23
1.6	<i>Conclusion</i>	24
2	ANALYSE ET CONCEPTION	26
2.1	<i>Introduction</i>	26
2.2	<i>Présentation du langage de modélisation</i>	26
2.3	<i>Diagramme de cas d'utilisation</i>	26
2.3.1	<i>Description de cas d'utilisation : Import File</i>	27
2.3.2	<i>Description de cas d'utilisation : Launch extraction</i>	28
2.4	<i>Diagramme de séquence boîte noire</i>	28
2.5	<i>Diagramme d'activité</i>	30
2.6	<i>Diagramme de classe</i>	30

2.7	<i>Conclusion</i>	31
3	PRETRAITEMENT	34
3.1	<i>Introduction</i>	34
3.2	<i>Analyse des documents</i>	34
3.2.1	<i>L'outil OCR</i>	34
3.2.2	<i>Reconnaissance de document</i>	34
3.2.3	<i>Processus de reconnaissance</i>	35
3.2.3.1	Acquisition	36
3.2.3.2	Prétraitement	36
3.2.3.3	Reconnaissance du contenu	36
3.2.3.4	Post-traitement	37
3.2.4	<i>La saisie de document</i>	37
3.2.5	<i>Identification de la langue et de la fonte</i>	37
3.2.6	<i>Processus de reconnaissance de caractères</i>	39
3.3	<i>Extraction de texte depuis des documents</i>	40
3.3.1	<i>Depuis une image de scène et une image</i>	40
3.3.1.1	East	40
3.3.1.2	East - Introduction	40
3.3.1.3	East - L'architecture du Réseau	41
3.3.1.4	East - Caractéristiques d'extraction	41
3.3.1.5	East - Branche de fusion	42
3.3.1.6	East - La couche de Sortie	42
3.3.1.7	East - La fonction de Perte	42
3.3.1.8	East - Suppression de la fusion maximale	43
3.3.1.9	Exigences - Image document / naturelle	43
3.3.2	<i>Depuis un document PDF</i>	44
3.3.2.1	Exigences - PDF	44
3.3.3	<i>Depuis un document Word</i>	44
3.3.3.1	Exigences - Word	44
3.3.4	<i>Depuis un document Excel</i>	45
3.3.4.1	Exigences - Excel	45
3.3.5	<i>Extraction des données à partir d'un CV</i>	45
3.3.5.1	Le NLP	46
3.3.5.2	Fonctionnement d'NLP	46
3.3.5.3	Données du CV	47
3.3.5.4	Exigences - CV	48
3.3.6	<i>Les Challenges et les difficultés d'extraction</i>	49
3.3.6.1	Complexité de la scène	49

3.3.6.2	Conditions d'éclairage inégal	49
3.3.6.3	La Rotation	50
3.3.6.4	Flou et dégradation	50
3.3.6.5	Environnements multilingues	50
3.3.6.6	Inclinaison (distorsion en perspective)	51
3.3.6.7	Les Fonts	51
3.3.7	<i>Les Erreurs de structuration</i>	51
3.3.7.1	Fusion	52
3.3.7.2	Fission	54
3.3.8	<i>Les Erreurs de reconnaissance de caractère</i>	55
3.3.8.1	Définition	55
3.3.8.2	Les erreurs	55
3.4	<i>Conclusion</i>	55
4	REALISATION DE LA SOLUTION INTELLIGENTE	57
4.1	<i>Introduction</i>	57
4.2	<i>Outils et Technologies utilisées</i>	57
4.2.1	<i>Environnement matériel</i>	57
4.2.1.1	PC portable LENOVO	57
4.2.2	<i>Environnement logiciel</i>	57
4.2.2.1	HTML5	57
4.2.2.2	CSS3	58
4.2.2.3	JavaScript	58
4.2.2.4	Jquery	59
4.2.2.5	Python	59
4.2.2.6	Flask	59
4.2.2.7	MySQL DataBase	60
4.3	<i>Présentation du WebSite</i>	60
4.3.1	<i>Page d'accueil</i>	60
4.3.2	<i>Extraction depuis catégorie « File »</i>	61
4.3.3	<i>Extraction depuis catégorie « CV File »</i>	61
4.3.4	<i>Manipulation des fichiers</i>	62
4.3.5	<i>Résultats</i>	62
4.4	<i>Conclusion</i>	65

Table des figures

1.1	Pyxicom Logo.	14
1.2	Exemple d'un texte structuré.	18
1.3	Exemple d'un texte non structuré.	19
1.4	Vue synthétique du processus Scrum.	21
1.5	Diagramme de Gantt.	23
2.1	Diagramme de cas d'utilisation.	27
2.2	Diagramme de séquence boîte noire.	29
2.3	Diagramme d'activité.	30
2.4	Diagramme de classe.	31
3.1	Système de reconnaissance des caractères multilingues.	35
3.2	Système de reconnaissance des caractères multilingues.	36
3.3	Différents niveaux de résolution.	38
3.4	Exemple d'application d'EAST.	40
3.5	L'architecture d'EAST.	41
3.6	L'architecture de VGG16.	42
3.7	Exemple de tokenization.	48
3.8	Ordre de lecture complexe dans un document multi-colonnes où la note de bas de page prolonge la première colonne à partir de sa moitié.	52
3.9	Erreur de segmentation - Fusion verticale de blocs. Les numéros donnent l'ordre de lecture des lignes.	53
3.10	Erreur de segmentation - Fusion horizontale.	53
3.11	Erreur de segmentation - Fission verticale.	54
3.12	Erreur de segmentation - Fission horizontale.	54
4.1	Logo de HTML5.	57
4.2	Logo de CSS3.	58
4.3	Logo de JavaScript.	58
4.4	Logo de JQuery.	59

4.5	Logo de Python.	59
4.6	Logo de Flask.	60
4.7	Logo de MySQL.	60
4.8	Page d'accueil du site web.	61
4.9	Extraction depuis « File ».	61
4.10	Extraction depuis « CV File ».	62
4.11	Aucun fichier sélectionné.	62
4.12	Fichier en cours d'extraction.	62
4.13	Enregistrement en tant que texte.	63
4.14	Exemple d'enregistrement en tant que texte.	63
4.15	Enregistrement en tant que csv.	63
4.16	Exemple d'enregistrement en csv.	64
4.17	Enregistrement en tant que sql database.	64
4.18	cv informatique dans le shell.	64

Abréviations

PME	Petites et Moyennes Entreprises
PMI	Petites et Moyennes Industries
OCR	Optical Character Recognition
EAST	Efficient Accurate Scene Text Detector
VGG16	Visual Geometry Group
PDF	Portable Document Format
FPS	Frames Per Second
CV	Curriculum Vitæ
NLP	Natural Language Processing
CSV	Comma Separated Values
UML	Unified Modeling Language

Introduction Générale

Pour acquérir une bonne et parfaite qualité, la formation théorique à elle seule ne suffit pas, il est donc nécessaire de suivre une démarche réelle permettant de voir comment se déroulent les tâches dans la vie professionnelle. A cet effet, et afin de valider mes études acquises au fil de cinq ans au sein de Faculté des Sciences et Techniques de SETTAT et en vue de l'obtention d'un Master Sciences et Techniques spécialisé en Informatique et Modélisation de Système Complexe, je suis amenée à effectuer un stage au siège de l'agence Pyxicom à Rabat d'une durée de 4 mois.

Le stage est considéré comme une occasion qui m'a permis le contact direct avec le marché du travail, il m'a aidé à renforcer la théorie par la pratique. Le sujet qui m'a été confié se résume comme suit : Le développement d'une solution intelligente qui a pour but la détection d'un texte dans des documents.

Offrir un service qui permet au client la reconnaissance automatisée de textes imprimés, et va être encore capable de "lire" le contenu d'un document. Un service qui permettra de dématérialiser n'importe quel document non modifiable en fichier bureautique. Cela permet ensuite de pouvoir les modifier et les exploiter. Le pouvoir même d'identifier une écriture manuscrite ou un texte dans un document image ou vidéo. C'est une question de gain de temps qui va justifier d'éviter la double saisie des informations et les erreurs de frappe.

Vu aussi que les entreprises organisent généralement des recrutements en sélectionnant les candidats parmi ceux qui ont déposé un dossier physique, ou électronique à travers le site internet de l'entreprise ou encore la messagerie électronique. Ces dossiers sont composés entre autres des demandes d'emploi, des CVs. Une des tâches de la direction des ressources humaines est alors d'identifier les compétences détenues par les candidats à la lecture de leur CV. Un service d'extraction automatique des compétences, d'email et du numéro du téléphone en plus de la photo du candidat va permettre de gagner le temps.

Certes, le bon fonctionnement de cette solution et le respect du cahier de charges sont très importants, la sécurisation de cette dernière est d'une importance majeure. Pour cela, elle a été prise en considération tout au long de réalisation. Ceci dit notre travail se devise en trois chapitres. Après avoir présenté dans le premier chapitre; l'agence Pyxicom, son historique, ainsi que la description du projet, la problématique traitée et le cahier de charge.

Le deuxième chapitre va présenter les différentes étapes de l'analyse et prétraitement.

Enfin, le troisième chapitre va présenter les différents outils utilisés lors de sa réalisation ainsi que les différentes interfaces de la solution intelligente.

CHAPITRE 1 : CADRE GENERAL DU PROJET

Chapitre 1

CADRE GENERAL DU PROJET

1.1 *Introduction*

Ce premier chapitre a pour objectif :

- D'une part, de présenter le lieu de stage : Pyxicom qui m'a accueilli comme stagiaire.
- D'autre part, de présenter notre problématique de stage qui consiste à proposer une solution intelligente pour Pyxicom. Pour cela nous allons détailler les besoins fonctionnels de cette dernière en se basant sur le cahier des charges.

1.2 *Présentation de l'organisme d'accueil*

1.2.1 *Environnement de stage*



FIGURE 1.1 – Pyxicom Logo.

Le développement réalisé tout au long de ces 20 années aussi bien au niveau des investissements qu'au niveau de performances est le signe de la bonne santé de l'agence.

L'essor escompté à travers les réalisations futures et la taille projetée feront certainement de Pyxicom l'agence la plus importante dans son secteur d'activité et par là-même le modèle à suivre.

Les autres activités de Pyxicom connaissent également des niveaux de performances assez remarquables, chacune dans son domaine, maniant habilement l'équilibre entre l'ambition, la sérénité, la sécurité et surtout l'enthousiasme.

Ces évolutions n'ont certes pu se concrétiser sans la volonté inébranlable du top management qui sait surmonter les difficultés tout en conjuguant les compétences individuelles et collectives.

Conscient des défis majeurs à venir, le management a toujours privilégié la mise en place et l'application des valeurs de bonne gestion : responsabilité et réalisation de soi ; travail d'équipe et solidarité ; enthousiasme et satisfaction des clients ; profitabilité et respect de l'environnement.

Autant de valeurs qui ont milité pour la réussite de leurs projets, autant de valeurs qui renforcent Pyxicom.

1.2.2 *Historique de Pyxicom*

Créée en Avril 2000, Pyxicom est née de la volonté de Patrick Raynaud d'intervenir en tant que conseil auprès des entreprises Marocaines. Si Pyxicom a progressivement intégré la plupart des composants métiers liés au développement d'applications Internet et multimédia, le Conseil reste le préalable au développement de ses relations avec ses clients et partenaires.

Désormais solidement installée au Maroc en ayant su gagner et conserver la confiance de grands comptes et de nombreuses PME-PMI de premiers plans, Pyxicom s'est développée aujourd'hui à l'international (projets Off Shore, délégation de personnel).

