Projet de Traitement Automatique des Langues : Classification par genre

Eleonore Bartenlian, Adrien Pavao

Octobre 2017

Introduction

Pour ce projet, nous avons travaillé sur le corpus "Pièces de théâtre classiques en français".

Plusieurs genres représentent le théâtre classique du XVIIème siècle en France ; très essentiellement la tragédie, mais également la comédie, ainsi que la tragi-comédie. Dans ce corpus, nous disposons de 50 genres différents, allant de la comédie au mélodrame en passant par le drame liturgique.

Nous avons fait de l'apprentissage sur ces textes et cela nous a permis de retrouver à quel genre appartient chaque texte dans un corpus de test à partir de données extraites des oeuvres du corpus et du fichier XML telles que la date, l'auteur, ou encore les mots les plus représentatifs.

Pour cela, nous avons effectué un pré-traitement sur les textes du corpus et parsé les fichiers XML pour récupérer les informations contenues dans le header et ainsi obtenir une liste d'attributs qui nous permettraient par la suite de faire un apprentissage sur le genre de chaque texte. Nous avons également analysé les mots du corpus dans son ensemble pour ciblé ceux qui représentent particulièrement bien un genre donné.

1 Description des données

Le corpus est constitué de 1030 textes de pièces de théatre classiques en français de 310 auteurs différents et 50 genres. Chacune est stockée dans un fichier au format XML.

```
Ballet
Mélodrame
Farce
Ambigu poissard
Proverbe
Bergerie
opuscule dramatique
Pastorale héroïque
Mascarade
Pantomime
Comédie
Dialogue
Piece
Pièce dramatique
Divertissement
Pièce épisodique
Tragédie
Tragi-comédie
Comédie héroïque
Opéra Bouffe
Saynette
Opéra
Comédie-ballet
dialogue
Divertissement spirituel
Tragédie lyrique
Opéra comique
Pastorale
Vaudeville
Tragédie en musique
Drame
drame liturgique
Prologue
Saynète
Monologue
Monologue lyrique
Parade
Trait historique
Intermède
A propos
Fait historique
Opuscule dramatique
Parodie
Comédie galante
Pièce
```

FIGURE 1 – Les 50 genres différents

Chaque fichier XML se divise en deux parties :

Le Header

On y trouve des méta-informations sur la pièce.

- Le titre
- L'auteur ainsi que diverses informations sur lui (son nom, sa date de naissance, son lieu de naissance, sa date de mort, son lieu de mort, ...)
- La date de publication
- Le genre
- La structure (nombre d'actes)
- La période
- L'inspiration
- Le type (vers, prose)
- etc.

Le Body

- le texte de la pièce
- L'acte, la scène, le lieu ou encore le moment (précisé dans des balises div)
- Les personnages présents sur scène et qui parle
- Un identifiant unique pour chaque ver
- La liste des personnages et leurs caractéristiques

Nous avons utilisé les données suivantes pour notre apprentissage :

- L'auteur,
- La taille moyenne des phrases,
- La longueur du texte,
- Le type (vers ou prose)
- L'inspiration (Histoires chrétiennes, etc.)
- La structure (nombre d'actes)
- La période (le siècle)
- Le nombre de personnages
- Le genre pour lequel les mots du texte ont le plus de correspondance.

L'auteur, l'inspiration, la structure, la longueur du texte, la période et le nombre de personnages sont facilement récupérables à partir des données du fichier XML et ne nécessitent pas ou peu de traitement ; c'est le genre pour lequel les mots ont le plus de correspondance qui a demandé le plus de pré-traitement. Nous expliquons dans la partie suivante en quoi cela consiste.

Nous avons également fait des statistiques sur ces attributs résumées dans les graphes ci-dessous.

La légende en haut à droite sur certains graphes correspond à la corrélation de Pearson et à une probabilité p.

Le coefficient de corrélation de Pearson mesure la relation linéaire entre deux ensembles de données. Strictement parlant, la corrélation de Pearson exige que chaque ensemble de données soit distribué selon une loi normale, ce qui n'est pas forcément le cas pour toutes nos données. Il varie entre -1 et +1 avec 0 signifiant qu'il n'y a aucune corrélation. Les corrélations de -1 ou +1 impliquent une relation linéaire exacte. Les corrélations positives impliquent que lorsque x augmente, il en est de même pour y. Les corrélations négatives impliquent que les augmentations x diminuent y.

La valeur p indique approximativement la probabilité qu'un système non corrélé produise des ensembles de données ayant une corrélation de Pearson au moins aussi élevée que celle calculée à partir de ces ensembles de données.

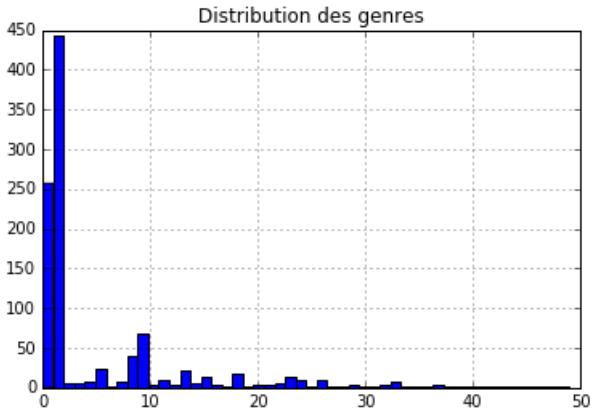


FIGURE 2 – Distribution des genres

On constate que deux genres sont particulièrement fréquents et représentent 50% des données : il s'agit de la tragédie et de la comédie.

