

République Tunisienne  
Ministère de l'Enseignement Supérieur  
et de la Recherche Scientifique

Université de Sfax

École nationale d'électronique  
et des télécommunications de Sfax



**Ingénieur en :**  
Génie Systèmes Électroniques  
et Communication

**Option :**  
Systèmes Électroniques Intelligents

N° d'ordre : GEC-SEI-3-23-14

# MEMOIRE

*présenté à*

**L'École Nationale d'Électronique et des  
Télécommunications de Sfax**

*en vue de l'obtention du*

**Diplôme National d'Ingénieur en :  
Génie Systèmes Électroniques et Communication**

**Option :  
Systèmes Électroniques Intelligents**

*par*

**Ghazi ABDALLAH**

**Conception et Réalisation d'une application de Parsing des  
CVs multilingues via NLP et IA**

Soutenu le 04/10/2023, devant la commission d'examen :

Mme.	Emna Kallel	Présidente
Mme.	Manel HENTATI	Examinatrice
M.	Amir GARGOURI	Encadrant
Mme.	Nesrine SANEI	Encadrant industriel



---

# DEDICATION

À mes chers parents, qui ont toujours cru en moi et m'ont soutenu tout au long des périodes les plus difficiles de ma vie, pour leur dévouement, leur amour inconditionnel et leurs sacrifices pour faire de moi la personne que je suis aujourd'hui. À mes chères sœurs et à mon frère, qui n'ont cessé de croire en moi et de me soutenir. À toutes les personnes qui ont croisé ma route, celles qui sont toujours présentes et celles que le temps a emportées de ma vie. À ceux que je chéris et que je chérirai toujours, à tous mes amis qui m'ont soutenu et encouragé.

À vous tous,

Je dédie cet humble travail, le fruit de mes efforts et de mes sacrifices au cours de toutes ces années d'études, à vous tous. Qu'ils y trouvent la reconnaissance de leur soutien indéfectible et l'expression profonde de ma gratitude..

**Ghazi ABDALLAH**



---

# REMERCIEMENT

Je souhaite exprimer ma profonde gratitude envers toutes les personnes qui ont contribué leur soutien et leur aide tout au long de la réalisation de ce rapport. Mes remerciements vont particulièrement à mon encadrant à l'École Nationale d'Électronique et des Télécommunications de Sfax, Monsieur **Amir Gargouri**, pour sa disponibilité malgré ses nombreuses responsabilités académiques, sa patience, ses encouragements, son encadrement de qualité, et ses conseils avisés tout au long de notre projet.

Je souhaite également exprimer ma profonde reconnaissance envers mon encadrante professionnelle, Madame **Nessrine Sanei**, et toute l'équipe de Rec-inov, qui m'ont donné l'opportunité de faire partie de cette équipe et de participer à ce projet passionnant dans le domaine de l'intelligence artificielle. Leurs conseils et leur soutien tout au long de ces derniers mois ont été inestimables.

Enfin, je tiens à remercier chaleureusement les membres du jury qui ont accepté d'examiner et d'évaluer mon travail. J'espère que ce rapport répondra à leurs attentes en termes de clarté et de contenu.



---

# TABLE DES MATIÈRES

<b>LISTE DES FIGURES</b>	<b>vii</b>
<b>LISTE DES TABLEAUX</b>	<b>viii</b>
<b>LISTE DES ABRÉVIATIONS</b>	<b>ix</b>
<b>INTRODUCTION GÉNÉRALE</b>	<b>xi</b>
<b>1 Étude préalable</b>	<b>2</b>
1.1 INTRODUCTION . . . . .	3
1.2 Présentation de l'organisme d'accueil . . . . .	3
1.2.1 Présentation générale . . . . .	3
1.2.2 Domaines d'activité . . . . .	4
1.3 Contexte général du projet . . . . .	4
1.3.1 Cadre du projet . . . . .	4
1.3.2 E-recrutement . . . . .	5
1.3.3 Impact de l'IA au processus du recrutement . . . . .	6
1.4 Problématique et étude de l'existant . . . . .	6
1.4.1 Problématique . . . . .	6
1.4.2 Etude de l'existant . . . . .	7
1.4.2.1 A l'échelle national . . . . .	7
1.4.2.2 A l'échelle international . . . . .	8
1.4.2.3 Critique de l'existant . . . . .	9
1.5 Objectifs et Solution proposée . . . . .	9
1.6 Notions générales . . . . .	10
1.6.1 Intelligence artificielle . . . . .	10
1.6.2 Apprentissage automatique "Machine Learning" . . . . .	11
1.6.3 Apprentissage profond "Deep Learning" . . . . .	11
1.6.4 Traitement automatique du langage naturel "NLP" . . . . .	12

1.7	Méthodologie de travail . . . . .	12
1.7.1	Méthodes Agiles . . . . .	13
1.7.2	Méthode SCRUM . . . . .	13
1.7.3	Planification du projet . . . . .	14
1.8	CONCLUSION . . . . .	15
<b>2</b>	<b>Etude conceptuelle</b>	<b>17</b>
2.1	INTRODUCTION . . . . .	18
2.2	Spécification des besoins . . . . .	18
2.2.1	Identification des acteurs . . . . .	18
2.2.2	Besoins fonctionnels . . . . .	19
2.2.3	Besoins non fonctionnels . . . . .	19
2.3	Modélisation avec UML . . . . .	20
2.3.1	Diagramme des cas d'utilisation . . . . .	20
2.3.2	Diagramme de classes . . . . .	23
2.3.3	Diagramme de séquence . . . . .	24
2.3.3.1	Diagramme de séquence du cas d'utilisation « S'authentifier »	25
2.3.3.2	Diagramme de séquence du cas d'utilisation « Déposer CV »	25
2.3.3.3	Diagramme de séquence du cas d'utilisation « Extraction des informations des CVs » . . . . .	26
2.4	Architecture de l'application . . . . .	27
2.5	Environnement de travail . . . . .	28
2.5.1	Outils de développements matériels . . . . .	29
2.5.2	Outils de développements logiciels . . . . .	29
2.6	CONCLUSION . . . . .	31
<b>3</b>	<b>Implémentation et Réalisation</b>	<b>32</b>
3.1	INTRODUCTION . . . . .	33
3.2	Construction de la base de données . . . . .	33
3.2.1	Web Scraping . . . . .	33
3.2.1.1	Technologies de Web Scraping . . . . .	34
3.2.1.2	Choix de langage et des technologies du Web scraping . . . . .	35
3.2.2	Formation de la base de données . . . . .	36
3.2.3	Traitement de la base de données . . . . .	36
3.2.4	Extraction du contenu textuel . . . . .	37
3.2.4.1	Outils d'extraction du contenu contextuel . . . . .	37

3.2.4.2	Comparaison entre les outils d'extraction du contenu contextuel . . . . .	38
3.2.5	Nettoyage de la base de données . . . . .	39
3.2.6	Annotation de la base de données . . . . .	39
3.2.6.1	Outils d'annotation des données . . . . .	40
3.2.6.2	Etude comparative des outils d'annotation des données . . . .	41
3.3	Implémentation des modèles d'extraction . . . . .	43
3.3.1	Implémentation du modèle de la bibliothèque RegEx . . . . .	43
3.3.1.1	Principe de fonctionnement . . . . .	44
3.3.1.2	Résultats d'implémentation . . . . .	44
3.3.2	Implémentation du modèle "Layout Parser" . . . . .	45
3.3.2.1	Architecture du modèle Layout Parser . . . . .	45
3.3.2.2	Résultats d'implémentation . . . . .	46
3.3.3	Implémentation du modèle Donut . . . . .	46
3.3.3.1	Architecture du modèle Donut . . . . .	47
3.3.3.2	Résultats d'implémentation . . . . .	47
3.3.4	Implémentation du modèle SpaCy . . . . .	48
3.3.4.1	Architecture du modèle SpaCy . . . . .	48
3.3.4.2	Apprentissage du modèle . . . . .	50
3.3.4.3	Résultats d'implémentation . . . . .	52
3.4	Conception de l'interface WEB . . . . .	55
3.4.1	Communication entre Nodejs et Python . . . . .	55
3.4.1.1	Child Process . . . . .	55
3.4.1.2	Hugging Face Model Hub . . . . .	55
3.4.1.3	TensorFlow.js . . . . .	56
3.4.1.4	Solution Choisie . . . . .	56
3.4.2	Interfaces conçues . . . . .	56
3.5	Conclusion . . . . .	59
<b>CONCLUSION GÉNÉRALE</b>		<b>60</b>
<b>BIBLIOGRAPHIE</b>		<b>61</b>

# LISTE DES FIGURES

1.1	Logo de Rec-Inov . . . . .	3
1.2	Logo Adecco . . . . .	7
1.3	Logo Indeed . . . . .	8
1.4	Logo CandidateZip . . . . .	8
1.5	Différentes parties de la solution . . . . .	10
1.6	Méthodologie SCRUM . . . . .	14
1.7	Planification des tâches des sprints . . . . .	15
2.1	Diagramme des cas d'utilisation . . . . .	21
2.2	Diagramme de classes . . . . .	24
2.3	Diagramme de séquence «S'authentifier» . . . . .	25
2.4	Diagramme de séquence «Déposer CV» . . . . .	26
2.5	Diagramme de séquence «Extraction des informations des Cvs» . . . . .	27
2.6	Architecture détaillée de notre application . . . . .	28
3.1	Comparaison entre deux CVs avant et après le traitement d'image . . . . .	37
3.2	les étiquettes d'annotation . . . . .	40
3.3	Interface Label-studio . . . . .	42
3.4	Sortie de l'annotation . . . . .	43
3.5	Architecture d'un réseau de neurones convolutif . . . . .	45
3.6	Résultat de la détection des régions avec Layout Parser . . . . .	46
3.7	Architecture du modèle Donut . . . . .	47
3.8	Architecture du modèle SpaCy . . . . .	49
3.9	Fonctionnement d'apprentissage du modèle SpaCy . . . . .	51
3.10	Organigramme du processus de conception du modèle d'extraction . . . . .	52
3.11	Courbe de précision du modèle . . . . .	53
3.12	Courbe de perte du modèle . . . . .	54
3.13	Résultat d'implémentation du modèle SpaCy, (A) : Cv d'entrée, (B) : Résultat de l'extraction du modèle SpaCy sous forme JSON . . . . .	54
3.14	Interface d'inscription du candidat . . . . .	57

3.15 Interface remplissage du Profil du candidat . . . . .	57
3.16 Interface d'authentification . . . . .	57
3.17 Interface du dépôt du CV du candidat . . . . .	58
3.18 Interface d'extraction des informations . . . . .	58





---

## LISTE DES TABLEAUX

1.1	Etude comparative des systèmes existants en terme de précision d'extraction . .	9
1.2	Tableau des sprints . . . . .	15
2.1	Description du cas d'utilisation "S'authentifier" . . . . .	22
2.2	Description du cas d'utilisation "Déposer CV" . . . . .	22
2.3	Description du cas d'utilisation "Extraction des informations des Cvs" . . . . .	22
2.4	Description du cas d'utilisation "Administrer les candidats" . . . . .	23
3.1	Etude comparative des langages de programmation pour le Web Scraping . . .	35
3.2	Etude comparative entre les différentes méthodes d'extraction du contenu textuelle . . . . .	38
3.3	Etude comparative des outils d'annotation . . . . .	41
3.4	Etude comparative des modèles pré-entraînés du SpaCy . . . . .	53



---

# LISTE DES ABRÉVIATIONS

**API** Application Programming Interface

**CNN** Convolutional Neural Network

**CPU** Central Processing Unit

**CSS** Cascading Style Sheets

**DB** Data Base

**Docx** Microsoft Word 2007/2010/2013 document

**DL** Deep Learning

**GPU** Graphics Processing Unit

**HTML** HyperText Markup Language

**HTTP** Hypertext Transfer Protocol

**IA** Intelligence Artificielle

**IDE** Integrated Development Environment

**JSON** JavaScript Object Notation

**ML** Machine Learning

**NER** Named Entity Recognition

**NLP** Natural Language Processing

**OCR** Optical Character Recognition

**PDF** Portable Document Format

**RAM** Random Access Memory

**RegEx** Regular Expressions

**RE** Relation Extraction

**UML** Unified Modeling Language

**UI** User Interface



---

# INTRODUCTION GÉNÉRALE

Le recrutement joue un rôle vital dans l'amélioration de la durabilité et de la réussite des entreprises. Les départements des ressources humaines constituent un élément central au sein de toute organisation, ayant pour responsabilité principale de sélectionner les membres adéquats pour rejoindre l'équipe. Cette tâche requiert un investissement considérable en temps pour examiner et pré-sélectionner chaque candidature. Malheureusement, il peut arriver que malgré ces efforts, aucun candidat qualifié ne se démarque, ce qui peut parfois aboutir à l'embauche de candidats peu adaptés.

Dans cet environnement concurrentiel intense, les recruteurs font face à l'examen de centaines, voire de milliers de CV pour un seul poste à pourvoir. De plus, les postulants ont tendance à envoyer des candidatures en masse, parfois sans prêter attention aux détails du poste ou aux qualifications requises mentionnées dans l'offre. Lorsqu'un candidat envoie son CV pour un poste, il ne parvient pas toujours à mettre en évidence sa pertinence spécifique pour le rôle, ce qui entraîne un grand nombre de refus et engorge la boîte de réception des ressources humaines. Il pourrait sembler judicieux de simplifier les étapes et les entretiens pour accélérer le processus de recrutement. Cependant, cette approche risque potentiellement de conduire à des embauches hâtives de personnes ne correspondant pas parfaitement aux exigences du poste.

La présélection des candidats représente une phase cruciale pour choisir les candidats appropriés pour les offres d'emploi. Cette étape de recrutement se base essentiellement sur la lecture et l'analyse des CV des candidats, en se concentrant sur les éléments pertinents tels que l'éducation, l'expérience et les certifications pour évaluer la correspondance du profil avec l'offre d'emploi.

Les CV des candidats peuvent adopter différents styles et formes de conception, ce qui rend la tâche de présélection difficile pour le département des ressources humaines. C'est pourquoi il est nécessaire de rechercher des outils et des logiciels de recrutement intelligents pour accélérer cette étape tout en maintenant un haut niveau de qualité dans la présélection.

Dans ce contexte, notre projet de fin d'études présenté par REC-INOV vise à développer une application basée sur l'intelligence artificielle pour l'analyse des CV multi-langues des candidats. Ce rapport est divisé en trois chapitres réparties comme suit :

- **Le chapitre 1** expose le contexte du projet et présente l'organisme qui accueille notre projet, il examine l'état actuel de la technologie en comparant les différentes solutions existantes pour identifier le problème à résoudre et proposer des solutions possibles. Ce chapitre se termine par la présentation de la méthodologie de travail et la planification du projet selon les principes de Scrum.

- **Le chapitre 2** se concentre sur l'analyse des besoins et les différentes architectures envisagées pour notre projet. Une attention particulière est portée aux diagrammes de cas d'utilisation, de séquences et de classes.

- **Le chapitre 3** décrit la dernière phase du processus de développement de notre solution. Après avoir spécifié nos besoins et notre conception, nous présentons également les différentes étapes pour la conception et le développement du projet.

En fin, notre rapport s'achève par une conclusion globale et quelques perspectives.

---

# Étude préalable

## Sommaire

---

<b>1.1</b>	<b>INTRODUCTION . . . . .</b>	<b>3</b>
<b>1.2</b>	<b>Présentation de l'organisme d'accueil . . . . .</b>	<b>3</b>
1.2.1	Présentation générale . . . . .	3
1.2.2	Domaines d'activité . . . . .	4
<b>1.3</b>	<b>Contexte général du projet . . . . .</b>	<b>4</b>
1.3.1	Cadre du projet . . . . .	4
1.3.2	E-recrutement . . . . .	5
1.3.3	Impact de l'IA au processus du recrutement . . . . .	6
<b>1.4</b>	<b>Problématique et étude de l'existant . . . . .</b>	<b>6</b>
1.4.1	Problématique . . . . .	6
1.4.2	Etude de l'existant . . . . .	7
<b>1.5</b>	<b>Objectifs et Solution proposée . . . . .</b>	<b>9</b>
<b>1.6</b>	<b>Notions générales . . . . .</b>	<b>10</b>
1.6.1	Intelligence artificielle . . . . .	10
1.6.2	Apprentissage automatique "Machine Learning" . . . . .	11
1.6.3	Apprentissage profond "Deep Learning" . . . . .	11
1.6.4	Traitement automatique du langage naturel "NLP" . . . . .	12
<b>1.7</b>	<b>Méthodologie de travail . . . . .</b>	<b>12</b>
1.7.1	Méthodes Agiles . . . . .	13
1.7.2	Méthode SCRUM . . . . .	13
1.7.3	Planification du projet . . . . .	14
<b>1.8</b>	<b>CONCLUSION . . . . .</b>	<b>15</b>

---

## 1.1 INTRODUCTION

Ce chapitre se concentre sur l'exposition de l'organisme d'accueil, suivie d'une mise en contexte générale du projet. Par la suite, nous conduirons une analyse approfondie des applications déjà présentes sur le marché. Nous terminons ce chapitre par une description détaillée de la planification prévue tout au long de la progression du projet.

## 1.2 Présentation de l'organisme d'accueil

Dans cette partie, nous aborderons l'introduction de l'organisme d'accueil par une vue d'ensemble de l'entreprise, suivi de près par ses objectifs, et enfin, nous explorerons ses différents domaines d'activité.

### 1.2.1 Présentation générale

« Rec-Inov » constitue une start-up labellisée qui se spécialise dans le secteur du recrutement en ligne. Implantée à Sfax, Tunisie, la société « Rec-Inov » a vu le jour en 2020 et s'intègre au sein de l'industrie des prestations de services liés à la technologie de l'information. L'effectif total se compose de 12 membres. Vous trouverez ci-dessous le logo de la start-up « Rec-Inov ».