Basée à Rabat, Pyxicom compte aujourd'hui dans son équipe plus de 100 collaborateurs.

Le capital social est passé à 543 400 DH en juin 2011.

Pyxicom est connu par son travail, auto-critique et esprit d'équipe.

Apporter des idées, faire utile et mettre en place une démarche progressive à la mesure de vos ambitions et de vos moyens sont les approches de Pyxicom.

1.2.3 *Activités de Pyxicom*

- Maintenance informatique :
 - *Maintenance de logiciels*
- Audit et conseil informatiques :
 - *Conseil en systèmes informatiques*
 - *Conseil en logiciels*
 - *Conseil en réseaux informatiques*
 - *Conseil en sécurité informatique*
 - *Services de technologies de l'information*
- Ingénierie informatique (SSII/ESN) :
 - *Conception de moteurs de recherche*
 - *Développement de logiciels libres (open source)*
- Sécurité informatique :
 - *Services de sauvegarde des données informatiques*
 - *Services de certification pour transactions sur Internet*
- Services Internet :
 - *Services de conception de sites Internet*
 - *Services de référencement de sites Internet*
 - *Production de catalogues électroniques (catalogues en ligne)*
- Production de constructions mécaniques :
 - *Services de fabrication assistée par ordinateur (FAO)*

1.3 *Cahier des Charges*

1.3.1 Étude préalable

Cette section a pour objectif l'étude et faire le tour parmi les solutions existantes sur le marché, elle permet de dégager les points forts et les points faibles de chacune de ces solutions.

1.3.2 Présentation de la problématique

De nombreux environnements de bureau, un temps considérable est consacré à des tâches inutiles telles que la saisie de données ou la recherche dans des piles de documents et de fichiers pour récupérer les informations nécessaires à l'exécution d'une tâche.

Dans cette ère de numérisation, de stockage, d'édition, d'indexation et de recherche d'informations dans un document numérique, il est beaucoup plus facile que de passer des heures à parcourir les documents imprimés / manuscrits / dactylographiés.

De plus, la recherche de quelque chose dans un grand document non numérique prend non seulement du temps, mais aussi, il est probable que nous manquions les informations lors du défilement manuel du document.

C'est pour cela qu'avec une grande partie de nos vies informatisées, il est extrêmement important que les machines et les humains puissent se comprendre et transmettre des informations dans les deux sens. La plupart du temps, les ordinateurs ont leur propre chemin - nous devons leur «parler» par le biais d'appareils relativement grossiers tels que les claviers et les souris afin qu'ils puissent comprendre ce que nous voulons qu'ils fassent. Mais quand il s'agit de traiter des informations plus humaines, comme un livre imprimé à l'ancienne ou une lettre griffonnée avec un stylo plume, les ordinateurs doivent travailler beaucoup plus dur. Et c'est heureusement pour nous que ces derniers s'améliorent de jour en jour pour accomplir les tâches que les humains pensaient être les seuls à pouvoir faire, souvent aussi mieux que nous.

1.3.3 Analyse de l'existant

Comme de nombreuses entreprises, notamment des institutions financières, les Capitaux disposent de milliers de documents à traiter, analyser et transformer afin de mener à bien ses opérations quotidiennes. Les exemples peuvent inclure des reçus, des factures, des formulaires, des relevés, des contrats et bien d'autres éléments de données non structurées, et il est important de pouvoir comprendre rapidement les informations intégrées dans des données non structurées telles que celles-ci.

L'intérêt porté par les applications manuscrites industrielles telles que le tri postal, la reconnaissance des montants de chèques, ou l'analyse de formulaires a favorisé l'affermissement des méthodes de reconnaissance, relevant des défis de plus en plus difficiles : écriture non contrainte par la forme du support et du

scripteur, utilisation d'un vocabulaire de plus en plus large, reconnaissance multi fonte, etc.

1.3.4 Solution Proposée

Pour résoudre notre problématique, heureusement les récents progrès de la vision par ordinateur nous permettent de faire de grands progrès pour alléger la charge de l'analyse et de la compréhension des documents.

C'est là que la reconnaissance optique de caractères entre en jeu. Programme permettant l'analyse automatique du texte imprimé et le transformer en une forme qu'un ordinateur peut traiter plus facilement. Les programmes d'analyse de l'écriture manuscrite sur les téléphones portables aux gigantesques machines de tri du courrier qui garantissent que tous ces millions de lettres parviennent à destination.

De nombreuses implémentations étaient disponibles avant même le boom de l'apprentissage en profondeur en 2012. Il reste un vrai challenge concernant en particulier les images de texte qui sont prises dans un environnement sans contrainte.

Arrière-plans principalement complexes, bruit, foudre, polices différentes et distorsions géométriques dans l'image. C'est dans de telles situations que les outils d'apprentissage automatique brillent.

Les défis du problème proviennent principalement de l'attribut des tâches en cours. On peut généralement diviser ces tâches en deux catégories :

- **Texte structuré** : Texte dans un document dactylographié. Dans un arrière-plan standard, une ligne appropriée, une police standard et surtout dense.

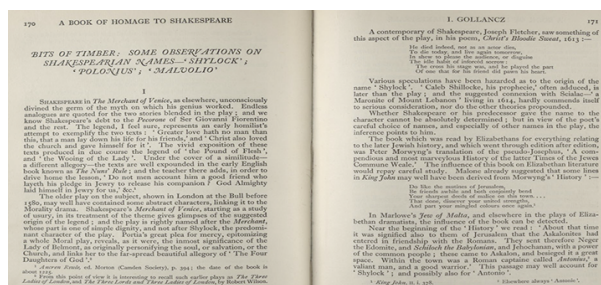


FIGURE 1.2 – Exemple d'un texte structuré.

- **Texte non structuré** : Texte à des endroits aléatoires dans une scène naturelle. Texte clairsemé, pas de structure de ligne appropriée, arrière-plan

complexe, à un endroit aléatoire dans l'image et pas de police standard.



FIGURE 1.3 – Exemple d'un texte non structuré.

Les systèmes réalisés permettront donc de répondre aux besoins de l'entreprise, ils doivent donc satisfaire à une solution intelligente qui a pour but de :

- Détecter le texte dans un document (Word, PDF, Excel)
- Détecter le texte dans une image
- Détecter le texte dans une image de scène.

C'était aussi décidé de ma part d'ajouter quelques fonctionnalités :

- Extraction des données d'un CV (prénom,email, numéro de téléphone, compétences techniques,etc) et les stocker dans un fichier csv ou dans une base de données MySQL selon l'offre demandée.

J'ai pris cette décision pour la dernière option vu que plusieurs candidats envoient leur demande d'emploi ou demande de stage aux entreprises. Alors c'est une occasion pour les sociétés de filtrer ces quantités d'informations et se contenter uniquement de ce qui est important et ceux le prénom,l'email, le numéro de téléphone ainsi que d'autres données importantes concernant le candidat.

1.3.5 Outils de développement

- HTML5 / CSS3 / JS.
- Python.

1.3.6 Phase de développement

- Sprint 1 : Application online.
- Sprint 2 : Application desktop sur Windows.

1.4 *Méthodologie de travail*

La gestion de projets joue un rôle crucial dans l'exécution et l'accomplissement des objectifs. Souvent, les entreprises ont des attentes énormes de leurs projets, mais en réalité les choses ne se réalisent pas comme prévu.

La méthodologie de gestion de projet nous aide à travers chaque étape d'un projet. Cela commence par nous aider à planifier, initier et mettre en oeuvre le projet. Les méthodologies facilitent même la clôture du projet. On peut utiliser ces modèles pour planifier nos tâches et atteindre nos objectifs.

Voici 5 méthodes de gestion de projet à reconnaître :

- La méthode en cascade.
- La méthode en cycle V.
- La méthode Prince2.
- La méthode Pert.
- Les méthodes agiles.

1.4.1 Principe des méthodes agiles

A l'origine des méthodes agiles il y a les 12 principes, et il est important de ne pas les perdre de vue.

- Prioriser la satisfaction du client.
- Accepter les changements.
- Livrer en performance des versions opérationnelles de l'application.
- Assurer le plus souvent possible une coopération entre l'équipe du projet et les gens du métier.
- Construire les projets autour de personnes motivées.
- Favoriser le dialogue direct.
- Mesurer l'avancement du projet en fonction de l'opérationnalité du produit.
- Adopter un rythme constant et soutenable par tous les intervenants du projet.
- Contrôler continuellement l'excellence de la conception et la bonne qualité technique.
- Privilégier la simplicité en évitant le travail inutile.
- Auto-organiser et responsabiliser les équipes.
- Améliorer régulièrement l'efficacité de l'équipe en ajustant son comportement.

1.4.2 L'agilité avec SCRUM

Scrum est une méthode de développement agile orientée projet informatique dont les ressources sont régulièrement actualisées.

La méthode Scrum tire son nom du monde du rugby, scrum = mêlée. Le principe de base étant d'être toujours prêt à réorienter le projet au fil de son avancement.

C'est une approche dynamique et participative de la conduite du projet. La mêlée est une phase de jeu essentielle au rugby. Elle permet au jeu de repartir sur d'autres bases. La réunion dans la méthode Scrum relaie la métaphore. Comme

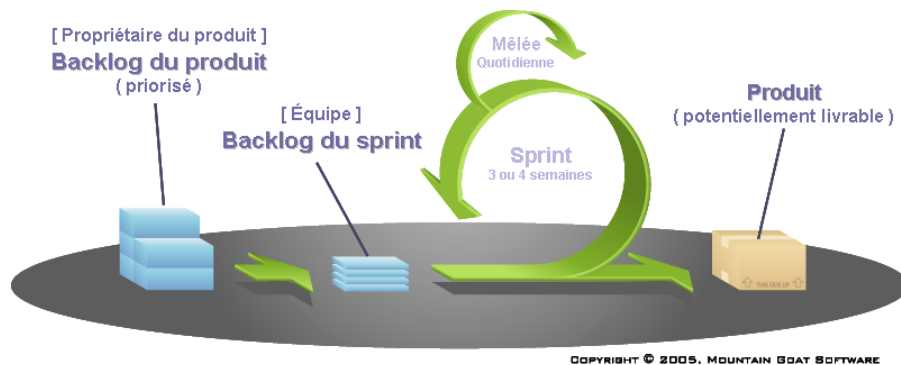


FIGURE 1.4 – Vue synthétique du processus Scrum.

le montre la figure ci-dessus, SCRUM s'impose comme un modèle d'efficacité, pronant l'expression : "aller à l'essentiel". Voici donc ce qu'il faut pour mettre en place correctement une méthodologie SCRUM efficace :

- Equipe responsable, en auto-organisation : l'équipe de développement se doit d'être autonome pour palier aux éventuels changements d'un client trop indécis.
- Avancement du produit par une série de « sprints » : les itérations sont courtes, 2 à 4 semaines au plus, pour permettre des interventions rapides en cas de problèmes.
- Exigences définies : en fait d'exigences, on parlera surtout d'éléments à mettre en oeuvre. Ces éléments sont regroupés sous l'appellation de "Backlog du produit".
- Pas de prescription de pratiques d'ingénierie : on reste dans un esprit d'autonomie de développement.
- Utilisation de règles génériques : on permet l'établissement d'un environnement agile propre au projet.

1.4.3 Les rôles SCRUM

Un équipage ne saurait mener à bien son navire sans que les rôles de chacun soient parfaitement définis. Il ne peut en effet y avoir plusieurs navigateurs, si tout le monde s'occupe des voiles et personne de la barre, l'embarcation n'ira pas loin et les risques de naufrage ne feront qu'augmenter.

