# Document de cadrage
# Projet de Fin d'Étude

Baptiste Hudyma - Martin Olivier

9 octobre 2019

# Table des matières

**1 Compréhension du problème** **3**
1.1 Contexte . . . 3
1.2 Analyse du problème . . . 3
1.3 Critères de satisfaction . . . 3

**2 Objectifs** **4**
2.1 Objectifs du projet . . . 4
2.2 Objectifs du groupe . . . 4

**3 Interlocuteurs** **5**
3.1 Contacts de la Prefecture de Mayenne . . . 5
3.2 Équipe responsable des projets ESIEA . . . 5
3.3 L'équipe du projet . . . 5
3.4 Parties prenantes du projet . . . 5

**4 Règles de fonctionnement** **6**
4.1 Fonctionnement de l'équipe avec le mentor . . . 6
4.2 Fonctionnement de l'équipe avec le commanditaire . . . 6
4.3 Fonctionnement interne de l'équipe . . . 6

**5 Moyens** **7**

**6 Méthode envisagée** **8**
6.1 Méthode de gestion du projet . . . 8
6.2 Réalisation Technique . . . 8

**7 Organisation de l'équipe** **9**

**8 Planning** **10**

# 1 Compréhension du problème

## 1.1 Contexte

Ce projet est proposé par la Prefecture de la Mayenne dans un contexte de dématérialisation des documents administratifs. En effet dans un effort de numérisation de leur base documentaire, il devient complexe de gérer la grande quantité/diversité de documents informatisés. De plus, le facteur humain de la classification non automatisée n'est pas une garantie d'archivage pérenne. En effet les employés administratifs changent, et d'une personne à l'autre la stratégie d'archivage peut et va varier.

C'est pourquoi la Prefecture de Mayenne envisage un projet de classification automatisée. La mise en oeuvre de ce projet permettra de faciliter la recherche de documents, accélérer les démarches et réduire la masse documentaire de l'État, tout en garantissant que le système d'archivage garde une logique constante, indépendante du facteur humain.

## 1.2 Analyse du problème

La Prefecture dispose d'un grand nombre de documents manuscrits qui nécessitent d'être numérisés et classés par un système indépendant de l'intervention humaine, source d'erreur. Ce travail laborieux est actuellement réalisé par des humains, ce qui rend cette tâche couteuse en temps et en ressources humaines.

La Prefecture envisage donc la mise en place d'un système automatisant toute la chaine de travail, de la classification d'un document à sa recherche. Afin de faciliter l'accès aux documents numérisés, la Prefecture désire aussi mettre en place un moteur de recherche afin de réduire le temps de recherche documentaire.

Le projet doit donc être un système automatisé capable d'analyser des documents numérisés et d'en tirer des informations caractéristiques. Ce logiciel doit ensuite se présenter sous la forme d'un moteur de recherche donnant la capacité à l'utilisateur de retrouver des documents de plusieurs façons différentes (mots clefs, ressemblance, ...).

Le projet est donc séparé en deux blocs principaux : Le classifier de documents et le moteur de recherche.

Le classifier doit être capable de lire n'importe quel document et d'en tirer des éléments caractéristiques internes, comme les parties prenantes, le type de document, la taxonomie, mais aussi les caractéristiques globales, comme par exemple décrire le document par un vecteur distinctif permettant de retrouver les documents similaires. Toutes ces informations seront stockées dans des métadonnées associées à chaque documents.

Le moteur de recherche doit se baser sur les métadonnées crées par le classifier pour retrouver des documents. On envisage des recherches par :

— taxonomie
— similarité apparente
— similarité de contenu
— documents associés à une personne
— documents associés à un document

## 1.3 Critères de satisfaction

TODO

# 2 Objectifs

## 2.1 Objectifs du projet

Les objectifs du projet on été définis avec le commanditaire lors de la première réunion. Pour que le projet soit fonctionnel, il doit :

— être open source
— analyser un document numérisé
— permettre la recherche d'un document dans le base de données

Ces objectifs ont été précisés ci dessous afin d'en tirer des étapes plus précises selon notre plan IVVQ.