FIGURE 1.1 – Logo de Rec-Inov

« Rec-Inov » poursuit deux objectifs fondamentaux, à savoir l'innovation et la réalisation.

- L'innovation se traduit par la mise en place de fondements et de perspectives pour l'équipe, favorisant l'élaboration de solutions novatrices en réponse aux besoins des clients actuels et futurs.

- La réalisation se matérialise par la création d'un environnement propice, permettant à l'équipe, répartie à l'échelle mondiale, de concrétiser leurs idées et de tenir les engagements envers leurs clients.

### 1.2.2 Domaines d'activité

Les domaines d'activité de la société « Rec-Inov » couvre un large spectre de domaines, notamment :

- **L'intelligence artificielle** : Cette dimension vise à garantir un matching intelligent, en ciblant précisément les profils les plus adaptés ainsi que l'extraction des informations pertinentes des CVs .

- **MERN Stack** : Le package MERN, composé des technologies MongoDB, Express, ReactJS et NodeJS, permet de concevoir une plateforme de recrutement en ligne, couvrant à la fois les aspects Frontend et Backend.

- **Blockchain** : Cette expertise englobe les domaines du Smart Contract et de la signature électronique, pour automatiser et sécuriser les informations liées aux contrats.

## 1.3 Contexte général du projet

### 1.3.1 Cadre du projet

Le projet achevé pendant cette période de stage de six mois répond aux nécessités de l'entreprise "Rec-Inov" qui cherche à incorporer des solutions d'intelligence artificielle sous le titre : "Élaboration et Réalisation d'une application pour le parsing de CVs multilingues de candidats via NLP et IA". Ce travail s'inscrit dans le contexte du projet de fin d'études au sein



de l'entreprise "Rec-Inov", visant l'obtention du diplôme d'ingénieur en génie électronique et communication à l'École Nationale d'Électronique et de Télécommunications de Sfax.

### 1.3.2 E-recrutement

L'e-recrutement, également connu sous les noms de recrutement électronique ou recrutement en ligne, constitue une véritable révolution dans le domaine du recrutement. Il se réfère à la dématérialisation des processus de recrutement grâce à Internet. Cette approche novatrice gagne en popularité au sein des entreprises, car elle offre un moyen plus rapide et pratique de définir et de gérer les processus de recrutement.

L'e-recrutement présente également de nombreux avantages pour les candidats, leur permettant d'augmenter le nombre de leurs candidatures grâce à la multitude d'offres disponibles en ligne.

Dans le domaine de l'e-recrutement, certains **avantages** sont à noter, notamment :

- Simplification des procédures de recrutement. Offre d'une solution économique, tant pour les recruteurs que pour les candidats, sans nécessité de déplacements physiques.
- Amélioration de la sélection des candidats possédant les compétences requises.
- Disponibilité 24 heures sur 24, 7 jours sur 7, d'une collection de CVs en ligne.

Cependant, il est important de prendre en considération certains **inconvenients** :

- En raison de la facilité de diffusion des offres d'emploi, cela peut entraîner un afflux important de candidatures à traiter.
- Du fait de la simplicité de candidature en réponse aux offres d'emploi, il peut y avoir une augmentation des candidatures en masse, ce qui peut conduire à une variété de candidatures, certaines étant moins qualifiées que d'autres.

### **1.3.3 Impact de l'IA au processus du recrutement**

L'impact des avancées technologiques sur les pratiques de recrutement est indéniable. Les entreprises adoptent de nouvelles approches pour sélectionner et embaucher des candidats, grâce à la collecte de vastes quantités de données sur ces derniers.

L'utilisation du big data et de l'intelligence artificielle permet aux entreprises de prendre des décisions plus éclairées, réduisant ainsi les erreurs de recrutement. Cette tendance vers l'intégration croissante du big data et de l'IA dans le recrutement semble inarrêtable.

Les plateformes de recrutement modernes adoptent des outils d'intelligence artificielle et des processus de langage naturel. Ces technologies permettent aux recruteurs d'obtenir des informations plus approfondies sur les candidats, et ce, à une vitesse qui était autrefois inimaginable.

## **1.4 Problématique et étude de l'existant**

Cette section débute par une exposition de la problématique, à laquelle succède une étude de l'existant.

### **1.4.1 Problématique**

L'e-recrutement couvre l'intégralité du processus de recrutement, depuis la publication de l'offre d'emploi jusqu'à la réception des candidatures, en incluant les entretiens d'embauche indispensables.

La présélection des candidats qualifiés et adaptés à une offre d'emploi donnée est une étape cruciale pour garantir le succès du processus de recrutement. Cette première étape implique l'analyse minutieuse des informations contenues dans les CV des candidats, en se concentrant particulièrement sur les sections relatives à leur éducation, leur expérience professionnelle et leurs certifications.

Cependant, de nos jours, les départements des ressources humaines sont souvent confrontés à un afflux massif de candidatures, parfois des centaines voire des milliers, pour un seul poste vacant ! Cette diversité de formats et de modèles de CV ainsi que sa langue d'écriture rend l'analyse du contenu de ces documents complexe et coûteuse, tant en termes de ressources matérielles que de temps.

Il est donc impératif de disposer d'un outil intelligent capable de structurer et d'ordonner ces informations de manière à les rendre rapidement analysables.

### 1.4.2 Etude de l'existant

Il est essentiel de réaliser une étude approfondie des solutions déjà présentes sur le marché avant de lancer un projet. Par conséquent, nous avons entrepris une analyse des solutions existantes pour l'extraction des informations des CVs des candidats.

#### 1.4.2.1 A l'échelle national

A l'échelle national, on trouve une plateforme en ligne qui assure cette fonctionnalité. On cite par exemple l'une des solutions les plus actives en Tunisie :

- **Adecco** : Adecco Tunisie RH, ou Adecco Tunisie Ressources Humaines, est une filiale d'Adecco Group, l'une des plus grandes entreprises de services en ressources humaines au monde. Adecco Tunisie RH est spécialisée dans le domaine du recrutement, de la gestion intérimaire, et des services RH en Tunisie.



FIGURE 1.2 – Logo Adecco

### 1.4.2.2 A l'échelle international

De nombreux logiciels de recrutement sont disponibles sur le marché, et certains d'entre eux proposent des fonctionnalités avancées pour extraire les informations pertinents des candidats. Parmi ces solutions, on peut identifier :

- **Indeed** : Indeed est un site web et un moteur de recherche d'emplois en ligne qui agit comme une plateforme mondiale de recrutement. Il permet aux chercheurs d'emploi de rechercher des offres d'emploi en fonction de leur emplacement géographique, de leur domaine de compétence, de leur niveau d'expérience et d'autres critères pertinents. Les employeurs peuvent également utiliser Indeed pour extraire les informations pertinents de leurs Cvs de manière structuré.



FIGURE 1.3 – Logo Indeed

- **CandidateZip** : CandidateZip est une plateforme web qui agit avec les recruteurs qui ont besoin de gérer facilement les CV et les offres d'emploi dans le cadre de leur processus d'embauche, sans aucune difficulté technique. Ils peuvent automatiser le processus et sélectionner rapidement le candidat le plus pertinent. Notre analyseur de CV automatise vos processus d'embauche et vous aide à trouver le candidat idéal.



FIGURE 1.4 – Logo CandidateZip

### 1.4.2.3 Critique de l'existant

Malgré la variété des services proposés par les solutions mentionnées précédemment, elles présentent plusieurs lacunes en ce qui concerne l'extraction des informations pertinentes des CVs. En se basant sur un ensemble des exemples des différents Cvs réelles des candidats, Voici une comparaison illustré par le tableau 1.1 ci-dessous :

**TABLE 1.1 – Etude comparative des systèmes existants en terme de précision d'extraction**

<b>Solution</b>	<b>Education</b>	<b>Expérience</b>	<b>Certification</b>
<b>Adecco</b>	0 %	10%	0 %
<b>Indeed</b>	45%	55 %	0%
<b>CandidateZip</b>	60 %	60%	0%

Après avoir effectué cette comparaison, il est évident que ces outils présentent des lacunes en ce qui concerne l'extraction des informations pertinentes des différentes sections d'un CV de candidat. Plus précisément, ils ont du mal à extraire de manière précise et complète les données des sections importantes d'un CV, et ils ne parviennent généralement pas à extraire correctement la section relative aux certifications du candidat.

## 1.5 Objectifs et Solution proposée

L'objectif central de ce projet est de mettre en œuvre des solutions avancées pour extraire avec précision les informations pertinentes des CVs des candidats. Ces données seront ensuite présentées de manière structurée en utilisant des technologies de pointe. Cette approche vise à intégrer harmonieusement ces informations dans la plate-forme de recrutement actuelle de Rec-Inov, tout en les associant au système de correspondance déjà en place.

Pour atteindre cet objectif, nous nous concentrons sur plusieurs tâches essentielles, notamment :

- **Téléchargement des CVs** : Nous envisageons de permettre aux utilisateurs de télécharger des CVs au format PDF, image ou Docx. Cette fonctionnalité simplifiera la collecte des informations pertinentes.

- **Extraction automatisée** : Nous développons un système automatisé basé sur l'intelligence artificielle (IA) pour extraire de manière précise les informations cruciales contenues dans les CVs. Il s'agit principalement des trois sections essentielles : éducation, expériences professionnelles et certifications. L'IA sera utilisée pour identifier et extraire ces données de manière cohérente et fiable.

- **Stockage et affichage structuré** : Les informations extraites seront sauvegardées de manière organisée dans une base de données ainsi que affichées sur une page web. Cela permettra un accès rapide et ordonné aux données, facilitant ainsi leur utilisation dans le processus de recrutement.

La figure 1.5 ci-dessous montre les différentes parties qui ont abouti à notre solution.

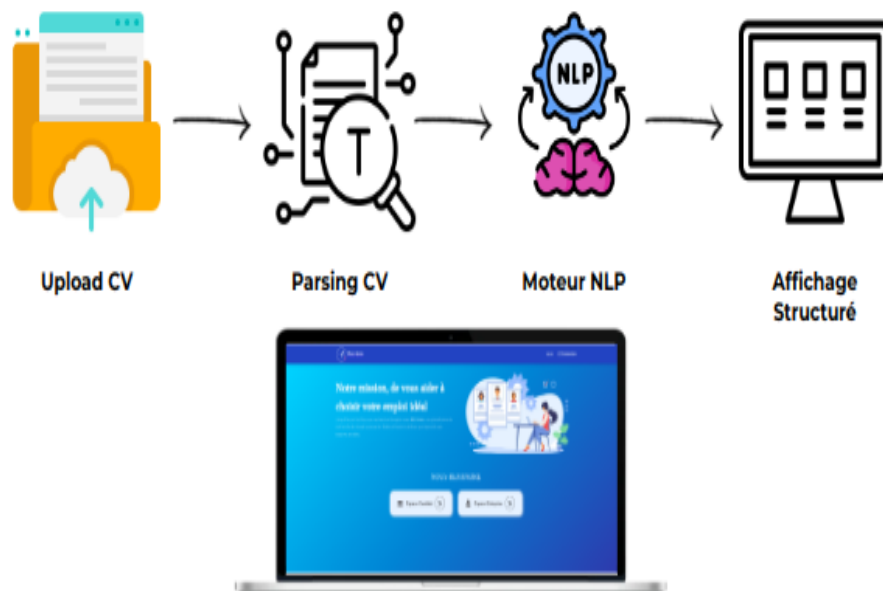


FIGURE 1.5 – Différentes parties de la solution

## 1.6 Notions générales

### 1.6.1 Intelligence artificielle

L'intelligence artificielle [1] est l'intelligence démontrée par les machines, par opposition à l'intelligence humaine ou animale. Elle est basée sur la création et l'application d'algorithmes exécutés dans un environnement informatique dynamique. L'objectif de l'IA est de permettre

aux ordinateurs de penser et d'agir comme des êtres humains, en imitant les capacités cognitives d'un être humain. Les tâches accomplies par l'IA incluent la reconnaissance vocale, la vision par ordinateur, la traduction entre langues naturelles, ainsi que d'autres mappages d'entrées. Pour se rapprocher le plus possible du comportement humain, l'IA a besoin d'une quantité de données et d'une capacité de traitement élevées. Le terme "intelligence artificielle" a été créé en 1955 par John McCarthy. Les machines capables d'agir de façon intelligente, d'assimiler des concepts abstraits et d'avoir une véritable conscience proche de celle des êtres humains appartiennent à la catégorie de l'IA forte.

### **1.6.2 Apprentissage automatique "Machine Learning"**

L'apprentissage automatique [2] (machine learning en anglais) est un champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'apprendre à partir de données, via des modèles mathématiques. Plus précisément, il s'agit du procédé par lequel les informations pertinentes sont tirées d'un ensemble de données d'entraînement. Les algorithmes d'apprentissage automatique posent des problèmes d'explicabilité globale du système. Il existe différents types d'apprentissage automatique : le supervisé, le non-supervisé et celui par renforcement. Les trois étapes essentielles de l'apprentissage automatique sont la récolte des données, la réconciliation (Data Wrangling) et l'entraînement du modèle.

### **1.6.3 Apprentissage profond "Deep Learning"**

L'apprentissage profond [3], également connu sous le nom de deep learning en anglais, est une technique d'apprentissage automatique qui repose sur des réseaux de neurones artificiels. Les réseaux de neurones sont des algorithmes capables d'imiter les comportements du cerveau humain. Dans le deep learning, ces réseaux sont composés de dizaines voire de centaines de couches de neurones, chacune recevant et interprétant les informations de la couche précédente. Cette technique permet d'apporter une plus grande complexité à l'établissement des règles. Les modèles de deep learning ont tendance à bien fonctionner avec une grande quantité de données.

### 1.6.4 Traitement automatique du langage naturel "NLP"

Le Traitement Automatique du Langage Naturel (TALN) [4], également connu sous le nom de Natural Language Processing (NLP en anglais), représente une discipline au sein de l'intelligence artificielle qui se concentre sur la capacité des machines à comprendre, manipuler et générer le langage naturel. Effectivement, cette discipline permet aux ordinateurs de saisir, interpréter et agir sur le langage humain en utilisant des méthodes issues de l'intelligence artificielle, de la linguistique informatique et de l'apprentissage automatique. Le TALN résout diverses problématiques, notamment la classification de texte, la reconnaissance des entités nommées dans le texte, la création automatique de résumés, la traduction automatique, les chatbots, les assistants intelligents, et bien d'autres applications. Les techniques exploitées dans le TALN englobent l'analyse syntaxique pour comprendre la structure grammaticale des phrases, l'analyse sémantique pour saisir le sens des mots et des phrases, l'analyse des émotions pour détecter les sentiments exprimés, ainsi que la traduction automatique pour convertir un texte d'une langue à une autre.

## 1.7 Méthodologie de travail

La réalisation d'un tel projet, quels que soient son envergure et ses objectifs, requiert l'établissement d'une structure de planification à travers toutes les phases de son cycle de vie. C'est à partir de ce besoin que s'est forgée la notion de méthodologie de travail. Malgré la variété des approches en analyse et en conception, on peut les classer en quatre catégories distinctes :

- Les méthodes cartésiennes ou fonctionnelles, comme la méthode SADT (Structured Analysis and Design Technique).
- Les méthodes systémiques, telles que MERISE et AXIAL.
- Les méthodes orientées objet, parmi lesquelles se détachent trois principales : la méthode OMT de Rumbaugh, la méthode BOOCH'93 de Booch et la méthode OOSE de Jacobson.
- Les méthodes agiles, dont font partie XP (eXtreme Programming), SCRUM, et autres.



### 1.7.1 Méthodes Agiles

Les méthodes agiles sont des approches méthodologiques principalement conçues pour la gestion de projets informatiques. Elles se traduisent par une augmentation de la productivité et un avantage compétitif pour les clients ainsi que les prestataires. L'essence des méthodes agiles repose sur des valeurs énoncées dans le Manifeste Agile :

- Donner la priorité aux individus et aux interactions plutôt qu'aux procédures et aux outils.
- Privilégier les applications opérationnelles plutôt qu'une documentation exhaustive.
- Encourager la collaboration avec le client plutôt que la négociation contractuelle.
- Favoriser l'acceptation du changement au lieu d'une planification rigide.

Ces quatre valeurs se posent en opposition avec les pratiques fréquemment associées aux méthodes traditionnelles : emphase sur les processus et les outils, mise en avant de la documentation, stricte adhésion au contrat initial et planification rigide.

### 1.7.2 Méthode SCRUM

Nous avons opté la méthode SCRUM [5] pour la mise en œuvre de notre projet, car elle présente plusieurs avantages par rapport aux autres approches de développement. En effet, elle garantit un déroulement harmonieux du projet au sein de l'équipe.

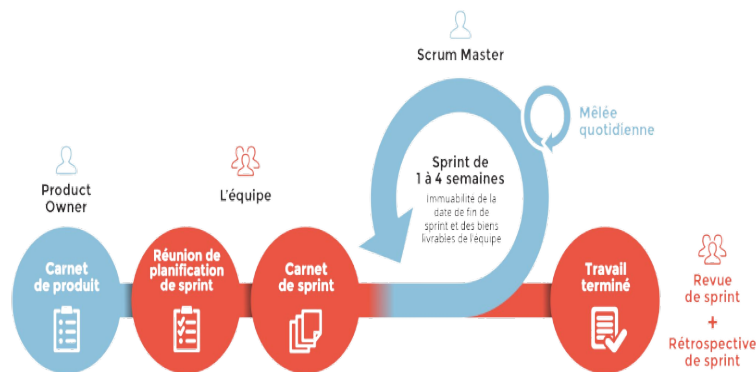
Scrum, une approche agile de développement, se focalise sur les projets informatiques avec des ressources constamment actualisées. Son nom est inspiré du monde du rugby (SCRUM = mêlée). Le principe sous-jacent est de rester toujours prêt à réajuster le projet au fur et à mesure de son avancée. Cette approche se veut dynamique et participative dans la gestion de projet, la mêlée de rugby étant une phase clé pour relancer le jeu sur de nouvelles bases. La métaphore de la mêlée est relayée dans la méthode Scrum.