Scrum définit donc trois rôles particuliers, chacun ayant une fonction bien précise.

- Le Product Owner qui porte la vision du produit à réaliser et travaille en interaction avec l'équipe de développement. Il s'agit généralement d'un expert du domaine métier du projet.
- L'Equipe de Développement qui est chargée de transformer les besoins exprimés par le Product Owner en fonctionnalités utilisables. Elle est pluridisciplinaire et peut donc encapsuler d'autres rôles tels que développeur, architecte logiciel, analyste fonctionnel, graphiste/ergonome, ingénieur système.
- Le Scrum Master qui doit maîtriser Scrum et s'assurer que ce dernier est correctement appliqué. Il a donc un rôle de coach à la fois auprès du Product Owner et auprès de l'équipe de développement. Il doit donc faire preuve de pédagogie. Il est également chargé de s'assurer que l'équipe de développement est pleinement productive. Généralement le candidat tout trouvé au rôle de Scrum Master est le chef de projet. Celui ci devra cependant renoncer au style de management « commander et contrôler » pour adopter un mode de management participatif.

1.5 *Pilotage du projet*

La clé principale de la réussite d'un projet est un bon planning. En effet, le planning aide à bien subdiviser le travail et séparer les tâches à réaliser, il offre une meilleure estimation et gestion de temps nécessaire pour chaque tâche. De plus, il donne assez de visibilité permettant d'estimer approximativement les dates d'achèvement des tâches.

1.5.1 Planification du projet

La planification de projet correspond à l'organisation des tâches à réaliser sur une période donnée. L'objectif de la planification est de déterminer le coût, les ressources mobilisées et la meilleure manière d'ordonnancer toutes les tâches à effectuer. D'avoir une vision claire de son projet et de le réaliser dans un minimum de temps en bref. Ces prévisions sont indispensables à une gestion de

projet efficace. Le planning élaboré se divise en quatre parties importantes :

- **Formation** : Cette phase contient la formation sur le framework de développement web en python Flask, ainsi que les formations en ligne et aussi les installations nécessaires pour le déroulement du stage.
- **Montée en compétences** : Cette phase a permis de se familiariser avec l'environnement de travail et de maîtriser les différents outils, résoudre des problématiques et aussi travailler sur des tâches qui ont été attribuées par nos encadrants.
- **Travail sur le projet** : Cette partie comprend quatre étapes :
 - Etude de l'existant.
 - Prétraitement et Processus.
 - Conception du projet
 - Développement du projet.
- **Rédaction du rapport** : Cette partie était intrinsèque aux autres parties.

1.5.2 Diagramme de Gantt

Dans notre projet, nous avons estimé de finaliser notre solution dans une durée approximative de 4 mois et 20 jours, comme le montre le diagramme suivant.



FIGURE 1.5 – Diagramme de Gantt.

1.6 *Conclusion*

Ce chapitre a été consacré à la présentation de l'organisme d'accueil, ensuite à la description de la problématique, puis ; il était question de définir les différents objectifs de l'application en faisant appel au cahier de charge. Enfin, la planification et le pilotage du projet avec la méthode SCRUM.

Dans le prochain chapitre, il s'agira de faire une étude conceptuelle du projet

CHAPITRE 2 : ANALYSE ET CONCEPTION

Chapitre 2

ANALYSE ET CONCEPTION

2.1 *Introduction*

Ce deuxième chapitre consiste à faire une étude conceptuelle. La phase de conception vient de répondre à la question "comment réaliser la solution?". Donc, il s'agit bien de la réponse aux exigences demandées et cela à travers une définition aux acteurs de l'application ainsi que les différents diagrammes d'UML élaborés afin de modéliser le système à mettre en place.

2.2 *Présentation du langage de modélisation*

Le langage UML ou langage de modélisation unifié a été pensé pour être un langage de modélisation visuelle commun, et riche sémantiquement et syntaxiquement. Il est destiné à l'architecture, la conception et la mise en œuvre de systèmes logiciels complexes par leur structure aussi bien que leur comportement. L'UML a des applications qui vont au-delà du développement logiciel, notamment pour les flux de processus dans l'industrie.

2.3 *Diagramme de cas d'utilisation*

Un diagramme de cas d'utilisation capture le comportement d'un système, d'un sous-système, d'une classe ou d'un composant tel qu'un utilisateur extérieur le voit. Il scinde la fonctionnalité du système en unités cohérentes, les cas d'utilisation, ayant un sens pour les acteurs. Les cas d'utilisation permettent d'exprimer le besoin des utilisateurs d'un système, ils sont donc une vision orientée utilisateur de ce besoin au contraire d'une vision informatique.

Le diagramme de cas d'utilisation de notre solution est présenté ci-dessous, l'acteur principal est représenté par le bonhomme à gauche. Il peut effectuer des actions spécifiques et bien définies dans le système, cet acteur est l'utilisateur de l'application.

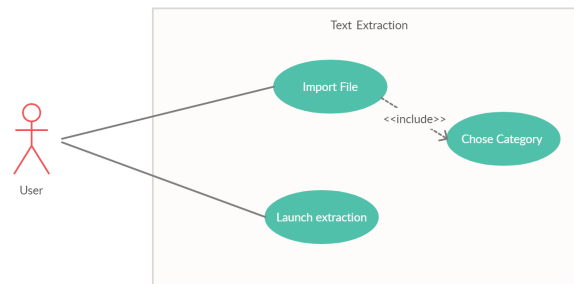


FIGURE 2.1 – Diagramme de cas d'utilisation.

2.3.1 Description de cas d'utilisation : Import File

Titre	Import File
Acteur(s)	Utilisateur
Description	Importer le fichier selon son type afin de lancer l'extraction.
Pré-Condition	Choisir la catégorie souhaitée.
Post-Condition	Lancer l'extraction.
Scénario normal	L'utilisateur choisit la catégorie souhaitée. Si la catégorie choisie est "File", il choisira le type de fichier, importera son fichier choisi. Si la catégorie choisie est "CV File", il importera un ou plusieurs fichiers CV en format pdf, choisira l'option de sauvegarde et sélectionnera le type d'offre souhaité.

2.3.2 Description de cas d'utilisation : Launch extraction

Titre	Launch extraction
Acteur(s)	Utilisateur
Description	Lancer le programme de reconnaissance, détection et extraction de texte.
Pré-Condition	Importer le fichier souhaitée.
Post-Condition	Téléchargement du fichier résultant.
Scénario normal	L'extraction est terminée et le fichier résultant est téléchargé.
Scénario d'exception	Erreur d'extraction et le fichier n'est pas téléchargé.

2.4 Diagramme de séquence boîte noire

Un diagramme de séquence est un diagramme d'interaction dont le but est de décrire comment les objets collaborent au cours du temps et quelles responsabilités ils assument. Il décrit un scénario d'un cas d'utilisation.

Un diagramme de séquence représente donc les interactions entre objets, en insistant sur la chronologie des envois de message. C'est un diagramme qui représente la structure dynamique d'un système car il utilise une représentation temporelle. Les objets, intervenant dans l'interaction, sont matérialisés par une « ligne de vie », et les messages échangés au cours du temps sont mentionnés sous une forme textuelle. La figure suivante présente le diagramme de séquence boîte noire du projet. Ce dernier met en évidence l'ensemble des interactions entre l'utilisateur et le système.

L'utilisateur dans un premier temps accède à l'application et choisit une catégorie, soit "File" ou bien "CV File".

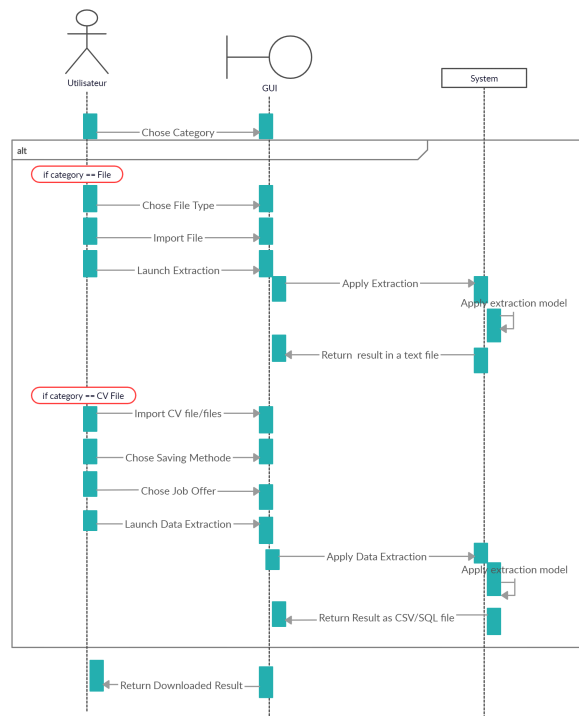


FIGURE 2.2 – Diagramme de séquence boîte noire.

Il est amené ensuite à choisir le type de fichier qui lui convient (image document, image naturelle, pdf, word, excel) si la catégorie était "File". Il doit encore importer le fichier et lancer l'exécution qui va charger le modèle et un fichier texte modifiable sera téléchargé.

Par contre, si la catégorie choisie était celle de "CV File", il sera amené à importer un ou plusieurs fichiers cv sous format pdf. Il choisira ensuite la méthode de sauvegarde soit en format csv ou bien dans une base de données MySQL. Il doit encore sélectionner le type d'offre d'emploi souhaité afin de filtrer les données extraites à partir du cv pour se contenter juste aux candidats répondant au type de cette offre. Après avoir lancé l'extraction, le modèle est chargé et un fichier sous format sql ou csv est téléchargé ayant les données filtrées des candidats.

2.5 *Diagramme d'activité*

Un diagramme d'activités (activités et transitions) est une variante du diagramme d'états-transitions (états et transitions). Les deux types de diagrammes permettent d'avoir deux vues différentes sur des automates donnés.

Un diagramme d'activités visualise un graphe d'activités qui modélise le comportement interne d'une méthode, ou plus généralement d'un processus impliquant un ou plusieurs classificateurs.

Un diagramme d'activités représente l'état d'exécution d'un mécanisme, sous la forme d'un déroulement d'étapes regroupées séquentiellement dans des branches parallèles de flots de contrôle. Il ne représente ni la collaboration ni le comportement des objets. Il est utile pour la représentation des processus métiers et les cas d'utilisation.

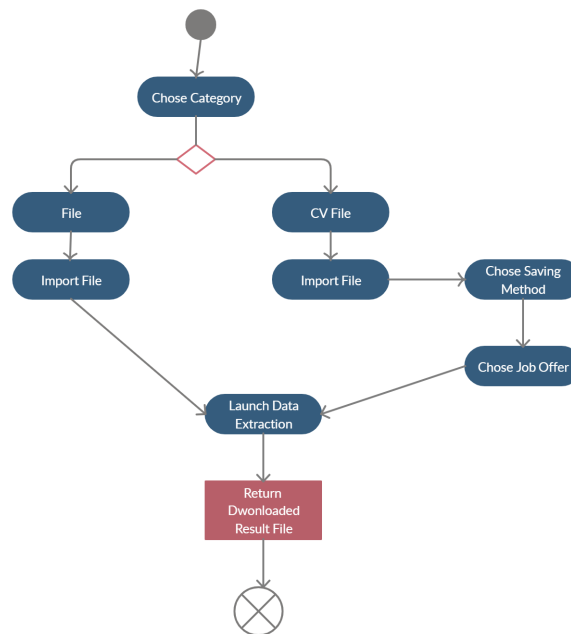


FIGURE 2.3 – Diagramme d'activité.

