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 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.

1.3 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

1.4 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

1.5 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

1.6 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").

  En outre, grâce à l'analyse des nuages de mots, nous constatons une utilisation quasi similaire du mot "plus'‘. Cette utilisation fréquente du mot "plus" par nos différents auteurs peut rendre difficile l'attribution précise de l'auteur en se basant uniquement sur la distance euclidienne intertextuelle, comme celle proposée par D. Labbé. Cela souligne l'importance

d'utiliser des métriques et des méthodes d'analyse supplémentaires pour une identification plus précise de l'auteur

1.7 interpretation dendrogramme

Lors de l'interprétation des résultats sur les dendrogrammes, nous constatons que certaines œuvres ne sont pas bien positionnées lorsque nous utilisons la distance euclidienne et la métrique du cosinus. En revanche, avec la métrique de Jaccard, nous observons des regroupements plus compacts, que ce soit pour les œuvres de Molière ou de Corneille. Cela suggère que la métrique de Jaccard est plus efficace pour capturer les similitudes et les différences entre les textes, grâce a son approche ensembliste qui le différencie des autres métriques. Ce qu'appuie notre étude, en considérons que l'approche et les choix des caractéristiques et des métriques sont plus pertinents dans la méthodologie développée par Mr Caffiero et Mr Camps, est qu'une étude stylométrique sur des textes littéraires si proches nécessite une approche plus fine et plus précise. La diffculté de l'attribution de l'auteur est d'autant plus grande que les auteurs sont proches stylistiquement.

D'après nos différentes expériences, nous pouvons conclure que le style d'écriture de Molière présente une unicité remarquable. Les caractéristiques spécifiques que nous avons observées dans ses œuvres, telles que l'utilisation fréquente de certains bigrammes et trigrammes, ainsi que des motifs récurrents et la fréquence des affixes, nous ont permis d'identifier et de distinguer son style d'écriture par rapport à celui des autres auteurs. Ces résultats renforcent l'idée que Molière possède un style d'écriture unique et reconnaissable.