Master Humanités Numériques Parcours Sciences des Données

Rapport de projet

Extraction de texte manuscrit dans deux corpus du XVIII^{e} siècle avec Kraken

Youna FROGER, Alice GYDÉ, Mathis QUAIS

Enseignants : Peter Stokes et Benjamin Kiessling

UE : 9.3.2 A - Fouille des Données et Machine Learning

université de TOURS | Centre d'études supérieures de la Renaissance

Introduction

Dans le cadre du cours de fouille de données, dispensé par monsieur Kiessling et monsieur Stokes aux étudiants de master 2 d'Humanités Numériques, il a été décidé de travailler sur la transcription automatique de manuscrits du XVIIIe siècle. Ce choix s'est présenté en lien avec les supports utilisés lors des différents cours que nous avons suivis, prenant en exemple des manuscrits anciens, et présentant les différents traitement informatique applicables sur ces derniers. Nous formons pour ce projet une équipe de trois étudiants ; Alice Gydé, Mathis Quais et moi-même. Nous avions pour objectif de réaliser des traitements (extraction de texte, amélioration de l'extraction, analyse d'image) sur des manuscrits anciens, et de réaliser un rapport présentant nos résultats afin de les questionner.

Notre choix s'est orienté vers un thème faisant un lien avec la partie Humanités de notre diplôme ; les cahiers de doléances de la révolution française. Ce sont des cahiers rédigés dans tous les bailliages de France entre janvier et avril 1789 en prévision des Etats Généraux convoqués par Louis XVI.

Les cahiers de doléances représentent une source très importante dans l'histoire de la Révolution française, et qui ont fait l'objet d'interprétations différentes selon les historiens. Pour rédiger les cahiers de doléances, des modèles circulaient, mais ils s'avèrent, dans la réalité, qu'ils peuvent être différents dans leur sujet ou dans leur forme en fonction de l'endroit où ils ont été rédigés. Ces documents ont été réalisés dans tous les départements de France, nous devons alors, dans un premier temps, faire une sélection des cahiers afin d'élaborer un corpus de travail. Après prospection de plusieurs sites d'archives départementales mettant à disposition leurs cahiers de doléances, notre choix s'est tourné vers deux départements français : le Finistère et les Yvelines.

Cette décision s'explique par la position géographique des deux départements, et l'accessibilité des documents sur les sites. En effet, puisque le projet nous impose un temps de travail plutôt court, et un matériel informatique limité, nous avions comme contrainte d'arriver a étudier une zone géographique large afin de faire des comparaisons sur le territoire français, sans pour autant traité un nombre très important de documents.

Le choix d'étudier des cahiers proches du centre royal de la France (les Yvelines), et ceux d'une région éloignée et bien différentes (le Finistère) s'est alors avéré intéressant. De plus, ces deux départements possèdent un site internet et une base de données d'archives accessible à tous et intuitive à utiliser. Enfin, nous notons que le site des archives du Finistère a mis en ligne, en plus des cahiers numérisés, une transcription correcte sous format jpg, détail qui nous paraissait intéressant, notamment si l'entraînement d'un modèle de transcription automatique s'imposait.

Le projet de transcription automatique de ce corpus permet d'embrasser deux des enseignements transmis lors de cette année de master 2 d'Humanités Numériques : tout d'abord, la partie fouille de données, objet de ce rapport, où nous nous pencherons sur l'extraction du texte

d'une image vers un texte exploitable, ainsi que l'analyse graphique des manuscrits. Enfin, ce projet nous permet de travailler sur la partie statistique (objet d'un second rapport), où nous étudierons, avec application de formules statistiques, le texte extrait lors de la première partie du projet, afin d'en extraire des thèmes récurrents en fonction des mots les plus utilisés.