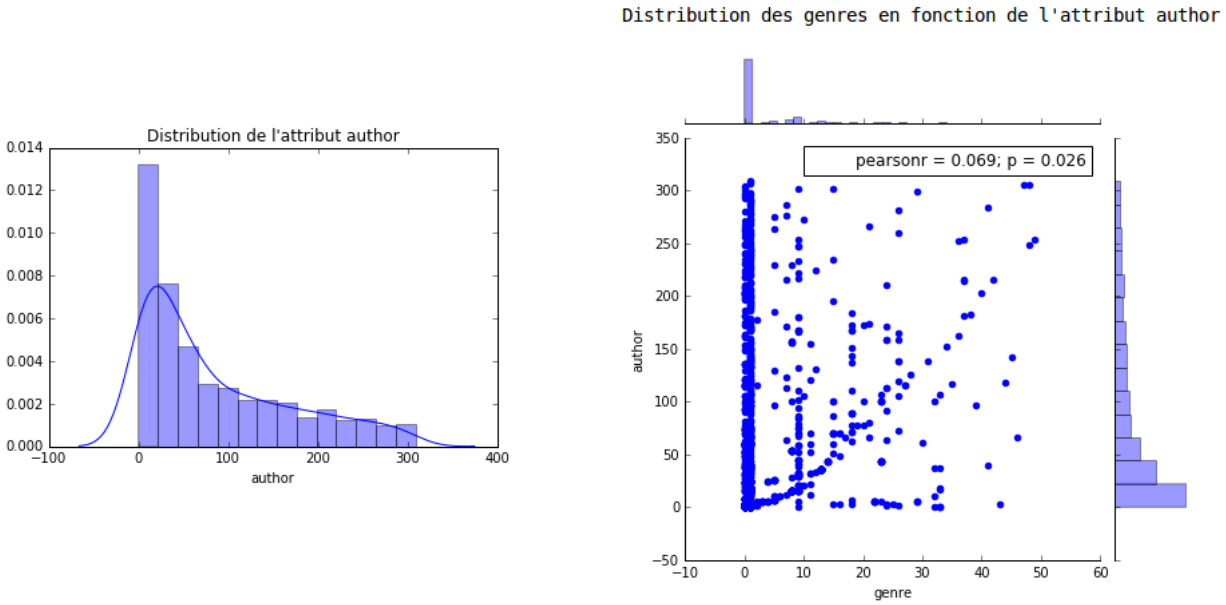


FIGURE 3 – Distributions des auteurs

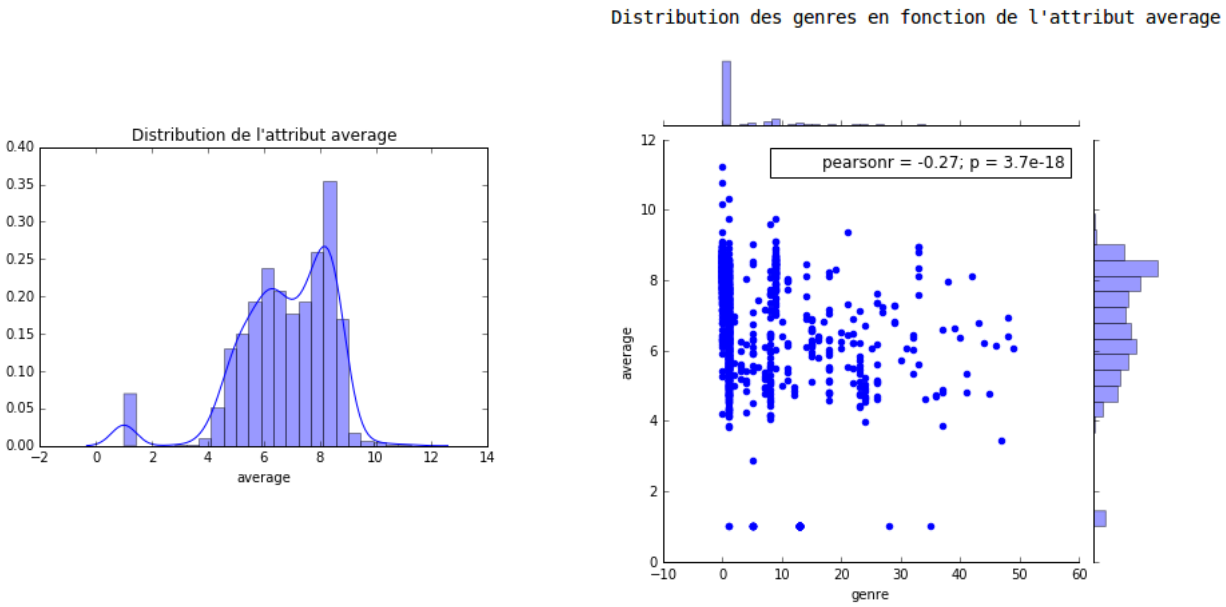


FIGURE 4 – Distributions des tailles moyennes des phrases (en mots)

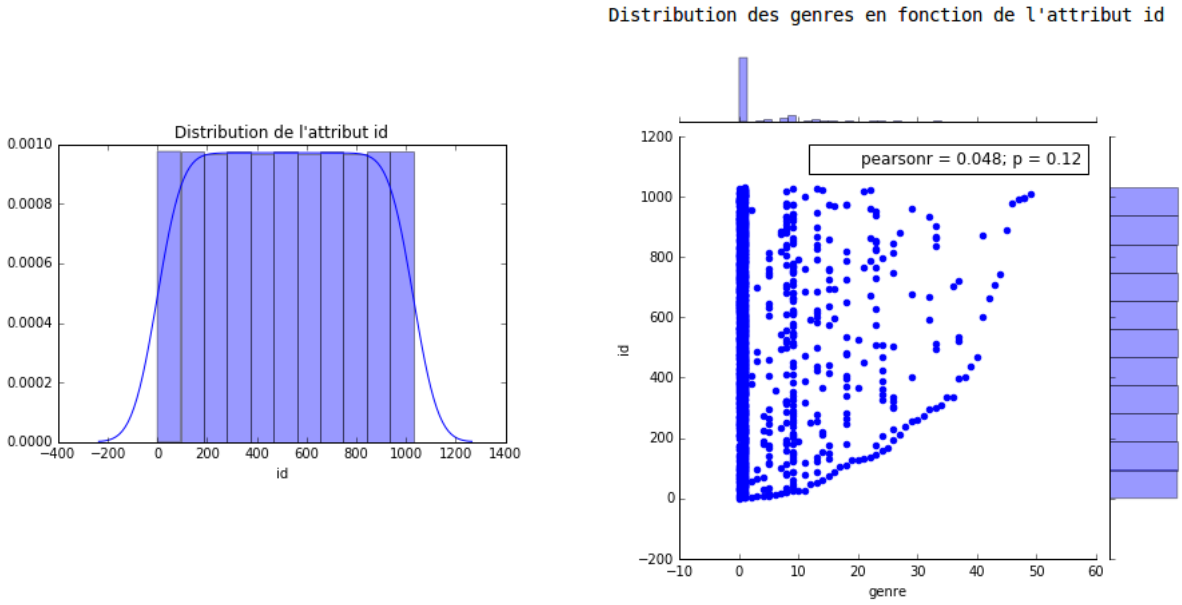


FIGURE 5 – Distributions des identificateurs uniques

L'attribut id est un simple identificateur que l'on a ajouté à chaque oeuvre. C'est pourquoi sa distribution est uniforme : il y a un id par oeuvre.

