Document de cadrage
Projet de Fin d'Étude

Baptiste Hudyma - Martin Olivier

7 octobre 2019

Table des matières

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

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

TODO : en rajouter
— Classer automatiquement des documents numérisés
— Mettre en place un moteur de recherche

3 Interlocuteurs

4 Règles de fonctionnement

5 Moyens

6 Méthode envisagée

Le projet est découpé en plusieurs parties principales :

— **Découpage du document en blocs caractéristiques** : Selon l'organisation du document, reperer les zones de texte caractéristiques

— **Reconnaissance des caractères** : Les blocs repérés lors de l'étape précédente doivent être analysés afin d'en retirer le texte exploitable.

— **Analyse du contenu des blocs** : Le contenu textuel des blocs de texte doit être analysé sémantiquement afin d'en retirer le sens principal du document et ses informations utiles.

— **Classification du document en fonction de son contenu** : Le contenu global du document doit être analysé afin de le classer.

— **Création d'une interface de recherche** : Une interface de recherche sera ensuite réalisée afin de faciliter l'accès aux documents classés.

7 Organisation de l'équipe

Notre équipe est composée de deux membres :
— Baptiste Hudyma : SPOC, représentant de l'équipe
— Martin Olivier : Spécialiste machine learning (auteur publié dans le domaine)

8 Planning