Les difficultés que nous risquons de rencontrer sont liées à la nature de nos documents. Puisqu'il s'agit de textes anciens manuscrits, il est probable que nous soyons confrontés à des documents à la lisibilité compromise, à un manque de clarté et à un état de conservation pas optimale. Une des étapes importantes de ce projet sera alors de trouver un modèle de transcription adapté à de l'écriture manuscrite française du XVIIIe siècle.

Pour conclure, le but de ce projet est alors d'utiliser des outils informatiques, via python et la librairie Kraken, afin de les appliquer à notre corpus pour essayer d'en extraire des informations permettant de faire la comparaison entre deux départements français différents.

Corpus et méthodologie

a. Description du corpus

À la fin du XVIIIe siècle, la monarchie française traverse une crise profonde. Le règne de Louis XVI est marqué par une situation économique désastreuse, aggravée par l'intervention française dans la guerre d'indépendance américaine, des mauvaises récoltes et une hausse du prix du pain. Malgré les tentatives répétées de ses ministres pour trouver des solutions, le roi échoue à redresser la situation. En 1789, dans un contexte de mécontentement généralisé, il décide de convoquer les États généraux, notamment pour lever de nouveaux impôts, qui doivent se réunir le 1er mai. Les sujets sont invités à exprimer leurs doléances dans des cahiers rédigés à l'échelle des paroisses, puis synthétisés au niveau des baillages ou des sénéchaussées.

Les cahiers de doléances constituent une source fondamentale pour l'histoire de la Révolution française. Dès le XIXe siècle, Alexis de Tocqueville les décrit dans *L'Ancien régime et la Révolution* comme « le testament de l'ancienne société française, l'expression suprême de ses désirs, la manifestation authentique de ses volontés ». Cependant, les historiens continuent de débattre : à quel point ces cahiers reflètent-ils fidèlement les aspirations de la population de l'époque ? Plusieurs facteurs peuvent nuancer leur authenticité. La faible alphabétisation, encore marquée en 1789 même si cela peut dépendre des régions, confère aux notables locaux un rôle important dans leur rédaction. Par ailleurs, des modèles de cahiers circulent, uniformisant parfois le contenu. En tout cas, des thèmes récurrents se dégagent, tels que la critique du poids de la fiscalité, la défense des privilèges locaux ou encore la demande de réformes dans le système de vote des États généraux.

Nous avons choisi deux départements qui présentent des différences notables. Les Yvelines sont au centre du royaume, et font partie du domaine royal depuis très longtemps. La langue française y est

bien implantée. Elle appartient aux pays d'élections. Le Finistère quant à lui présente plusieurs particularités qui seront sans doute mises en lumière par l'analyse des cahiers. La Bretagne a été intégrée assez tardivement dans le royaume, avec son annexion en 1532. La population y parle encore le breton, et tient à ses privilèges, comme le fait de ne pas payer l'impôt sur le sel, la gabelle. Ce choix de département a aussi été motivé par la disponibilité et la qualité du site de leurs archives. Celui fait par les archives départementales du Finistère est très complet, en offrant à la fois une numérisation des cahiers mais aussi une transcription. Ceux des Yvelines sont également de bonnes qualités, même s'il faut compter plus de cahiers illisibles car trop abîmés, mais surtout l'absence pour ce département de transcription.

En tout cas, nous semblent faire de notre sujet un exemple pratique particulièrement pertinent pour notre formation en humanités numériques, en alliant questionnements historiques, exploitation des ressources patrimoniales mais aussi l'utilisation d'outils informatiques.

b. Méthodologie

Dans un premier temps, nous avons configuré notre environnement de travail en groupe, afin de se partager les documents et le script pour que chacun puisse intervenir dessus. Après une réunion de préparation en groupe, la solution d'utiliser un drive Google partagé et coder sur un notebook Google Colab s'est avéré être la plus pratique pour notre cas. Nous avions conscience de la limite imposée par la puissance de Google Colab, c'est pourquoi nous nous sommes accordés sur le fait de travailler certaines parties du traitement des manuscrits en local sur nos machines respectives qui proposent une puissance supérieure à celle proposée par Google.

