Master Humanités Numériques
Parcours Sciences des Données

Rapport de projet

Statistiques et data mining de textes manuscrits dans deux corpus du XVIII^{e} siècle

Youna FROGER, Alice GYDÉ, Mathis QUAIS

Enseignant : Carlos-Emiliano Gonzales-Gallardo

UE : 9.3.1 – Statistiques et Data Mining

1 Introduction

Dans le cadre du cours de Statistiques et Data Mining, dispensé par Monsieur Gonzales-Gallardo aux étudiants de master 2 d'Humanités Numériques, il a été décidé de travailler sur le même corpus d'études que pour le cours de Fouille des données. Ce choix est lié aux supports utilisés dans les différents cours, qui prenaient pour exemple des manuscrits anciens et illustraient les traitements informatiques applicables à ces documents. Nous formons pour ce projet une équipe de trois étudiants ; Alice Gydé, Mathis Quais et Youna Froger. Notre objectif était d'effectuer des analyses statistiques (reconnaissance des corpus d'origine, clusterisation des images et du texte) sur des manuscrits anciens et de présenter nos résultats dans un rapport accompagné de nos analyses.

Notre choix s'est orienté vers un thème faisant un lien avec la partie Humanités de notre diplôme : les cahiers de doléances de 1789. Ce sont des cahiers rédigés dans tous les bailliages de France entre janvier et avril 1789 en prévision des Etats généraux convoqués par le roi 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. D'une part, Nous souhaitions étudier une zone géographique large afin de comparer différentes régions françaises. D'autre part, il ne fallait pas non plus traiter un trop grand nombre de documents puisque le projet nous impose un temps de travail plutôt court, et un matériel informatique limité.

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 d'un premier rapport), où nous nous penchons 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, 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 loin d'être optimal.

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

2 Corpus et méthodologie

2.1 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.

2.2 Méthodologie

Dans un premier temps, nous avons configuré notre environnement de travail en groupe afin de partager les documents et le script, permettant à chacun d'intervenir efficacement. Après une réunion de préparation, nous avons décidé d'utiliser un Google Drive partagé comme solution centrale. De même, travailler sur un notebook Google Colab s'est révélé être la solution la plus pratique. Cependant, étant conscients des limites de puissance de Google Colab, nous avons convenu de traiter certaines parties des données en local sur nos machines personnelles, qui offrent des capacités supérieures.

Une fois cet environnement de travail mis en place, nous avons commencé à travailler sur le corpus d'images en appliquant les méthodes vues en cours de fouille de données, afin de préparer le matériel nécessaire. Nous avons collecté les images des cahiers ainsi que les transcriptions fournies par le département du Finistère. Ces deux ensembles constituent nos corpus : 84 documents pour les Yvelines et 154 pour le Finistère. L'étape suivante consistait à appliquer des traitements de transcription automatique sur ces deux corpus.

Pour effectuer cette tâche, Monsieur Kiessling nous a proposé d'utiliser un nouveau modèle de transcription intitulé FoNDUE-GD. Ce modèle s'est montré relativement performant sur nos documents, mais les transcriptions comportaient encore quelques erreurs. Faute de temps, nous n'avons pas pu entraîner le modèle pour améliorer ces résultats.

Par ailleurs, lors de notre cours de statistique et de data-mining, nous avons exploré la bibliothèque scikit-learn, utilisée pour appliquer des traitements statistiques sur des données textuelles. Cela nous a inspirés à tester une fonctionnalité sur nos données d'images. Nous avons cherché à déterminer automatiquement des clusters à partir de nos deux corpus mélangés, en nous appuyant uniquement sur des critères visuels comme la couleur et la texture des images. Pour ce faire, nous avons adapté les codes vus en cours afin qu'ils prennent en

entrée nos images au lieu de textes. Les résultats obtenus (accessibles sur GitHub) ont permis de constater une certaine différenciation des deux corpus.

Une fois les textes transcrits et stockés dans un fichier, nous avons envisagé leur analyse. Nous avons suivi les méthodes vues en cours : tokenisation des chaînes de mots, suppression des stopwords à l'aide de la bibliothèque NLTK, puis inclusion dans un corpus unique. Les textes ont ensuite été vectorisés avec TF-IDF, réduits en dimensions avec PCA, et enfin regroupés en clusters grâce à KMeans. Au départ, nous fixions notre nombre de clusters de manière arbitraire.

L'étape suivante a été d'identifier les thèmes présents dans les textes. Nous avons récupéré les colonnes de la matrice TF-IDF converties en tableau, puis calculé la moyenne des scores TF-IDF pour chaque mot. Cette méthode a permis d'extraire les dix mots les plus récurrents. Enfin, nous affichons les textes par clusters. Cette partie est plus intéressante pour ceux du Finistère, car nous avons retenu les noms des paroisses.