Cet attribut n'est pas censé avoir de corrélation avec le genre, ni avoir d'influence sur le résultat de la classification. On remarque cependant que la distribution des genre en fonction de l'id n'est pas uniforme.

Les id ont été attribués suivant l'ordre alphabétique des oeuvres. Le nom des fichiers commençant par le nom de l'auteur, on peut donc supposer que c'est l'influence de l'auteur sur le genre qui provoque ces corrélations.

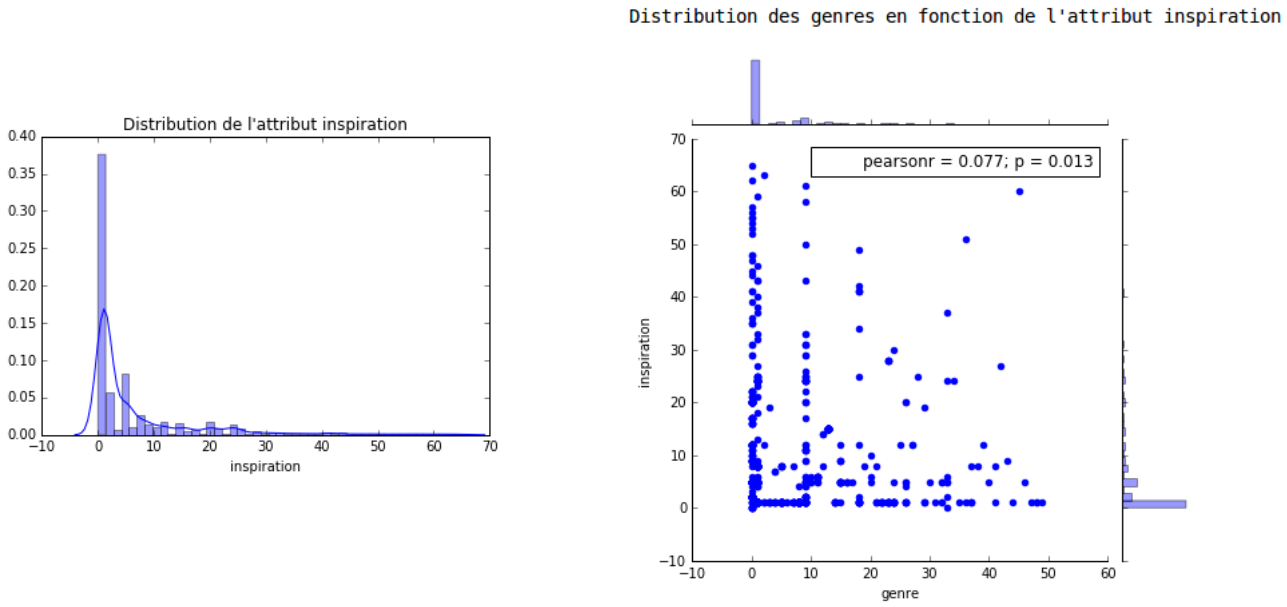


FIGURE 6 – Distributions des inspirations

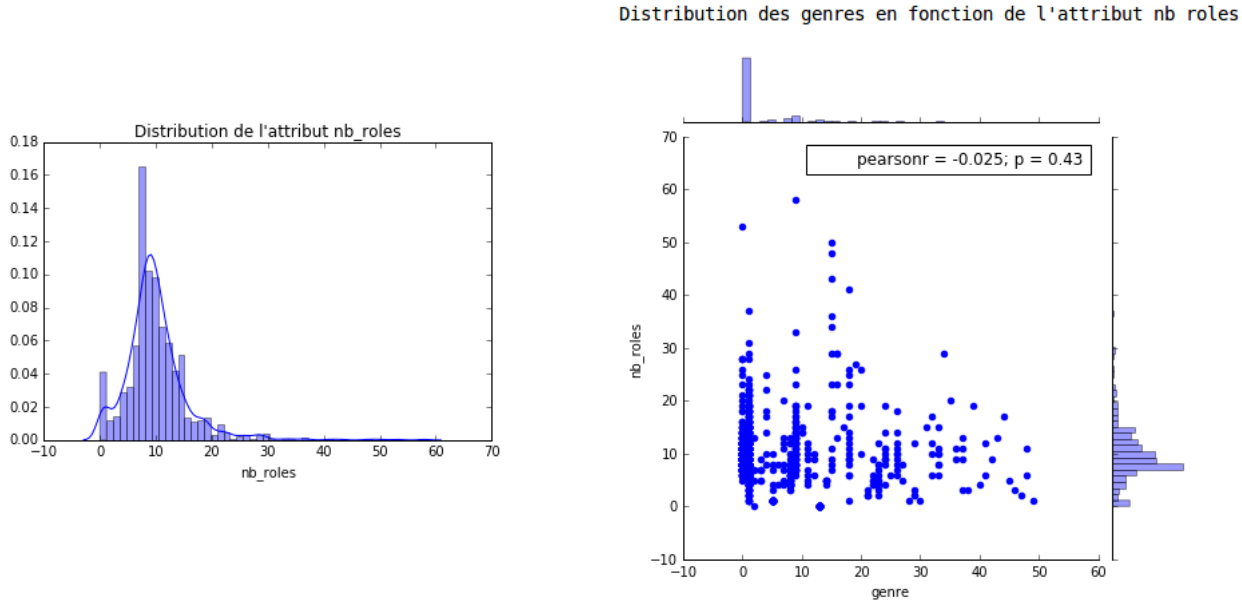
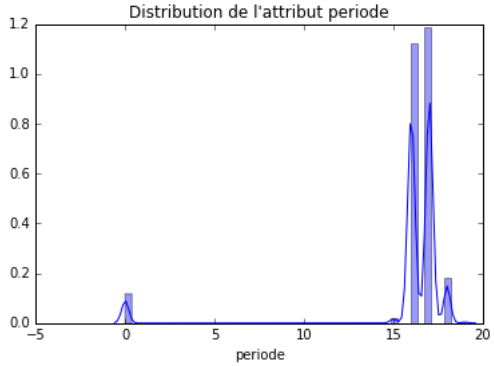


FIGURE 7 – Distributions des nombres de personnages

Les distributions des inspirations et des nombres de personnages semble être des Gaussiennes.