Une fois notre environnement de travail mit en place, il s'agit désormais de constituer le corpus d'image sur lequel nous allons travailler. Après quelques prospections sur la possible utilisation de IIIF par les deux sites d'archives, il semble que ce ne soit pas le cas, et que la solution la plus efficace sur le moment est de télécharger les images à la main. Nous nous sommes partagé cette mission afin de ne pas perdre trop de temps. Les images ont donc été téléchargées et renommées pour le Finistère du nom du bailliage correspondant, afin de faire correspondre l'image du manuscrit à l'image de la transcription, en vue d'un éventuel entraînement de notre modèle de transcription. Pour le département des Yvelines, nous nous sommes contenté de garder le nom du fichier de base, puisqu'aucune transcription correcte n'est fournie par les archives du département.

Afin d'exploiter les transcriptions fournies par le département du Finistère, il était nécessaire d'appliquer un traitement OCR à nos documents puisque ces dernières sont des images au format JPG, et non pas des fichiers texte exploitable. Pour cela, nous avons utilisé Tesseract, en local sur une de nos machines, en rédigeant un script shell, et qui a permit de traiter les 194 documents transcrits en quelques minutes.

Nos deux corpus forment deux dossiers d'image, représentant 84 documents pour les Yvelines, et 154 pour le département du Finistère. Il s'agit désormais d'appliquer les traitements de transcription automatique sur nos deux corpus. Pour cela, nous passons d'abord l'ensemble de nos documents à un traitement de l'image, visant a améliorer la transcription. En reprenant les codes réalisés lors d'un cours présenté par monsieur Stokes, nous transformons nos images en les passant en niveau de gris, et en opérant une binarisation en appliquant le filtre threshold_otsu, qui effectue un seuillage automatique à partir de la forme de l'histogramme de notre image en niveau de gris.

Une fois nos images traitées, dans un premier temps, nous avons testé un modèle de reconnaissance de texte dans une image vue en cours. Il s'agit du modèle McCATMuS, qui est un modèle de transcription spécifique pour les documents manuscrits datant du XVIe au XXIe siècle, en français. D'après la description de ce modèle, il nous paraissait comme un choix parfait pour notre cas. Les premiers résultats n'étant pas forcément concluants, et après en avoir discuté avec monsieur Kiessling, nous tentons d'essayer un autre modèle de transcription, nouvellement sortit, intitulé FoNDUE-GD. Après comparaison des résultats, ce second choix s'avère être plus efficace sur nos documents, mais les transcriptions présentent encore quelques erreurs. La solution serait donc d'entraîner le modèle, grâce aux transcriptions correctes des archives du Finistère, que nous avons en format texte. Nous nous sommes appuyé sur la documentation de la librairie Kraken, section training, pour élaborer un script d'entraînement, à partir de nos images et de notre texte correspondant. La partie entraînement de ce projet n'ayant pas abouti, elle sera développée dans la partie résultat du présent rapport.

Lors de notre cours de statistique et data-mining, nous avons appris a utiliser la librairie scikit-learn de Python, qui permet d'appliquer des traitements statistiques à des jeux de données (textuel en ce qui concerne le cours directement), mais nous avons pensé à tester une fonctionnalité en particulier sur notre jeu de données d'image. En effet, il était question ici de faire déterminer automatiquement des clusters parmi nos deux corpus mélangés, afin de voir si la fonction arrive à différencier, sur seul critère : les éléments graphiques, les deux corpus en deux groupes distincts. Pour cela, nous avons utilisé les codes rédigés dans le cadre de notre cours, et nous l'avons adapté pour qu'il prenne en entrée nos images plutôt que du texte, se basant alors sur la couleur et texture de nos images.

Résultats

Partie 1 extraction du texte

Notre projet nous a amenés à travailler avec deux modèles de transcription différents. Les premiers résultats avec le modèle McCATMuS n'étaient pas forcément concluants. En effet, même

si certains mots sont reconnaissables, cela ne constituait pas une base assez solide pour effectuer par la suite sur le traitement statistique sur ces données.

