Classification de documents avec Forêt aléatoire

Pascal Brissette

14 novembre 2022

Résumé

Ce rapport retrace les étapes du processus de classification d'un ensemble de documents publiés dans différents médias québécois de langue française entre 2005 et 2020. La classification des documents de langue anglaise, pour la même période, a fait l'objet d'un rapport distinct par Lisa Teichmann.

1 Objectif général de la tâche

Utilisation d'un classifieur documentaire pour repérer, dans un ensemble d'articles de presse moissonnés depuis Eurêka et divers sites de journaux, des documents portant sur l'alimentation humaine.

2 Chaine de traitement

— Annotation manuelle de 4349 documents par le directeur du projet et deux assistantes de recherche ;
— Prétraitement des documents et repérage des algorithmes les plus performants ;
— Sélection des variables les plus significatives ;
— Entraînement et comparaison des modèles
— Réglage des hyperparamètres du modèle de forêt aléatoire ;
— Évaluation du modèle avec le sous-ensemble de test
— Classification finale avec le modèle testé
— Examen à la pièce d'un échantillon de 100 documents classifiés par le modèle

3 Annotation manuelle de documents

La classification de documents est une tâche courante du traitement automatique des langues naturelles qui consiste à confier à un algorithme la tâche d'apposer des étiquettes ou catégories à des documents de manière automatique après l'avoir entraîné avec un lot de documents classifiés. La première étape consiste, pour l'équipe de chercheurs, à établir une série de règles de classification, aussi précises que possible, et à apposer manuellement, en fonction de ces règles, des étiquettes à un lot de documents choisis au hasard.

Dans le cadre de ce travail, 4349 documents ont été étiquetés manuellement par l'équipe de recherche. On a apposé aux documents répondant aux critères de pertinence l'étiquette '2keep' et à ceux qui n'y répondaient pas, l'étiquette '2rmv'. La distribution des documents par étiquette est la suivante :

Étiquette	Nombre de documents
2keep	2424
2rmv	1925

Table 1 – Nombre d'articles pour chaque étiquette (première étape)

4 Prétraitement

La classification manuelle décrite ci-dessus a été faite avec les textes originaux. Les annotateurs avaient accès à l'intégralité des textes. Les algorithmes de classification ne peuvent assimiler et traiter le texte sous cette forme. Il faut pratiquer une série d'opérations qui vont transformer le texte en une représentation matricielle composée de chiffres. On profite de cette transformation pour éliminer les documents ou éléments à l'intérieur des documents qui génèrent du « bruit » ou alourdissent inutilement la structure de données. Les opérations suivantes ont été faites dans cette perspective :

— Élimination des documents dont le contenu textuel était inférieur à 85 caractères ;
— Conversion de tous les caractères en minuscules ;
— Suppression de la ponctuation et des symboles ;
— Composition de ngrammes à partir de noms de quartiers (ex. « hochelaga_maisonneuve »), d'organismes (ex. : « moisson_montréal ») ou d'expressions significatives compte tenu de la problématique de l'alimentation humaine (ex. : « food_truck »). ;
— Suppression des mots fonctionnels (« car », « qui », etc.).

Pour corriger le déséquilibre entre les deux sous-ensembles ("2keep"/"2rmv"), un échantillon aléatoire de doublons de documents "2rmv" a été créé et ajouté aux documents de ce type. L'ensemble de documents annotés était ainsi composé de 2424 x 2 documents = 4848 documents.

Cet ensemble a été mélangé de façon aléatoire, puis divisé en deux sous-ensembles, l'un destiné à entrainer le modèle (80% du lot), l'autre à évaluer son efficacité (20%).

Une matrice a été formée sur la base de la fréquence lexicale. Les tests effectués avec normalisation (L2) n'ont pas permis d'améliorer les résultats, de sorte que le modèle final repose sur un calcul de fréquences uniquement pondérées (tf-idf).