Ci dessous, les objectifs précis ont été redéfinis :

— être open source et réalisé avec des technologies open source
— récupérer le contenu dun document numérisé
— analyser le contenu sémantique dun document
— représenter un document sous une forme vectorielles
— classer les documents selon une taxonomie établie
— ajouter à chaque document numérisé des métadonnées en format CEDAF
— trouver les versions mineures dun document déjà stocké
— permettre la recherche dun document par taxonomie
— permettre la recherche dun document par similarité apparente
— permettre la recherche dun document par similarité du contenu
— permettre la recherche des documents associés à une personne
— permettre la recherche des documents associés à un document

Ces objectifs ont été définis lors de la rédaction du plan IVVQ et ont été validés avec le commanditaire. Ils seront à la base des vérification que nous effectuions avec le commanditaire à la fin de la l'année. Ces objectifs sont aussi sujet à évolution car nous travaillons avec les méthodes agiles avec notre commanditaire.

## 2.2 Objectifs du groupe

Notre groupe à plusieurs objectifs pour ce projet. Tout d'abord nous voulons aborder ce projet comme une entreprise le ferait. C'est pourquoi nous allons mettre en place un plan IVVQ (Intégration Vérification Validation Qualification) à présenter et faire valider par notre commanditaire. Ce plan sera la référence que nous utiliserons pour vérifier et valider notre réalisation avec le commanditaire à la fin de l'année, et nous sera aussi utile pour le planning prévisionnel et la marche à suivre lors de la production. Nous avons également pour objectif d'utiliser les méthodes agiles dans un contexte réel. Jusqu'ici l'application de ces méthodes à été pour nous un exercice qu'il était facile de tordre pour respecter la consigne. Dans ce projet, il à été défini avec le commanditaire que nous travaillerons avec lui avec la méthode agile.

D'un point de vue technique, nous voulons profiler nos compétences en machine learning, surtout appliqué au domaine de l'analyse naturelle de texte (NLP). Nous pourrons aussi expérimenter avec les technologies de moteur de recherche, la deuxième partie importante du projet.

A la fin du projet, nous voulons avoir produit un POC fonctionnel qui sera satisfaisante pour le commanditaire et qui lui donnera envie de revenir travailler avec l'ESIEA l'année prochaine.

# 3 Interlocuteurs

## 3.1 Contacts de la Prefecture de Mayenne

Francois Xavier ROCHE, Directeur de projet
Frédéric ARRIGHI, Commanditaire du projet et lien entre la prefecture et l'ESIEA
Cyril BEIDÉ, Expert en archivage de documents

## 3.2 Équipe responsable des projets ESIEA

Aurelien BURGET, référent entreprenaria de l'ESIEA
Guillaume LEMESLE, responsable des PST ESIEA
Patricia BESSET-VEZIAT, mentor du projet

## 3.3 L'équipe du projet

Baptiste HUDYMA, SPOC de l'équipe
Martin OLIVIER, Expert marchine learning
Staberlin VALENTAIN, Responsable management

## 3.4 Parties prenantes du projet

Le projet est destiné à la Prefecture de la Mayenne, qui pourra l'utiliser pour gérer sa masse documentaire.
Il pourra aussi concerner le service des archives de la Mayenne.

# 4 Règles de fonctionnement

## 4.1 Fonctionnement de l'équipe avec le mentor

Notre première réunion avec notre mentor, Patricia Besset-Véziat, nous à permis de mettre au point nos règles de fonctionnement. Nous avons convenu d'une réunion de suivi par semaine afin de vérifier l'avancement et la bonne conduite du projet. Cette réunion n'est pas nécessairement en face à face et pourra être réalisée en visioconférence.