Transcription du cahier Audierne_Cahiers avec McCATMuS (image de base)	**Transcription du cahier Audierne_Cahiers avec McCATMuS (image traitée avec ostu)**	**Transcription correcte des archives**
Lannone No Gufue 5 24. V Napvoeur et CominaEn Cette perrie Cesuitut Des Délibirations Det communuer D'andioine Assimoues a L'Entravrdinaire Et Couroguées scton des formen Cnonie Dana le procés verbat De L'assemblée Tesue ce Jour Treisse avit mi sept cent quatre Vingt neuf. de Dit Nésutiat sonnant Le cahier Des Plaister Doléanien Et resnontrancen qu'ilus Entesdent faire a Sa majisté sur tont ce qui peut- intérenen de Dien de L'etat. Dassemblie Aprés avoir Délibéré a Drretté 1o quit soit Donne a ses réprésentants aux Etats généraux Des pouvvirs Suffisantr pour promirer à La frame uné coustitution heureuse. 2^o que conformément à ce porivipe its soient chargés De Demander le retour periodique Des Etats généraux Et que L'oidre Du tiers y soit toujours couvoqué En nomor Egut a Jamvous 7 N a Dunanimité. 3 n reolr	[[AnsHbIL No1 9ulie d 2 34 Hil N roc e commei so En Cette parsye q Wesuitut Des Délibirations Des commuuser d'andisini Arssimblier à L' eotravidisaire. Et]][[ ]] Couroquées Scton Lei foinier cnonie Danse le pororie ierbât= Di l'ussenibru Tisini c jour Treine prut muz Jept ient quetre Vingt neuf Le Dit Nésusal 'onnant le cassier des Plaisiter Doléanien Et resnontranien qu'ilns Titesdent ffaire a Sa majesté Sur tout ce qui poeut itereaue Bien De Aetat. Sanemble Aprres avoir Dílidere a Drretté a D'unanimité. 1^e guil Soit Donne a Sis representants aux N3 Etâts yéniraux Des pouvvirs Seftisantn pour porduirer a da frame uni constitution hereuse. E que conformément à ce porimipe its Seient charges D. Demandir le retoior periodique Dei Chrti généraux Et que L'ordu Du tierx y soit toujourx couvoque En nomor Cezat a LarGEAAEi. 7 de jressioi p.t 267 L itn ----	AUDIERNE Cahier de doléances Résultat des délibérations des communes d'Audierne assemblées à l'extraor- dinaire et convoquées selon les formes énoncées dans le procès-verbal de l'as- semblée tenue ce jour 13 avril 1789, lesdits résultats formant le cahier des plaintes, doléances et remontrances qu'elles entendent faire à Sa Majesté sur tout ce qui peut intéresser le bien de l'Etat. L'assemblée, après avoir délibéré, a arrêté à l'unanimité : 1° Qu'il soit donné à ses représentants aux Etats généraux des pouvoirs suffisants pour procurer à la France une constitution heureuse. 2° Que, conformément à ce principe, ils soient chargés de demander le retour périodique des Etats généraux et que l'ordre du Tiers y soit toujours convo- qué en nombre égal à celui des deux autres ordres réunis. 3 Que Sa Majesté soit suppliée d'abolir ces distinctions humiliantes qui avilirent les communes

		aux derniers Etats généraux.

Alors, nous avons testé avec un autre modèle de transcription, appelé FoNDUE-GD, qui s'est révélé être davantage efficace que notre premier modèle de test.