5 Sélection des variables

La matrice (« Document Term Matrix ») issue de l'opération précédente, après filtrage par seuil et plafond de fréquences (doc.prop.max : 0.6 ; term.freq.min : 20), comprend 7380 colonnes, correspondant chacune à un lexème. Une sélection des lexèmes statistiquement significatifs a ensuite été faite sur la base du coefficient « Mean Decrease Gini », fourni par le modèle de forêt aléatoire. 24 seuils (de 0.25 à 1.4) ont été testés. Le tableau ci-dessous montre que le modèle est le plus efficace avec 379 éléments/lexèmes (« Mean Decrease Gini» supérieur à 1.15) :

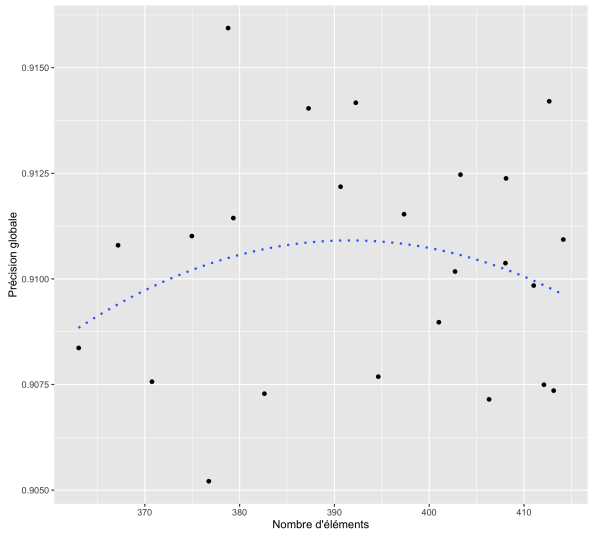

FIGURE 1 – Efficacité comparée d'une forêt aléatoire avec différents seuils « Mean Decrease Gini ».

6 Entrainement et comparaison des modèles

Dans la phase initiale d'entrainement, nous avons testé quatre différents algorithmes en vue de choisir le plus performant : arbre de décision (CART), méthode des k plus proches voisins (KNN), allocation de Dirichlet latente (LDA), machine à support de vecteurs (SVM), forêt aléatoire (RF).

Le classifieur utilisant une forêt aléatoire a donné les meilleurs résultats, avec une efficacité moyenne de 93,5

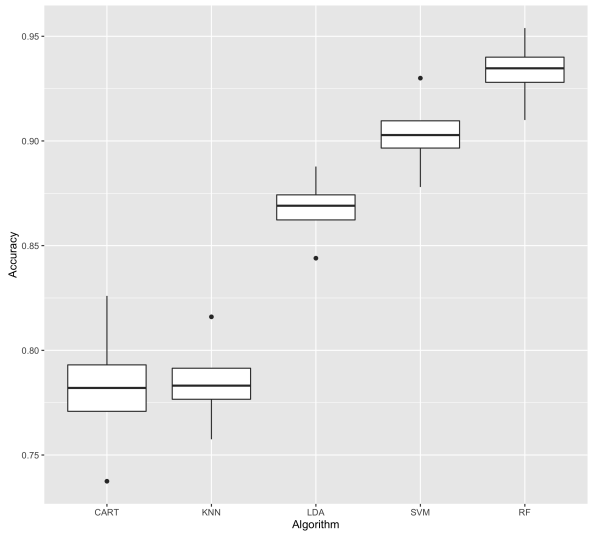

FIGURE 2 – Efficacité comparée de cinq algorithmes utilisés pour la classification de documents (sans réglages préalables). La forêt aléatoire présente les meilleurs résultats.

Avant de lancer le processus de classification, nous avons repéré les paramètres optimaux de la méthode, soit le nombre d'éléments à utiliser pour chaque arbre des forêts en fonction du taux d'erreur (‘mtry‘), ainsi que le nombre d'arbres à utiliser (‘ntree‘) dans le processus. Plus de 150 combinaisons ont été testées (‘mtry‘ entre 1 et 50 ; ‘ntree‘ de 500, 1000 et 1500). Des classifieurs ont été entrainés avec les cinq meilleures combinaisons, puis testés à l'aveugle pour savoir lequel comportait le moins d'erreur. Comme on le verra dans le graphique suivant, les deux paramètres ont des impacts significatifs sur l'efficacité du modèle. La combinaison idéale, pour notre modèle, est un ‘mtry‘ de 34 et 1500 arbres décisionnels.