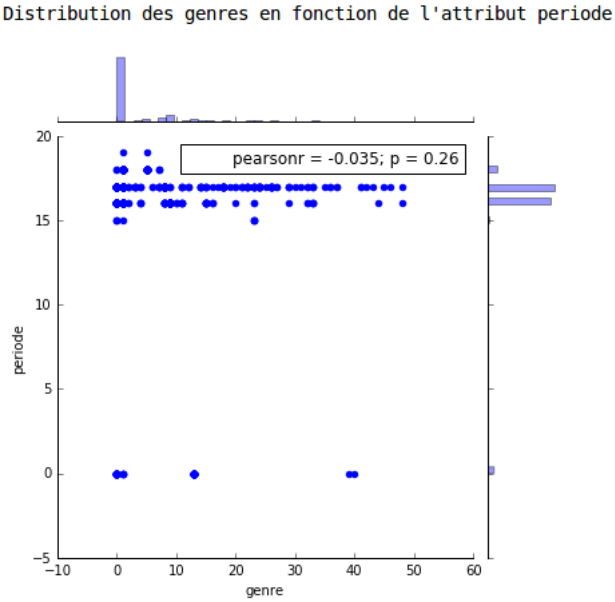
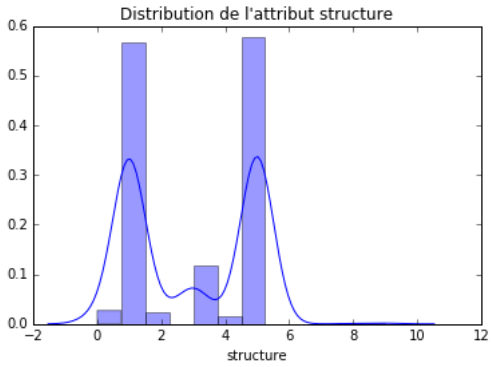


FIGURE 8 – Distributions des périodes (siècles), 0 correspondant à l'absence d'information