Les intervenants majeurs dans le cycle SCRUM sont :

- **Le Propriétaire du Produit** : il incarne le client et reste en contact avec le Scrum Master. Dans notre contexte, il s'agit de Mme Jamila KHLIF.

- **Le Scrum Master** : il est en charge de l'application quotidienne des principes Scrum, créant un environnement de travail optimal pour l'équipe. Dans notre cas, Mme Nesrine SANEI assume ce rôle.

- **L'équipe de travail** : regroupant de deux à neuf membres, chaque individu doit, lors des sessions Scrum, identifier les tâches achevées et celles en cours dans un « backlog de produits » propre à chaque sprint. Notre équipe est constituée des employés et stagiaires. Le processus de la méthodologie Scrum est illustré dans la figure 1.6 suivante.



**FIGURE 1.6 – Méthodologie SCRUM**

### 1.7.3 Planification du projet

En accord avec les principes de la méthodologie Scrum, nous avons organisé notre projet en suivant les sprints et leurs tâches tels que décrits dans le tableau 1.2 ainsi que dans la figure 1.7.

TABLE 1.2 – Tableau des sprints

Sprint	Description	Durée	Priorité
0	Etude de l'art	2 semaines	Haute
1	Formation et conversion de la base du données	4 semaines	Haute
2	Annotation de la base du données	4 semaines	Haute
3	Conception et réalisation du modèle IA pour extraire les informations pertinents des CVs	4 semaines	Haute
4	-Développement des interfaces Web de login et d'inscription des candidats. -Développement d'une interface pour le téléchargement des Cvs	2 semaines	Moyenne
5	Déploiement du modèle et affichage des résultats dans une interface web	2 semaines	Haute
6	-Développement de l'interfaces d'affichage des informations des candidats	2 semaines	Haute
7	Amélioration	4 semaines	Haute

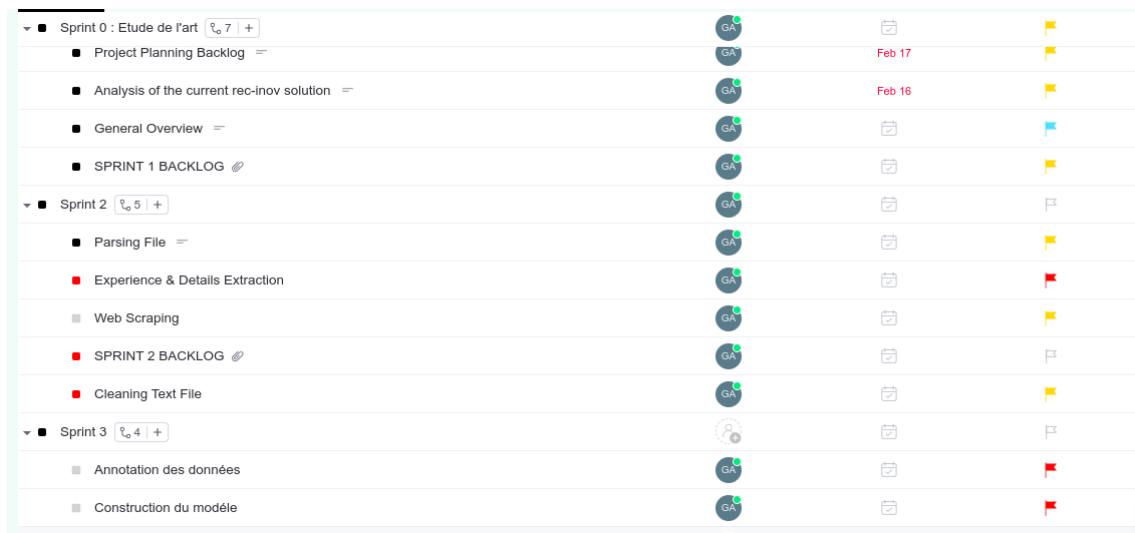


FIGURE 1.7 – Planification des tâches des sprints

## 1.8 CONCLUSION

Dans ce premier chapitre introductif, nous avons exposé l'entreprise qui accueille notre projet de fin d'études. Ensuite, nous avons fourni un aperçu du cadre de notre projet en mettant en évidence quelques termes clés. Par la suite, nous avons procédé à une analyse et à une évaluation des solutions déjà existantes. En conséquence, nous avons formulé notre proposi-

tion de résolution et nous avons identifié les défis à relever, afin de détailler notre approche méthodologique et la planification du déroulement du projet.

Dans le chapitre à venir, nous abordons les exigences spécifiques de notre projet et nous procédons à une étude conceptuelle approfondie.

---

# Etude conceptuelle

## Sommaire

---

<b>2.1</b>	<b>INTRODUCTION . . . . .</b>	<b>18</b>
<b>2.2</b>	<b>Spécification des besoins . . . . .</b>	<b>18</b>
2.2.1	Identification des acteurs . . . . .	18
2.2.2	Besoins fonctionnels . . . . .	19
2.2.3	Besoins non fonctionnels . . . . .	19
<b>2.3</b>	<b>Modélisation avec UML . . . . .</b>	<b>20</b>
2.3.1	Diagramme des cas d'utilisation . . . . .	20
2.3.2	Diagramme de classes . . . . .	23
2.3.3	Diagramme de séquence . . . . .	24
<b>2.4</b>	<b>Architecture de l'application . . . . .</b>	<b>27</b>
<b>2.5</b>	<b>Environnement de travail . . . . .</b>	<b>28</b>
2.5.1	Outils de développements matériels . . . . .	29
2.5.2	Outils de développements logiciels . . . . .	29
<b>2.6</b>	<b>CONCLUSION . . . . .</b>	<b>31</b>

---

## 2.1 INTRODUCTION

La qualité du point de départ joue un rôle déterminant dans la réussite de toute étude. Lancer un tel à partir de zéro, sans une architecture préétablie, implique, entre autres, la nécessité d'identifier les besoins et de définir la solution adéquate. C'est dans cette optique que nous avons élaboré ce chapitre pour conduire une analyse approfondie de notre application.

Ce chapitre abrite une analyse fonctionnelle qui requiert l'identification des éléments constitutifs du projet tels que les acteurs et les exigences fonctionnelles. Cela est complété par les exigences non fonctionnelles ainsi que les divers scénarios d'utilisation. Subséquemment, nous présentons en détail la conception du projet, mettant en avant ses composants essentiels, ainsi que les scénarios généraux et spécifiques qui le caractérisent.

## 2.2 Spécification des besoins

Dans cette section, nous allons présenter les parties prenantes du système ainsi que les besoins fonctionnels et non fonctionnels de notre application.

### 2.2.1 Identification des acteurs

Les parties prenantes actives au sein de l'application se déclinent comme suit :

- **Le Recruteur (Administrateur) :** Cette entité incarne un responsable des ressources humaines ou toute autre personne ayant pour but de repérer des collaborateurs dans les domaines de l'informatique, du commerce et des affaires. Ce rôle lui confère l'accès pour consulter et gérer les informations de l'ensemble des candidats.
- **Le Candidat (Utilisateur) :** Ce participant représente une personne en quête d'opportunités professionnelles dans les secteurs mentionnés précédemment. Il est en mesure d'effectuer l'envoi de son CV pour qu'il soit analysé (parsé).

### 2.2.2 Besoins fonctionnels

Les besoins fonctionnels définissent les différentes fonctionnalités qu'un logiciel doit offrir à ses utilisateurs afin de répondre à des besoins spécifiques. Voici la description des différents besoins fonctionnels de notre projet :

- Télécharger des CVs aux formats autorisés.
- Effectuer l'analyse syntaxique des CVs.
- Extraire automatiquement les informations pertinentes.
- Effectuer un stockage structuré des informations extraites.

Maintenant, nous exposons les besoins fonctionnels de chaque acteur :

#### **Candidat :**

- Créer son compte.
- Se connecter à son compte.
- Compléter son profil.
- Télécharger son CV .
- Consulter les informations pertinentes de son CV.
- Modifier ou supprimer les résultats obtenues.

#### **Administrateur :**

- Se Connecter à son compte.
- Gérer les profils des candidats ( ajouter, modifier, supprimer).

### 2.2.3 Besoins non fonctionnels

Cette application doit répondre aux besoins non fonctionnels, qui englobent les exigences garantissant le bon déroulement de l'application. Les besoins non fonctionnels associés à notre application sont les suivants :

- Sécurité : L'authentification est essentielle pour permettre à l'utilisateur de gérer ou de modifier les données appropriées.
- Performance : L'application doit présenter une réactivité optimale, c'est-à-dire que le temps de réponse doit être minutieusement optimisé.
- Modularité : la solution doit être modulaire pour garantir la souplesse et l'évolutivité.

## 2.3 Modélisation avec UML

La modélisation revêt une importance cruciale dans le cycle de vie de tout produit. Pour ce faire, il est essentiel de sélectionner la méthode de modélisation qui correspond le mieux à nos besoins et exigences dans le but de révolutionner le processus de recrutement.

En vue de satisfaire ces impératifs, notre choix s'est porté sur l'utilisation du langage UML (Unified Modeling Language ou Langage de Modélisation Unifiée en français) pour concevoir la modélisation de notre solution. L'approche UML s'intègre harmonieusement dans le cadre de la modélisation applicative orientée objet. Grâce à ses divers schémas, il offre une flexibilité remarquable qui permet d'appréhender les multiples facettes de l'application. Notre préférence pour ce langage est également justifiée par le fait qu'UML est devenu la norme de modélisation adoptée par l'ensemble des applications orientées objet.

L'UML [6] représente un langage graphique de modélisation informatique. De nos jours, il fait office de référence pour la modélisation objet et la programmation orientée objet. Ce langage rassemble des éléments modélisés du monde réel (architectures, objets, individus, signaux, parties du corps, etc.) et virtuel (temps, prix, compétences, etc.) en une gamme d'entités informatiques baptisées "objets" .

### 2.3.1 Diagramme des cas d'utilisation

Un diagramme de cas d'utilisation est un outil qui permet de visualiser les fonctionnalités d'un système du point de vue de l'utilisateur, également appelé "acteur" en UML. Il s'agit de la première étape dans l'analyse UML, permettant de :



- Modéliser les besoins des utilisateurs.
- Identifier les principales fonctionnalités et les limites du système.
- Représenter les interactions entre le système et ses utilisateurs.

Le diagramme de cas d'utilisation global illustré par la figure 2.1, expose les principales fonctionnalités que doit offrir notre application au candidat et à l'admin .

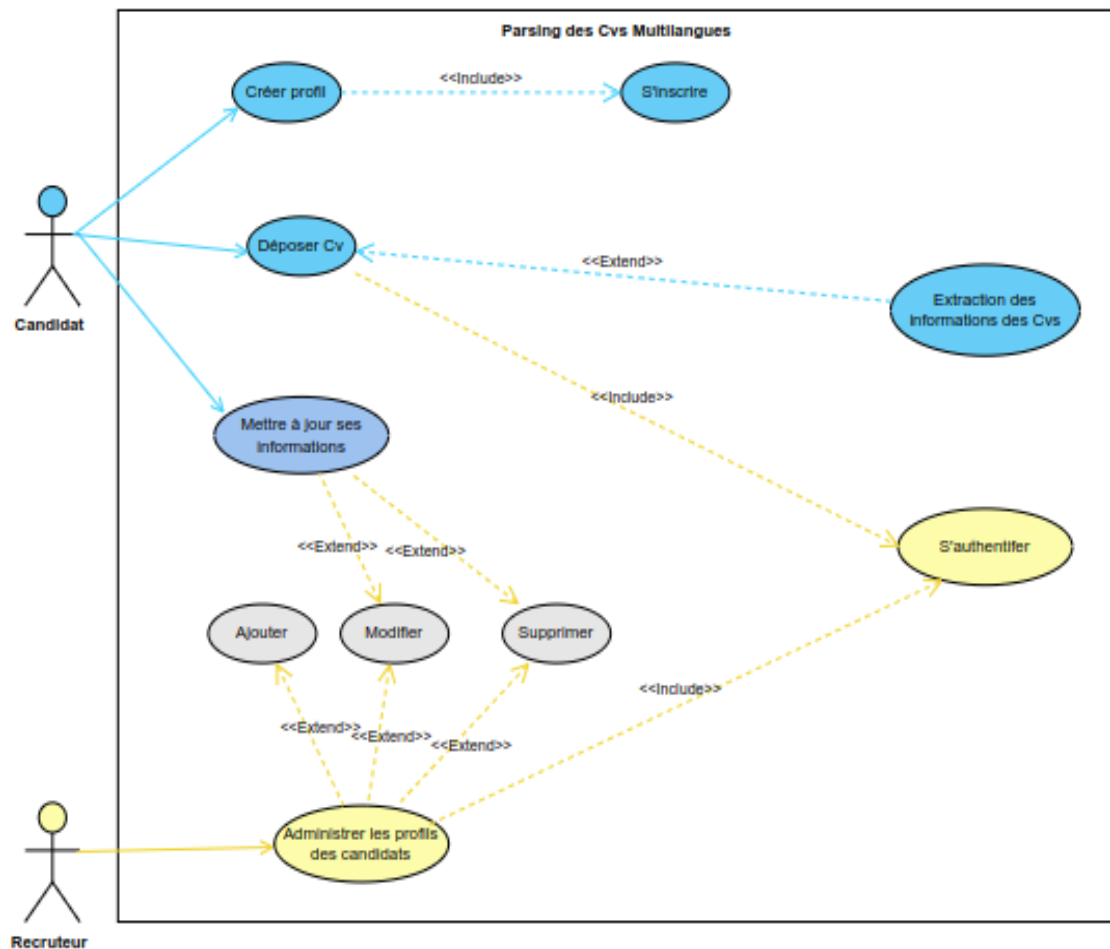


FIGURE 2.1 – Diagramme des cas d'utilisation

Ensuite, nous allons procéder à une analyse détaillée des cas d'utilisation les plus importants. Le tableau 2.1 ci-dessous présente une description de cas d'utilisation « **S'authentifier** ».

**TABLE 2.1 – Description du cas d'utilisation "S'authentifier"**

Cas d'utilisation	S'authentifier
Acteur	- Candidat.
Conditions	- Le nom du candidat ainsi que son mot de passe doivent être corrects.
Description	-Le candidat doit saisir son nom et son mot de passe pour accéder à l'application.
Résultat	- Accès à l'application.

Le candidat dispose la possibilité de déposer son CV. Le tableau 2.2 suivant décrit le cas d'utilisation « **Déposer CV** ».

**TABLE 2.2 – Description du cas d'utilisation "Déposer CV"**

Cas d'utilisation	Déposer CV
Acteur	- Candidat
Conditions	- Le candidat doit s'authentifier - Le Cv doit être sous format Docx, image ou PDF. - La taille du Cv ne doit pas dépasser 5 MB.
Description	- Le candidat doit importer son Cv en respectant les conditions prédéfinis.
Résultat	- Dépôt du Cv du candidat.

Le candidat peut déclencher l'extraction des informations pertinentes de son CV déposé en cliquant sur le bouton "Parse". Ces données sont ensuite enregistrées dans la base de données et affichées sur l'interface.

Le tableau suivant 2.3 décrit le cas d'utilisation « **Extraction des informations des Cvs** ».

**TABLE 2.3 – Description du cas d'utilisation "Extraction des informations des Cvs"**

Cas d'utilisation	Parse CV
Acteur	- Candidat
Conditions	- Le candidat doit s'authentifier. - Le candidat doit déposer son Cv.
Description	- L'utilisateur clique sur le bouton "parse" pour extraire les informations pertinentes. - Le candidat consulte les résultats obtenus.
Résultat	- Les informations pertinentes sont extraites de manière structurée à partir du CV du candidat.

Les informations pertinentes extraites des CV des candidats ainsi que leurs informations sont, ensuite, mises à la disposition du recruteur. Le tableau 2.4 décrit le cas d'utilisation « **Administrer les profils des candidats** »

**TABLE 2.4 – Description du cas d'utilisation "Administrer les candidats"**

<b>Cas d'utilisation</b>	<b>Administrer les candidats</b>
<b>Acteur</b>	- Recruteur.
<b>Conditions</b>	- Le recruteur doit s'authentifier.
<b>Description</b>	<ul style="list-style-type: none"><li>- Le recruteur consulte les informations stockées dans la base des données.</li><li>- Le recruteur ajoute les profils des candidats.</li><li>- Le recruteur modifie les profils des candidats.</li><li>- Le recruteur supprime les profils des candidats.</li></ul>
<b>Résultat</b>	- Le recruteur gère les profils des candidats.

### **2.3.2 Diagramme de classes**

Les diagrammes de classes sont parmi les outils les plus essentiels en développement orienté objet. Ils sont utilisés pour définir la structure des entités manipulées par les utilisateurs dans un système. En conception, ces diagrammes servent à représenter la structure sous-jacente d'un code orienté objet. La figure 2.2 montre le diagramme de classes de notre application.

