Qui a écrit Molière ?

CHIKHANI Charles ZAPFACK MESSIAN Jasen Steve

Encadrant : M. Hervé Fournier

Université Paris Cité

02 Juin 2023

Abstract

Contents

1 Introduction

1.1 Présentation du sujet : la remise en question de l'attribution des pièces de Molière à l'auteur lui-même

Depuis plusieurs décennies, une question persiste dans le domaine de la littérature : l'attribution des pièces de Molière à l'auteur lui-même est-elle remise en question ? Cette controverse a été initiée par Pierre Louÿs, un romancier du XXe siècle, qui a suggéré que Pierre Corneille aurait pu être l'auteur véritable des pièces de Molière. Cependant, cette théorie repose sur des fondements fragiles et ne bénéficie d'aucune preuve concrète. Malgré cela, la rumeur persiste, alimentée par des éléments tels que l'éclosion tardive de Molière en tant qu'auteur, son prétendu manque d'éducation et de culture, ainsi que l'absence de preuves manuscrites permettant de réfuter directement cette hypothèse.

1.2 Importance de l'analyse textuelle et statistique dans la résolution de cette question

Au début des années 2000, Cyril et Dominique Labbé, deux chercheurs, ont avancé l'idée selon laquelle Corneille aurait écrit pour Molière. Leur méthode consiste à mesurer une "distance inter-textuelle" qui évalue la différence de lexique entre les textes des deux auteurs. Si cette distance ne dépasse pas un certain seuil, les deux pièces sont considérées comme écrites par le même auteur. Ils se basent également sur le fait que de nombreux dramaturges de l'époque signaient leurs œuvres sous le nom de "comédien poète", ce qui permettait aux véritables auteurs de rester anonymes tout en bénéficiant de la promotion et de la représentation de leurs pièces par les acteurs.

Cependant, cette méthodologie a été contestée par d'autres chercheurs. Certains ont souligné que l'implémentation de la méthode de Cyril et Dominique Labbé pourrait "lisser artificiellement les différences entre les auteurs", en utilisant une distance euclidienne qui accorde trop de poids aux lemmes fréquents, réduisant ainsi la disparité entre les fréquences observées de différentes formes. D'autres approches ont été proposées pour résoudre le problème de l'attribution des comédies de Molière. Certaines méthodes utilisent une analyse textuelle et statistique pour comparer les styles d'écriture, tandis que d'autres adoptent des approches plus qualitatives en examinant les intrigues, la versification et les sujets choisis dans les pièces.

1.3 Hypothèses et problématique

Dans ce rapport, nous examinerons deux hypothèses qui remettent en question la paternité des œuvres de Molière.

- La première hypothèse suggère que Molière aurait fourni des brouillons à Pierre Corneille, qui aurait ensuite versifié les pièces, peut-être avec l'aide de son frère. Selon cette hypothèse, Molière aurait créé les intrigues, mais la versification aurait été réalisée par Pierre Corneille, sans recevoir un crédit explicite. La deuxième hypothèse soutient que Molière n'aurait ni écrit les intrigues ni les vers de ses pièces, et qu'il n'aurait été qu'un nom célèbre utilisé pour promouvoir les pièces et dissimuler le véritable auteur.

- La deuxième hypothèse, suggère que Molière n'aurait pas écrit ni les intrigues ni les vers de ses pièces, et qu'il n'aurait été qu'un nom célèbre utilisé pour aider à promouvoir la pièce, pour satisfaire l'ego de l'acteur principal/metteur en scène et pour dissimuler le nom de l'auteur réel. Selon cette hypothèse, les sujets choisis dans les pièces de Molière, comme les Précieuses Ridicules, auraient été plus proches des intérêts habituels de P. (ou T.) Corneille et ne refléteraient aucune influence de Molière. Si cela était vrai, tous les indicateurs devraient montrer que le vocabulaire et le style de Molière n'existent pas, et les pièces de Molière devraient être confondues avec celles de P. Corneille selon chacun des six critères évalués dans cette étude.

2 Compréhension de l'article de Florian Cafiero et Jean-Baptiste Camps

Le but de nos chercheurs sera de réfuter ces 2 hypothèse, nous allons plus en parler de la méthodologie et la procédure pour le faire

2.1 Méthodologie

cela sera constituer de 3 sets de d'oeuvres que nous appellerons corpus.(def. un corpus désigne une collection importante et structurée de textes ou de documents utilisée pour l'analyse linguistique) et cela peut comprend généralement variété de textes telle que des livres, des articles et etc.

- le première le corpus exploratoire est constitué d'un large échantillon de comédies en vers. Cet échantillon comprend des pièces d'au moins 5000 mots, pour les auteurs ayant écrit au moins trois comédies. Il inclut des pièces de théâtre de 12 auteurs.