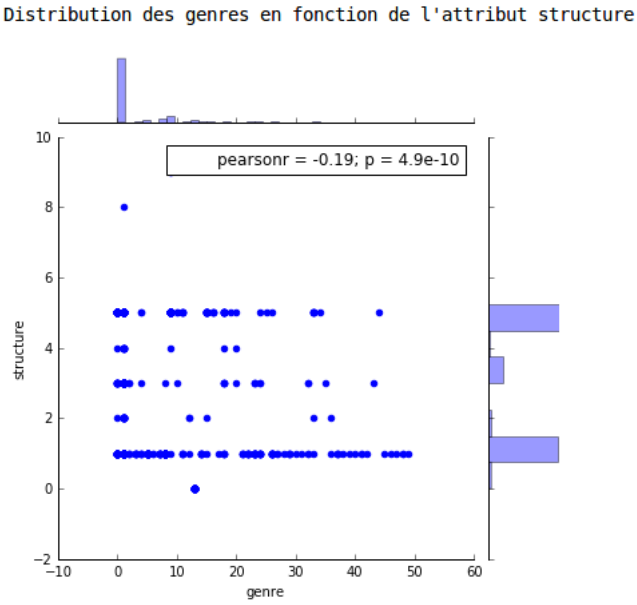


FIGURE 9 – Distributions du nombre d'actes

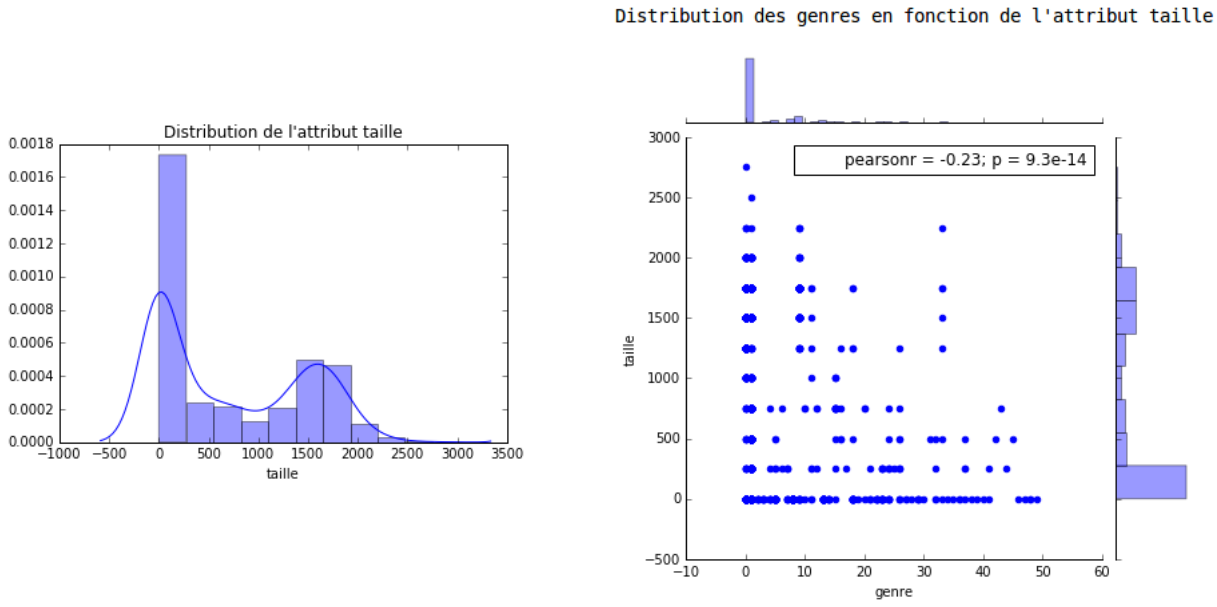


FIGURE 10 – Distributions de la taille

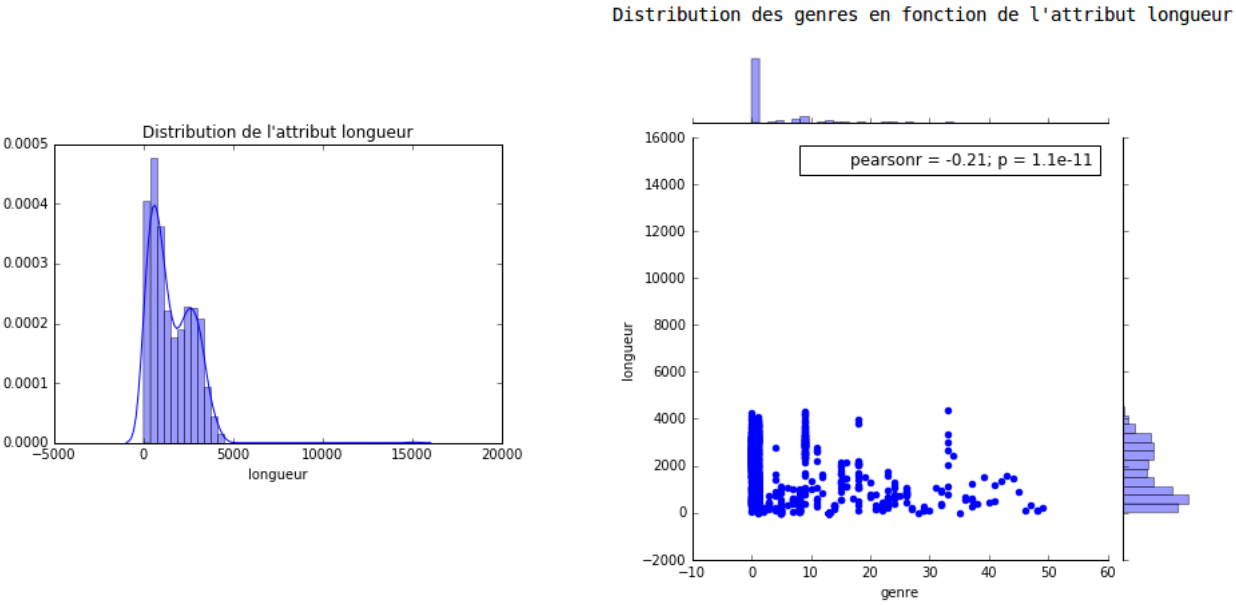


FIGURE 11 – Distributions de la longueur des textes

La longueur correspond au nombre de tokens du texte. On remarque que certains genres semblent être caractérisé par des textes souvent longs, et d'autres par des textes souvent courts. Nous verrons si cela se confirme en analysant les résultats.

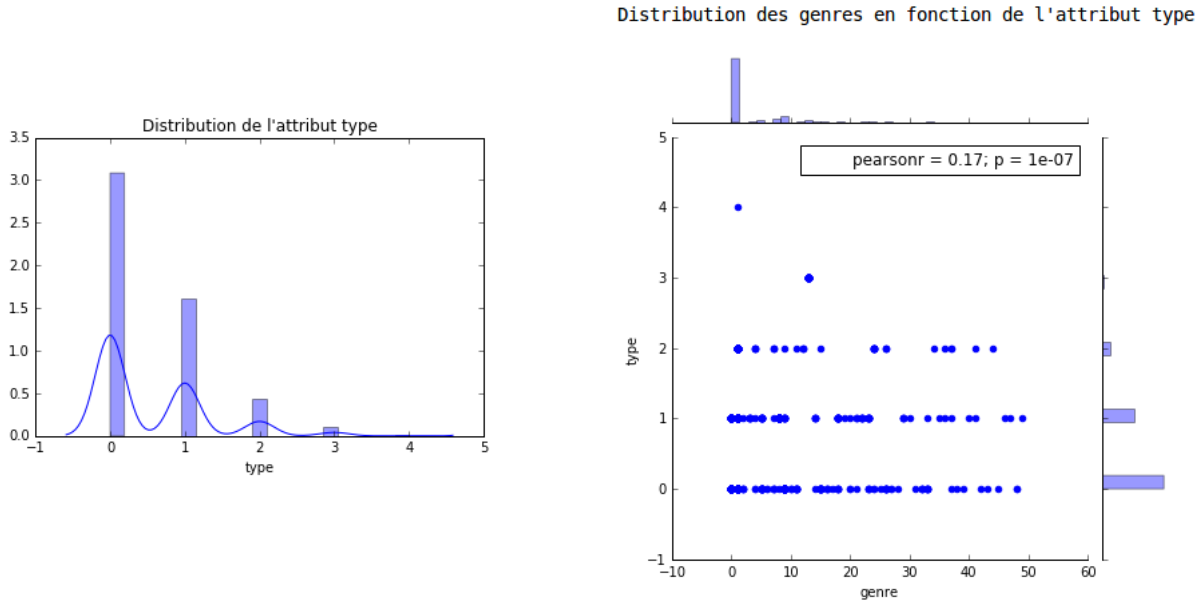


FIGURE 12 – Distributions du type des textes

Le type d'un texte est son style d'écriture : prose, vers ou mixte. La valeur 4 est à ignorer puisqu'elle correspond à la faute de frappe 'vers ' au lieu de 'vers'. Sur le graphique 0 représente l'écriture en vers, 1 l'écriture en prose et 2 l'écriture mixte.

2 Prétraitement

Dans un premier temps, nous modifié certains fichiers XML contenant des erreurs de syntaxe (DURANT_TELMAITRETELVALET.xml, GOUGES_MOLIERENINON.xml, GOULARD_AGIS.xml, GOUGES_EXCLAVAGEDESNOIRS.xml).

Nous avons ensuite parsé les fichiers XML à l'aide de lmxl [1] avec etree.