2.6 *Diagramme de classe*

Le diagramme de classes est considéré comme le plus important de la modélisation orientée objet, il est le seul obligatoire lors d'une telle modélisation.

Alors que le diagramme de cas d'utilisation montre un système du point de vue des acteurs, le diagramme de classes en montre la structure interne. Il permet de fournir une représentation abstraite des objets du système qui vont interagir pour réaliser les cas d'utilisation. Il est important de noter qu'un même objet peut très bien intervenir dans la réalisation de plusieurs cas d'utilisation. Les cas d'utilisation ne réalisent donc pas une partition des classes du diagramme de classes. Un diagramme de classes n'est donc pas adapté (sauf cas particulier) pour détailler, décomposer, ou illustrer la réalisation d'un cas d'utilisation particulier. Pour notre application, les principales classes qui jouent un rôle

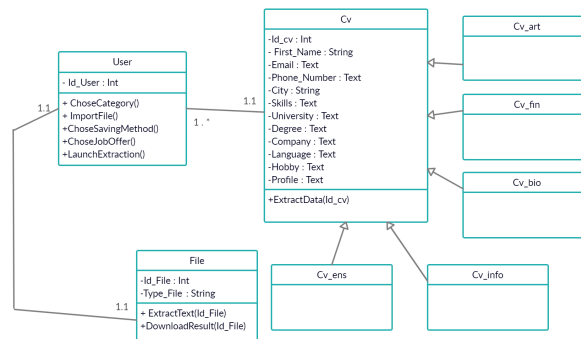


FIGURE 2.4 – Diagramme de classe.

sont Utilisateur, File et CV. Le diagramme de la figure ci-dessus décrit le diagramme de classe utilisé dans cette application, on voit que les deux classes Utilisateur et File et même Utilisateur et CV font relation. Les classes Cvart (CV d'art), Cvfin (CV de finance/commerce), Cvbio (CV de biologie/santé), Cvinfo (CV d'informatique) et Cvens (CV d'éducation/enseignement) héritent de la classe CV, évitant les redondances des attributs et aussi pour faciliter la maintenance. C'est l'utilisateur qui effectue les opérations de choisir la catégorie, type de fichiers ...

Le résultat est téléchargé sous la forme d'un fichier exploitable s'il s'agissait d'un fichier de la classe File, sinon, il est sous la forme d'un fichier CSV ou SQL selon le type d'offre choisi s'il s'agissait d'un CV.

2.7 Conclusion

Dans ce chapitre, nous avons décrit la conception globale de notre système, à travers les différents diagrammes d'UML.

Dans le prochain chapitre, il s'agira de faire un prétraitement de projet.

CHAPITRE 3 : PRETRAITEMENT

Chapitre 3

PRETRAITEMENT

3.1 *Introduction*

Ce troisième chapitre consiste à faire une étude analytique. Il s'agit, d'une part, de définir la stratégie avec laquelle on a démarré notre projet. D'autre part les différentes approches et les systèmes utilisés pour l'extraction du texte ou des caractères dans des documents différents pour développer notre solution intelligente.

3.2 *Analyse des documents*

3.2.1 *L'outil OCR*

L'objectif d'un système OCR est de reconnaître le texte et puis le convertir en une forme modifiable. La reconnaissance optique de caractères implique la traduction du texte dans l'image en codes de caractères modifiables tel que l'ASCII.

Les systèmes OCR proposés par les différents chercheurs sont composés d'un ensemble de modules. L'architecture du système varie d'un système à un autre en fonction des besoins.

Le système suivant peut être une généralisation de tous les systèmes proposés.

3.2.2 *Reconnaissance de document*

La reconnaissance de documents ou plutôt l'analyse d'images de documents concerne tout le processus de conversion de l'image. Ce processus est relatif à



FIGURE 3.1 – Système de reconnaissance des caractères multilingues.

toutes les questions autour du langage écrit et sa transformation numérique : reconnaissance de caractères, formatage du texte, structuration du contenu et accès à l'information pour des applications d'indexation.

S'agissant souvent d'un processus de rétro conversion d'une structure existante, le processus de reconnaissance est guidé par un modèle explicite ou implicite de la classe étudiée. Le modèle décrit les éléments composant le document et leurs relations. Cette description peut être physique, relatant le format de mise en page, logique décrivant l'enchaînement des sous-structures, ou sémantique portant sur le sens affecté à certaines parties. C'est un processus encodant évidemment les caractères et participe de manière très active à la reconnaissance de la structure.

Ce processus serait sans doute clair et "simple" s'il ne s'agissait que de documents textuels pour lesquels on dispose d'une structure éditoriale hiérarchique ; le problème est beaucoup plus délicat pour d'autres classes de documents où l'information n'est pas très organisée et le contenu est hétérogène (comprenant un mélange d'imprimé, de manuscrit et de graphique), comme c'est le cas pour les formulaires, les documents postaux ou techniques, les magazines, etc. Dans ce cas, il n'existe pas de modèle direct pour décrire la composition du document et l'on a souvent recours à un mélange de techniques de traitement d'images et du langage pour extraire l'information.

Le monde économique s'est emparé très tôt de cette technologie. Il a finalisé les premiers travaux sur la reconnaissance optique des caractères.

3.2.3 *Processus de reconnaissance*

Le but de notre travail consiste à élaborer un système de reconnaissance des caractères dans un document multilingue contenant du texte. La vue globale

du système est présentée sur la figure ci-dessous. L'image est entrée dans le

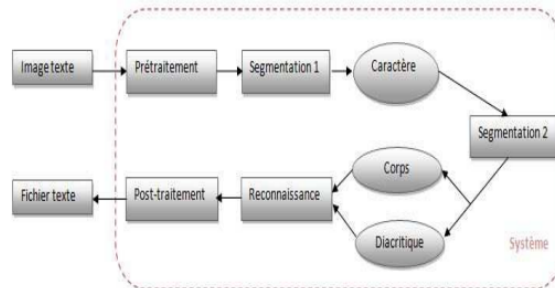


FIGURE 3.2 – Système de reconnaissance des caractères multilingues.

système et subie un ensemble de prétraitements. Elle est segmentée d'abord en lignes et en caractères. Ensuite, une extraction du corps et du diacritique d'un caractère est effectuée. Cette étape aura comme sortie le diacritique et le corps ainsi que la position du diacritique par rapport au corps pour le cas du caractère avec diacritique et retournera l'image d'entrée dans le cas d'un caractère sans diacritique. L'étape de reconnaissance consiste à reconnaître les éléments extraits. Le post-traitement est composé de deux étapes : d'abord une combinaison entre le diacritique et le corps reconnus pour former le caractère final en se basant sur la position du diacritique puis l'étape de vérification par l'utilisateur.

3.2.3.1 Acquisition

Permettant la conversion du document papier sous la forme d'une image numérique (bitmap). Cette étape est importante car elle se préoccupe de la préparation des documents à saisir, du choix et du paramétrage du matériel de saisie (scanner), ainsi que du format de stockage des images.

3.2.3.2 Prétraitement

Dont le rôle est de préparer l'image du document au traitement. Les opérations de prétraitement sont relatives au redressement de l'image, à la suppression du bruit et de l'information redondante, et enfin à la sélection des zones de traitement utiles.

3.2.3.3 Reconnaissance du contenu

Celui qui conduit le plus souvent à la reconnaissance du texte et à l'extraction de la structure logique. Ces traitements s'accompagnent le plus souvent d'opé-

rations préparatoires de segmentation en blocs et de classification des médias (graphiques, tableaux, images, etc.).

3.2.3.4 Post-traitement

La reconnaissance en vue de valider l'opération de numérisation. Cette opération peut se faire soit automatiquement par l'utilisation de dictionnaires et de méthodes de correction linguistiques, ou manuellement au travers d'interfaces dédiées.

3.2.4 *La saisie de document*

La saisie du document est opérée par balayage optique. Le résultat est rangé dans un fichier de points, appelés pixels, dont la taille dépend de la résolution. Les pixels peuvent avoir comme valeurs : 0 (éteint) ou 1 (actif) pour des images binaires, 0 (blanc) à 255 (noir) pour des images de niveau de gris, et trois canaux de valeurs de couleurs entre 0 et 255 pour des images en couleur. La résolution est exprimée en nombre de points par pouce (ppp). Les valeurs courantes utilisées couramment vont de 100 à 400 ppp. Par exemple, en 200 ppp, la taille d'un pixel est de 0,12 mm, ce qui représente 8 points par mm. Pour un format classique A4 et une résolution de 300 ppp, le fichier image contient 2 520 × 3 564 pixels. Il est important de noter que l'image n'a à ce niveau qu'une simple structure de lignes de pixels qu'il faudra exploiter pour retrouver l'information. La Figure ci-dessous montre différents niveaux de résolution utilisés pour le même document. On peut remarquer la dégradation occasionnée par 75 ppp, l'insuffisance des 300 ppp pour le graphique, et l'inutilité des 1200 ppp pour l'ensemble. La technicité des matériels d'acquisition (scanner) a fait un bond considérable ces dernières années. On trouve aujourd'hui des scanners pour des documents de différents types (feuilles, revues, livres, photos, etc.). Leur champ d'application va du "scan" de textes au "scan" de photos en 16 millions de couleurs (et même plus pour certains). La résolution est classiquement de l'ordre de 300 à 1200 ppp selon les modèles.

3.2.5 *Identification de la langue et de la fonte*

Plusieurs modules ont été généralisé pour faciliter l'adaptation et la conduite des documents d'une manière spécifique. Cette réduction est apportée par l'identification du langage et de la fonte qui peuvent varier dans un même document. La connaissance de la langue permet d'adapter les modèles au vocabulaire spécifique. L'identification de la fonte permet de réduire le nombre d'alternatives des formes pour une classe donnée de caractères, conduisant pratiquement à du mono fonte.

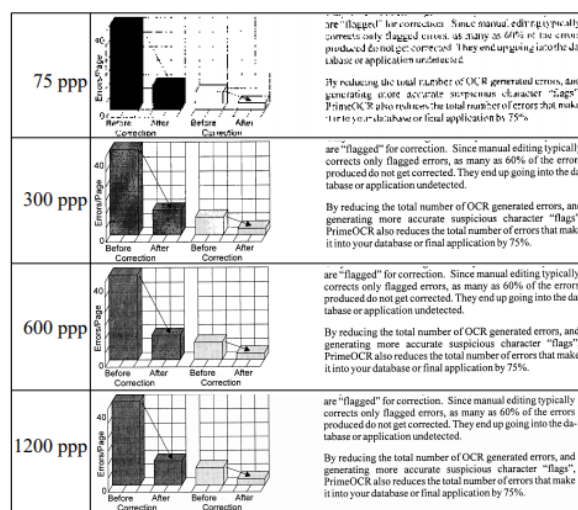


FIGURE 3.3 – Différents niveaux de résolution.

Ces deux mesures peuvent également servir dans des opérations d'indexation et d'interprétation.

Pour l'identification de la langue, Spitz, un des pionniers du domaine, a proposé une méthodologie permettant de classer cinq langues différentes dans un même document. Il différencie d'abord les langues latines des langues asiatiques en utilisant l'écart type de la position verticale des concavités par rapport à la ligne de base. Ces concavités sont situées à la limite de la ligne de base pour le Latin, tandis qu'elles sont uniformément distribuées pour le Chinois, le Japonais et le Coréen. Ensuite, les trois langues asiatiques sont séparées par examen de l'histogramme de distributions de leurs points.