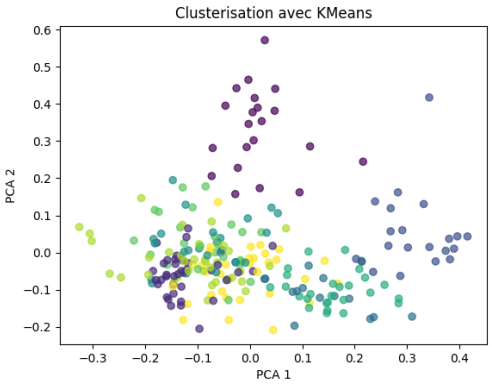
Bien que notre pipeline fonctionnait correctement, les résultats ne nous semblaient pas totalement satisfaisants. Nous avons donc introduit plusieurs ajustements. En observant les mots les plus fréquents dans chaque cluster, nous avons enrichi notre liste de stopwords en y ajoutant des termes non pertinents. Nous avons également affiné la vectorisation TF-IDF en excluant les termes apparaissant dans plus de 85 % des documents ou dans moins de cinq documents. Enfin, nous avons décidé de déterminer automatiquement le nombre optimal de clusters en utilisant le score de silhouette, abordé en TD.

Malgré ces ajustements, les résultats sont restés assez similaires malgré quelques améliorations. De même, en retirant les noms propres, tels que les prénoms, de la liste des stopwords nous avons constaté des clusteurs plus significatifs. En effet, une partie des cahiers commence par une liste des participants à leur rédaction, tandis que pour d'autres, cette liste se trouve à la fin, voire est absente. Nous avons donc exclu les termes tels que "Jean", "Yves", ou "François". En relaçant le code plusieurs fois, nous avons fini par constituer une liste des noms et prénoms les plus usuels, permettant ainsi de bien les supprimer.

En plus de cette version du code où les deux corpus sont mélangés, nous avons avons produits une version de l'analyse séparée pour les Yvelines et le Finistère. Enfin, étant donné que nous disposions également des transcriptions faites à la main des premières pages des cahiers du Finistère, nous les avons aussi soumises à ces analyses.

3 Résultats et analyse

3.1 Affichages des clusters

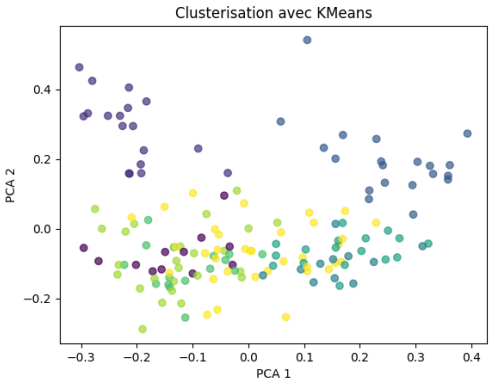


(a) Clusteration

```
Thèmes identifiés par cluster :
Cluster 0 : tous, droits, anne, mariage, egale, mars, demandes, contrat, suivant, bretagne
Cluster 1 : plus, paroisse, dont, partie, tout, bois, roy, tous, fait, etats
Cluster 2 : leon, mars, monsieur, lettres, lieu, jour, reglements, conformement, dits, dernier
Cluster 3 : lettres, lecture, cahier, royal, tenue, cloche, faite, 1789, faites, versailles
Cluster 4 : tous, plaintes, paroisse, sire, si, habitans, sans, meme, cahier, tout
Cluster 5 : yres, paroisse, jour, corentin, procureur, corps, deliberations, lieu, dimanche, ditte
Cluster 6 : soyent, toutes, droits, fait, 50, demandons, celle, sans, ee, supression
Cluster 7 : art, comme, droit, tout, sans, demandent, toutes, paroisse, article, suppression
Cluster 8 : dela, tiers, personne, lu, roy, faire, etat, voix, tous, communes
```

