Rapport de stage

Étude préliminaire pour le tri de texte de loi par apprentissage automatique

Étudiants *:*
Hakim Chekirou
Mohamed Ouaguenouni

Encadrants *:*
Anne Gaucher
Jamy Chahal

Septembre 2020

Table des matières

1 Introduction

L'objectif du stage est de réaliser une étude préalable pour le développement d'un programme d'apprentissage automatisé permettant d'approfondir des résultats de recherche d'une étudiante en première année de thèse.

La thèse s'intitule *« L'indépendance fonctionnelle de la magistrature au XIXe siècle. Une autre histoire de la séparation des pouvoirs »*, et a pour objet l'étude des décisions rendues par les juges tout au long du XIX^{me} siècle. Elle repose, pour partie, sur la lecture de recueils de jugements, afin d'appréhender les manifestations d'autonomie des juges lorsqu'ils rendent la justice, par rapport aux pouvoirs exécutif et législatif.

Il s'agit, en premier temps, de procéder à une reconnaissance de texte sur des recueils de jurisprudence et d'extraire les arrêts et jugements. La deuxième tache est de concevoir un algorithme, capable de repérer dans le corpus des arrêts, des mots et expressions qui se rapprochent de ceux, déjà identifiés par l'étudiante en thèse. L'idée étant de trouver une récurrence, précise ou simplement proche, des termes déjà identifiés pour avoir un traitement plus rapide de recueils numérisés. Ce travail, en grande partie de linguistique, permet de mettre à jour des réflexions systématisées dans l'art de rendre la justice.

2 Travaux effectués

2.1 Récupération des données

Pour réaliser le stage, on a exploité deux types de documents. Le premier, *Le recueil général des lois et des arrêts : en matière civile, criminelle, commerciale et de droit public, par J.-B. Sirey*, est un recueil de lois et d'arrêts publié jusqu'à 4 fois par an, chaque recueil contenant entre 300 et 1200 pages. Ils sont disponibles sous forme PDF, JPEG ou texte. La version texte a été générée de façon automatique par un programme de reconnaissance optique de caractères (OCR) de la bibliothèque nationale de France, la qualité de la reconnaissance n'est pas assez bonne pour l'utiliser telle quelle et on cherche à trouver une meilleure technique d'extraction. Nous avons traité cette base de 1791 jusqu'en 1899.

La deuxième source, *la gazette des tribunaux* est un journal de jurisprudence et de débats judiciaires publié presque quotidiennement. Il n'est disponible que sous format PDF et nous traitons ce document de 1870 jusqu'en 1899.

Pour les deux types de documents, nous avons écrit des scripts de scrapping en python, qui explorent les sites de la BNF et de la médiathèque de l'école d'administration pénitentiaire pour récupérer les documents ainsi que d'autres informations.

2.2 Reconnaissance du texte

Le but de cette partie est de pouvoir parcourir les documents sous forme de texte. Nous avons pour cela fait une comparaison de l'état de l'art des modèles libre de droit de reconnaissance de texte. Nous avons comparé les trois modèles suivant : Tesseract, PyOCR et textract. Nous avons aussi exploré l'utilisation de l'outil PDFMiner qui permet de récupérer du texte représenté sous format pdf.
Pour la première base, aucun des modèles n'atteint des résultats significativement supérieur à la reconnaissance déjà réalisée par la BNF. En conséquence, on s'est concentré sur l'amélioration de la version texte déjà existante. On a supprimé des blocs de textes inutiles à notre travail comme des notes de fins de page et des entêtes. Certains caractères étant mal reconnus, on a corrigé l'orthographe avec l'outil Symspellpy.

Pour la deuxième base, les pdfs ont été enrichies par un programme d'OCR. Le texte a été récupéré en utilisant l'outil PDFminer. Une fois le texte obtenu, nous avons fait du post-traitement, c'est a dire retirer tout ce qui n'est pas utile à notre travail.

2.3 Extraction des arrêts et jugements

Pour extraire les décisions de justice, nous avons parcouru le texte en appliquant des expressions régulière (REGEX) dessus. Pour chaque type de documents, les arrêts sont codifiés sous différents formats. Nous avons aussi extrait la juridiction et la date de chaque arrêt. En tout, plus de 80000 arrêts sont obtenus de la première base et près de 10000 pour la deuxième. Tous les arrêts sont stockés dans des fichiers CSV.

2.4 Recherche d'information

Une fois le corpus des arrêts construit, l'objectif est de retourner les arrêts intéressants en suivant les critères identifiés dans la thèse. Certains critères sont réguliers tel que utilisation de termes comme abrogation ou évidence dans la décision de justice, d'autres sont plus compliqués comme des expressions qui indique l'application d'un pouvoir créateur, ces critères ne sont pas réguliers et nécessitent un algorithme d'apprentissage pour les détecter. Cependant, nous disposons d'une base de seulement 229 exemple d'arrêts labélisés, cette quantité n'étant pas suffisante, on a limité la recherche aux critères réguliers. Pour ce faire, nous avons traité chaque arrêt en identifiant les occurrences de chaque critères.

2.5 Développement de l'application

Pour permettre aux chercheurs de travailler avec cette base, nous avons réalisé une application web en utilisant le cadre de développement open source Django.
On a intégré tous les arrêts extraits au format CSV dans une base SQL incluse dans l'application. On a ensuite construit une interface capable de prendre en compte la période de la recherche, les motifs et expressions à détecter et comment combiner ces cratères (conjonction ou disjonction). L'interface visualise les arrêts avec toutes les informations qui y sont rattachées ainsi qu'un lien pour visualiser la page exacte du pdf où l'arrêt apparaît. On a aussi rajouté une fonctionnalité de sauvegarde des arrêts significatifs pour ensuite les télécharger dans un unique fichier.

3 Conclusion et perspectives

3.1 Conclusion

Ce travail est une étude préliminaire sur les possibles applications des techniques de reconnaissance de caractères, de traitement du texte et d'apprentissage automatique en général pour l'accélération et la validation des recherches en histoire du droit.
Nous avons numérisé, extrait et traité près de cent mille arrêts sur une période de plus de cent ans. Nous avons développé un programme de tri du texte, puis une application permettant de faire une recherche personnalisée des arrêts. La lecture de chaque recueil et la retranscription manuelles de tous ces arrêts aurait pris plusieurs années. Notre contribution représente donc une accélération substantielle des travaux reposant sur ce type de documents. Il est évident que ce programme ne remplace pas le travail d'un chercheur, mais le gain de temps pour valider une hypothèse est considérable.

3.2 Perspectives

Bien que notre travail soit utile en soit, il y a un plusieurs idées que nous n'avons pas eu le temps d'explorer et qui pourrait être le sujet de futurs travaux. La continuation directe de ce stage serait d'étendre la détection aux expressions subtiles et non régulières en utilisant une base d'arrêts labélisé plus conséquente et un algorithme d'apprentissage automatique. Une autre route qui pourrait être empruntée est de faire du "Topic Modeling" ou modèle de sujet, qui consiste en un modèle probabiliste permettant de déterminer des sujets ou thèmes abstraits dans un document. Pour des réalisations plus vastes, on pourrait envisager une application intégrant toutes les étapes du traitement sans besoin d'un informaticien pour manuellement définir l'extraction des jugements. Une telle application recevrait un document en entrée, extrairait le texte et prendrait de la part du chercheur le format des arrêts à extraire, ensuite le chercheur serait libre de chercher et analyser le corpus selon ses besoins.