La multiplication des fontes s'ajoute à la multiplication des langues dans un document. Les fontes sont classées en fonction de la police, du style (gras, italique) et du corps. Avec Anigbogu, nous avons proposé dans le cadre de sa thèse une méthode structurale utilisant les mêmes primitives du module de reconnaissance pour identifier la fonte majoritaire dans un bloc de texte. Zramdini a proposé le système ApOFIS capable de distinguer plus de 280 fontes différentes en combinant 10 polices, 7 corps et 4 styles. La fonte est identifiée avec 97% de précision, tandis que le style, le corps et la pente sont identifiés avec une précision s'échelonnant entre 97.5 et 99.9%.

3.2.6 *Processus de reconnaissance de caractères*

Un texte est une association de caractères appartenant à un alphabet, réunis dans des mots d'un vocabulaire donné. On doit retrouver ces caractères, les reconnaître d'abord individuellement, puis les valider par reconnaissance lexicale des mots qui les contiennent. Cette tâche n'est pas triviale car si on doit apprendre à distinguer la forme de chaque caractère dans un vocabulaire de taille souvent importante, on doit en plus être capable de la distinguer dans chacun des styles typographiques (polices), chaque corps et chaque langue, proposés dans le même document. Cette généralisation omni fonte et multilingue n'est pas toujours facile à cerner et reste génératrice de leurs principales erreurs. Un système de reconnaissance de caractères est composé de plusieurs modules : segmentation, apprentissage, reconnaissance et vérification lexicale.

- La segmentation permet d'isoler les éléments textuels, mots et caractères, pour la reconnaissance. Elle se base sur des mesures de plages blanches (interlignes et inter caractères) pour faire la séparation. La multiplicité des polices et la variation des justifications empêchent de stabiliser les seuils de séparation, conduisant à la génération de blancs inexistantes ou au contraire à l'ignorance de blancs séparateurs de mots. Ce type d'erreur est très fréquent, d'après une récente étude réalisée par Nagy et al.
- La reconnaissance de caractères permet de se prononcer sur l'identité d'un caractère à partir d'un apprentissage de sa forme. Cette étape nécessite une étape préalable de paramétrisation de la forme, définissant des données, des mesures, ou des indices visuels sur lesquels s'appuie la méthode de reconnaissance. Suivant la nature de ces informations, il existe plusieurs catégories de méthodes : syntaxique (description par une grammaire), structurelle (description par un graphe), ou statistique (description par partitionnement de l'espace). Ces dernières ont de loin le plus grand intérêt avec les méthodes à base de réseaux de neurones, ou de modèles stochastiques. La complexité de la tâche vient de l'apprentissage qui nécessite, pour sa stabilité, d'un très grand nombre d'échantillons par classe, et de la recherche d'indices visuels discriminants, ce qui n'est pas aisé dans un contexte omni fonte comme celui concerné par la numérisation automatique. Pour accélérer la reconnaissance, on s'appuie sur la similarité entre une forme reconnue et les formes étudiées.
- Le post-traitement est effectué quand le processus de reconnaissance aboutit à la génération d'une liste de lettres ou de mots possibles, éventuellement classés par ordre décroissant de vraisemblance. Le but principal est d'améliorer le taux de reconnaissance en faisant des corrections orthographiques ou morphologiques à l'aide de dictionnaires de digrammes, trigrammes ou n-grammes. Quand il s'agit de la reconnaissance de phrases

entières, on fait intervenir des contraintes de niveaux successifs : lexical, syntaxique ou sémantique.

3.3 *Extraction de texte depuis des documents*

3.3.1 *Depuis une image de scène et une image*

- Image d'un document Structuré
- Image Naturelle

3.3.1.1 East

Avant l'introduction de l'apprentissage automatique dans le domaine de la détection de texte, il était difficile pour la plupart des approches de segmentation de texte de fonctionner sur des scénarios difficiles. Les approches conventionnelles utilisent des fonctionnalités conçues manuellement tandis que les méthodes d'apprentissage approfondi apprennent des fonctionnalités efficaces à partir des données de formation. Ces approches conventionnelles sont généralement en plusieurs étapes, ce qui se termine par des performances globales légèrement inférieures. Nous apprendrons un algorithme basé sur l'apprentissage profond (EAST) qui détecte le texte avec un seul réseau de neurones avec l'élimination des approches multi-étapes.



FIGURE 3.4 – Exemple d'application d'EAST.

3.3.1.2 East - Introduction

L'algorithme EAST utilise un seul réseau de neurones pour prédire un mot ou un texte de niveau ligne. Il peut détecter du texte dans une orientation arbitraire avec des formes quadrilatères. En 2017, cet algorithme a surpassé les méthodes de pointe. Cet algorithme se compose d'un réseau entièrement convolutif avec un état de fusion de suppression non maximale (NMS). Le réseau entièrement

convolutif est utilisé pour localiser le texte dans l'image et cette étape NMS est essentiellement utilisée pour fusionner de nombreuses zones de texte détectées imprécises en une seule zone de délimitation pour chaque région de texte (mot ou texte de ligne).

3.3.1.3 East - L'architecture du Réseau

L'architecture EAST a été créée en tenant compte de différentes tailles de régions de mots. L'idée était de détecter de grandes régions de mots qui nécessitent des fonctionnalités de l'étape ultérieure du réseau neuronal tout en détectant de petites régions de mots qui nécessitent des fonctionnalités de bas niveau à partir des étapes initiales. Pour créer ce réseau, les auteurs ont utilisé trois branches se combinant en un seul réseau neuronal.

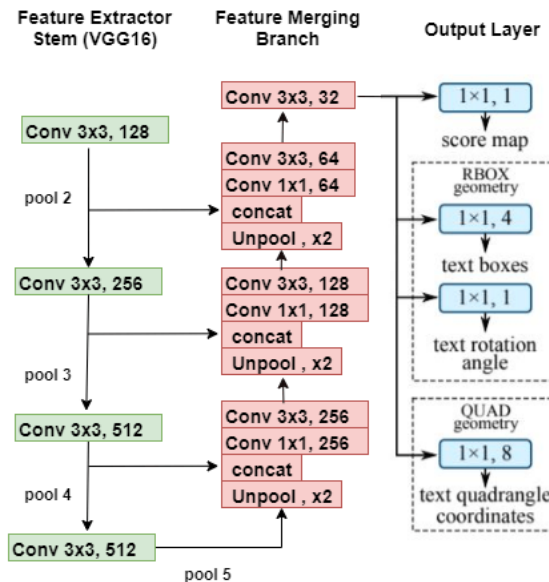


FIGURE 3.5 – L'architecture d'EAST.

3.3.1.4 East - Caractéristiques d'extraction

Cette branche du réseau est utilisée pour extraire des entités de différentes couches du réseau. Cette tige peut être un réseau convolutionnel préformé sur l'ensemble de données ImageNet. Les auteurs de l'architecture EAST ont utilisé PVANet et VGG16 pour l'expérience. Nous verrons l'architecture EAST avec le réseau VGG16 uniquement. Pour la tige de l'architecture, il prend la sortie du modèle VGG16 après les couches pool2, pool3, pool4 et pool5.

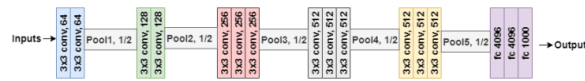


FIGURE 3.6 – L'architecture de VGG16.

3.3.1.5 East - Branche de fusion

Dans cette branche du réseau EAST, il fusionne les sorties de fonction d'une couche différente du réseau VGG16. L'image d'entrée est transmise via le modèle VGG16 et les sorties de quatre couches différentes de VGG16 sont prises. La fusion de ces cartes d'entités coûtera des calculs. C'est pourquoi EAST utilise une architecture U-net pour fusionner progressivement les cartes d'entités. Premièrement, les sorties après la couche pool5 sont sur échantillonnées à l'aide d'une couche déconvolutionnelle. Maintenant, la taille des entités après cette couche serait égale aux sorties de la couche pool4 et les deux sont ensuite fusionnées en une seule couche. Ensuite, Conv 1×1 et Conv 3×3 sont appliqués pour fusionner les informations et produire la sortie de cette étape de fusion.

De même, les sorties d'autres couches du modèle VGG16 sont concaténées et enfin, une couche Conv 3×3 est appliquée pour produire la couche finale de la carte d'entités avant la couche de sortie.

3.3.1.6 East - La couche de Sortie

La couche de sortie se compose d'une carte de score et d'une carte de géométrie. La carte de score nous indique la probabilité de texte dans cette région tandis que la carte de géométrie définit la limite de la zone de texte. Cette carte de géométrie peut être une boîte pivotée ou un quadrilatère. Une zone pivotée se compose des coordonnées en haut à gauche, de la largeur, de la hauteur et de l'angle de rotation de la zone de texte. Alors que le quadrangle se compose des quatre coordonnées d'un rectangle.

3.3.1.7 East - La fonction de Perte

La fonction de perte utilisée dans cet algorithme EAST comprend à la fois la perte de carte de score et la fonction de perte de géométrie. Dans la formule ci-dessus, les deux pertes sont combinées avec un poids. Ce sert à donner de l'importance à différentes pertes. Ce poids est souvent égal à 1.

$$L = L_s + \lambda_g L_g$$

3.3.1.8 East - Suppression de la fusion maximale

Les géométries prédites après un réseau entièrement convolutionnel passent par une valeur de seuil. Après ce seuillage, les géométries restantes sont supprimées à l'aide d'un NMS sensible à la localité. Un NMS naïf s'exécute en $O(n^2)$. Mais pour exécuter ceci dans $O(n)$, les auteurs ont adopté une méthode qui utilise la suppression ligne par ligne. Cette suppression ligne par ligne prend également en compte la fusion itérative de la dernière fusionnée. Cela rend cet algorithme rapide dans la plupart des cas, mais la pire complexité temporelle reste $O(n^2)$.

3.3.1.9 Exigences - Image document / naturelle

- Détecter le texte et l'extraire à partir d'une image document par effectuer un seuil et des contours.
- Enregistrer texte dans un fichier txt.
- La première couche est notre activation sigmoïde de sortie qui nous donne la probabilité d'une région contenant du texte ou non.
- La deuxième couche est la carte d'entités en sortie qui représente la «géométrie» de l'image - nous serons en mesure d'utiliser cette géométrie pour dériver les coordonnées du cadre de délimitation du texte dans l'image d'entrée.
- Saisir les dimensions du volume des partitions puis initialiser deux listes.
- Parcourir chacun des index de colonne pour notre ligne actuellement sélectionnée.
- Filtrer les détectations de texte faibles en ignorant les zones qui n'ont pas une probabilité suffisamment élevée.
- Le détecteur de texte EAST réduit naturellement la taille du volume lorsque l'image passe à travers le réseau.
- Appliquer une suppression non maximale à nos boîtes englobantes pour supprimer les boîtes englobantes se chevauchant faiblement.
- Afficher les prédictions de texte résultantes.
- Enregistrer le mask et le texte résultant et modifiable dans un fichier txt.

3.3.2 *Depuis un document PDF*

Dans cette section, on va extraire le texte à partir d'un document PDF, on va être capable de :

- Extraire les informations du document
- Division des documents page par page
- Fusionner des documents page par page
- Crypter et décrypter des fichiers PDF
- Extraire les tableaux
- Enregistrer la sortie dans un fichier .txt

3.3.2.1 Exigences - PDF

- Lire le fichier pdf.
- Appliquer une boucle for pour lire toutes les pages du fichier pdf et en extraire le texte.
- Enregistrer le texte dans un fichier txt modifiable.

3.3.3 *Depuis un document Word*

Les données du fichier source peuvent être tabulaires, contenues dans les champs de formulaire d'un formulaire protégé ou, si on utilise Word 2007 ou version ultérieure, contenues dans des contrôles de contenu. En utilisant un fichier source comme feuille de données de cette manière, on peut créer une collection de modèles qui dessinent ou extraient des données à partir d'un fichier de données source commun.