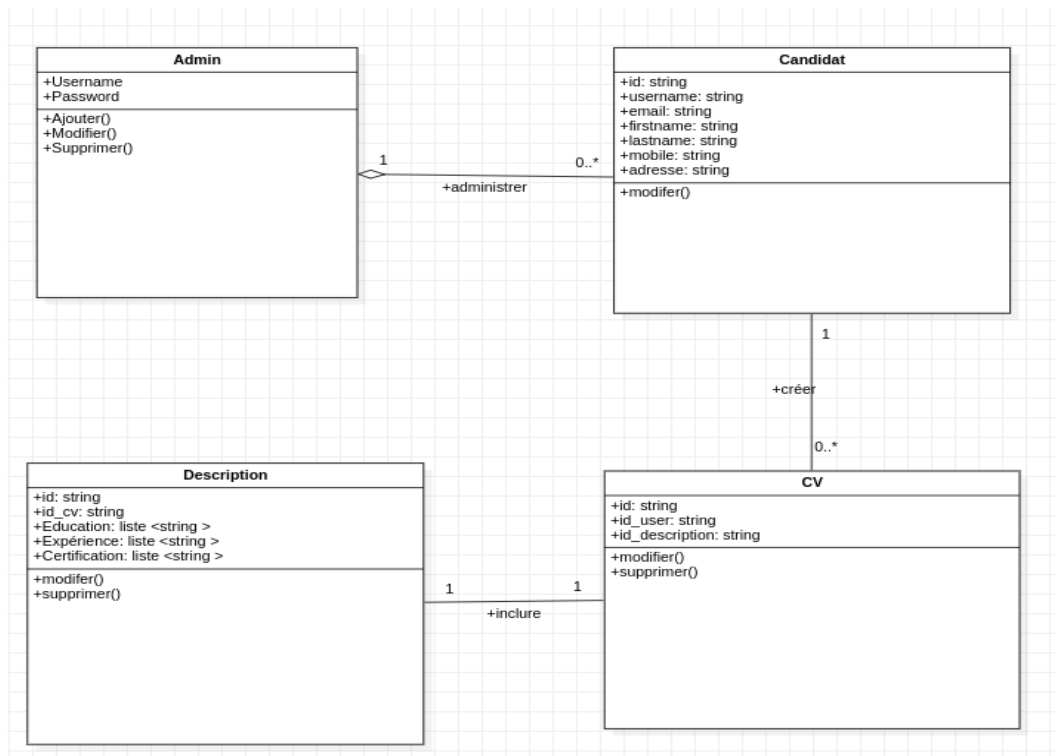


FIGURE 2.2 – Diagramme de classes

### 2.3.3 Diagramme de séquence

Un diagramme de séquence est une représentation graphique des interactions entre les objets d'un système, organisées chronologiquement. Dans cette section, nous exposons les diagrammes de séquence cruciaux de notre application.

Chaque diagramme de séquence est composé de deux dimensions principales :

- La dimension temporelle : elle est représentée par un axe vertical qui illustre la progression ou la durée des actions dans le temps.
- La dimension des objets : cette dimension est représentée par un axe horizontal qui met en évidence l'implication et les interactions entre les différents éléments du système.

### 2.3.3.1 Diagramme de séquence du cas d'utilisation « S'authentifier »

La première étape à faire lors de l'ouverture de notre application est l'authentification de l'utilisateur. La figure 2.3 montre le diagramme de séquence «S'authentifier».

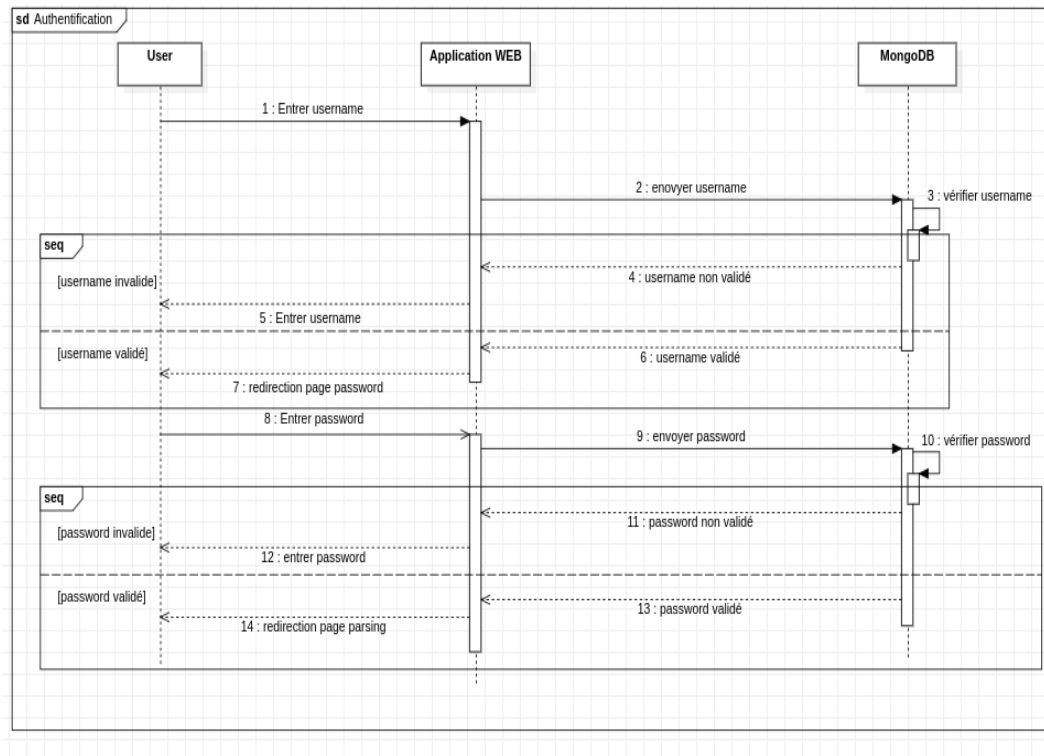


FIGURE 2.3 – Diagramme de séquence «S'authentifier»

### 2.3.3.2 Diagramme de séquence du cas d'utilisation « Déposer CV »

Une fois l'authentification réussie, le candidat a la possibilité de déposer son CV. La figure 2.4 montre le diagramme de séquence «Déposer CV».

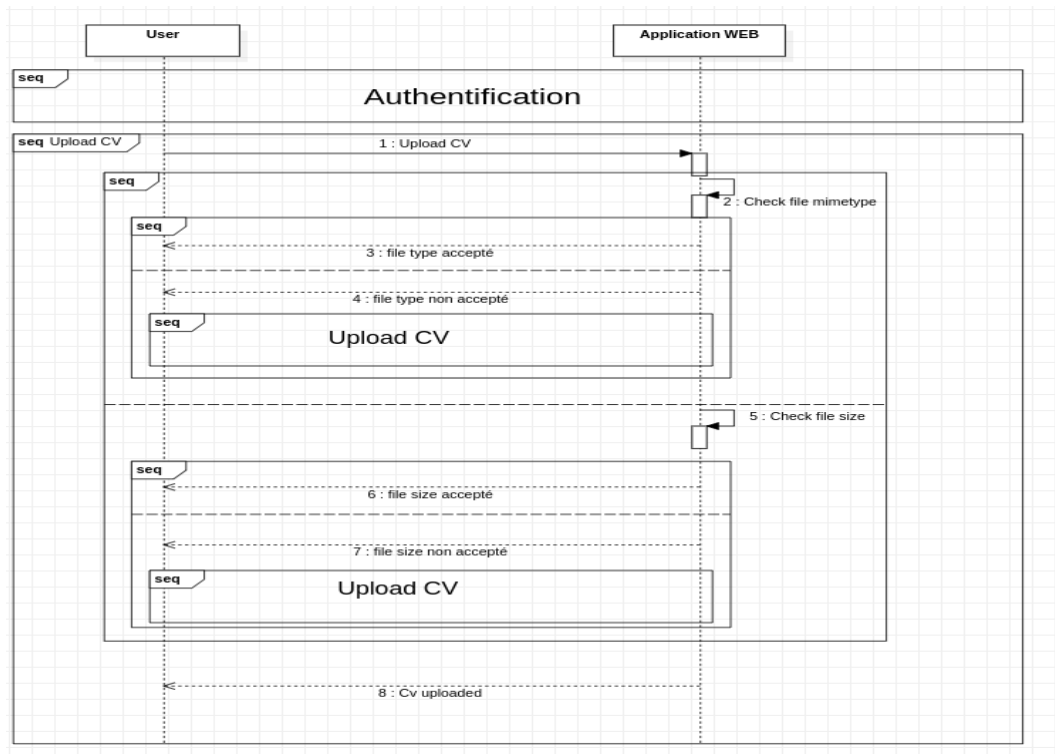


FIGURE 2.4 – Diagramme de séquence «Déposer CV»

### 2.3.3.3 Diagramme de séquence du cas d'utilisation « Extraction des informations des CVs »

Dès qu'un CV est déposé, un API Flask entre en action pour le traiter. Il génère alors un fichier JSON contenant les résultats, qui sont ensuite transmis à notre application WEB, afin d'être affichées. La figure 2.5 montre le diagramme de séquence du «Extraction des informations des CVs ».

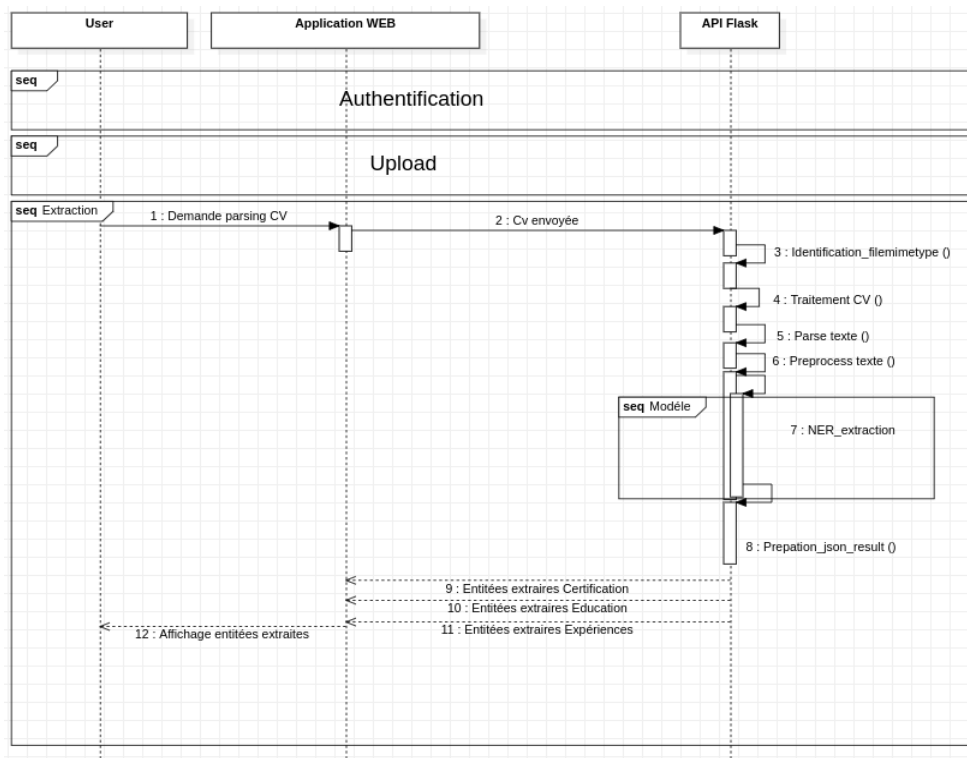


FIGURE 2.5 – Diagramme de séquence «Extraction des informations des Cvs»

## 2.4 Architecture de l'application

L'architecture d'un système à concevoir demeure un élément déterminant pour la performance et la possibilité de la réutilisation. En tenant compte des exigences précédemment évoquées, notre solution doit garantir des niveaux élevés de performance et de flexibilité pour chacun de ses éléments constitutifs.

Pour mettre en œuvre notre application, nous avons adopté l'architecture à trois niveaux. Également appelée architecture trois tiers ou architecture à trois couches, elle constitue une approche client-serveur où coexistent et sont maintenus des modules distincts, permettant d'organiser les applications en trois niveaux distincts : le niveau Présentation, le niveau Application et le niveau Données.

L'architecture de notre application est composée des éléments suivants :

- **Client** : Cette partie repose sur React et assure la gestion de la logique de navigation. Elle se compose d'une combinaison de HTML, CSS et JavaScript, interprétée par le navigateur pour présenter l'aspect visuel et gérer la logique des données.
- **Serveurs** : Cette partie englobe la fonctionnalité de l'application, c'est-à-dire celle qui met en place la logique métier et décrit les opérations effectuées sur les données en fonction des règles métier. Elle répond aux demandes de l'utilisateur émises par l'interface utilisateur.
- **Base de données** : Ce niveau est dédié à la gestion de l'accès aux données de l'application, stockant et gérant les données associées à celle-ci.

La figure 2.6 présente l'architecture générale de notre application.

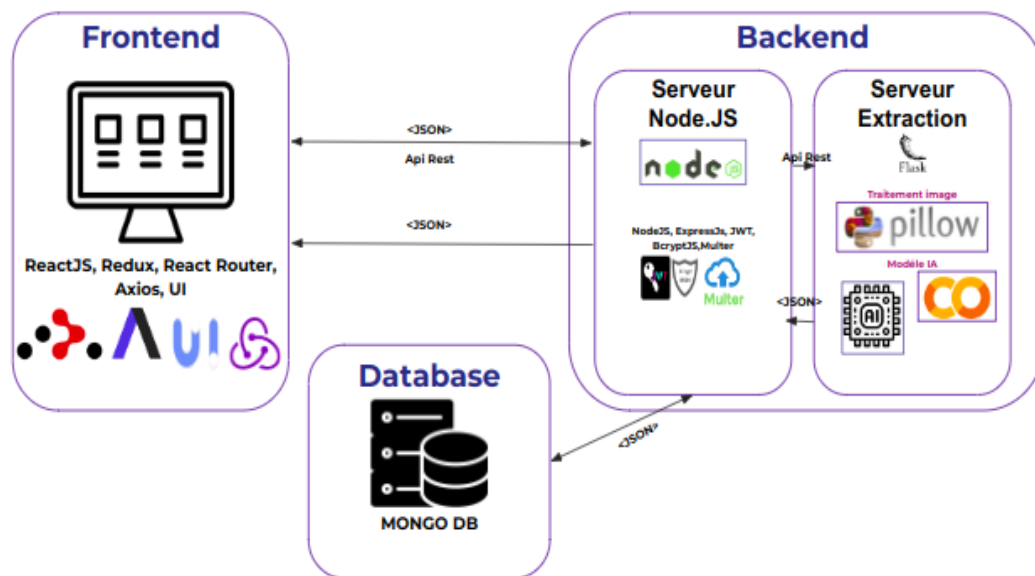


FIGURE 2.6 – Architecture détaillée de notre application

## 2.5 Environnement de travail

Dans cette section, nous allons répertorier les équipements et les logiciels utilisés dans le processus de développement d'une solution visant à soutenir les professionnels des RH dans la recherche des meilleurs talents.



### 2.5.1 Outils de développements matériels

Dans notre projet nous avons utilisé :

- Une machine "Lenovo Ideapad 3 15IML05" équipée d'un processeur "Intel Core i5-10210U" 10ème génération, 2.8 Ghz up to 4.1 Ghz, 6 Mo de cache, d'une mémoire 8 Go, aussi d'une carte graphique "Geforce MX130" avec 2 Go de mémoire GDDR5.
- Un système d'exploitation Ubuntu 20.04 avec une valeur totale de disque dur "1To" en HDD.

### 2.5.2 Outils de développements logiciels

La réalisation de ce projet a nécessité l'utilisation d'un ensemble d'outils et de technologies, énumérés ci-dessous :

- **Python** : Python [7] est le langage de programmation le plus répandu dans le domaine de l'intelligence artificielle. Il est orienté objet et se distingue par sa relative facilité d'apprentissage. Python jouit d'une grande popularité au sein de la communauté scientifique, notamment dans le domaine de l'intelligence artificielle.
- **Google Colab** : Google Colab ou Colaboratory [8] est un service cloud, offert par Google, basé sur Jupyter Notebook qui permet d'entraîner des modèles de Machine Learning directement dans le cloud.
- **Vscode** : Visual Studio Code [9] est un éditeur de code développé par Microsoft qui est compatible avec Windows, Linux et MacOS. Il offre de nombreuses fonctionnalités telles que le débogage, la mise en évidence de la syntaxe, l'achèvement intelligent du code, les extraits de code, la refonte du code et une intégration Git. C'est un outil largement utilisé par les développeurs pour sa polyvalence et sa richesse en fonctionnalités qui facilitent la programmation et la gestion de projets logiciels.
- **Nodejs** : Node.js [10] est un environnement d'exécution open source et multi-plateforme pour JavaScript, principalement utilisé côté serveur. Il permet aux développeurs d'exécuter du code JavaScript en dehors du navigateur web, ouvrant ainsi la voie à des applica-

tions back-end performantes et évolutives. Node.js est largement adopté dans le développement web pour sa capacité à gérer des opérations asynchrones de manière efficace, ce qui le rend particulièrement adapté aux applications en temps réel et aux services WEB.

- **Reactjs** : Reactjs [11] est une bibliothèque JavaScript open source utilisée pour le développement d'interfaces utilisateur côté client, notamment les interfaces web. Elle a été développée par Facebook et est largement utilisée pour la création d'applications web interactives et réactives. React se distingue par son approche de la construction d'interfaces utilisateur modulaires basées sur des composants, ce qui facilite la réutilisation du code et la maintenance des applications. Il est souvent utilisé en conjonction avec d'autres technologies front-end et back-end pour créer des applications web modernes et dynamiques.
- **Expressjs** : Express.js [12], souvent simplement appelé Express, est un framework d'application web back-end conçu pour Node.js. Il est distribué en tant que logiciel libre et open source, ce qui signifie qu'il est accessible gratuitement et que sa source est ouverte pour la modification et l'amélioration par la communauté de développeurs.

Express est particulièrement réputé pour sa simplicité et sa flexibilité, ce qui en fait un choix populaire pour la création d'applications web et d'API. Il facilite la gestion des routes, des requêtes HTTP et des vues dans les applications Node.js, ce qui accélère le développement web. Express est souvent considéré comme le framework de serveur standard pour Node.js en raison de sa popularité et de son adoption généralisée.