Nous avons tout d'abord extrait toutes les informations qui nous étaient nécessaires dans le header ainsi que le texte. Nous avons ensuite séparé ce texte en token, et nous avons enlevé les apostrophes ainsi que les ponctuations. Nous avons ensuite retiré les mots de trois lettres ou moins, retirant ainsi bon nombre de mots peu utile lors de l'extraction d'informations.

3 Description des expériences menées

Classification et évaluation des attributs

Pour la classification, nous avons principalement utilisé un Random Forest Classifier. Elles permettent d'éviter le problème de surapprentissage auquel font face les Decision Trees [2]. C'est également grâce à cela que nous avons pu caractériser les importances de chaque attribut.

1. `http://lxml.de/`
2. `https://web.stanford.edu/~hastie/Papers/ESLII.pdf`, Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome (2008). *The Elements of Statistical Learning*, pages 587–588

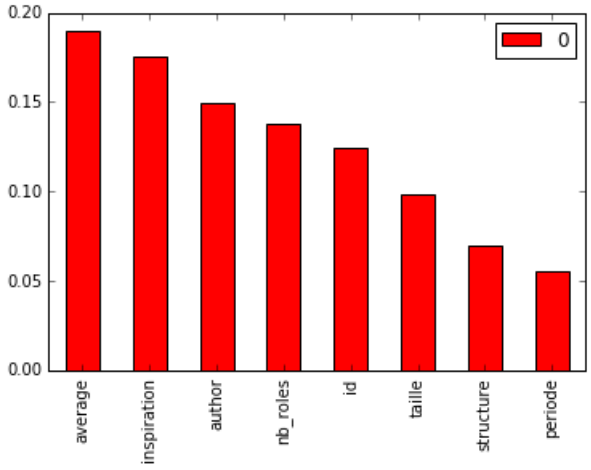


FIGURE 13 – Importance des attributs

Vous trouverez ci-dessus le classement de l'importance des attributs d'après le Random forest classifier. On constate que c'est **le nombre moyen de mots par phrase qui apporte le plus d'informations**, suivi de près par l'inspiration.

Au contraire, la structure n'est pas parlante, ce qui peut paraître surprenant : en effet, la tragédie française est composée de 5 actes, et le corpus contient beaucoup de tragédies ; on aurait pu penser que cette information était pertinente.

L'attribut période, situé en dernier dans l'ordre d'importance, ne décrit que le siècle d'écriture de l'oeuvre. Une date plus précise constituerait une information plus précieuse.

La liste des lemmes est plus porteuse de sens que la liste des tokens. En effet, on a regroupé ensemble les formes d'un mots pouvant être analysées comme un seul objet ; cela permet d'avoir plus de mots ayant du sens dans la liste et de faciliter le traitement.

Un score de probabilité d'appartenance aux genres

Chacun des textes possède un score qui correspond à une sorte de probabilité d'appartenance à un genre que nous avons calculé.

Pour cela, nous avons commencé par calculer les 70 mots les plus fréquents par genre mais peu fréquent dans le reste du corpus sur les données d'apprentissage uniquement. Pour des raisons évidentes, nous ne pouvions pas utiliser l'information du genre sur les données de test. Nous avons regroupé les textes par genre (comme si on n'avait plus que 50 textes) et avons calculé les TF-IDF pour chaque mot de ces textes et ainsi avons obtenu les mots les plus représentatifs de chaque genre.

Le TF-IDF (de l'anglais *term frequency-inverse document frequency*) est une méthode de pondération utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus.

Si une requête contient le terme T, un document a d'autant plus de chances d'y répondre qu'il contient ce terme : la fréquence du terme au sein du document est grande. Néanmoins, si le terme T est lui-même très fréquent au sein du corpus, c'est-à-dire qu'il est présent dans de nombreux documents, il est en fait peu discriminant. C'est pourquoi le schéma propose d'augmenter la pertinence d'un terme en fonction de sa rareté au sein du corpus (fréquence du terme dans le corpus IDF élevée). Ainsi, la présence d'un terme rare de la requête dans le contenu d'un document fait croître le "score" de ce dernier. [3]

3. `https://en.wikipedia.org/wiki/Tf-idf`

Le TF-IDF nous a donc permis de supprimer les mots non porteurs de sens tels que les articles ou les déterminants qui ne contiennent a priori pas d'information qui puisse nous aider à catégoriser les textes.

Nous avons commencé par tenter d'utiliser le TF-IDF de sklearn mais ça n'a pas bien marché ; c'est pourquoi nous avons préféré implémenter notre propre TF-IDF.

```
Genre : Tragédie
Mots représentatifs du genre : ['vous', 'pour', 'plus', 'bajazet', 'mais', 'seigneur', 'dans', 'nous', 'romains', 'votre'
, 'rome', 'sans', 'elle', 'tout', 'madame', 'point', 'fait', 'coeur', 'dieux', 'sang', 'amour', 'titus', 'porus', 'marius
', 'bien', 'avec', 'crime', 'romain', 'trop', 'mort', 'germanicus', 'fils', 'faut', 'sort', 'peut', 'pyrrhus', 'haine', '
attale', 'cette', 'gloire', 'tous', 'golo', 'taxile', 'leur', 'massinisse', 'dont', 'quel', 'encor', 'camp', 'darius', 's
partacus', 'ciel', 'quoi', 'contre', 'arsace', 'princesse', 'prince', 'empereur', 'perdiccas', 'pouvoir', 'enfin', 'faire
', 'rien', 'moins', 'scipion', 'ninus', 'sylla', 'quand', 'voir', 'tullie']
```

FIGURE 14 – Mots représentants la classe "Tragédie"

Les dénouement des tragédies sont souvent malheureux, en général, c'est la mort. La tragédie aborde des thèmes comme la passion, la vengeance et l'héroïsme : en effet, on constate que les mots "mort", "crime" ou encore "haine" sont présents dans cette liste.

Cette méthode nécessite cependant de grandes quantités de données en entrée, ce dont nous ne disposions pas forcément pour toutes les classes.