Dans cette section, on va extraire le texte à partir d'un document Word, on va être capable de :

- Extraire des données stockées dans une table de document source
- Extraire les notes de bas de page et les notes de fin
- Convertir les puces et les listes numérotées en ascii avec indentation
- Couverture complète des tests et documentation pour les développeurs
- Enregistrer la sortie dans un fichier .txt

3.3.3.1 Exigences - Word

- Lire le fichier word.
- Appliquer une boucle for pour lire toutes les pages du fichier word et en extraire le texte.
- Appliquer la notion du dataframe pour extraire les tableaux contenus dans le fichier.

- Enregistrer le texte dans un fichier txt modifiable.

3.3.4 *Depuis un document Excel*

On peut lire les données des cellules en spécifiant une feuille spécifique et les cellules à partir desquelles On souhaite extraire les données.

On a normalement des lignes séquencées par des nombres et des colonnes séquencées par des caractères alphabétiques. Avec `xlsx`, les lignes et les colonnes sont spécifiées par des en-têtes et des numéros, la première ligne commençant par les en-têtes et la première colonne commençant par 1.

Dans cette section, on va extraire le texte à partir d'un document Excel, on va être capable de :

- Extraire les headers d'un fichier excel
- Extraire tous les données du fichier excel (lignes + colonnes)
- Les colonnes seront numérotées commençant par un 0.
- Enregistrer le résultat en tant que fichier `.txt` sous forme d'un tableau.

3.3.4.1 Exigences - Excel

- Utiliser la notion du dataframe pour convertir un fichier excel en un fichier modifiable.
- Enregistrer le résultat sous fichier txt avec un affichage tabulaire.

3.3.5 *Extraction des données à partir d'un CV*

Notre époque est de plus en plus influencée par l'utilisation des données intelligentes (smart data) et du web sémantique. Le processus de recrutement n'en est pas pour autant facile, en particulier en matière de recherche de profils et de talents, car les approches actuelles se limitent à la recherche par mots-clefs.

Différentes personnes de différents domaines et d'horizons différents ont des personnalités variées. De même, leur modèle de rédaction de CV fluctue également. Ils ont travaillé dans différents types de projets et chacun d'entre eux possède un style varié de rédaction. Rendant ainsi chaque CV unique en soi.

Chaque jour, les entreprises exploraient des centaines de CV sur Internet. Après avoir rassemblé les CV, les dirigeants appelants avaient l'habitude de résumer le CV, d'entrer des détails spécifiques dans leur base de données, puis d'appeler le candidat pour un conseil en emploi. Un cadre a pris environ 10 à 15 minutes par

CV pour le résumer et entrer les détails dans la base de données. La solution donc est d'automatiser ce processus.

3.3.5.1 Le NLP

Le NLP est une branche de l'intelligence artificielle qui s'occupe particulièrement du traitement du langage écrit aussi appelé avec le nom français TALN (traitement automatique du langage naturel). C'est tout ce qui est lié au langage humain et au traitement de celui-ci par des outils informatiques.

Le traitement du langage naturel aide les ordinateurs à communiquer avec les humains dans leur propre langue et met à l'échelle d'autres tâches liées au langage. Par exemple, il permet aux ordinateurs de lire du texte, d'entendre la parole, de l'interpréter, de mesurer le sentiment et de déterminer quelles parties sont importantes.

Les machines d'aujourd'hui peuvent analyser plus de données basées sur le langage que les humains, sans fatigue et de manière cohérente et impartiale. Compte tenu de la quantité stupéfiante de données non structurées générées chaque jour, des dossiers médicaux aux médias sociaux, l'automatisation sera essentielle pour analyser efficacement les données de texte et de parole.

3.3.5.2 Fonctionnement d'NLP

Le traitement du langage naturel comprend de nombreuses techniques différentes pour interpréter le langage humain, allant des méthodes statistiques et d'apprentissage automatique aux approches basées sur des règles et algorithmiques. Nous avons besoin d'un large éventail d'approches, car les données textuelles et vocales varient considérablement, tout comme les applications pratiques.

Les tâches de base de la NLP comprennent la tokenisation et l'analyse, la lemmatisation / stemming, le balisage d'une partie du discours, la détection du langage et l'identification des relations sémantiques.

En termes généraux, les tâches de NLP décomposent le langage en morceaux élémentaires plus courts, essayent de comprendre les relations entre les morceaux et explorent comment les morceaux fonctionnent ensemble pour créer du sens.

Ces tâches sous-jacentes sont souvent utilisées dans des capacités NLP de niveau supérieur, telles que :

- **Catégorisation du contenu** : Un résumé de document basé sur la

langue, y compris la recherche et l'indexation, les alertes de contenu et la détection de duplication.

- **Découverte et modélisation de sujets** : Capturer avec précision le sens et les thèmes dans les collections de texte et appliquez des analyses avancées au texte, comme l'optimisation et les prévisions.
- **Extraction contextuelle** : Extraire automatiquement des informations structurées à partir de sources textuelles.
- **Analyse des sentiments** : Identifier l'humeur ou les opinions subjectives dans de grandes quantités de texte, y compris le sentiment moyen et l'exploration d'opinion.
- **Conversion parole-texte et texte-parole** : Transformer les commandes vocales en texte écrit et vice versa.
- **Résumé du document** : Génération automatique d'un synopsis de grands corps de texte.
- **Traduction automatique** : Traduction automatique de texte ou de discours d'une langue à une autre.

3.3.5.3 Données du CV

spaCy est une bibliothèque gratuite et open-source pour le traitement du langage naturel (NLP) en Python avec de nombreuses fonctionnalités intégrées. Il devient de plus en plus populaire pour le traitement et l'analyse des données en NLP. Les données textuelles non structurées sont produites à grande échelle, et il est important de traiter et de tirer des informations des données non structurées. Pour ce faire, on doit représenter les données dans un format compréhensible par les ordinateurs.

Notre programme a pour intérêt l'extraction des informations suivantes :

- L'image du candidat
- Le prénom du candidat
- L'email du candidat
- Le numéro du téléphone du candidat
- La ville du candidat
- Les compétences techniques du candidat
- L'université / la faculté du candidat
- Le diplôme obtenu du candidat
- La société / l'entreprise du candidat
- Les langues maîtrisées par le candidat
- Les loisirs du candidat
- Le profile du candidat
- Enregistrer le résultat dans un fichier csv.

- Enregistrer le résultat dans une Base de données MySQL.
- Filtrer les candidats selon le type d'offre d'emploi/ de stage demandé par rapport au profile du candidat.
- Téléchargement de la base de données en format sql ou le téléchargement du fichier csv.

3.3.5.4 Exigences - CV

- Lire le fichier en tant que pdf et extraire l'image du candidat.
- Utiliser les opérations à bas d'expressions rationnelles pour l'extraction de l'email et le numéro de téléphone du candidat.
- Détecter les parties colorantes dans le fichier pdf indiquant l'image du candidat (si = rgb,cmyk,gray).
- Filtrer l'object image dans le pdf et obtenir la hauteur et la longueur de l'image.
- Récupérer l'image,la lire/enregistrer avec pillow.
- Exploiter le NLP (natural language processing) pour l'extraction du prénom,la ville,les compétences techniques,la faculté / l'université,le diplôme obtenu,l'entreprise / la société, les langues maîtrisées,les loisirs ainsi le profile du candidat à l'aide d'une tokenization (processus de segmentation de texte en phrases et en mots qui sert à couper le texte en morceaux appelés jetons tout en jetant quelques caractères comme la ponctuation ...) comme le montre la figure ci-dessous.

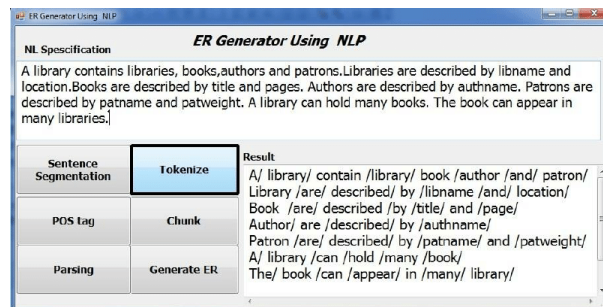


FIGURE 3.7 – Exemple de tokenization.

- Gratter d'abord les mots-clés de chaque section. Si par exemple, je souhaite extraire le nom de l'université. Par conséquent, je stock plusieurs noms d'universités dans un fichier csv. Ensuite, j'utilise la tokenization pour vérifier si ce nom d'université se trouve parmi ceux dans le fichier csv. S'il est trouvé, cette information sera extraite du CV. (même procédure pour extraction de la ville,nom,profile,diplôme...)

- Enregistrer les données extraites dans un fichier csv.
- Enregistrer les données extraites dans une Base de données MySQL.
- Filtrer les candidats selon le type d'offre demandé par rapport au profile du candidat. Si le profile est identique à l'offre. Le candidat est selectionné.
- Les offres demandées correspondent à : Art/Média, Biologie/Santé, Internet/Informatique, Finance/Commerce/Comptabilité et Education/Enseignement.
- Enregistrer le résultat dans une table sql/fichier csv.
- Télécharger la base de données sous forme d'un fichier sql, sinon le fichier csv.

3.3.6 *Les Challenges et les difficultés d'extraction*

Pour une reconnaissance de caractères de bonne qualité et de haute précision, on attend toujours des images de haute qualité ou à haute résolution avec certaines propriétés structurelles de base telles qu'un texte et un arrière-plan très différents. La façon dont les images sont générées est un facteur important et déterminant la précision et le succès, car cela affecte souvent considérablement la qualité des images. Généralement, avec des images produites par des scanners donne une grande précision et de bonnes performances. En revanche, les images produites par les caméras ne sont généralement pas aussi performantes en entrée que les images numérisées à utiliser en raison de facteurs liés à l'environnement ou à la caméra. De nombreuses challenges peuvent apparaître, qui sont clarifiées comme suit.

3.3.6.1 Complexité de la scène

Dans un environnement régulier, nous pouvons voir un grand nombre d'objets artificiels qui sont inclus dans des images prises par la caméra telles que des peintures, des bâtiments et des symboles. Ces objets ont des structures et des apparences comparées au texte, ce qui rend la reconnaissance de texte très difficile dans l'image traitée. Le texte lui-même est régulièrement présenté pour encourager le déchiffrement. Le défi avec la complexité de la scène est que la scène environnante rend difficile la séparation du texte du non-texte.

3.3.6.2 Conditions d'éclairage inégal

Souvent, la prise d'images dans des environnements naturels entraîne un éclairage et des ombres inégaux. Cela pose un défi pour car il dégrade les caractéristiques souhaitées de l'image et entraîne donc des résultats de détection, de segmentation et de reconnaissance moins précis. Cette condition avec un éclairage inégal est ce qui distingue une image numérisée d'une image produite avec

un appareil photo. L'absence de telles disparités dans l'éclairage et les ombres rend les images numérisées préférées aux images de caméra pour leurs meilleures caractéristiques et qualité. Bien que l'utilisation d'un flash intégré à l'appareil photo puisse éliminer de tels problèmes avec un éclairage inégal, il présente de nouveaux défis.

3.3.6.3 La Rotation

Pour les systèmes de reconnaissance optique de caractères, le point de vue de l'image d'entrée provenant de la caméra d'un appareil portable ou d'autres gadgets utilisés pour l'image prise n'est pas fixé comme une entrée de scanner, ce qui pourrait fausser les lignes de texte à partir de leur orientation unique, observée. De très mauvais résultats seront observés lorsqu'une telle image asymétrique est envoyée au classificateur. De nombreuses techniques disponibles pour le réalignement des documents images, telles que le profil de projection, l'algorithme RAST, les méthodes de transformation de Fourier, etc.