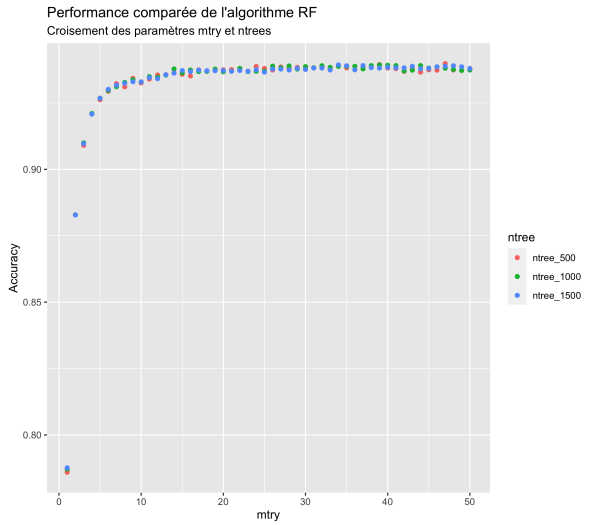

FIGURE 3 – Performances comparées de l'algorithme RF avec divers hyperparamètres.

7 Évaluation du modèle avec le sous-ensemble de test

Un modèle recourrant à ces paramètres optimaux est encore loin de la perfection, mais il est satisfaisant du point de vue de la précision globale (0,924). Cet équilibre se reflète dans les résultats de la classification à l'aveugle faite par le modèle sur le sous-ensemble réservé pour le test.

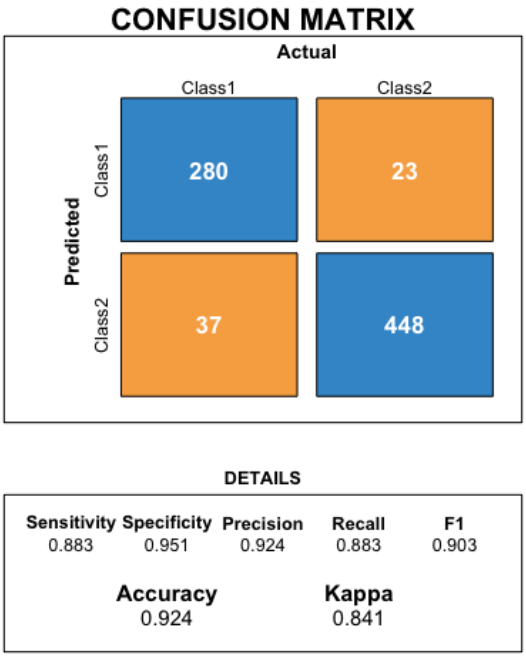

FIGURE 4 – Matrice de confusion : résultats de la classification réalisée à l'aveugle par le modèle optimisé.

8 Classification finale avec le modèle testé

Les étapes précédentes conduisent à celle-ci, soit à l'annotation automatique de tous les documents qui n'ont pas déjà fait l'objet d'une lecture rapprochée. Ces documents doivent être prétraités exactement comme l'ont été ceux qui ont servi à entraîner le modèle, puis on crée une matrice ayant le même nombre de dimensions (nombre de colonnes) que le lot d'entrainement.

La classification finale a donné les résultats ci-dessous :

Class	Number of Articles
2keep	22 417
2rmv	6 257

TABLE 2 – Nombre d'articles retenus et rejetés par le modèle final

Un échantillon aléatoire de 100 documents étiquetés «2keep» a été transféré dans Recogito (https ://recogito.pelagios.org/) et relu par deux membres de l'équipe. Six documents étaient des faux positifs, ce qui tend à confirmer la performance du classifieur.

Références

[1] Helfin I. Rhys (2020) *Machine Learning with R, the tidyverse and mlr*, Shelter Island (NY), Manning Publications.

[2] Emil Hvitfeldt et Julia Silge (2022) *Supervised Machine Learning for Text Analysis in R*, Boca Raton (FL), CRC Press.

[3] Max Kuhn et Kjell Johnson (2016) *Applied Predictive Modeling*, New York, Springer.

[4] Jean-Herman Guay (2014) *Statistiques en sciences humaines avec R*, Sainte-Foy, Presses de l'Université Laval.

[5] Julia Silge et David Robinson (2017) *Text Mining with R. A Tidy Approach*, Sebastopol (CA), O'Reilly.