Une fois que nous avions les mots les plus courants de chaque genre, nous avons calculé les scores (la liste de taille 50) pour chaque oeuvre. On compte pour chaque genre le nombre d'occurrences des mots de sa liste représentative. Enfin, on divise chacun de ces nombres par le nombre total de mots du texte afin d'obtenir une probabilité entre 0 et 1 et de ne pas avoir de trop grandes différences dans nos attributs (pour certains textes, le nombre d'occurrences varie entre 10 et 50 alors qu'il est entre 600 et 1200 pour d'autres).

```
Genre : Comédie
Comédie : 0.0543855497713
Tragi-comédie : 0.0422385234264
Tragédie : 0.0414103170847
Comédie héroïque : 0.0393989588263
Pastorale : 0.0390834516485
Drame : 0.0367565862123
Pastorale héroïque : 0.036519955829
Opéra : 0.0313535257927
Opéra comique : 0.030170373876
Tragédie en musique : 0.028750591576
```

```
Genre : Tragi-comédie
Tragi-comédie : 0.0535467298953
Comédie : 0.0524270565764
Tragédie : 0.0497266679839
Comédie héroïque : 0.0450503852993
Pastorale : 0.0428769017981
Pastorale héroïque : 0.0427451755253
Opéra : 0.0372126720674
Opéra comique : 0.0327339787921
Tragédie en musique : 0.0326681156557
Drame : 0.0326681156557
```

FIGURE 15 – Scores sur deux textes d'exemples et leurs genres réels (comédie et tragi-comédie)

On voit que dans ces deux exemples le genre ayant obtenu le meilleur score est le bon. C'est le cas à chaque fois pour les genres les plus représentés (comédie, tragédie), et bien reconnaissable (tragi-comédie).

```
Genre : Dialogue
Comédie : 0.0439360457559
Tragi-comédie : 0.0371766541011
Tragédie : 0.0357467827895
Comédie héroïque : 0.0350968412843
Drame : 0.034446899779
Pastorale héroïque : 0.0334069933706
Pastorale : 0.031847133758
Pièce dramatique : 0.029507344339
Dialogue : 0.0289873911348
Opéra comique : 0.0289873911348
```

```
Genre : Opéra
Comédie : 0.048281861679
Tragédie : 0.0480643758156
Opéra : 0.0466507177033
Tragi-comédie : 0.0446933449326
Pastorale : 0.0433884297521
Comédie héroïque : 0.0413223140496
Tragédie en musique : 0.0391474554154
Pastorale héroïque : 0.0387124836886
Monologue lyrique : 0.0332753371031
Drame : 0.0320791648543
```

FIGURE 16 – Scores sur deux textes d'exemples et leurs genres réels (dialogue et opéra)

On remarque ici que parfois ce n'est pas la bonne classe qui obtient le meilleur score. Il est toutefois extrêmement rare que la classe ne se situe pas dans les 10 meilleurs scores. Les informations

extraites à l'aide de TF-IDFs concernant les genres peuvent donc avoir un impact sur la décision d'un classifieur.

Par la suite, nous avons décidé de ne passer au classifieur que l'indice du meilleur genre car les 50 attributs en plus n'amélioraient pas le score.

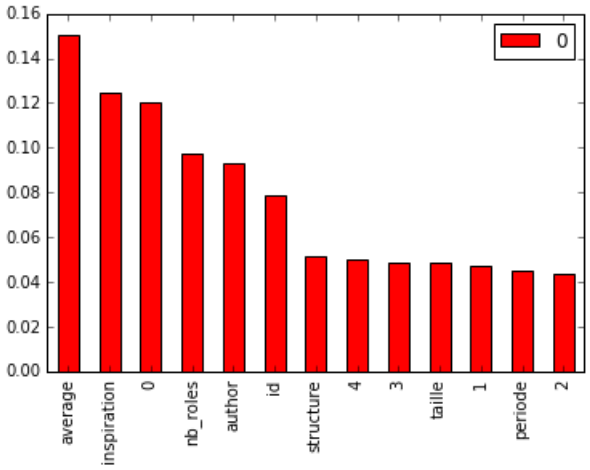


FIGURE 17 – Importance des attributs avec en plus les 5 classes les plus proches par rapport aux scores

Les chiffres de 0 à 4 correspondent aux 4 classes les plus proches par rapport aux mots fréquents (0 correspond à la classe la plus proche et 4 à la 5ème plus proche). On constate bien que seul le premier apporte réellement une information importante.

On obtenait 72% de réussite avec les 5 genres les plus probables et nous sommes passé à 77% en ne gardant que le premier.

Nous avons ensuite décidé d'enlever de notre analyse tous les mots de taille strictement inférieur à 4. Précédemment, nous ne traitions pas les mots de taille strictement inférieur à 3, mais nous nous sommes rendu compte que les mots de taille 3 étaient souvent impertinents. Suite à cette modification, l'attribut 0, c'est-à-dire le genre le plus probable d'après notre calcul de score, est passé de troisième à deuxième attribut le plus déterminant lors de la classification par Random Forest, comme le montre le graphique suivant :

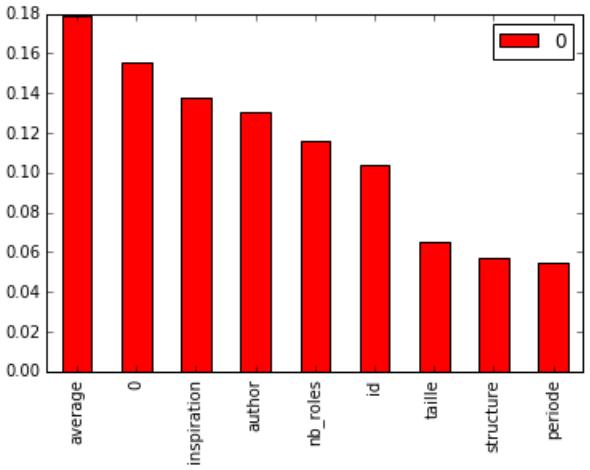


FIGURE 18 – Importance des attributs

Cette information nous démontre que cette modification a été favorable.