- **Flask** : Flask [13] est un micro-framework open-source destiné au développement web en Python. Il est considéré comme un micro-framework car il offre un ensemble de fonctionnalités de base, tout en étant légèrement plus léger et moins intrusif que d'autres frameworks web Python plus complets comme Django. Flask a pour philosophie de maintenir un noyau simple et extensible, ce qui signifie qu'il fournit des fonctionnalités essentielles pour la création d'applications web, mais permet aux développeurs d'ajouter des extensions et des fonctionnalités supplémentaires au besoin. Cette approche modulaire en fait un choix populaire pour les projets de développement web Python de petite à moyenne envergure, ainsi que pour les développeurs qui préfèrent une plus grande liberté dans la structure de leurs applications.

- **MongoDB** : MongoDB [14] est une base de données NoSQL (non relationnelle) populaire qui stocke les données sous forme de documents au format JSON, offrant ainsi une grande flexibilité et évolutivité pour la gestion et le stockage de données.
- **Gitlab** : GitLab [15] est une plateforme de gestion de dépôts de code source qui permet aux équipes de développement de collaborer, de suivre les versions de leur code, de gérer des projets et d'automatiser les processus de développement logiciel. Elle offre des outils de gestion de code source, d'intégration continue et de déploiement continu (CI/CD) ainsi que des fonctionnalités de suivi de problèmes et de gestion de projets, le tout intégré en une seule solution.
- **ClickUP** : ClickUP [16] est une plateforme de productivité et un logiciel de gestion de projet qui vise à remplacer plusieurs applications par une seule. Il comprend des fonctionnalités telles que des tâches, des documents, des objectifs, un chat et une gestion de calendrier. Il est conçu pour fonctionner pour tous les types d'équipes, leur permettant de planifier, d'organiser et de collaborer en un seul endroit.
- **Microsoft Teams** : Microsoft Teams [17] est une plateforme de collaboration et de communication en temps réel pour les organisations. Elle permet de travailler en équipe, de partager des fichiers et des applications, de planifier des réunions et de discuter par messagerie instantanée. Elle intègre également des fonctionnalités de visioconférence et de chat vidéo, ainsi que des outils de stockage et de transfert de fichiers avec SharePoint et de prise de notes avec OneNote.

## 2.6 CONCLUSION

Ce chapitre a été dédié à l'analyse des besoins fonctionnels et non fonctionnels de l'application, ainsi qu'à la présentation des divers acteurs impliqués. De plus, nous avons exposé les différents diagrammes UML. En parallèle, nous avons approfondi les concepts sous-jacents à l'architecture globale du projet ainsi les outils que nous avons sélectionné pour réaliser notre application.

---

# Implémentation et Réalisation

## Sommaire

---

<b>3.1</b>	<b>INTRODUCTION . . . . .</b>	<b>33</b>
<b>3.2</b>	<b>Construction de la base de données . . . . .</b>	<b>33</b>
3.2.1	Web Scraping . . . . .	33
3.2.2	Formation de la base de données . . . . .	36
3.2.3	Traitement de la base de données . . . . .	36
3.2.4	Extraction du contenu textuel . . . . .	37
3.2.5	Nettoyage de la base de données . . . . .	39
3.2.6	Annotation de la base de données . . . . .	39
<b>3.3</b>	<b>Implémentation des modèles d'extraction . . . . .</b>	<b>43</b>
3.3.1	Implémentation du modèle de la bibliothèque RegEx . . . . .	43
3.3.2	Implémentation du modèle "Layout Parser" . . . . .	45
3.3.3	Implémentation du modèle Donut . . . . .	46
3.3.4	Implémentation du modèle SpaCy . . . . .	48
<b>3.4</b>	<b>Conception de l'interface WEB . . . . .</b>	<b>55</b>
3.4.1	Communication entre Nodejs et Python . . . . .	55
3.4.2	Interfaces conçues . . . . .	56
<b>3.5</b>	<b>Conclusion . . . . .</b>	<b>59</b>

---

### 3.1 INTRODUCTION

Au fil de ce dernier chapitre, nous amorçons par la mise en exergue de notre environnement de travail, et nous exposons les multiples outils de développement qui ont été employés. Subséquemment, nous démêlons les choix techniques effectués, lesquels se fondent sur des expériences et des comparaisons pertinentes. Pour clore, nous mettons en lumière les étapes de réalisation de notre travail, en nous basant sur des captures d'écran des interfaces qui mettent en relief les différentes fonctionnalités intégrées dans l'application.

### 3.2 Construction de la base de données

La construction d'une base de données est une étape cruciale dans la mise en œuvre d'un projet. La collecte de données, qui consiste à rassembler et à mesurer des informations provenant de diverses sources, est un processus essentiel. Pour exploiter efficacement les données que nous collectons dans le développement de solutions d'intelligence artificielle (IA) pertinentes, il est impératif de les collecter et de les stocker de manière à ce qu'elles aient une signification en lien avec le problème commercial spécifique que nous cherchons à résoudre.

Cette section se focalise à détailler la stratégie adoptée pour la collecte des CVs dans une base de données destinée à être utilisée dans le développement de notre système d'extraction des informations pertinentes.

#### 3.2.1 Web Scraping

Le web scraping, également appelé extraction de données web, est une technique informatique qui consiste à extraire automatiquement des informations à partir de sites web en utilisant des programmes ou des scripts. Elle permet de récupérer divers types de données non structurées, tels que du texte, des images, des vidéos et des liens. Dans notre projet, le web scraping est particulièrement utile pour enrichir notre base de données.

### 3.2.1.1 Technologies de Web Scrapping

On distingue différents technologies et bibliothèques appliqués en Web Scrapping :

- **Beautiful Soup** : Beautiful Soup [18] une bibliothèque Python qui facilite la navigation et l'extraction de données à partir du code HTML. Elle est souvent utilisée en conjonction avec des bibliothèques telles que Requests pour récupérer des pages web. Beautiful Soup est appréciée pour sa facilité d'utilisation et sa popularité dans le domaine de l'analyse web et du web scraping.

- **Requests** : Requests [19] est un module Python qui permet d'envoyer des requêtes HTTP et de récupérer des données à partir de pages web. Elle est largement utilisée en raison de sa simplicité d'utilisation et de sa flexibilité. Requests peut être utilisé en conjonction avec BeautifulSoup, une bibliothèque de parsing HTML, pour extraire des informations spécifiques à partir des pages web téléchargées.

- **Scrapy** : Scrapy [20] est un framework de web scraping en Python qui offre une flexibilité et une puissance considérables pour extraire des données à partir de sites web. Il est particulièrement adapté pour gérer des projets de grande envergure et collecter d'importantes quantités de données de manière structurée. Scrapy simplifie le processus d'extraction de données en fournissant des outils pour naviguer dans les pages web, extraire des informations spécifiques et stocker les données collectées dans un format souhaité, comme une base de données ou un fichier CSV.

- **Selenium** : Selenium [21] est une bibliothèque qui permet d'automatiser le navigateur web, ce qui est particulièrement utile pour simuler des interactions avec un site web. Cette bibliothèque permet de contrôler un navigateur comme Chrome, Firefox, ou d'autres, et de reproduire des actions telles que le clic sur des boutons, la saisie de texte dans des formulaires, et la navigation entre les pages. Cela peut être précieux pour le web scraping de sites web complexes qui nécessitent une authentification ou des interactions utilisateur pour accéder aux données souhaitées.

### 3.2.1.2 Choix de langage et des technologies du Web scraping

Depuis de nombreuses années, Python est largement reconnu comme le langage de prédilection pour le web scraping. Python se distingue comme l'outil le plus efficace et polyvalent pour aborder une variété de tâches liées à l'extraction de données web.

Le tableau 3.1 ci-dessous présente une comparaison des langages de programmation couramment utilisés pour le web scraping.

**TABLE 3.1 – Etude comparative des langages de programmation pour le Web Scraping**

Langage	Nbre de pages /5 min	Limitations
<b>Python</b>	805	
<b>C++</b>	362	- La mise en place d'un système de web scraping en utilisant C++ peut s'avérer coûteuse.
<b>Nodejs</b>	200	- Non adapté aux projets de données de grande envergure. - Manque de fiabilité et d'expérience.
<b>Php</b>	500	- Déconseillé pour les projets de données à grande envergure, cela entraîne de multiples problèmes.

Python est une solution exceptionnelle car elle offre un accès à des bibliothèques de premier plan conçues pour la collecte de données rapide et efficace. Parmi les bibliothèques les plus couramment utilisées, on trouve "Scrapy", "Beautiful Soup" et "Selenium".

Scrapy, en particulier, est un framework Python conçu pour le scraping web à grande échelle. Il met à disposition tous les outils nécessaires pour extraire des données à partir de sites web de manière efficace, les traiter selon nos besoins et les stocker dans la structure et le format de notre choix. De plus, Scrapy est très optimisé et consomme moins de ressources informatiques que BeautifulSoup et Selenium.

#### **Avantages de Scrapy :**

- Spécialement conçu pour les projets de scraping de grande envergure.
- Moins coûteux en termes d'utilisation de la mémoire et du processeur (CPU).
- Vitesse supérieure à celles du BeautifulSoup et du Selenium, grâce à son fonctionnement asynchrone et sa capacité à traiter des requêtes simultanément.

### 3.2.2 Formation de la base de données

Grâce aux technologies de web scraping, nous avons réussi à collecter automatiquement 300 CVs à partir du site de recrutement très réputé, LinkedIn. Cette opération a considérablement enrichi notre base de données de CVs multilingues, la portant à un total de 560 CVs.

### 3.2.3 Traitement de la base de données

Avant toute exploitation de notre base de données, et afin de garantir les meilleurs résultats possibles pour notre projet, nous devons tout d'abord effectuer une conversion de l'ensemble de la base de données au format image. Cette étape initiale est cruciale, car elle nous permettra d'entreprendre des opérations et des traitements visant à améliorer la qualité des documents en entrée et à clarifier leur contenu. Cette clarification est essentielle pour que le processus de "parsing" s'effectue correctement.

La conversion des données présentée en images constitue la première étape de notre processus. Elle permettra de représenter visuellement les documents, ce qui facilitera la détection et la correction des éventuelles imperfections ou anomalies dans le texte. Cette conversion peut être réalisée en utilisant des techniques de numérisation et de traitement d'image, garantissant ainsi une représentation fidèle des documents originaux.

Une fois que les données sont au format image, nous pouvons entreprendre diverses opérations de traitement. Cela peut inclure la suppression de bruits, l'amélioration de la résolution, la correction de l'éclairage et la binarisation des données.

L'objectif principal de ces traitements est d'obtenir des images de haute qualité qui serviront de base solide pour notre processus de parsing ultérieur. La figure 3.1 montre l'amélioration apportée à un exemple de CV après le traitement d'image.





FIGURE 3.1 – Comparaison entre deux CVs avant et après le traitement d'image

## 3.2.4 Extraction du contenu textuel

Après avoir amélioré la qualité des CVs d'entrée, l'étape suivante consiste à extraire le contenu textuel des CVs téléchargés. Pour déterminer la meilleure approche d'extraction de texte, nous avons mené une comparaison entre différents outils disponibles. Comme mentionné précédemment, notre objectif est d'extraire le contenu textuel des CVs qui peuvent se présenter sous trois formats principaux : PDF, image ou Docx. Pour ce faire, nous avons essentiellement deux méthodes à envisager : soit réaliser l'extraction du contenu textuel pour chaque format individuellement, soit commencer par convertir le CV d'entrée en format image avant d'extraire son contenu.

### 3.2.4.1 Outils d'extraction du contenu contextuel

- **PyPDF 2** : PyPDF2 [22] est une bibliothèque Python qui permet de manipuler des fichiers PDF. Elle offre des fonctionnalités pour lire, écrire, copier, fusionner, diviser, recadrer, transformer, chiffrer et déchiffrer des fichiers PDF. Elle peut également extraire du texte et des métadonnées à partir de fichiers PDF. PyPDF2 est une bibliothèque purement en Python qui ne nécessite aucune dépendance externe autre que la bibliothèque standard de Python.

- **Tesseract OCR** : Tesseract OCR [23] est un moteur de reconnaissance optique de caractères open source qui extrait le texte des images et des documents sans couche de texte et

produit le document dans un nouveau fichier texte, PDF ou dans la plupart des autres formats populaires. Il a été développé à l'origine par Hewlett-Packard comme logiciel propriétaire dans les années 1980, mais il a été publié en open source en 2005 et le développement a été sponsorisé par Google depuis 2006. Tesseract OCR est hautement personnalisable et peut fonctionner avec la plupart des langues, y compris les documents multilingues et le texte vertical.

- **Docxpy** : Docxpy [24] est une bibliothèque Python qui permet d'extraire le texte, les hyperliens et les images des fichiers docx. Elle est basée sur le projet python-docx2txt et ajoute une nouvelle fonctionnalité pour extraire les hyperliens et leurs textes correspondants.

- **Texttract OCR** : Texttract OCR [25] est un service de reconnaissance optique de caractères (OCR) fourni par Amazon Web Services qui utilise l'apprentissage automatique pour extraire automatiquement le texte, les tableaux et les formulaires des documents numérisés, tels que les fichiers PDF et les images. Il peut extraire du texte imprimé, du texte manuscrit et des données structurées à partir de différents types de documents, tels que les relevés financiers, les formulaires fiscaux, les résultats d'étudiants, etc.

### 3.2.4.2 Comparaison entre les outils d'extraction du contenu contextuel

Afin de choisir la méthode d'extraction du contenu textuelle adéquat pour notre application, nous illustrons dans le tableau 3.2 ci-dessous illustrent une comparaison entre les différentes méthodes d'extraction, en termes du temps d'exécution, consommation mémoire et consommation CPU Suite à une comparaison des moyennes du temps d'exécution, de la consommation

**TABLE 3.2 – Etude comparative entre les différentes méthodes d'extraction du contenu textuelle**

Méthodes	Temps d'exécution	Consommation mémoire	Consommation CPU
PDF parse + Tesseract OCR + doctotext	0.22 s	42 MB	0.312 %
PyPDF2 + Tesseract OCR + docxpy	0.29 s	40 MB	0.397 %
Amazon TEXTTRACT	0.33 s	35 MB	0.324 %
Conversion en image + OCR	0.23 s	30 MB	0.261 %

mémoire et de la consommation CPU entre différentes méthodes, nous avons pu constater que la méthode Conversion en image + OCR se révèle être la plus performante. En effet, les trois premières méthodes qui consistent à identifier le type de chaque CV d'entrée puis à effectuer la récupération du contenu en utilisant la bibliothèque appropriée, ont été confrontées à la dernière méthode qui consiste à convertir d'abord les données en images, puis à extraire le contenu textuel grâce à une bibliothèque OCR. Nous avons constaté également que la méthode qu'on a choisi est nettement plus rapide et plus efficace en termes de consommation des ressources. De plus, elle nous offre la possibilité de réaliser des traitements d'images pour améliorer la qualité globale du processus.

### **3.2.5 Nettoyage de la base de données**

Le nettoyage de la base de données, qui est au format textuel, est une étape cruciale pour garantir la qualité de nos données. Cette étape essentielle nous permet d'éliminer efficacement les caractères spéciaux indésirables, les caractères non pertinents, ainsi que les retours à la ligne inutiles. De plus, nous pouvons également supprimer les formats inconnus, garantissant ainsi la cohérence et la qualité de nos données. Cette démarche facilite considérablement leur utilisation dans nos futures analyses et applications.

### **3.2.6 Annotation de la base de données**

Une fois que les données textuelles sont prêtes sous la forme désirée, l'étape suivante de notre processus est l'annotation des données. En effet, l'annotation des données consiste à attribuer des étiquettes à chaque élément d'un ensemble de données afin de spécifier le résultat attendu que le modèle d'apprentissage automatique supervisé doit prédire. Ce processus englobe la classification et l'étiquetage des données, ce qui implique que chaque point de données disponible est manuellement catégorisé pour que le modèle d'apprentissage automatique puisse l'utiliser.

Pour commencer, il est essentiel de définir la méthode d'étiquetage ainsi que les libellés associés. Dans notre cas, notre objectif est de réaliser la reconnaissance des entités nommées

dans les trois sections pertinentes des CVs, à savoir l'éducation, la certification et l'expérience. Nous avons identifié principalement 11 entités nommées, comme illustré dans la figure 3.2.



FIGURE 3.2 – les étiquettes d'annotation

### 3.2.6.1 Outils d'annotation des données

#### - Prodigy :

Prodigy [26] est un outil d'annotation téléchargeable pour les tâches de traitement du langage naturel et de vision par ordinateur. Il utilise l'apprentissage actif pour sélectionner les exemples les plus utiles pour l'annotation humaine, ce qui permet de créer des ensembles de données de haute qualité avec moins d'efforts. Prodigy est extensible et personnalisable pour différents types de tâches et de formats de données. Il fournit une interface utilisateur simple et efficace pour l'annotation, ainsi qu'une API pour l'intégration dans des pipelines de traitement de données.

#### - Doccano :

Doccano [27] est un outil d'annotation de texte open source qui permet de créer des jeux de données de NLP pour les modèles. Il offre une interface utilisateur intuitive pour l'annotation, ainsi que des fonctionnalités de personnalisation pour les types d'annotations et les formats de données. Doccano est relativement facile à déployer et à prendre en main. Il est un outil

fantastique pour les développeurs web et permet de réaliser des projets de manière rapide et efficace.

### - SpaCy-annotator :

Spacy-annotator [28] est une extension de l'API de spaCy qui permet d'annoter des textes directement dans l'interface de spaCy. Il offre des fonctionnalités de personnalisation pour les types d'annotations et les formats de données. Spacy-annotator est un outil pratique pour les utilisateurs de spaCy qui souhaitent annoter des textes sans avoir à passer par un autre outil d'annotation.