## 4.2 Fonctionnement de l'équipe avec le commanditaire

Nous avons convenu avec le commanditaire d'une méthode de travail agile. Pour cela, nous envisageons une réunion de suivi toutes les deux semaines, et une réunion de pilotage tous les mois. La réunion de suivi sera faite par visioconférence ou par échange de mail informatifs du travail réalisé. La réunion de pilotage sera réalisée en face à face, à l'ESIEA Laval ou à la Prefecture de Mayenne. Cette réunion à pour objectif de présenter l'avancement du projet et de discuter des éventuels changements désirés par le commanditaire, conformément à l'interaction client de la méthode agile.

## 4.3 Fonctionnement interne de l'équipe

Le fonctionnement interne du groupe est axé autour de la méthode agile. Nous allons faire une réunion de planning une fois par semaines au début de la semaine afin de déterminer le programme de travail de la semaine à venir, et une réunion à la fin de la semaine pour faire le point sur les tâches effectuées. Au cour de la semaine, nous resterons en communication le plus possible malgré la distance Paris Laval qui limite les interactions directes.

# 5 Moyens

La prefecture met à notre disposition un grand nombre de documents. Cependant, le commanditaire ne met aucunes autres ressources physiques à notre disposition.

Afin d'obtenir les ressources nécessaires à la réalisation du projet, nous avons sollicité l'association robotique de Paris, la DTRE, qui nous à donné l'accès à distance à un ordinateur ayant des performances suffisantes pour réaliser notre projet.

# 6 Méthode envisagée

## 6.1 Méthode de gestion du projet

Au sein de l'équipe, nous avons adoptés une méthode de fonctionnement agile avec une organisation du projet autour d'un plan IVVQ. Le plan IVVQ a été validée par le client, et résume les caratéristiques principales du projet ainsi que les tests techniques et fonctionnels qui seront réalisés pour assurer le fonctionnement final du projet.

Nous avons décidé de séparer le projet en trois blocs principaux. Chaque membre de l'équipe est responsable d'un bloc du projet, c'est à dire de la façon dont il va être réalisé. Toutes fonction d'un bloc principal peut être confiée à un autre membre du groupe qui est plus à l'aise avec l'approche technique du problème.

La rédaction des documents de rendu comme celui ci sera réalisée avec le langage LaTeX.

L'intégralité du code et des documents LaTeXsera géré sur GitHub, à l'adresse `https://github.com/smallito/PFE`. Les documents LaTeXcompilés seront ensuite placés sur le sharepoint.

## 6.2 Réalisation Technique

Nous avons séparé le projet en plusieurs parties principales :

— **Reconnaissance des caractères** : Le document doit être analysé afin d'en retirer le contenu textuel.

— **Analyse du texte** : Le contenu textuel du document doit être analysé sémantiquement de placer le document dans une ou plusieurs taxonomie. Ces informations seront ajoutées aux métadonnées du document.

— **transformation vectorielle du document** : Le texte du document sera ensuite passé par un word2vec, qui transformera le contenu sémantique en vecteur de nombres flottants, qui sera utilisé pour retrouver les documents avec des contenus identiques. Le document sera aussi passé dans un encodeur pour obtenir un vecteur permettant de retrouver les documents par ressemblance visuels par la suite.

— **Extraction d'informations utiles** : Le document sera ensuite passé par une suite d'analyseurs pour en retirer des informations que le commanditaire aura jugé utiles. Par exemple, on pourra trouver le nom des personnes mentionnées dans ce document, les dates de validité, les signatures, si le document est un faux, ... Ces informations seront ajoutées aux métadonnées du document.

— **Création d'une interface de recherche** : Une interface de recherche sera ensuite réalisée afin de faciliter l'accès aux documents classés. La méthode de recherche sera basée sur les informations présentées précédemment.

# 7 Organisation de l'équipe

Notre équipe est composée de trois membres :

— **Baptiste Hudyma** : SPOC, représentant de l'équipe, responsable de la partie principale du moteur de recherche

— **Martin Olivier** : Spécialiste machine learning (auteur publié dans le domaine), responsable de la partie d'analyse sémantique.

— **Staberlin Valentain** : Responsable management de l'équipe, responsable de la partie d'analyse des documents.

# 8 Planning