Transcription du cahier Audierne_Cahiers avec FoNDUE-GD (image traitée avec ostu)	**Transcription correcte des archives**
Att Ms Gresez + 34. 8beocce et comme1 En Cettre partye esuitut Des Délibérations Des communes D'audiine dessembües a l'Eatravidinaire Et nt Convognées seton Les formes Enoncées Daux le proies virbut Di l'assemblee Tériue ce jour treille arril mil sept ient quatre virngt neuf. Le Dits résutloit conant le cuitier Des Plaisites Dotéanies Et resnontrances qu'eles Esitendent faire a sa majisti sur tout ce qui pent intéresson Le Dien De L'Etat. Laremoie après avoir Délibère a Arretté a L'unanimité. 1^o quil soit Donne a ses répresentants aux . Etats généraux Des pouvvirs suffisants pour procurer a La france une constitiution heurent. 2e que conformément à ce principe ils soients charges De Demander le retoier périodique Des etuts généraux Et que Lordre Du tierx y soit toujours couvoqué En nombre Egut a f LE.fCLi.	AUDIERNE Cahier de doléances Résultat des délibérations des communes d'Audierne assemblées à l'extraor- dinaire et convoquées selon les formes énoncées dans le procès-verbal de l'as- semblée tenue ce jour 13 avril 1789, lesdits résultats formant le cahier des plaintes, doléances et remontrances qu'elles entendent faire à Sa Majesté sur tout ce qui peut intéresser le bien de l'Etat. L'assemblée, après avoir délibéré, a arrêté à l'unanimité : 1° Qu'il soit donné à ses représentants aux Etats généraux des pouvoirs suffisants pour procurer à la France une constitution heureuse. 2° Que, conformément à ce principe, ils soient chargés de demander le retour périodique des Etats généraux et que l'ordre du Tiers y soit toujours convo- qué en nombre égal à celui des deux autres ordres réunis.

Nous remarquons aussi que les cahiers provenant du département des Yvelines semblent être davantage transcrits correctement, témoignant peut-être d’une écriture plus lisible pour cette zone géographique.

Transcription du cahier depot-fonds-images03-collectionnum-serieb-13b-13b11-2-frad078-000-246-13b11-2-001 avec FoNDUE-GD (image traitée avec ostu)

Cahier des plaintes et doteances des habitans
de la Paroisse de Viroflay.
Les habitans de Viroflay pour satisfaire aux désirs du Roi se sont
assemblés en Communauté, et ont arrêté d'une voix unamne Les Plaintes
Doléances et demandes qui suivent.
Se Plaignent Lesdits habitans.
9°o_ Que les bésoins de L'état soient devenus si énormes, par toutes ses depenses, étrangeres
qui n'ont point un rapport direct, ni a sa conservation, ni a l'utilité commune.
par les grandes charges etemplois auxqu'els sont attachés des revenus immenses.
par la solde d'un militaire si nombreux, et qui paroit si peu néces faire en tems
de poix
par les pensions, graces, et bienfaits que la noblesse, et tant d'autres personnes
attachées a la Cour, tirent continuellement des manns du Souverain; Ce qui
Séleve a une somme infiniment disproportionnelle, a celle que cette multitude d'homme
paye à l'état, quoique ces hommes possédent les plus grandes propriétés.
par toutes ces compagnies de Traitane qui partaget si considérablement les revenue
du Souverain.
par la vénalité des charges de toute espece, dons les acquéreurs épuisent le trésor de
l'état, au moyen des gros intérêts qu'ils savent retires de leurs avances.
Toutes ces grandes parties de depense ne sont poiint nécessaires a L'état; elles
sont au contraire une cause d'acroissement de ses bésonts et d'augmentation d'impon
pour le Peuple.
2e Que les habitans des Campagnes, qui ont tres peu de propriété; qui ne sont
la plupart que fermiere des terres qu'ils euttirent, et pour lesquelles ils payent
de forts loyers; qui n'ont généralement aucuns moyens de se faire quelques
profits avantageux, partent seuts la plus grande charge des impôts, qu'ils
en soient accablés, et ce, en grande partie pour les causes enoner en l'articte
etdessus.
Prinire par aftrud O Vavitur. Par Ro
Atmaut

Nous pouvons en conclure que la transcription avec le modèle FoNDUE-GD, offre un résultat presque exploitable, mais des erreurs persistent. Nous avons la possibilité de comprendre globalement le texte mais cela n'est pas parfait.