- le deuxième corpus le corpus final est construit Pour obtenir un résultat plus lisible et moins biaisé, ils vont se concentrée sur les sous-genre. Afin d'éviter les biais liés aux sous-genres, ils vont exclure les comédies héroïques et les courtes farces comiques. Pour éliminer le bruit ajouté par de nombreux phénomènes (co écriture, plagiat, attribution incertaine, etc.) sans

rapport avec les hypothèses présenter plus tôt, ils choisissent de concentrer uniquement sur cinq auteurs majeurs de l'époque. Ce corpus final comprend 37 pièces de T. et P. Corneille, Molière, Rotrou et Scarron

- le troisième corpus qui les sert de test pour vérifier la précision de leur approche et consiste de comédie en vers écrit après la mort de P.Corneille et Molière

2.2 les caractéristique d'étude de texte (studies features)

sur chacun de ces corpus , nos chercheurs on appliquer les caractéristiques qui suit

- Lexicon: un lexicon désigne l'ensemble des mots et des unités lexicales d'une langue, ainsi que leurs sens, leurs formes grammaticales et leurs relations. Des exemples sont Maison, Chien, Arbre.
- Rhyme Lexicon: fait référence à un lexique spécifique aux rimes.Il s'agit d'une liste qui répertorie les mots et les expressions en fonction de leurs sonorités et de leurs similarités phonétiques, en particulier en ce qui concerne la dernière syllabe ou les sons finaux des mots. exemple rat, chat, chapeau, bateau, plateau tous les mots se terminent par le son "-o" ou "-au".
- Word forms
- Affixes
- Morphosyntactic sequences
- Mot Fontionelle/ mot-outils (function words): sont des mots grammaticaux qui ont principalement un rôle syntaxique ou grammatical dans une phrase plutôt qu'un sens lexical spécifique.Les mots fonctionnels sont souvent des prépositions, des conjonctions, des pronoms, des déterminants, des adverbes de liaison et des particules grammaticales. Comme les prépositions : de, à, dans, sur, sous etc.

2.3 Choisir la caractéristique

En Générale, La sélection des caractéristiques les plus fiables et informatives pour l'analyse stylistique de texte est une question qui a fait l'objet de nombreuses contributions. Afin d'augmenter la fiabilité des analyses, dans un corpus contenant des textes de longueurs variables, ils ont décidé de sélectionner des caractéristiques avec une approche statisticienne en fonction du niveau de confiance et de la marge d'erreur que nous pouvions obtenir même pour le plus petit échantillon disponible dans notre corpus.

La taille minimale de l'échantillon ,n, a été calculée en utilisant la formule suivante où p est la probabilité moyenne de la caractéristique dans notre corpus,

utilisée comme estimation de la probabilité de la population pie, z est le niveau de confiance et e est la marge d'erreur de l'estimation de probabilité. Nous avons fixé z de manière à obtenir un niveau de confiance supérieur à 90 et e = 2s, où s est l'écart-type de la caractéristique dans le corpus.

$$n = p(1-p)(z/e)^2$$

mais pour cela les caractéristique doivent suivre une distributions gaussiennes

2.4 Algorithme de Clusterization

Dans une approche d'analyse statisticien dans l'attribution d'auteur grâce au machine Learning , Algorithme de clusterization hiérarchique est appliqué a chacun des corpus cité plus haut. cette algo est un type spécifique d'algorithme de regroupement utilisé en apprentissage automatique et en analyse de données. Il s'agit d'une approche ascendante où chaque point de données est considéré initialement comme un cluster séparé, puis fusionné de manière itérative en fonction de leur similarité

pour un rappel un cluster dans le contexte de l'analyse de regroupement, désigne un groupe de points de données partageant des similitudes ou présentant des schémas lorsqu'ils sont comparés à d'autres points de données. Ces points de données sont regroupés en fonction de certains critères, tels que la proximité dans l'espace des caractéristiques ou la similarité dans les valeurs des attributs. Les clusters sont formés en fonction des mesures de similarité ou de dissimilarité utilisées dans l'algorithme de regroupement.

les mesures de similarité et de dissimilarité sont calculer grâce a des métrique

2.5 Métrique de algorithme

Le choix de la mesure de distance et du critère de liaison (par exemple, liaison complète, simple ou moyenne) détermine la manière dont la similarité entre les clusters est mesurée lors du processus de fusion.

- la métrique de distance entre point utiliser dans cette algo la similarité ou la dissimilarité entre les points de données ou les clusters. Elle détermine comment l'algorithme quantifie la distance ou la dissimilarité entre les observations afin de former des clusters. exemple de métrique couramment utilisées , la distance euclidienne, la distance de Manhattan, métrique de cosinus.