### - Label Studio :

Label Studio [29] est un outil d'annotation open source pour la création de jeux de données de NLP et de vision par ordinateur. Il offre une interface utilisateur intuitive pour l'annotation, ainsi que des fonctionnalités de personnalisation pour les types d'annotations et les formats de données. Label Studio est un outil polyvalent pour la création de jeux de données de haute qualité pour l'entraînement de modèles d'IA.

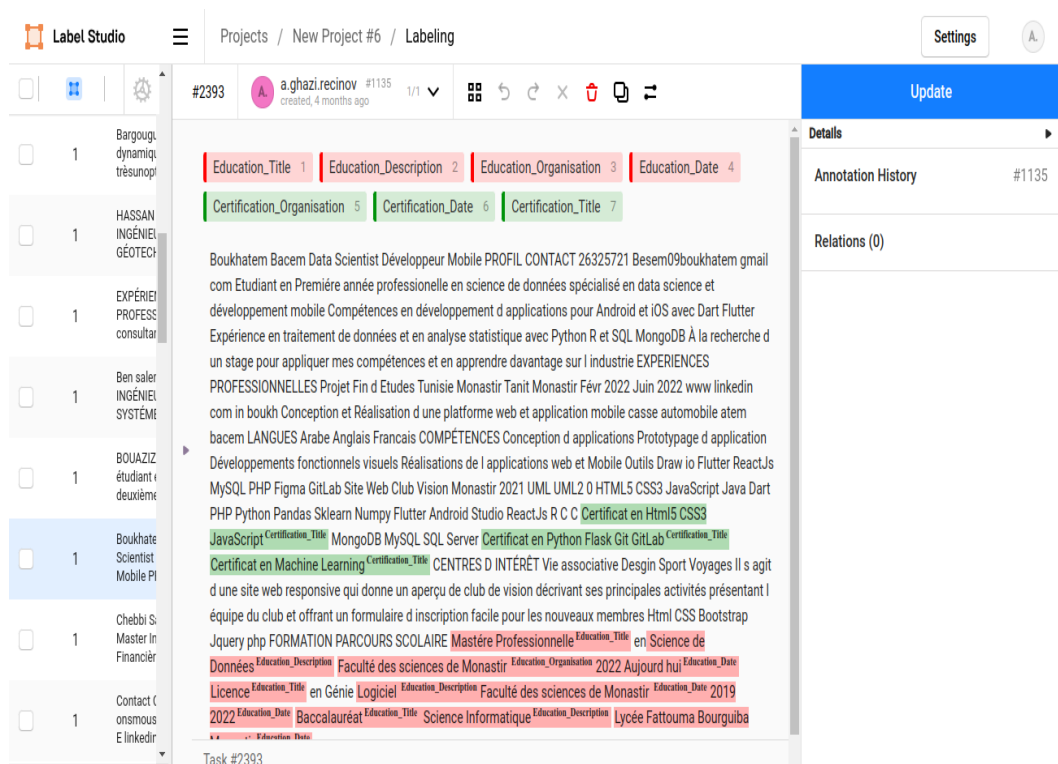
### 3.2.6.2 Etude comparative des outils d'annotation des données

Dés que les étiquettes sont bien définies, on doit choisir un outil d'annotation performant pour effectuer un label à chaque entité, pour cela on a du faire une étude comparative entre les outils d'annotations comme illustré dans le tableau 3.3.

**TABLE 3.3 – Etude comparative des outils d'annotation**

Outils d'annotations	Fonctionnalités	Cout	Automatisation	Raccourcis clavier
Label Studio	Annotation de texte, d'images, de fichiers audio et vidéo	Gratuit	Oui	Oui
Prodigy	Annotation de texte et de vision par ordinateur	Payant	Oui	Non
Doccano	Annotation de texte	Gratuit	Non	Non
Spacy-annotator	Extension de l'API de spaCy pour l'annotation de texte	Gratuit	Non	Non

En termes d'automatisation de l'annotation, Label Studio est l'outil le plus performant, car il offre des fonctionnalités d'apprentissage actives pour sélectionner les exemples les plus utiles pour l'annotation humaine. En ce qui concerne les raccourcis clavier, Label Studio offre des raccourcis clavier personnalisables pour faciliter l'annotation rapide. Prodigy offre également des fonctionnalités d'automatisation de l'annotation, mais ne propose pas de raccourcis clavier personnalisables. Doccano et Spacy-annotator ne proposent pas d'automatisation de l'annotation ou de raccourcis clavier. Pour toutes ces raisons là le choix s'est mis sur l'outil d'annotation Label Studio. La figure 3.3 montre l'interface de l'annotation via label studio, alors que la figure 3.4 présente la sortie après l'annotation.



**FIGURE 3.3 – Interface Label-studio**

```

1 |-DOCSTART- -X- 0
2 BOUKTHIR -X- _ 0
3 DHIA -X- _ 0
4 EDDIN -X- _ 0
5 STAGE -X- _ 0
6 OUYRIER -X- _ 0
7 060 -X- _ 0
8 EXPERIENCES -X- _ 0
9 PROFESSIONNELLES -X- _ 0
10 INFINITY -X- _ 0
11 SOLUTIONS -X- _ 0
12 STAGE -X- _ 0
13 OUVRIER -X- _ 0
14 & -X- _ 0
15 ENGINEERING -X- _ 0
16 EDUCATION -X- _ 0
17 LYCEE -X- _ B-Education_Organisation
18 15 -X- _ I-Education_Organisation
19 BAC -X- _ B-Education_Title
20 TECHNIQUE -X- _ B-Education_Description
21 NOVEMBRE -X- _ 0
22 1995 -X- _ 0
23 ISSAT -X- _ B-Education_Organisation
24 SOUSSE -X- _ I-Education_Organisation
25 CYCLE -X- _ B-Education_Title
26 PREPARATOIRE -X- _ I-Education_Title
27 + -X- _ 0
28 2 -X- _ 0
29 ANS -X- _ 0
30 CYCLE -X- _ B-Education_Title
31 INGENIEUR -X- _ I-Education_Title
32 COMPETENCES -X- _ 0
33
34 R -X- _ 0
35 A -X- _ 0
36 NI -X- _ 0
37 A -X- _ 0
38 R -X- _ 0

```

FIGURE 3.4 – Sortie de l’annotation

### 3.3 Implémentation des modèles d’extraction

Maintenant que nos données sont prêtes, nous pouvons passer à l’étape de la modélisation. Nous avons construit plusieurs modèles pour aborder notre problème spécifique, en utilisant diverses techniques comme décrit en détail dans les parties qui suivent.

#### 3.3.1 Implémentation du modèle de la bibliothèque RegEx

La bibliothèque RegEx [30] (expressions régulières) est un ensemble d’outils informatiques permettant de rechercher et de manipuler des motifs de texte complexes dans des chaînes de caractères, facilitant ainsi la recherche et l’extraction d’informations spécifiques dans un texte.

### 3.3.1.1 Principe de fonctionnement

Pour récupérer des données relatives à l'éducation, aux certifications et à l'expérience, on utilise des modèles de recherche basés sur des motifs de texte prédéfinis. Lorsque nous utilisons Regex pour extraire des sections spécifiques telles que l'éducation, nous commençons par définir des expressions régulières spécifiques à cette section. Ces expressions sont conçues pour correspondre aux schémas attendus dans la section de l'éducation, tels que des mots-clés comme "diplôme", "université", "année de graduation", etc.

Une fois que nous avons défini nos expressions régulières, nous les appliquons au texte du CV. La bibliothèque Regex parcourt ensuite le document à la recherche de correspondances avec nos expressions régulières. Lorsqu'une correspondance est trouvée, la bibliothèque extrait automatiquement les informations pertinentes en fonction de notre modèle. Par exemple, elle peut extraire le nom de l'institution, le type du diplôme, la spécialité, la date d'obtention du diplôme, et d'autres détails similaires.

### 3.3.1.2 Résultats d'implémentation

Après de nombreuses essais effectuées, on a conclut que cette solution n'est pas générique pour être adaptée à l'analyse de CVs, car le contenu textuel de ces documents n'est souvent ni structuré ni uniforme. Les formats et la disposition des informations varient considérablement d'un CV à un autre, et les mots-clés ne sont pas toujours alignés de manière cohérente. Cette approche de recherche de motifs fonctionne mieux pour des textes structurés et précis, mais elle peut rencontrer des limitations significatives lorsqu'elle est appliquée à un grand nombre de CVs variés. Dans de tels cas, des approches plus avancées, telles que l'utilisation de techniques de traitement du langage naturel (NLP) ou de méthodes d'apprentissage automatique, peuvent être nécessaires pour extraire efficacement des informations pertinentes.



### 3.3.2 Implémentation du modèle "Layout Parser"

Layout Parser [31] est une bibliothèque Python qui fournit des modèles d'apprentissage en profondeur pré-entraînés pour détecter la mise en page des images de documents. Il permet d'extraire des informations de mise en page telles que les zones de texte, les tableaux, les images et les graphiques, ainsi que de convertir ces informations en données structurées. Layout Parser est facile à utiliser et fournit une interface simple et propre pour la détection de mise en page précise. Il est également extensible et peut être utilisé pour entraîner des modèles personnalisés pour des tâches spécifiques.

#### 3.3.2.1 Architecture du modèle Layout Parser

Layout Parser utilise des modèles d'apprentissage en profondeur pour détecter la mise en page des images de documents. Les modèles sont basés sur des réseaux de neurones convolutifs (CNN) [32] et sont pré-entraînés sur des ensembles de données de documents pour détecter les zones de texte, les tableaux, les images et les graphiques. Les modèles sont ensuite utilisés pour prédire les zones de chaque type du contenu dans une image de document donnée. Layout Parser utilise également des algorithmes de post-traitement pour améliorer la précision de la détection de la mise en page. Une fois que les zones du contenu ont été détectées, Layout Parser peut extraire les informations de chaque zone et les convertir en données structurées. Les informations extraites peuvent être utilisées pour effectuer des tâches telles que la reconnaissance optique de caractères (OCR), la classification de documents et l'extraction d'informations. La figure 3.5 décrit l'architecture d'un réseau de neurones convolutif.

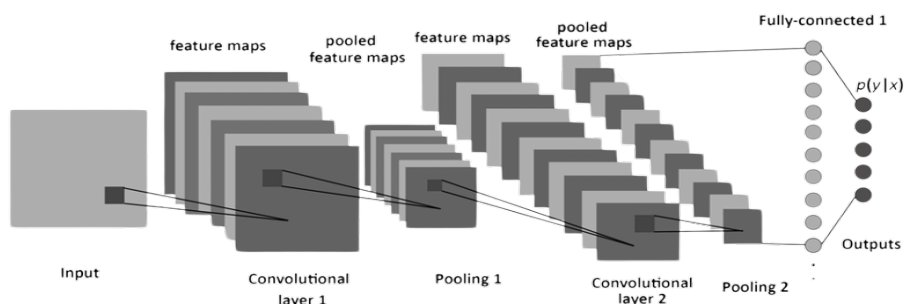


FIGURE 3.5 – Architecture d'un réseau de neurones convolutif

## 3.3.2.2 Résultats d'implémentation

Afin d'utiliser le modèle Layoutparser pour détecter les sections "éducation", "certification" et "expériences" dans des images de documents, nous avons tout d'abord finetuner le modèle en utilisant l'outil Label Studio pour marquer manuellement ces sections. Ensuite nous avons entraîné le modèle CNN pré-entraîné pour détecter ces régions pendant 1000 epochs, mais nous avons obtenu des résultats médiocres en termes de précision, avec un score de 38 %. La figure 3.6 montre le résultat de l'implémentation du modèle Layoutparser pour un cv donné.



FIGURE 3.6 – Résultat de la détection des régions avec Layout Parser

## 3.3.3 Implémentation du modèle Donut

Le modèle Donut [33] est un modèle de compréhension de documents qui a été proposé dans [34] le cadre de l'article "OCR-free Document Understanding Transformer" par Geewook Kim et al. Il s'agit d'un modèle end-to-end qui utilise un encodeur Transformer d'image et un décodeur Transformer de texte autoregressif pour effectuer des tâches de compréhension de documents telles que la classification de documents et l'extraction d'informations. Contrairement aux modèles traditionnels de reconnaissance optique de caractères (OCR), Donut ne nécessite pas de moteur OCR pour traiter les documents numérisés. Le modèle est pré-entraîné sur des ensembles de données de documents pour détecter les zones de texte, les tableaux, les images

et les graphiques, et peut être finetuné pour des tâches spécifiques. Donut utilise une architecture de réseau de neurones convolutifs (CNN) pour encoder les images de documents et une architecture de Transformer pour décoder le texte.

## 3.3.3.1 Architecture du modèle Donut

L'architecture du modèle Donut est basée sur un encodeur Transformer d'image et un décodeur Transformer de texte autoregressif [35]. Le modèle est pré-entraîné sur des ensembles de données de documents pour détecter les zones de texte, les tableaux, les images et les graphiques. Le modèle utilise une architecture de réseau de neurones convolutifs (CNN) pour encoder les images de documents et une architecture de Transformer pour décoder le texte. Le modèle est conçu pour effectuer des tâches de compréhension de documents telles que la classification de documents et l'extraction d'informations. Contrairement aux modèles traditionnels de reconnaissance optique de caractères (OCR), Donut ne nécessite pas de moteur OCR pour traiter les documents numérisés. Le modèle est end-to-end, ce qui signifie qu'il peut traiter des images de documents brutes sans prétraitement. Le modèle peut être finetuné pour des tâches spécifiques en utilisant des ensembles de données annotées. La figure 3.7 montre l'architecture du modèle Donut.

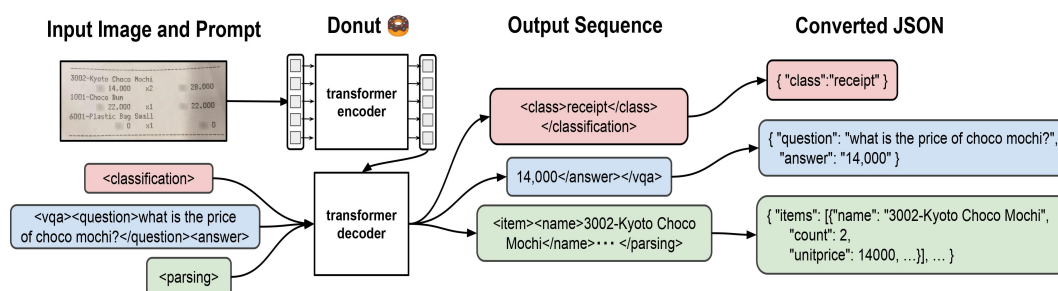


FIGURE 3.7 – Architecture du modèle Donut

## 3.3.3.2 Résultats d'implémentation

Après avoir utilisé le modèle Donut pour extraire les informations pertinentes des CVs en utilisant l'outil d'annotation Sparrow pour préparer 300 CVs et les fournir en format adéquat au modèle Donut, nous avons fait le "finetuning" du modèle. Cependant, nous avons rencontré

des problèmes lors de l'entraînement en termes d'utilisation de RAM et de GPU, vue que ce modèle nécessite des machines puissantes. Il est important de noter que l'entraînement de modèles d'apprentissage en profondeur peut être un processus complexe et nécessite souvent des ressources importantes. Malgré les difficultés rencontrées, l'utilisation du modèle Donut pour extraire des informations pertinentes des CVs est une approche prometteuse pour automatiser le processus de compréhension de documents.

Pour améliorer d'avantage le système d'extraction des informations pertinentes des CVs multilingues des candidats, on a eu recours d'utiliser une autre méthode qui réside dans le modèle de traitement du langage naturel "SpaCy" qui sera détaillé dans la partie suivante.

### 3.3.4 Implémentation du modèle SpaCy

SpaCy [36] est une bibliothèque open-source de traitement du langage naturel en Python. Elle fournit des fonctionnalités telles que la reconnaissance d'entités nommées (NER), l'étiquetage grammatical (POS), l'analyse de dépendance, les vecteurs de mots et bien plus encore. SpaCy fournit également des modèles pré-entraînés pour plusieurs langues, y compris l'anglais, le français, l'espagnol, l'allemand, le portugais, l'italien, le néerlandais et le danois. Les modèles pré-entraînés peuvent être utilisés pour effectuer des tâches de traitement du langage naturel telles que la classification de documents, l'extraction d'informations et la traduction automatique.

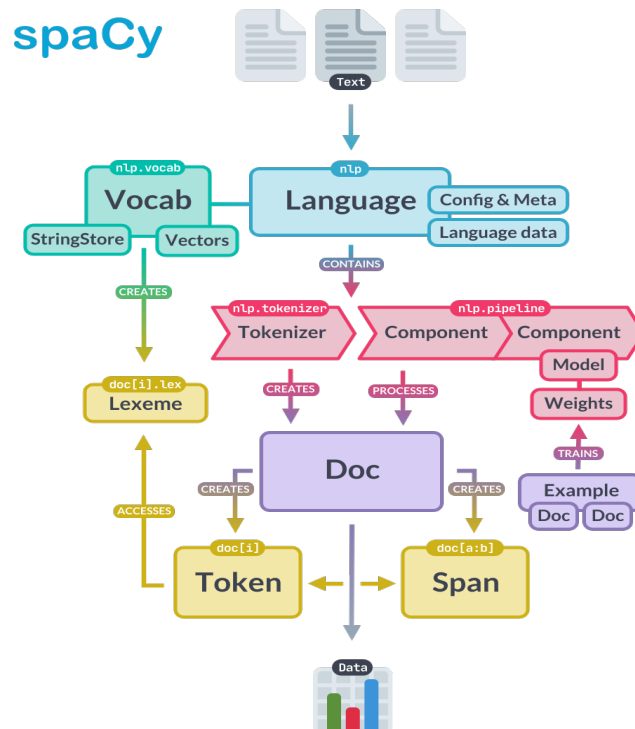
SpaCy fournit également des outils pour entraîner des modèles personnalisés sur des ensembles de données spécifiques. Les modèles personnalisés peuvent être entraînés pour effectuer des tâches spécifiques telles que la reconnaissance d'entités nommées pour des domaines spécifiques ou la classification de documents pour des types de documents spécifiques.