3.3.6.4 Flou et dégradation

Étant donné que le travail sur une variété de distances est destiné à de nombreux appareils photo numériques, la mise au point de l'appareil photo est un facteur important. Pour la meilleure précision de reconnaissance et de segmentation des caractères, une netteté des caractères est requise. À de grandes ouvertures et de courtes distances, une mise au point inégale peut être observée lorsqu'un petit point de vue change. Pour la plupart liés à la photographie, il existe deux types d'obscurité : l'obscurité floue et l'obscurité du mouvement. Au moment d'attraper un élément en mouvement, lorsque le taux d'ombre de la caméra n'est pas suffisamment élevé, le capteur se présente à une scène en constante évolution. En conséquence, un flou sera observé dans les parties en mouvement.

3.3.6.5 Environnements multilingues

Même si une grande partie des langues du latin ont de nombreux caractères, par exemple, le japonais, le chinois et le coréen, ont un grand nombre de classes de caractères. Les caractères connectés existent dans les langues arabes, qui selon le contexte, changent de forme d'écriture. En hindi, les syllabes représentent en combinant des lettres alphabétiques en des milliers de formes. Dans les situations multilingues, c'est plus gênant.

3.3.6.6 Inclinaison (distorsion en perspective)

Les images de document obtenues par les scanners sont constamment parallèles au plan du capteur, mais cela ne peut pas être observé à tout moment pour une image enregistrée obtenue par une caméra portable, qui peut ne pas être généralement parallèle au plan de forme. Par conséquent, les lignes de texte qui sont éloignées de l'appareil photo semblent plus petites que celles qui sont plus proches de l'appareil photo, ce qui semble plus grand. Cette situation provoque des images inclinées. L'observation d'une distorsion de perspective est claire si le dispositif de reconnaissance n'est pas intolérant à la perspective, ce qui entraîne un taux de reconnaissance et une précision inférieurs. Les téléphones portables ont un avantage avec les capteurs d'orientation. Ils peuvent reconnaître si l'appareil est incliné et en cas de torsion, ils peuvent interdire aux clients de prendre des photos. Cette fonctionnalité permet également à l'utilisateur d'aligner le plan du formulaire avec celui de la caméra. Les capteurs d'orientation peuvent donc garantir que les images produites satisfont un certain degré de régularité.

3.3.6.7 Les Fonts

Le style italique et les polices de script des caractères peuvent se chevaucher, ce qui rend difficile l'exécution de certains des principaux processus tels que la segmentation. Les caractères de différentes polices présentent de grandes variations intra-classe et forment de nombreux sous-espaces de motif, ce qui rend difficile une reconnaissance précise lorsque le nombre de classes de caractères est important.

3.3.7 *Les Erreurs de structuration*

Une erreur possible, en lien direct avec la segmentation est l'inversion de certaines zones par rapport à l'ordre logique de lecture du document. Sur l'exemple ci-dessous le paragraphe en bas à gauche correspond à des notes de bas de page. Si on regarde le document de loin, sans prendre en compte son contenu, l'ordre logique de lecture qui apparait est de lire la colonne de gauche puis la colonne de droite.

On voit clairement que la disposition physique joue un rôle important dans la définition de l'ordre de lecture mais qu'elle peut être trompeuse. C'est par le contenu que nous pourrons faire la différence et retrouver le bon ordre.

THE RAS TAFARI MOVEMENT IN JAMAICA 167

it attempts to operationally delineate the concept of awareness of group hostility. Second, it suggests a technique for the measurement of awareness. Third, the technique can be applied to the measurement of awareness of other social problems. Fourth, in a given community the relative positions of awareness to different social problems could be ascertained. Finally, the items of an awareness instrument should be tested for the scalability of these items.

THE RAS TAFARI MOVEMENT IN JAMAICA: A STUDY OF RACE AND CLASS CONFLICT*

GEORGE EATON SIMPSON

Ottawa College

THE contra-acculturative aspects of Messianic cults and nativistic movements have long been of interest to anthropologists and sociologists.¹ Ras Tafari, a Jamaican cult which originated in 1930, is violently anti-white on the verbal level. Its members regard Haile Selassie (Ras Tafari), Emperor of Abyssinia, as the living God, see no hope for black men in the British West Indies, and look forward to an early return to Ethiopia.

The "Rasta" people consider Marcus Garvey, revered founder of the Universal Negro Improvement Association, as the forerunner of their movement. They claim that Garvey, "the world's greatest statesman," was sent by Ras Tafari "to cut and clear."² Garvey advocated a mass migration to Africa, and his slogans "Africa for the Africans—At Home and Abroad" and "One God!

One Aim! One Destiny!" are proclaimed at every Ras Tafari meeting.

In the early days of the movement, opposition came from both the ordinary Jamaicans and the police. Lower class Jamaicans stored spears, dashed banners, and smashed lamps at street meetings. An active early leader of the cult was arrested, jailed, and tried seven times, but never convicted, on charges of disorderly conduct, ganja (marijuana) smoking, and larceny. Open hostility to the movement has declined to some extent in recent years due, in part, to the well-disciplined control of members during meetings. Middle and upper class Jamaicans, as well as foreigners, still fear the Ras Tafarians, but available evidence does not support the widespread belief that they are bearded hoodlums.

Western Kingston and Eastern St. Andrew constitute the center of the Ras Tafari movement, but groups have been formed in other parts of the island. Participants are lower class Jamaicans, many of them unemployed or underemployed, who reside in crowded, blighted areas.

At present, twelve or fifteen Ras Tafari groups operate in Kingston and St. Andrew, with memberships ranging from twenty-five to one hundred and fifty or more. Groups form, split, and dissolve, and some individuals accept cult beliefs without attaching themselves to an organization. In contrast to a Revivalist group, which is dominated by a leader, a Ras Tafari band is extremely democratic. Everyone who wishes to speak must be heard, often at some length, and no action is taken without a vote of the membership, or, at the least, the executive committee. Names of these groups include: Ethiopian Coptic League, United Ethiopian Body, Ethiopian Youth Cosmic Faith,

* With the support of a grant from the American Philosophical Society. I am indebted to Mr. Arthur Bethune, of Kingston, Jamaica, for assistance in the collection of data on the Ras Tafari movement. Paper read at the annual meeting of the American Sociological Society, September, 1954.

¹ See James Mooney, "The Ghost Dance Religion and Sioux Outbreak of 1890," *Bureau of American Ethnological Reports*, 14, part 2 (1892); A. H. Gayton, "The Ghost Dance of 1870 in South-Central California," *University of California Publications in Archaeology and Ethnology*, 28 (1938); Bernard Barber, "Acculturation and Messianic Movements," *American Sociological Review*, 6 (1941), pp. 465-499; Ralph Linton, "Nativistic Movements," *American Anthropologist*, 45 (1943), pp. 230-240; M. J. Herskovits, *Man and His Works* (New York: Knopf, 1948), pp. 531-532.

² This expression is used in the Jamaican Revivalist cults (Pocomania and Revival Zion) to refer to the process of removing evil spirits by ritualistic means.

FIGURE 3.8 – Ordre de lecture complexe dans un document multi-colonnes où la note de bas de page prolonge la première colonne à partir de sa moitié.

Les erreurs de segmentation se situent majoritairement au niveau des lignes et des blocs. Les erreurs de segmentation des mots sont plus rares en général. En ce qui concerne les caractères, le cas est trop rare pour être envisagé. Il peut survenir dans le cas de documents de mauvaise qualité où les caractères ont été abîmés, coupés etc. On distingue 5 types d'erreur de segmentation :

3.3.7.1 Fusion

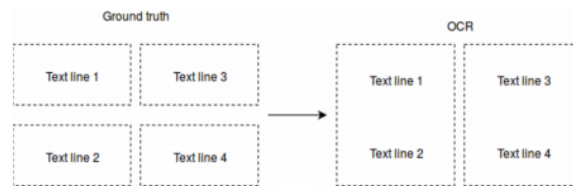


FIGURE 3.9 – Erreur de segmentation - Fusion verticale de blocs. Les numéros donnent l'ordre de lecture des lignes.

Verticale : Cette erreur ne modifie pas l'ordre de lecture en général mais elle peut le faire si l'ordre est particulier. Si cette erreur se produit sur un bloc, alors elle est indétectable au niveau des lignes (les lignes ne sont pas modifiées). Si elle se produit sur une ligne alors la reconnaissance va sûrement échouer puisqu'on tentera de reconnaître une ligne dans une image en contenant deux. Le premier cas est bénin tandis que le dernier cas est grave puisqu'il est quasiment certain qu'on aura une mauvaise reconnaissance.



FIGURE 3.10 – Erreur de segmentation - Fusion horizontale.

Horizontale : Cette erreur conduit à la fusion des lignes adjacentes horizontalement. Elle modifie l'ordre de lecture du document. C'est une erreur de segmentation grave car elle associe deux zones séparées physiquement. De plus elle peut être difficilement détectable car elle dépend du contenu qui lui aussi peut être erroné.

3.3.7.2 Fission

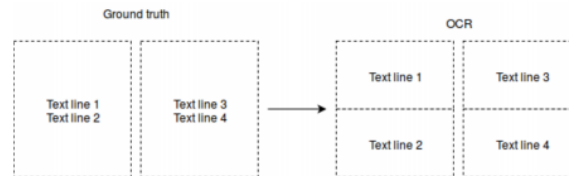


FIGURE 3.11 – Erreur de segmentation - Fission verticale.

Verticale : La fission verticale des blocs au même titre que la fusion verticale est bénigne. Elle est aussi non visible au niveau des lignes (lorsqu'elle est sur les blocs). Si elle s'effectue sur des lignes alors cette erreur est grave car on devra reconnaître une ligne en ayant la moitié ou un morceau. Dans ce dernier cas, on est quasiment sûr que la reconnaissance échouera.



FIGURE 3.12 – Erreur de segmentation - Fission horizontale.

Horizontale : Si on se place au niveau des lignes, la fission horizontale découpera la ligne concernée en au moins 2 morceaux. On peut détecter cette erreur si la fission est contenue dans un bloc et que les lignes avant et après font la taille du bloc. Sinon, une vérification par le contenu est nécessaire pour la détecter. On peut aussi trouver des mots coupés en 2 le plus souvent, ce qui peut poser problème lors d'une tâche requérant une comparaison de contenu. Elle ne modifie pas l'ordre de lecture lorsqu'elle est encadrée par deux lignes, cependant on peut penser à un paragraphe coupé entièrement en deux, ce qui provoquera une vraie perturbation de l'ordre.

3.3.8 *Les Erreurs de reconnaissance de caractère*

3.3.8.1 Définition

La reconnaissance de caractères consiste à identifier la forme à partir d'image. Certaines méthodes ne vont pas reconnaître les caractères indépendamment les uns des autres mais au contraire, reconnaître des mots entiers (dans le cas où le nombre de mots à reconnaître est réduit).

3.3.8.2 Les erreurs

Les erreurs de reconnaissance se situent donc au niveau des caractères. On distingue 3 types d'erreurs :

- **Délétion** : caractère manquant. $word \rightarrow wrd$
- **Substitution** : un caractère remplacé par un autre (voire plusieurs).
 $word \rightarrow w0rd$ $word \rightarrow ivord$
- **Insertion** : caractère inséré dans le texte $word \rightarrow wordsd$

La plupart des erreurs de reconnaissance de caractères proviennent d'une mauvaise interprétation de la morphologie des caractères à reconnaître conduisant parfois à une confusion avec d'autres caractères proches. Par exemple : "M" peut être compris comme un ensemble de lettre qui serait "l V l" qui correspondent au découpage de la lettre d'origine. Ces erreurs sont donc fonction de la police du texte et bien sûr de la connaissance de cette police.