(b) 10 mots les plus présents dans chaque clusters

FIGURE 1 – Résultats pour les corpus des Yvelines et du Finistère

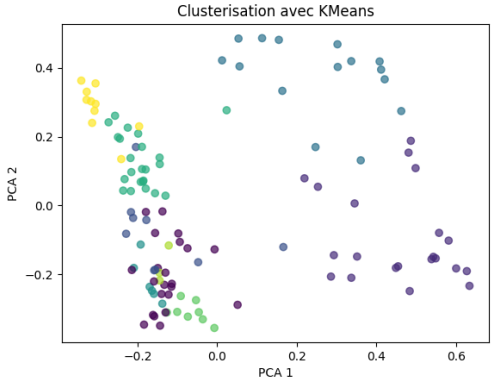


(a) Clusteration

```
Thèmes identifiés par cluster :
Cluster 0 : soyent, demandons, grandes, trop, sans, celle, plus, seuls, droits, empots
Cluster 1 : tous, droits, egale, demandes, bretagne, anne, contrat, mariage, autres, droit
Cluster 2 : leon, mars, lettres, reglements, jour, monsieur, conformement, dits, lieu, paroisse
Cluster 3 : yres, procureur, corentin, paroisse, germain, registre, vincent, corps, extrait, roi
Cluster 4 : paroisse, faite, reglement, lettres, yres, convocation, dite, habitans, lecture, tous
Cluster 5 : plus, sire, si, roy, plaintes, personne, france, porter, droit, paroisse
Cluster 6 : tous, ordres, autres, fait, faire, paroisse, sans, nombre, charges, toutes
Cluster 7 : etat, lu, dela, lieu, droit, deliberations, comme, tout, jour, lettre
```

(b) 10 mots les plus présents dans chaque clusters

FIGURE 2 – Résultats pour le corpus du Finistère

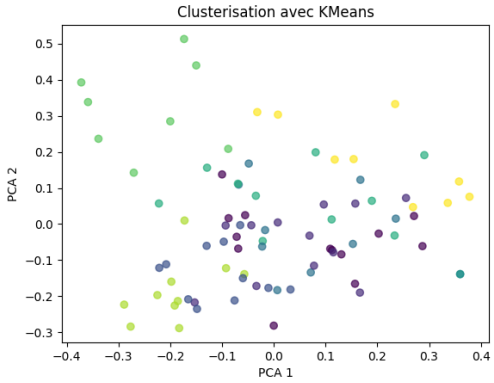


(a) Clusteration

```
Thèmes identifiés par cluster :
Cluster 0 : plus, sire, si, ni, faire, trop, paroisse, roi, majesté, général
Cluster 1 : droits, tous, bretagne, supprimer, vassaux, conserver, duchesse, anne, titres, franchises
Cluster 2 : tiers, ville, etat, quimper, etats, députés, généraux, générale, concarneau, habitants
Cluster 3 : règlements, fera, léon, sénéchal, lettres, droits, paroisse, bretagne, jour, signé
Cluster 4 : tous, déclarons, citoyens, biens, charges, distinction, obéissance, sans, convoqués, clergé
Cluster 5 : paroisse, jour, ladite, procureur, convocation, délibérations, règlement, corentin, général, habitants
Cluster 6 : demandons, brest, supprimés, suppression, capitation, plus, villes, grands, cet, corvée
Cluster 7 : voix, ordre, assemblées, quelconques, pourvu, ordres, admis, égal, tête, tiers
Cluster 8 : lecture, publication, messe, déclaré, faites, paroisse, monsieur, accoutumée, ainsi, connaissance
```

(b) 10 mots les plus présents dans chaque clusters

FIGURE 3 – Résultats pour le corpus du Finistère déjà transcrits



(a) Clusteration

```
Thèmes identifiés par cluster :
Cluster 0 : article, articles, paroisse, nation, comme, bailliage, versailles, habitans, ainsi, ladite
Cluster 1 : plus, biens, tout, tous, pouvoir, roy, ordres, etats, celle, chaque
Cluster 2 : terre, paroisse, publique, plus, dont, non, proprists, peut, partie, st
Cluster 3 : plaintes, lieu, dit, paris, ditte, ville, droit, jour, cahier, etre
Cluster 4 : rendre, ans, roy, tous, doivent, datte, execution, rediger, representer, tems
Cluster 5 : toutes, cahier, habitans, tous, si, somme, payer, faire, plaintes, ceux
Cluster 6 : tout, sans, art, toutes, comme, suppression, classes, aucun, aides, demandent
Cluster 7 : paroisse, gibier, plus, bois, partie, grand, depuis, laintes, beaucoup, doleances
Cluster 8 : etats, royal, art, lettres, present, etat, ordre, plaintes, ordres, droit
```

(b) 10 mots les plus présents dans chaque clusters

FIGURE 4 – Résultats pour le corpus des Yvelines

Notre script nous donne 8 ou 9 clusters dans chacune de ces configurations. Comme le montre les représentations graphiques, la plupart des clusters ne sont pas très éloignés des uns des autres. A chaque fois, les 10 mots les plus présents dans chacun d'entre eux nous indiquent quel pourrait être leur thème. Si pour certains clusters ils sont plus ou moins vides de sens, ce n'est pas le cas pour tous.