#### 3.3.4.1 Architecture du modèle SpaCy

L'architecture de SpaCy [37] est basée sur des modèles constitués des fonctions qui connectent une instance de modèle, que l'on peut ensuite utiliser dans un composant de pipeline ou comme couche d'un réseau plus grand. Les modèles pré-entraînés de SpaCy sont disponibles pour plu-

sieurs langues, y compris l'anglais, le français, l'espagnol, l'allemand, le portugais, l'italien, le néerlandais et le danois. Les modèles pré-entraînés peuvent être utilisés pour effectuer des tâches de traitement du langage naturel telles que la classification de documents, l'extraction d'informations et la traduction automatique. La figure suivante illustre l'architecture du modèle SpaCy.

La bibliothèque est conçue pour être utilisée en production et permet de construire des applications qui traitent et comprennent de grands volumes de texte. Les objets centraux de SpaCy sont la classe Language, le Vocab et l'objet Doc. La classe Language est utilisée pour traiter un texte et le transformer en un objet Doc. Le Vocab est un ensemble de tables de recherche qui rendent les informations courantes disponibles dans tous les documents. L'objet Doc est l'un des objets les plus importants de l'architecture de SpaCy et possède la séquence de jetons ainsi que toutes leurs annotations. La structure de données de SpaCy centralise les chaînes, les vecteurs de mots et les attributs lexicaux, ce qui permet d'économiser la mémoire en évitant de stocker plusieurs copies des données. La figure 3.8 montre l'architecture du modèle SpaCy.



**FIGURE 3.8 – Architecture du modèle SpaCy**

Pour mettre en œuvre le modèle SpaCy, nous devons utiliser deux concepts clés dans le domaine du traitement du langage naturel qui sont :

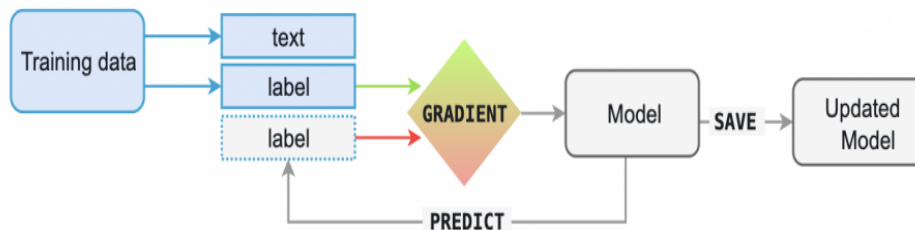
- **La reconnaissance d'entités nommées (NER)** : C'est un processus de traitement du langage naturel qui consiste à identifier et extraire des informations spécifiques à partir d'un texte, telles que les noms de personnes, les noms de lieux, les noms d'organisations, les dates, les montants d'argent, etc. La NER [38] est souvent utilisée pour extraire des informations structurées à partir de textes non structurés, tels que des articles de presse, des rapports financiers, des CVs, etc. La NER est un élément clé de nombreuses applications de traitement du langage naturel, telles que la classification de documents, l'extraction d'informations et la traduction automatique.

- **La relation extraction (RE)** : C'est un processus de traitement du langage naturel qui consiste à identifier les relations sémantiques entre des entités nommées dans un texte non structuré. La RE est souvent utilisée en conjonction avec la reconnaissance d'entités nommées (NER) pour extraire des informations structurées à partir de textes non structurés, tels que des articles de presse, des rapports financiers, des CVs, etc. La RE peut être utilisée pour extraire des relations entre des personnes, des organisations, des lieux, des événements, etc. La RE [39] est un élément clé de nombreuses applications de traitement du langage naturel, telles que la classification de documents, l'extraction d'informations et la traduction automatique. La RE peut être entraînée à l'aide de modèles d'apprentissage en profondeur tels que les réseaux de neurones convolutifs (CNN) et les Transformers.

### 3.3.4.2 Apprentissage du modèle

L'apprentissage [40] est un processus itératif dans lequel les prédictions du modèle sont comparées aux annotations de référence afin d'estimer le gradient de la perte. Le gradient de la perte est ensuite utilisé pour calculer le gradient des poids par rétropropagation. Les gradients indiquent comment les valeurs des poids doivent être modifiées pour que les prédictions du modèle se rapprochent davantage des étiquettes de référence au fil du temps. Lors de l'entraînement d'un modèle, nous ne voulons pas seulement qu'il mémorise nos exemples, nous voulons

qu'il élabore une théorie qui puisse être généralisée à des données inédites. La figure qui suit montre un schéma détaillé qui explique l'apprentissage du modèle SpaCy. La figure 3.9 montre le fonctionnement d'apprentissage du modèle SpaCy.



**FIGURE 3.9 – Fonctionnement d'apprentissage du modèle SpaCy**

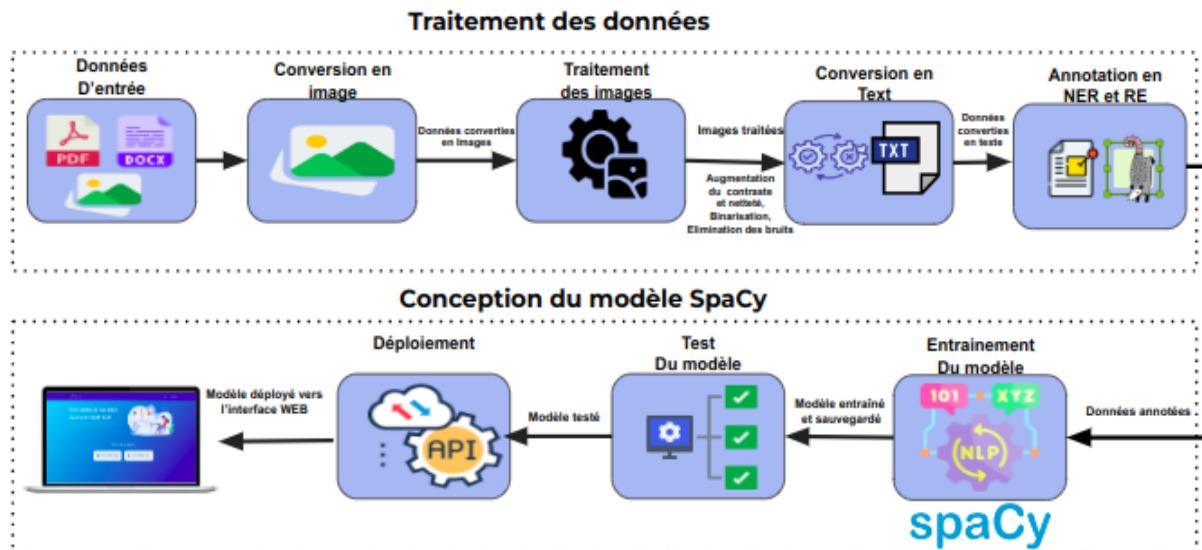
On distingue différents modèles pré-entraînés du SpaCy, à savoir :

**XX ent wiki sm** : Il s'agit d'un modèle multilingue de reconnaissance d'entités nommées (NER) entraîné sur des données de Wikipedia. Il peut être utilisé pour extraire des entités nommées dans plusieurs langues, y compris l'anglais, le français, l'espagnol, l'allemand, le portugais, l'italien, le néerlandais et le danois.

**FR core news sm** : Il s'agit d'un modèle pré-entraîné pour le français qui fournit des fonctionnalités telles que la reconnaissance d'entités nommées (NER), l'étiquetage grammatical (POS), l'analyse de dépendance, les vecteurs de mots et plus encore. Il est entraîné sur des données de presse écrite et fournit une précision élevée pour les tâches de traitement du langage naturel en français.

**EN core web sm** : Il s'agit d'un modèle pré-entraîné pour l'anglais qui fournit des fonctionnalités telles que la reconnaissance d'entités nommées (NER), l'étiquetage grammatical (POS), l'analyse de dépendance, les vecteurs de mots et plus encore. Il est entraîné sur des données de texte écrit sur le web (blogs, actualités, commentaires) et fournit une précision élevée pour les tâches de traitement du langage naturel en anglais.

L'organigramme 3.10 suivant décrit le processus suivi pour former le modèle SpaCy et générer les résultats d'extraction souhaités.



**FIGURE 3.10 – Organigramme du processus de conception du modèle d'extraction**

Pour construire notre modèle d'extraction des informations pertinentes des Cvs nous avons suivi les étapes suivantes :

- 1- Fournir les données d'entrée.
- 2- Convertir les données en formats images.
- 3- Traiter les images converties afin de les binariser et augmenter leurs contrastes.
- 4- Convertir les images traitées en formats textuelles.
- 5- Assurer l'annotation en NER et RE.
- 6- Entraîner le modèle SpaCy.
- 7- Tester les performances du modèle SpaCy.
- 8- Déployer le modèle SpaCy vers l'interface.

### 3.3.4.3 Résultats d'implémentation

Dans le but d'évaluer les performances des trois modèles SpaCy conçus, nous avons entraîné sur notre ensemble de données annotées, nous avons utilisé également une taille de lot (batch size) de 16 et 1000 epochs pour chaque modèle. Nous avons utilisé le score F1 pour



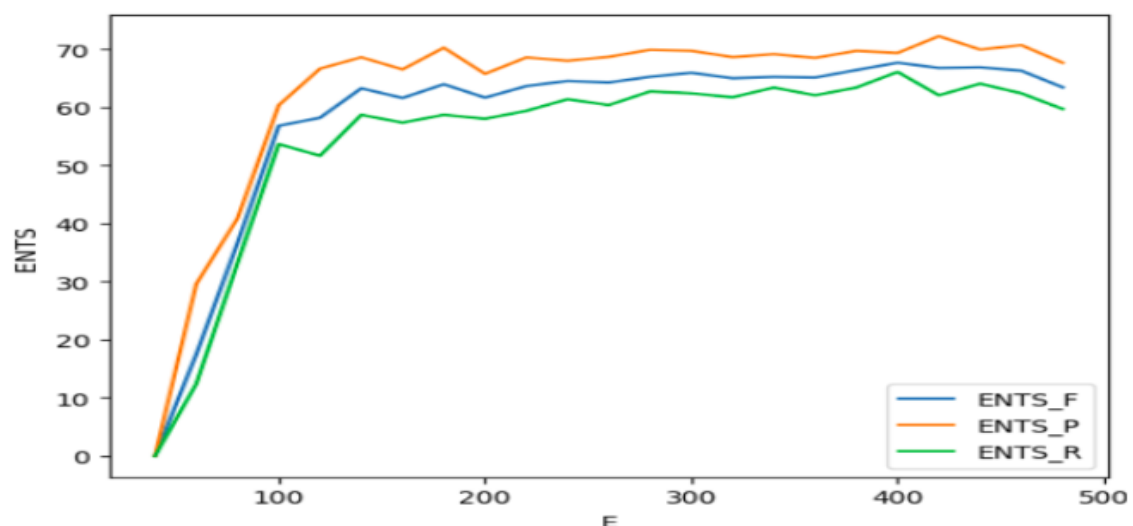
évaluer la reconnaissance d'entités nommées (NER) et la perte (loss) pour mesurer la qualité de la prédiction comme montré par le tableau 3.4.

**TABLE 3.4 – Etude comparative des modèles pré-entraînés du SpaCy**

Modèle	xx ent wiki	en core web sm	fr core news sm
Données	560 Cvs	560 Cvs	560 Cvs
Batch size	16	16	16
Epochs	1000	1000	1000
F1 Score	70	47	50
Loss NER	48	144	120
Mémoire occupée	4 MB	49 MB	38 MB

D'après les résultats d'implémentation, on distingue que le modèle **xx ent wiki** est le plus performant en terme de précision, ainsi qu'il est le moins couteux en terme de mémoire occupée, ce qui rend ce modèle le plus adapté à notre application d'extraction des informations pertinentes des Cvs multilanuges.

Pour montrer d'avantage les performances et la capacité du modèle à apprendre, nous présentons la progression de la précision et de la perte au cours de la phase d'apprentissage, respectivement par la figure 3.10 et la figure 3.11.



**FIGURE 3.11 – Courbe de précision du modèle**

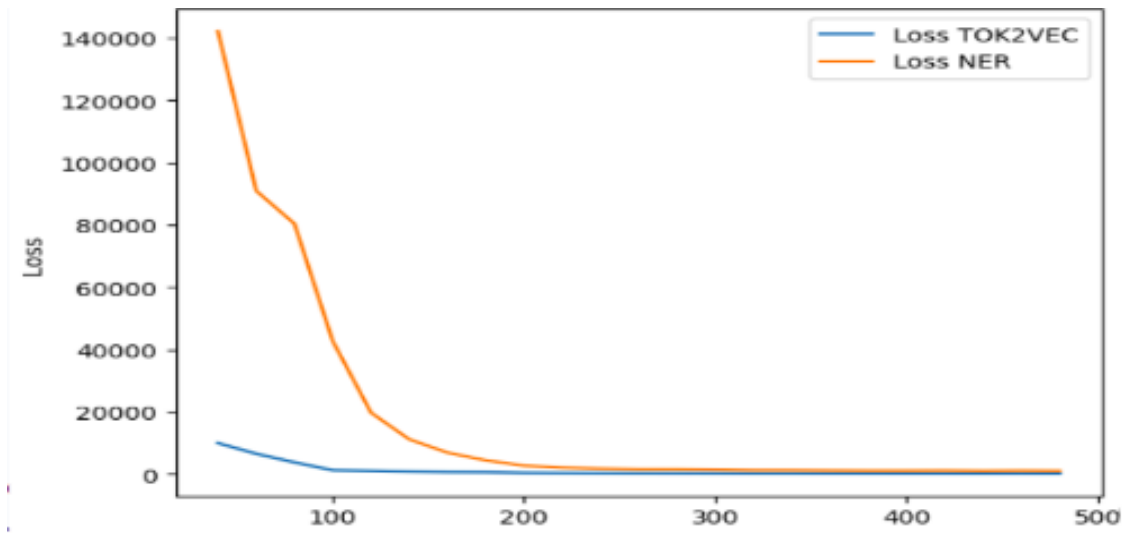


FIGURE 3.12 – Courbe de perte du modèle

La figure 3.13 présente un exemple d'extraction des informations pertinentes d'un CV fourni par notre modèle implémenté sous format JSON.

### GHAZI ABDALLAH

ETUDIANT EN INGENIERIE DES SYSTEMES ELECTRONIQUES ET COMMUNICATION

A LA RECHERCHE D'UN PREMIER EMPLOI

+216 64770502 | ghabdallah@gmail.com | ghabdallah99@sfax.tn

#### EDUCATION

- Ecole Nationale d'Electronique et des Telecommunications Sfax 09/20 - Present  
Génie des systèmes électroniques et de communication.  
Option: Systèmes électroniques intelligents.
- Institut préparatoire aux études d'ingénieurs Kairouan 09/18 - 07/20  
Option: Physique-Technique.

#### EXPERIENCES

- RECINOV Startup Sfax 09/20 - Present  
Stage ingénieur PFE : Système d'extraction des données des CVs de manière automatisée pour faciliter le processus de recrutement au département RH.  
Développement d'un système d'extraction des données des CVs de manière automatisée pour faciliter le processus de recrutement au département RH.  
Mots clés : NLP, Python, Spacy, RegEx, Data Mining, REIN STACK framework.
- Centre de recherche en numérique Sfax 09/20 - Present  
Stage ingénieur : Classification des maladies des tumeurs.  
Développement d'un modèle de classification des maladies des tumeurs des tumeurs des tumeurs.  
Mots clés : DL, Python, TensorFlow, Keras.
- Tunisie Telecom Mahdia 09/20 - 07/20  
Stage senior : Architecture des réseaux.

#### CERTIFICATIONS

- Data Scientist with python back - DataCamp
- C Programming with Linux - Coursera
- Participation in AIML Hackathon - TunisiaAttack22

#### VIE ASSOCIATIVE

- Club Adromodisme Enst'com 09/20 - 09/22  
- Président
- Evénement national ENET AERO CUP 1.8 07/20 - 07/21  
- Membre actif
- Manipulation des cartes électroniques.

#### PROJETS

- PPA  
- Détection de matricule  
Construction d'un modèle VGG16 basé sur dataset qui détecte la matricule d'une voiture.  
Mots clés : Python, DL, YOLOv3, OpenCV, Raspberry pi, AWS Stream.
- Projets académiques  
- Quadricoptère  
Réalisation d'un quadricoptère de course "F1" à base d'un contrôleur de vol F405.  
Mots clés : Système embarqué, F405, Bessflight.
- Radioamateur  
Réalisation d'une radioamateur pour le contrôle des mini avions RC à travers des joystick.  
Mots clés : Arduino, C, AVR, joystick.