En ajoutant ensuite les attributs 'longueur' et 'type', voici le graphe des importances :

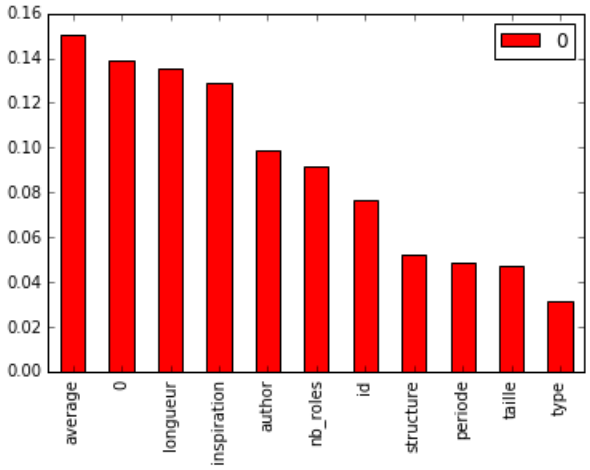


FIGURE 19 – Importance des attributs

On confirme donc l'hypothèse de départ selon laquelle le genre a une influence sur la longueur de l'oeuvre. Cependant, le type (vers ou prose) semble très peu utile, de façon assez surprenante.

Il est intéressant de remarquer que les deux attributs les plus utiles sont ceux qui mettent en jeu de l'analyse du texte des pièces de théâtre.

Ensuite, nous avons transformé les données non numériques (l'auteur, l'inspiration, le type et le genre) en entiers pour pouvoir les passer au classifieur à l'aide d'un mapping utilisant des ensembles.

Nous avons pris 80% des textes pour l'apprentissage sélectionnés aléatoirement, et les 20% restant pour le test. A chaque nouvel apprentissage, ces ensembles étaient sélectionnés aléatoirement à nouveau.

	author	average	genre	id	inspiration	longueur	nb_roles	periode	structure	taille	type
0	0	8.500000	0	0	0	3187	9	16	5	1750	0
1	1	7.193516	1	1	1	3782	13	16	5	1750	0
2	2	5.016667	2	2	1	173	7	17	1	0	0
3	3	8.268537	0	3	2	2207	12	17	5	1500	0
4	4	6.962500	1	4	1	883	15	16	1	500	0
5	1	5.335731	1	5	1	1126	15	16	3	0	1
6	5	5.548780	3	6	1	725	5	17	1	0	1
7	6	5.082267	4	7	1	1388	22	16	5	0	2
8	7	5.586207	5	8	1	76	1	18	1	0	1
9	8	7.960818	0	9	2	2635	8	16	5	1500	0
10	9	8.146903	0	10	2	2569	11	16	5	1500	0
11	10	8.545455	5	11	1	121	1	18	1	0	1
12	0	5.054496	1	12	1	1013	6	17	1	500	2
13	11	7.434109	6	13	1	1041	8	17	1	750	0
14	12	5.388889	7	14	1	55	7	17	1	0	1

FIGURE 20 – Attributs après avoir converti les valeurs non flottantes en entiers.

	author	average	id	inspiration	longueur	nb_roles	periode	structure	taille	type	0
957	6	5.317073	957	63	125	0	16	3	0	0	2.0
267	23	8.777542	267	5	2164	10	17	5	1500	0	0.0
863	41	7.768116	863	12	2806	7	16	5	1500	0	9.0
342	167	6.202454	342	1	986	9	18	5	0	1	1.0
416	10	5.142857	416	1	14	2	18	1	0	1	1.0

FIGURE 21 – Ensemble d'entraînement : le genre n'y figure plus, mais on retrouve l'attribut '0' (genre probable).

4 Analyse des résultats

Notre problème comporte 50 classes. Un classifieur complètement aléatoire obtiendrait donc 2% de réussite en moyenne.

En réalisant l'apprentissage sur des données fortement incomplètes (seulement les id par exemple), on obtient 20% de réussite. En effet, cela est dû au fait que les distributions entre les classes ne soit pas équivalentes : certaines sont très représentées.

Nous avons commencé par faire un premier test de classification par genre pour voir si le projet tenait la route et si on pourrait faire quelque chose. Nous n'avions que deux attributs : l'auteur et l'inspiration.

Le classifieur utilisé pour ce premier test était un Decision Tree Classifier. Nous obtenions **60%** de réussite, ce qui paraît être un bon résultat en considérant le nombre de classes (50) et d'exemples (1030) à notre disposition.

Après avoir ajouté de nouvelles informations (toutes sauf la structure, la période, le nombre de personnages et le score) et utilisé un Decision Tree Classifier, on obtenait **70%** de réussite.

Ensuite, en utilisant un Random Forest Classifier et avec le genre le plus probable ajouté aux attributs, le taux de réussite variait entre 74% et **77%** selon le nombre d'estimators spécifié en paramètres.

Enfin, après avoir rajouté les attributs "longueur" et "type" on parvient à atteindre **79%** de réussite avec les bons paramètres du Random Forest Classifier et une répartition entre les ensembles d'entraînement et de test de 90% et 10%.

5 Conclusion et Pistes d'améliorations

Nous obtenons un résultat final de 79%. Certains attributs sont plus utiles que d'autres, mais c'est l'ensemble de ces attributs qui nous permet d'obtenir un aussi bon score malgré le grand nombre de classes et le petit nombre de textes composant le corpus.

Afin d'améliorer notre classifieur, on pourrait rajouter au corpus des textes pris de sources extérieures comme par exemple des site tels que `http://libretheatre.fr` ou encore `https://www.leproscenium.com/` qui répertorient les oeuvres théâtrales du domaine public en français.

On pourrait également utiliser un classifieur plus efficace tel qu'un SVM ; même si les données se prêtent bien aux Random Forest Classifier, les SVM restent en règle générale plus efficaces. Malheureusement, le coût est élevé et le temps de calcul peut vite devenir très long.

Enfin, on pourrait rechercher des caractéristiques plus spécifiques à chaque genre ; par exemple, classer automatiquement une pièce avec un seul personnage dans "monologue" ou "monologue lyrique".

Répartition des tâches

Globalement, nous avons avancé ensemble lors de ce projet. Les idées nous venaient en discutant et nous avons travaillé ensemble presque systématiquement.

Pour essayer tout de même de dégager une répartition des tâches, nous dirions qu'Eléonore s'est plus penchée sur le rapport tandis qu'Adrien était plus axé sur le code.

En particulier, Adrien a plus fait le parsing XML et les calculs de TF-IDFs et Eléonore s'est plus occupée de la classification.