3.4 *Conclusion*

Ce chapitre a présenté les différents aspects analytiques de notre solution intelligente ainsi que les différents prétraitements intervenants dans cette application tels que la reconnaissance de caractères, l'extraction du texte depuis différents documents et les erreurs rencontrées lors du développement, etc.

Dans le chapitre suivant, nous abordons la mise en place de la solution, ainsi que les outils de technologies utilisées.

*CHAPITRE 4 :
REALISATION
DE LA SOLUTION
INTELLIGENTE*

Chapitre 4

REALISATION DE LA SOLUTION INTELLIGENTE

4.1 *Introduction*

Ce chapitre présente l'étape de réalisation en montrant les principales interfaces de l'application et la plateforme utilisée ainsi que les différents outils et technologies intervenants dans cette réalisation.

4.2 *Outils et Technologies utilisées*

4.2.1 *Environnement matériel*

4.2.1.1 PC portable LENOVO

- AMD A4-9125 RADEON R3, 4 COMPUTE CORES 2C+2G (2.30 GHz).
- Mémoire RAM 4GO.

4.2.2 *Environnement logiciel*

4.2.2.1 HTML5



FIGURE 4.1 – Logo de HTML5.

L'HTML (Hypertext Markup Language) est un langage informatique pour créer des pages web. Ce langage permet de réaliser de l'hypertexte à base d'une structure de balisage. L'HTML5 est le successeur de l'HTML 4.01, ça veut dire qu'il s'agit toujours du HTML à la différence de quelques nouvelles balises. A cette raison, il sera très facile d'intégrer du contenu multimédia et graphique pour le Web sans utiliser le flash et plugins tiers.

4.2.2.2 CSS3



FIGURE 4.2 – Logo de CSS3.

Les feuilles de styles (en anglais "Cascading Style Sheets", abrégé CSS) sont un langage qui permet de gérer la présentation d'une page Web. Les styles permettent de définir des règles appliquées à un ou plusieurs documents HTML. Ces règles portent sur le positionnement des éléments, l'alignement, les polices de caractères, les couleurs, les images de fond, etc.

4.2.2.3 JavaScript



FIGURE 4.3 – Logo de JavaScript.

JavaScript est un langage de script incorporé dans un document HTML. Ce langage est un langage de programmation qui permet d'apporter des améliorations au langage HTML en permettant d'exécuter des commandes du côté client, c'est-à-dire au niveau du navigateur et non du serveur web.



FIGURE 4.4 – Logo de JQuery.

4.2.2.4 JQuery

JQuery est une bibliothèque JavaScript libre qui porte sur l'interaction entre JavaScript (comprenant Ajax) et HTML, et a pour but de simplifier des commandes communes de JavaScript.

4.2.2.5 Python

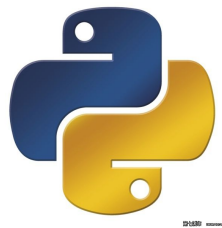


FIGURE 4.5 – Logo de Python.

Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes.

4.2.2.6 Flask

Flask est un micro-framework python facile et simple qui permet de faire des applications web évolutives. Flask dépend de la boîte à outils WSGI de Werkzeug et du moteur de templates Jinja.

Un micro framework est un framework qui tente de fournir uniquement les composants absolument nécessaires à un développeur pour créer une application. Dans le cas des frameworks d'applications Web, un micro framework peut être



FIGURE 4.6 – Logo de Flask.

spécifiquement conçu pour la construction d'API pour un autre service ou une autre application.

4.2.2.7 MySQL DataBase



FIGURE 4.7 – Logo de MySQL.

MySQL est un système de gestion de bases de données relationnelles (SGBDR). Il est distribué sous une double licence GPL et propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle, PostgreSQL et Microsoft SQL Server.

4.3 *Présentation du WebSite*

4.3.1 *Page d'accueil*

La page d'accueil permet aux utilisateurs d'accéder à l'application, en utilisant un seul click sur le lien redirigeant vers deux autres pages, « File » pour l'extraction du texte à partir d'un fichier PDF, Document Word, Document d'image, Document d'Excel ou d'une image de scène, ou « CV File » qui permet l'extraction des données à partir d'un CV sous forme de fichier PDF.

Les comptes Facebook, Twitter et LinkedIn sont liées aux pages officielles de la société.

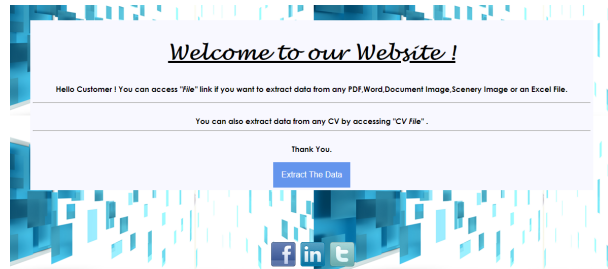


FIGURE 4.8 – Page d'accueil du site web.

4.3.2 *Extraction depuis catégorie « File »*

Interface d'extraction de texte depuis « File » :

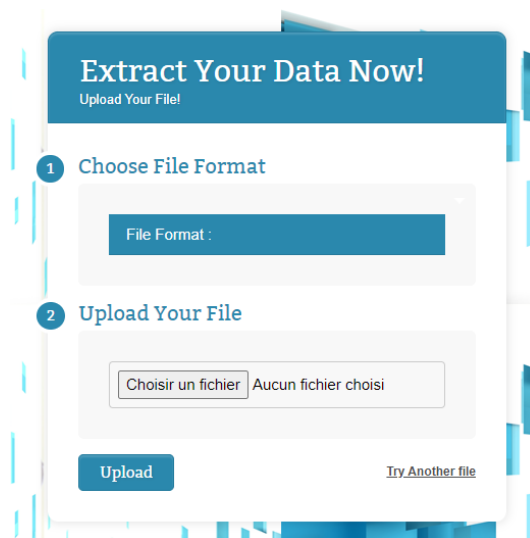


FIGURE 4.9 – Extraction depuis « File ».

4.3.3 *Extraction depuis catégorie « CV File »*

Interface d'extraction de données d'un CV depuis « CV File » :

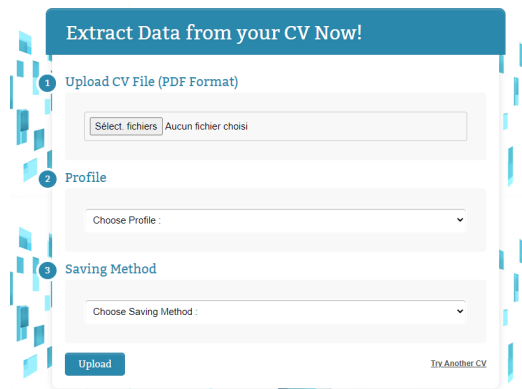


FIGURE 4.10 – Extraction depuis « CV File ».

4.3.4 Manipulation des fichiers

Si aucun fichier n'est sélectionné, un message s'affiche : Sinon, le fichier est

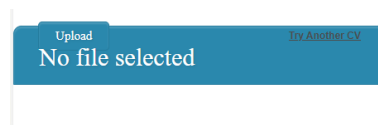


FIGURE 4.11 – Aucun fichier sélectionné.

en cours d'extraction :

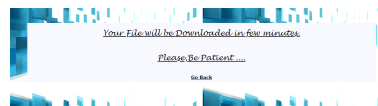


FIGURE 4.12 – Fichier en cours d'extraction.

4.3.5 Résultats

En cas d'extraction depuis catégorie « File », le résultat enregistré sous le même nom du fichier sous forme .txt :

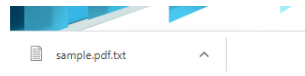


FIGURE 4.13 – Enregistrement en tant que texte.

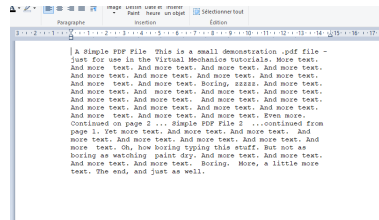


FIGURE 4.14 – Exemple d'enregistrement en tant que texte.

En cas d'extraction depuis catégorie « CV File », si la méthode d'enregistrement était CSV, le résultat enregistré est sous forme .csv indiquant même le type d'offre choisi (ici info) :



FIGURE 4.15 – Enregistrement en tant que csv.

4.4 *Conclusion*

Ce chapitre a décrit l'architecture technique du futur système et cité les différents outils et framework adopté, on a présenté un aperçu sur quelques interfaces réalisées permettant l'interaction entre l'utilisateur et le système.

Conclusion Générale

Au bout de notre cursus en Master Informatique et Modélisation des Systèmes Complexes, nous avons été chargés de réaliser un projet de fin d'études. Notre travail s'est basé sur le développement d'une solution intelligente pour la reconnaissance et la détection de texte depuis des documents.

En effet, ce projet était une étape très importante dans mon cycle de formation vu qu'il était une occasion très intéressante et bénéfique pour savoir comment appliquer sur le plan pratique des connaissances théoriques déjà acquises, il m'a aussi permis d'acquérir de nouvelles connaissances techniques.

En même temps, j'ai appris l'importance de la recherche et de la communication afin de s'informer. Ainsi que l'importance de la gestion du temps et de la planification des tâches pour le bon déroulement des travaux. Et grâce à un environnement favorable pour le travail et la coordination d'efforts, j'ai pu réaliser le projet demandé.

Ce travail peut s'étendre encore plus, mais le fait d'être arrivée ce stade dans le projet me donne plus de confiance en soi-même et m'encourage à continuer, vu les problèmes que j'ai confronté pour apprendre des nouveaux langages et outils de travail.

Au cours de la phase de réalisation de notre application, on a pu implémenter la majeure partie des fonctionnalités du système cible. Pour atteindre cet objectif, une étude de l'existant a été réalisée, et les besoins fonctionnels auxquels le système futur devrait répondre ont été identifiés. On a ensuite entamé une étude conceptuelle globale selon la méthodologie SCRUM. L'étude technique nous a permis de décider les outils à utiliser lors de la réalisation de la solution.

En termes de perspectives, j'ai proposé d'améliorer l'application en optimisant le temps d'exécution, ainsi qu'améliorer la structure des textes extraits à partir d'un fichier word et pdf.

Webographie

- <https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada>
- <https://medium.com/capital-one-tech/learning-to-read-computer-vision-methods-for-extracting-text-from-images-2ffcdae11594>
- <https://www.tutorialspoint.com/flask/index.htm>
- <https://viveksb007.github.io/2018/04/uploading-processing-downloading-files-in-flask>
- <https://www.sqlshack.com/how-to-backup-and-restore-mysql-databases-using-the-mysqldump-command/>
- <https://stackoverflow.com/questions/27339064/export-mysql-database-to-sql-file-with-python-os-system>
- <https://www.daniweb.com/programming/databases/threads/450368/error-1054-42s22-unknown-column-firstname-in-field-list>
- <https://dev.mysql.com/doc/connector-python/en/connector-python-example-ddl.html>
- <https://www.looklinux.com/mysql-warning-using-a-password-on-the-command-line-interface-can-be-insecure/>
- <https://www.sas.com/enus/insights/analytics/what-is-natural-language-processing-nlp.html>