#### COMPÉTENCES

- Langages de programmation  
Python, JavaScript, C, C++, Java.
- Outils de développement  
Visual Code, Jupyter notebook, Arduino IDE, STM32CubeIDE.
- Intelligence Artificielle  
Machine Learning, Deep Learning, Natural Language Processing, Data Mining
- Cartes électroniques  
Raspberry, Arduino, STM32, ESP32
- Méthodologies  
Gitlab, Scrum, UML

#### LANGUES

- Français - Professionnelle
- Anglais - Courant
- Arabe - Langue maternelle

#### LOISIRS

Football, Camping, Natation

(A)

```
[('02 23 - En cours', 'Experience_Date'), ('Stage ingénieur PFE', 'Experience_Title'), ('RECINOV Startup Sfax', 'Experience_Organisation'), ('06 22 - 08 22', 'Experience_Date'), ('Tunisie Télécom', 'Experience_Organisation'), ('Stage ingénieur', 'Experience_Title'), ('Centre de recherche en numérique Sfax', 'Experience_Organisation'), ('génie', 'Education_Titre'), ('Systèmes électroniques intelligents', 'Education_Description'), ('09 18 - 07 20', 'Education_Date'), ('Institut préparatoire', 'Education_Organisation'), (' préparatoire', 'Education_Titre'), ('Physique-Technique', 'Education_Description'), ('09 20 - Present', 'Education_Date'), ('Ecole nationale d'Electronique et des telecommunications sfax', 'Education_Organisation'), ('Data Scientist with python trak', 'Certification_Titre'), ('Datacamp', 'Certification_Organisation')]
```

(B)

FIGURE 3.13 – Résultat d'implémentation du modèle SpaCy, (A) : Cv d'entrée, (B) : Résultat de l'extraction du modèle SpaCy sous forme JSON

## 3.4 Conception de l'interface WEB

### 3.4.1 Communication entre Nodejs et Python

Dans le but de garantir la montée en charge, nous avons conçu un module d'extraction d'informations en utilisant le langage de programmation Python. Le défi majeur consiste à établir une liaison entre ce modèle et un serveur Node.js. C'est pourquoi nous présentons ci-dessous une comparaison des différentes méthodes de communication entre un modèle d'intelligence artificielle développé en Python et un serveur Node.js :

#### 3.4.1.1 Child Process

- **Description** : Le module Child [41] Process de Node.js permet d'exécuter des scripts ou des commandes dans des langages autres que JavaScript, tel que Python. Il autorise l'exécution d'un script Python au sein de l'application Node.js et la gestion des données d'entrée/sortie entre le script Python et l'application Node.js.

- **Inconvénients** : Cette méthode est déconseillée pour l'implémentation d'algorithmes d'apprentissage automatique ou d'apprentissage profond en raison de sa lenteur et des problèmes potentiels qu'elle peut entraîner.

#### 3.4.1.2 Hugging Face Model Hub

- **Description** : Cette approche implique l'envoi d'une requête au "Hugging Face Model Hub" [42] pour utiliser des modèles d'IA disponibles dans le cloud, sans avoir besoin de télécharger des bibliothèques volumineuses ni d'utiliser des ressources informatiques locales.

- **Inconvénients** : Cette solution n'est pas gratuite, et les comptes gratuits sont limités à 30 000 caractères d'entrée par mois.

### 3.4.1.3 TensorFlow.js

- **Description** : TensorFlow.js [43] est une bibliothèque JavaScript utilisée pour l'apprentissage automatique. Bien que nous utilisions PyTorch pour l'apprentissage, il est possible de convertir le modèle en TensorFlow à l'aide de "Hugging Face" pour pouvoir l'utiliser en JavaScript.

- **Inconvénients** : Il existe une bibliothèque de tokenization requise pour Node.js qui n'est pas encore très stable.

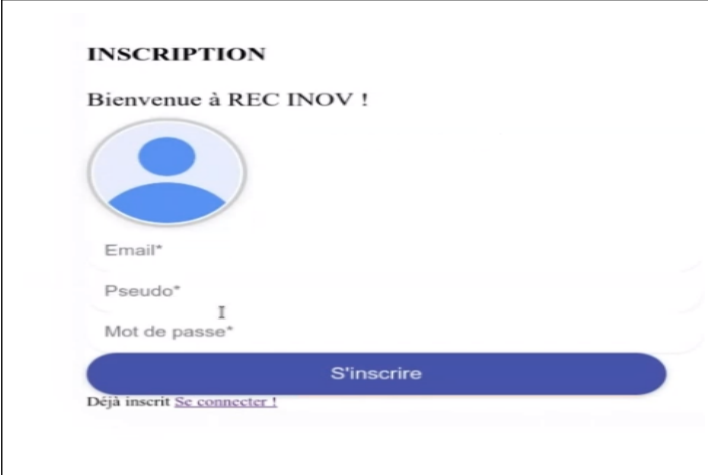
### 3.4.1.4 Solution Choisie

#### Utilisation du serveur **Flask**

Flask, en tant que micro-framework, offre une manière simple et rapide de configurer un serveur en Python pour gérer les requêtes entre Node.js et Python. Cette approche implique la configuration d'un serveur pour chaque langage et le partage de données au format JSON via des requêtes GET et POST.

## 3.4.2 Interfaces conçues

Dans cette section, nous illustrons différentes interfaces qui ont été développées au cours de ce projet pour chaque type d'utilisateur. Nous commencerons par les interfaces destinées aux candidats. Les trois figures ci-dessous illustrent les étapes d'inscription et de connexion des candidats.



**INSCRIPTION**

Bienvenue à REC INOV !



Email\*

Pseudo\*

Mot de passe\*

[I](#)

**S'inscrire**

Déjà inscrit [Se connecter !](#)

**FIGURE 3.14 – Interface d’inscription du candidat**



**Profil**

Vous pouvez modifier vos informations !



prénom

nom

numéro de tél

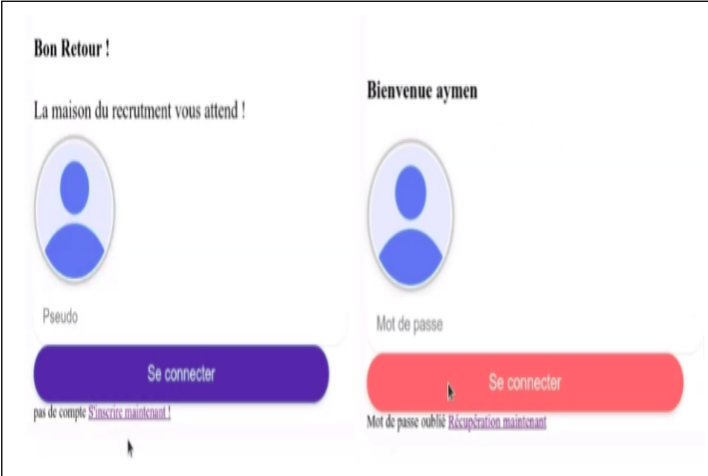
email\*

adresse

**Mettre à jour**


[Se déconnecter](#)

**FIGURE 3.15 – Interface remplissage du Profil du candidat**



**Bon Retour !**

La maison du recrutement vous attend !




Pseudo

**Se connecter**

[pas de compte S'inscrire maintenant !](#)

**Bienvenue aymen**



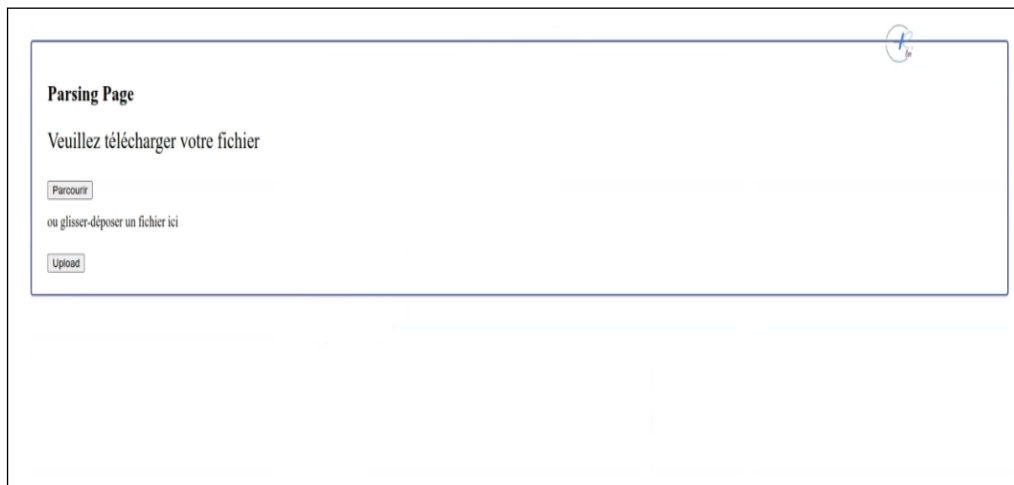
Mot de passe

**Se connecter**

Mot de passe oublié [Récupération maintenant](#)

**FIGURE 3.16 – Interface d’authentification**

Une fois que le candidat est connecté avec succès, il sera redirigé vers la page de l'upload de son CV, comme présenté par la figure suivante.



**FIGURE 3.17 – Interface du dépôt du CV du candidat**

Le candidat a maintenant la possibilité d'ajouter son CV soit par le bouton "Upload" soit en glissant son CV. Une fois le CV déposé respecte les conditions en termes de taille et de type, le candidat pourra parser les informations de son cv en appuyant sur le bouton parse.



**FIGURE 3.18 – Interface d'extraction des informations**

### 3.5 Conclusion

Dans ce chapitre, nous avons présenté la phase de réalisation de notre projet, en expliquant les modèles choisis et certaines interfaces graphiques de l'application, afin de donner un aperçu clair du travail effectué. D'après les résultats d'implémentation il s'était avéré que le modèle SpaCy est le plus performant en terme d'efficacité et de rapidité.



---

# CONCLUSION GÉNÉRALE

Ce projet constitut un travail accompli pendant mon stage de fin d'études au sein de la startup < REC-INOV >. L'objectif principal de ce stage était la conception et la réalisation d'une application de "Parsing" de CVs multilingues en utilisant la technologie NLP (Traitement du Langage Naturel) et l'Intelligence Artificielle (IA).

Ce projet a exigé de nombreux mois de travail acharné, à la fois sur le plan de la recherche théorique et du développement technique. Ces efforts ont finalement abouti à une solution qui non seulement répond à nos besoins, mais qui se distingue également par son efficacité et ses performances. La mise en œuvre de notre projet s'est articulée autour de trois parties majeures :

**1- Formation de la Base de Données :** Dans cette première étape, nous avons entrepris la collecte de données en utilisant la technique de Web Scraping. Nous avons ainsi constitué une base de données diversifiée contenant des CVs multilingues et sous différents formats, couvrant les types de CV les plus couramment utilisés aujourd'hui.

**2- Développement d'un Parseur Intelligent :** Cette deuxième étape a été consacrée à la création d'un parseur intelligent. Ce parseur est essentiel car il est chargé d'extraire de manière précise et efficace les informations pertinentes des CVs, ce qui constitue un critère essentiel pour la présélection des candidats.

**3- Conception de l'Interface Utilisateur et Administrateur :** La troisième partie de notre projet a concerné le développement de l'interface utilisateur. Ce volet nous a permis d'envisager un nouveau domaine d'application c'est celui du développement WEB.

Pour aboutir à construire notre système d'extraction des informations des CVs, nous avons conçu quatre modèles différents, ainsi, d'après les résultats d'implémentation, on a pu constater que le modèle Spacy est le plus performant en termes d'efficacité et d'occupation en mémoire.



Nous avons été confrontés à divers défis tout au long de ce travail, mais nous avons réussi à les surmonter avec succès. Ce projet a constitué une opportunité exceptionnelle d'approfondir nos compétences en matière d'intelligence artificielle et de traitement du langage naturel. De plus, nous avons pu observer de près la méthodologie Agile et son rôle dans la gestion d'une équipe de développement.

En termes de perspectives d'avenir, nous envisageons d'améliorer d'avantage les performances de notre application, d'améliorer l'aspect esthétique de notre interface WEB et de développer une application mobile pour notre système.



---

# NETOGRAPHIE

- [1] **Artificial Intelligence Definitions by Professor Christopher Manning, September 2020, Stanford University** : <https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf>
- [2] **apprentissage automatique** <https://www.cnil.fr/fr/definition/apprentissage-automatique>
- [3] **apprentissage profond** <https://datascientest.com/deep-learning-definition>
- [4] **Traitement automatique du langage naturel** <https://www.deeplearning.ai/resources/natural-language-processing/>
- [5] **SCRUM** <https://www.linkedin.com/pulse/what-scrum-when-use-advantages-disadvantages-deepak-mohan-kumar>
- [6] **UML** <https://www.geeksforgeeks.org/unified-modeling-language-uml-introduction/>
- [7] **Python** <https://www.bocasay.com/fr/meilleurs-langages-programmation-ia-2022/>
- [8] **Google Colab** <https://ledatascientist.com/google-colab-le-guide-ultime/>
- [9] **Visual Studio Code** <https://code.visualstudio.com/docs/editor/whyvscode>
- [10] **Node.js** <https://www.freecodecamp.org/news/node-js-basics/>
- [11] **Reactjs** <https://kinsta.com/knowledgebase/what-is-react-js/>
- [12] **Express.js** <https://kinsta.com/knowledgebase/what-is-express-js/>
- [13] **Flask** [https://en.wikipedia.org/wiki/Flask\\_\(webframework\)](https://en.wikipedia.org/wiki/Flask_(webframework))
- [14] **MongoDB** <https://www.techtarget.com/searchdatamanagement/definition/MongoDB>
- [15] **Gitlab** <https://www.blogdumoderateur.com/tools/gitlab/>
- [16] **ClickUP** <https://www.blogdumoderateur.com/tools/clickup/>

- [17] **Microsoft Teams** <https://support.microsoft.com/en-us/office/get-started-with-microsoft-teams-b98d533f-118e-4bae-bf44-3df2470c2b12>
- [18] **Beautiful Soup** <https://www.topcoder.com/thrive/articles/web-scraping-with-beautiful-soup>
- [19] **Requests** <https://www.digitalocean.com/community/tutorials/how-to-work-with-web-data-using-requests-and-beautiful-soup-with-python-3>
- [20] **Scrapy** <https://scrapy.org/>
- [21] **Selenium** <https://www.browserstack.com/guide/python-selenium-to-run-web-automation-test>
- [22] **PyPDF2** <https://nanonets.com/blog/pypdf2-library-working-with-pdf-files-in-python/>
- [23] **Tesseract OCR** <https://tesseract-ocr.github.io/>
- [24] **Docxpy** <https://towardsdatascience.com/how-to-extract-data-from-ms-word-documents-using-python-ed3fbb48c122>
- [25] **Textract OCR** <https://aws.amazon.com/textract/>
- [26] **Prodigy** <https://prodi.gy/>
- [27] **Doccano** <https://towardsdatascience.com/doccano-a-tool-to-annotate-text-data-to-train-custom-nlp-models-f4e34ad139c3>
- [28] **Spacy-annotator** <https://github.com/ieriii/spacy-annotator>
- [29] **Label Studio** <https://labelstud.io/guide/labeling.html>
- [30] **RegEx** <https://docs.python.org/fr/3/howto/regex.html>
- [31] **Layout Parser** <https://layout-parser.github.io/>
- [32] **(CNN)** <https://layout-parser.readthedocs.io/en/latest/example/deeplayoutparsing/index.html>
- [33] **Donut** <https://huggingface.co/docs/transformers/modeldoc/donut>
- [34] **OCR-free Document Understanding Transformer** by Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park, 6 Oct 2022, Cornell University <https://arxiv.org/abs/2111.15664>

- [35] **Papers Explained 20 : Donut**, by **Ritvik Rastogi**, 7 Feb 2023, **DAIR.AI**  
<https://medium.com/dair-ai/papers-explained-20-donut-cb1523bf3281>
- [36] **SpaCY** <https://spacy.io/usage/spacy-101>
- [37] **Architecture SpaCy** <https://spacy.io/api/architectures>
- [38] **NER** <https://www.processmaker.com/fr/blog/named-entity-recognition-challenges-and-solutions/>
- [39] **RE** <https://paperswithcode.com/task/relation-extraction>
- [40] **apprentissage spacy** <https://spacy.io/usage/training>
- [41] **Child Process** <https://www.geeksforgeeks.org/run-python-script-node-js-using-child-process-spawn-method/>
- [42] **Hugging Face Model Hub** <https://huggingface.co/docs/hub>
- [43] **TensorFlow.js** <https://www.tensorflow.org/js>

# Conception et Réalisation d'une application de Parsing des CVs multilingues via NLP et IA

---

**Ghazi Abdallah**

---

## **Résumé :**

Ce projet vise à extraire les informations pertinentes des CV multilingues des candidats. Une fois le CV téléchargé, un algorithme de traitement du langage naturel est chargé d'extraire ces informations. Les informations extraites seront ensuite affichées à l'utilisateur à travers une interface WEB et stockées dans la base de données. L'objectif de ce projet est d'automatiser le processus de recrutement de candidats en ligne. .

**Mots clés :** Apprentissage profond, développement web, traitement automatique du langage naturel, Python, Nodejs, React, Express, MongoDB, Flask, HuggingFace, SpaCy, web scraping.

## **Abstract :**

This project aims to extract relevant information from candidates' multilingual CVs. Once the CV has been downloaded, a natural language processing algorithm is tasked with extracting this information. is responsible for extracting this information. The extracted information will then be displayed to the user and stored in the database. The aim of this project is to automate the online candidate recruitment process.

**Key-words :** Deep learning, web development, natural processing language, Python, Nodejs, React, Express, MongoDB, Flask, HuggingFace, SpaCy, web scraping. scraping

## **تلخيص :**

يهدف المشروع إلى استخراج المعلومات ذات الصلة من السيرة الذاتية المتعددة اللغات للمرشح، وبمجرد تحميل السيرة الذاتية، تكون خوارزمية معالجة اللغة الطبيعية مسؤولة عن استخراج المعلومات. المعلومات المستخرجة ستكون بعد ذلك معروضة في صفحة ويب و مسجلة في قاعدة البيانات. الهدف من هذا المشروع هو أتمتة عملية توظيف المرشحين عبر الإنترنت.

**الكلمات المفتاحية :** التعلم العميق, تطوير الويب, لغة المعالجة الطبيعية, بايثون, تجريف الويب.