  Nos chercheurs ont quand a eux utilisés la distance de Burrow's delta et le min-max. Nous détaillerons la distance de Burrow.

  cette distance calcule la distance de Manhattan entre les scores z des fréquences de ces caractéristiques dans les textes de deux auteurs. Elle mesure la dissimilarité entre leurs styles d'écriture en prenant en compte les différences dans les fréquences normalisées de ces caractéristiques.

rappel le Z-score est une notion de statistique qui quantifie le nombre d'écart-types une observation ou un point de données est éloigné de la moyenne d'une distribution.

la formule de calcule de la distance de Burrow :

$$delta(A, B) = \sum_{i=1}^{n} abs((A_i - B_i)/\sigma i)$$

où les Ai et Bi sont des fréquence de mots dans le texte. Le sigma i est la variance de utilisation du mot

- la métrique d'union(linkage) une méthode spécifique utilisée dans le regroupement hiérarchique pour déterminer la distance entre les clusters lors du processus de fusion.Par exemple le ward linkage , single linkage etc

  le calcul de distance se base sur cette formule, prenons l'exemple le cluster C1 et C2, G1 et G2 leurs centroides respectifs, n1 et n2 le nombre d'individus dans les clusters respectifs. La distance d entre les clusters, à minimiser, est définie par l'équation suivante :

$$d^2(C1, C2) = n1 * n2/n1 + n2.d^2(G1, G2)$$

2.6 Dendrogramme

Après l'application de l'Algorithme de clusterization sur chacun des corpus. On a comme résultat, un dendrogramme pour chacun des caractéristique

de cette interprétation , on en ressort avec une réfutation des hypothèse énonce plus haut car comme indiquer sur la photo on peut voir une distinction du cluster de Molière comparer au autre auteur.

3 Notre expérience

3.1 Description de l'expérience que nous avons menée

Dans le cadre de cette étude, nous avons mené une expérience visant à déterminer la paternité des textes de Molière. L'objectif principal était de développer une méthodologie pour identifier les caractéristiques distinctives du style d'écriture de Molière.Nous comparons ensuite les textes de Moliere et ces de Corneille.

3.2 Choix des outils d'analyse textuelle et statistique utilisés

Pour mener a bien notre expérience, nous avons opté pour la bibliothèque **NLTK** (*Natural Language Toolkit*) de Python,très Utiliser dans la discipline du **NLP** (*Natural Language Processing*).Ensuite nous utilisons aussi l'agorithme open-source de Facebook, **fasttext** qui nous permets de vectoriser chacun de notre text

pour notre approche Machine learning, nous avons utilisé la librairie *scikit-learn* de python qui offre des fonctionnalités avancées pour l'analyse de texte et la classification telle que algorithme de clusterisation hierachique et le K-means

3.3 Nos etapes

- Notre corpus est constituer des Oeuvres de Corneile en format pdf, et celle de Moliere en format xml, que nous avons prises depuis le depot git des chercheurs
- le prétraitement de notre corpus se base sur la caractéristique des affixes dans notre textes. Cette étape nous fournira des fichier texte des œuvres de molières et de corneille contenant la listes de tout les mots sans suffixes, la complexité en temps globale de cette étape est de $O(n^2)$
- la vectorisation des fichiers txt.
- Nous avons appliqué nos algorithmes de clustering, qui comprennent le k-means utilisant la distance euclidienne et la métrique du cosinus, ainsi que l'agglomerative Clustering utilisant la distance de Jaccard. Les mesures d'agrégation des clusters que nous avons utilisées sont "complete" et "average". En ce qui concerne la complexité de nos algorithmes, elle dépend de plusieurs facteurs tels que la taille des données et le nombre de clusters.
- on a la les dendrogramme correspondant a chaque métrique avec leurs critère d'union

3.4 Observation

Comme nous l'avons constaté lors de l'application de l'algorithme hiérarchique sur des données textuelles, l'utilisation de la distance euclidienne, telle qu'utilisée par D. Labbé, ou la métrique du cosinus, ont montré une moindre fiabilité et conduisent à des conclusions différentes de celles obtenues en utilisant une métrique de Jaccard, qui se base davantage sur une approche ensembliste

3.5 Nos différents résultas

Pendant nos expériences pratiques, nous avons eu l'occasion de tester différentes méthodes d'algorithme de regroupement (clusterisation) de diverses manières, en nous appuyant sur différentes caractéristiques. Par exemple :

- Après avoir analysé les digrammes et les trigrammes, nous avons comparé les œuvres de Molière et de Corneille. Les résultats se présentent sous la forme d'un fichier texte répertoriant toutes les paires de mots différents fréquemment associés ainsi que les groupes de trois mots (trigrammes) qui apparaissent ensemble avec leur fréquence respective.

  Une observation intéressante est que le digramme ("peut", "être") est plus fréquent, avec 235 occurrences, dans les œuvres de Molière, tandis que chez Corneille, c'est le digramme ("après", "avoir") qui revient le plus souvent, avec 117 occurrences.

  En ce qui concerne les trigrammes, le triplet de mots le plus fréquent chez Molière est ("monsieur", "oui", "monsieur"), tandis que chez Corneille, c'est ("plus", "puis", "faire").

4 Conclusion

À partir de nos différentes expériences, nous pouvons conclure que le style d'écriture de Molière présente une unicité distincte. Les caractéristiques spécifiques que nous avons observées dans ses œuvres, telles que l'utilisation fréquente de certains digrammes et trigrammes, ainsi que des motifs récurrents, ont contribué à identifier et distinguer son style d'écriture des autres auteurs. Ces résultats renforcent l'idée que Molière possède un style d'écriture unique et reconnaissable.