Partie 2 entraînement du modèle

Au vu des résultats plutôt concluants pour le département des Yvelines, mais toujours fragiles pour le département du Finistère, nous nous sommes penchés sur la question de l'entraînement des modèles de transcription. Pour cela, la lecture de la documentation de Kraken a été un point important de ce projet. Nous avons compris que, pour entraîner un modèle, nous avions besoin de fichiers XML afin de faire le lien entre l'image, son alignement et le texte correct.

Après avoir rassemblé une partie de notre corpus (en se basant sur les cahiers du Finistère puisqu'ils possèdent une transcription correcte.), avec un dossier d'image et un dossier de texte au même nom, il fallait désormais faire le lien entre les deux. Le problème de la génération du fichier

XML par cahiers s'est posé rapidement, puisqu'il n'est pas possible de la réaliser avec Kraken. En effet, d'après la documentation de la librairie, il est nécessaire d'utiliser des outils externes afin de réaliser ce travail, des outils comme Escriptorium. Sous les conseils de nos enseignants, nous avons entrepris de les contacter, malheureusement, sans réponse.

Alors, à l'aide d'une IA, nous avons tenté de générer ses fichiers XML au moyen d'un script Python. Cette tentative s'est avérée peu concluante, et nous avons fait face à tout un tas de message d'erreur, notamment lors de l'utilisation de la commande train de Kraken :

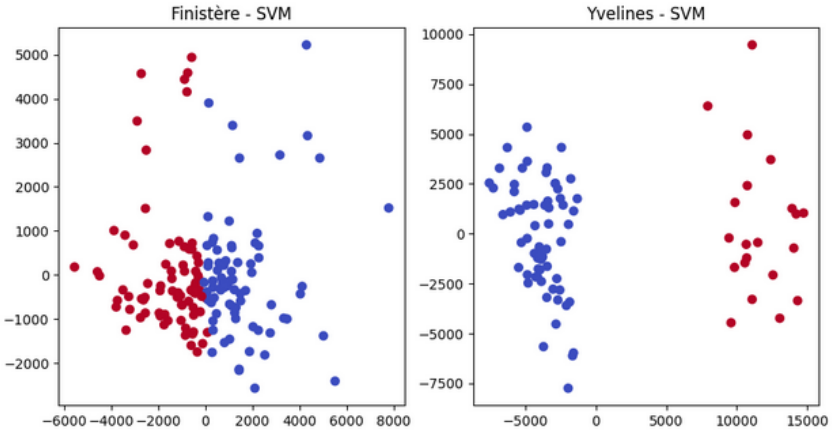
```
ketos train train_images/*.jpg train_xml/*.xml --output mon_model --epochs 10
```

L'erreur indiquait qu'il n'arrivait pas à trouver le chemin de l'image. C'est donc sans entraînement de notre modèle que nous avons généré, au moyen d'un script complet, la transcription de nos documents. Cette opération aurait pris plusieurs heures sur notre espace Google Colab, c'est pourquoi nous avons utilisé une de nos machines pour la réaliser en seulement quelques dizaines de minutes.