3.2 Thèmes

Plusieurs clusters nous permettent bien d'identifier les thèmatiques que nous évoquions dans notre présentation du contexte historique de la période. Il y

a des points communs entre les deux départements. Nous avons par exemple évidemment des clusters relatifs à l'écriture même des cahiers (avec des mots comme "convocation", "rédaction"...), avec aussi des mentions adminstratives (les "baillages") et à la géographie locale. Les grandes villes de ces régions sont mentionnées comme Brest pour le Finistère pu Versailles pour les Yvelines.
Pour le Finistère, nous retrouvons un thème très important pour les Bretons de l'époque, le maintien des privilèges de leur région. Dans le cluster 1, on retrouve mention de la "Bretagne", de "Anne, du "mariage et du "contrat". Il s'agit ici du mariage de Anne de Bretagne avec le roi de France Charles XVIII en 1491, puis en 1499 avec Louis XII. Même si à terme le duché est intégré au sein du royaume de France, il conserve ses privilèges, en matière fiscale par exemple. Dans le cluster correspondant dans le corpus déjà transcrit du Finistère, cela est encore plus précisé : il faut "conserver" les "franchises", c'est-à-dire les privilèges ou "libertés" dont jouissent les Bretons".

Dans ce même corpus, nous retrouvons une revendication centrale de l'époque, la réforme du système de vote aux Etats généraux, ce qui semble être le thème du cluster 7. On y retrouve les "voix", une "assemblée", les "ordres", le "tiers", "égal" et "tête". Le tiers état revendique à l'époque un vote par tête au états généraux, contrairement à la noblesse et au clergé qui veulent le maintien du vote par ordre, qui les avantagent. Cette revendication est largement partagé à l'époque. Un autre thème important est la pression fiscale et les corvées, qu'on retrouve dans le cluster 6 avec des mots comme "suppression", "supprimés", "corvée" et "capitation". Les corvées correspondent au travail gratuit que devait effectuer les memebres du tiers état pour les privilégiés, par exemple participer à refaire les routes. La capitation est un impôt direct qui pèse sur les sujets du royaume, à part pour les nobles et les clercs. En bref, il s'agit bien de revendications sociales, pour plus d'égalité.

De manière générale, les thèmes qui ressortent pour les cahiers du Finistère semblent plus revendicatifs que ceux des Yvelines. Dans ce département, il n'est pas directement fait mention de ces éléments, même si des termes comme "suppression" ou "aide" peuvent peut-être y faire mention. Néanmoins, un sujet important pour la population majoritairement rurale de l'époques peut se trouver dans le cluster 7 de ce département avec des "bois" et "gibier". Il s'agit sans doute de demandes liées à l'exploitation des forêts. Les membres du tiers états demandent de pouvoir récolter le petit bois, notamment pour allumer les feux, et de chasser, ce qui est encore réservé à la noblesse.

4 Conclusion

L'analyse statistique des corpus que nous transcrits par HTR nous a bien permis de mettre en relief différentes revendications populaires à la veille de la Révolution, qu'il s'agisse du maintien de privilèges locaux ou d'une plus grande égalité sociale. Nous pouvons nous demander si l'absence de certaines revendica-

tions dans les clusters pour le corpus des Yvelines est significatif d'une attitude plus passive vis-à-vis de la monarchie (du fait de la proximité avec Paris et de l'intégration plus ancienne), ou bien simplement à des problèmes lors de la transcription.

Pour cette partie du travail, notre corpus a pu nous poser problèmes à cause de la structure des cahiers. De fait, nous n'avions pas pensé que certains étaient signés sur la première page, ce qui fait que de nombreux noms et prénoms parasitent notre analyse. Nous en avons exclu une bonne partie au fur et à mesure. Malgré tout, les revendications principales indiquées dans l'historiographie ressortent bien dans les clusters. Cela témoigne aussi que sans doute les rédacteurs des cahiers commençaient par évoquer leurs demandes les plus pressantes. De même, bien que les résultats de nos transcriptions avec Kraken soient convenables, les différences avec le texte déjà transcrit restent notables. Avec des modèles mieux entraînés, nous aurions probablement obtenu des résultats différents.

Pour avoir une étude plus significative sur le sujet, il faudrait modifier différents éléments, par exemple prendre plusieurs pages des cahiers, voire leur intégralité. De même, peut-être que des méthodes statistiques ou de traitement du language plus avancées pourraient révéler d'autres éléments, comme par exemple l'analyse des sentiments.