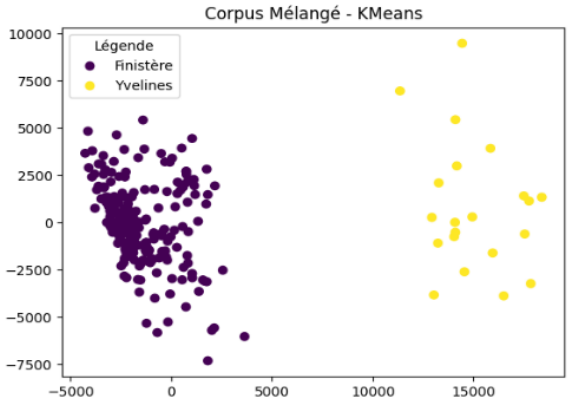
Partie 3 : clusterisation des images

Afin de pallier le problème de l'entraînement de notre modèle, nous avons voulu tenter d'appliquer des formules statistiques à notre corpus d'images. Nous avons utilisé deux algorithmes de clusterisation : K-MEANS et SVM. Les résultats se sont avérés identiques pour les deux essais. Dans un premier temps, nous avons réalisé une clusterisation au sein des deux corpus séparés, pour voir si des différences graphiques ressortent à l'intérieur même d'un corpus d'une même aire géographique. Nous remarquons que pour le corpus du Finistère, les documents forment un groupe assez uni, et il n'existe pas de cluster bien défini. Les documents du département du Finistère semblent être uniformes dans leurs caractéristiques graphiques, avec quelques documents s'écartant du groupe principal.

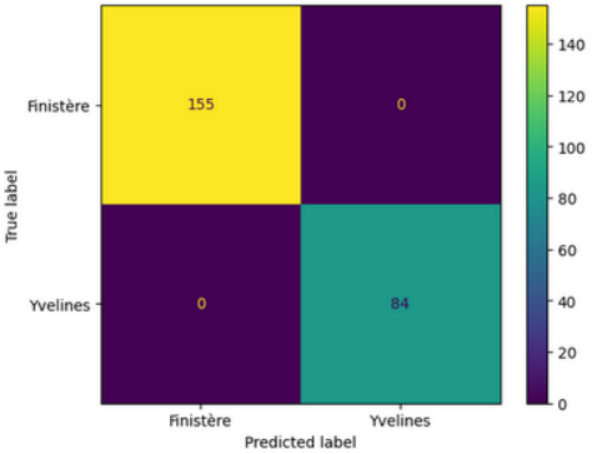
En ce ce qui concerne le département des Yvelines, la clusterisation met en avant deux groupes bien distincts. Il semblerait que dans le lot de document de cette zone géographique, deux types aux caractéristiques visuelles communes existent et ressortent.



Enfin, pour la clusterisation des deux corpus mélangé, les résultats sont plutôt concluant, avec deux clusters bien délimités.



La matrice de confusion générée renforce notre étude en montrant que la clusterisation arrive a bien former les deux groupes sans se tromper dans les corpus d'origines. En effet, on apprend que les documents Finistère ont correctement été détectés comme étant ceux du Finistère, et ceux des Yvelines de la même manière pour ce département.



Conclusion

Pour ce projet de fin d'année, nous avons réalisé sur nos corpus trois traitements. Le premier, servait à uniformiser nos images en les convertissant en niveau de gris et en effectuant une binarisation, censé améliorer les résultats de transcriptions. Ces derniers ont, en effet, été améliorés, bien que la différence ne soit pas réellement flagrante. Cette étape de traitement des images à tout de même retenue notre attention a figure parmi le script de transcription générale.

En ce qui concerne la transcription, après essai avec deux modèles différents, FoNDUE-GD s'est imposé comme le plus efficace, et a été sélectionné pour notre script principal. Les quelques erreurs subsistantes nous ont amenées à tenter d'améliorer le modèle avec un jeu de donnée vérifié, ce qui n'a finalement pas été possible du fait du manque de ressources à notre disposition. C'est pourquoi nos résultats ne sont pas parfaits, mais exploitables pour la partie étude statistique de ce projet.

Une application d'algorithme de clusterisation sur nos images a permit de mettre en avant quelques points intéressants et nous a fait réfléchir sur la partie historique de notre projet :

La clusterisation confirme les résultats de transcription puisque le groupe Finistère forme un groupe très dispersé, alors que le groupe des Yvelines forme deux clusters bien déterminés montrant qu'il y a sûrement deux grands modèles de cahiers pour ce département. Les cahiers des Yvelines sont certainement plutôt bien codifiés, puisqu'ils ont été correctement placés dans le bon groupe sans erreur.