Loin d'être le privilège des plus grands établissements, l'intelligence artificielle peut être mobilisée dans toute bibliothèque universitaire. Illustration pratique avec l'expérience du SCD de l'université Côte d'Azur.

Pour une approche décomplexée de l'IA

UNIVERSITÉ CÔTE D'AZUR

Si l'on devait résumer les quelques sujets qui nous ont collectivement animés ces dernières années, l'intelligence artificielle (IA) en est curieusement absente, ou du moins marginalement traitée : transition bibliographique, Web de données, *open access*, ou encore services aux chercheurs ont la part belle, mais point ou peu de *machine learning* et autres réseaux de neurones. En même temps que les applications et services à base d'IA se déploient couramment dans notre vie quotidienne, alors même que ces « nouvelles » technologies de traitement de la donnée commencent concrètement à infuser dans les projets de recherche scientifique (notamment en humanités numériques), force néanmoins est de constater que l'IA tarde à se manifester dans nos quotidiens de bibliothécaires. Sans doute en raison du fait que, pour paraphraser Philippe Le Pape à propos de la transition bibliographique, il n'y aura pas de grand soir de l'IA en bibliothèque mais une succession de petits matins au gré des expérimentations et retombées des projets menés, notamment, à l'Abes et à la Bibliothèque nationale de France. Cela dit, même si ces opérateurs paraissent les mieux placés en terme de compétences techniques et de maîtrise des données pour aborder ces problématiques d'IA, sur le papier rien ne s'oppose à ce que tout un chacun dans son établissement se forme et s'essaie à tester, expérimenter, prototyper, voire utiliser en production des algorithmes d'apprentissage machine, ceux-ci étant largement documentés et leurs implémentations disponibles dans des librairies *open source*.

L'IA OUTIL DE PRÉDICTION

Fondamentalement, les dispositifs d'IA s'appuient sur des logiques inhabituelles pour des bibliothécaires férus de modélisation de haut niveau, dont découle la norme qui encadre la production de données. Inversement, la spécificité des techniques d'apprentissage machine est d'être totalement empirique et de partir des données afin d'en induire les motifs implicites et corrélations sous-jacentes. Sous réserve de disposer d'assez de données et de leur appliquer les traitements adéquats à l'aide de techniques statistiques d'échantillonnage et d'outils de *data mining*, l'apprentissage automatique ou profond consiste à découvrir des éléments de structuration invisibles et à modéliser les *patterns* implicites contenus dans ces observations d'apprentissage afin d'en extraire de l'information et de faire des prédictions. À la base du *machine learning*, rien de magique donc, mais l'application balisée et rigoureuse de techniques statistiques et géométriques connues depuis longtemps pour effectuer de la régression, classification ou « clusterisation » sur des jeux de données simples. Pour la réalisation de tâches plus complexes et non linéaires, telles que la reconnaissance vocale ou la reconnaissance de formes dans des corpus d'images, les algorithmes de *deep learning* s'appuient, eux, sur les technologies à base de réseaux de neurones pour modéliser des transferts auto-apprenants d'informations de plus en plus abstraites entre couches de neurones, à l'image de ce qu'il se produit dans les cerveaux des mammifères.

EXPLORER LES COLLECTIONS D'UN POINT DE VUE NOUVEAU

Afin d'illustrer des possibilités de prises en main dans le cadre de problématiques simples d'établissement, voici trois exemples, dont deux ont été abordés dans une démarche de preuve de concept sans objectif opérationnel, tandis que le dernier cas s'intègre, lui, dans un circuit de traitement ayant donné lieu à la mise en production concrète d'une application (le baromètre Science ouverte de l'université Côte d'Azur développé par le SCD). La première expérimentation consiste à prototyper un système de recommandations de documents intégré avec l'Opac[1] afin d'évaluer dans quelle mesure l'enrichissement des résultats fournis par le moteur de recherche du catalogue grâce à des notices « proches » complémentaires favorise la découvrabilité de contenus au cours de l'expérience usager. Pour ce faire, le cœur du processus consiste à travailler sur un jeu de données bibliographiques et un jeu de données d'usage des documents afin d'opérer des calculs de :

- Similarité entre contenus pour mesurer la proximité entre chaînes de caractères de chaque paire de notices à partir d'une analyse type NLP[2] sur les champs jugés les plus signifiants.

[1] https://bibliotheques.wordpress.com/2019/11/04/ia-et-opac-mettre-en-place-un-moteur-de-recommandation-dans-un-opac-1-4-quelques-bases-conceptuelles

[2] Traitement automatique du langage naturel

• Similarité par l'usage qui permet de déduire des associations de documents à partir des prêts effectués par les lecteurs (les lecteurs qui ont emprunté ceci ont aussi emprunté cela).
On est ici dans le cas d'un apprentissage basé sur des observations à partir duquel l'algorithme est ensuite généralisable à de nouveaux exemples en calculant des proximités à partir des données apprises. Ainsi ce prototype[3] d'application Web, qui requête à la fois l'API de l'Opac et le modèle obtenu, permet d'afficher, pour chaque notice de la liste de résultat, les notices similaires calculées.

Le deuxième exemple proposé part d'une idée toute simple et illustre explicitement l'objectif fondamental de l'IA – faire des prédictions à partir de jeux de données –, cette fois-ci dans le cadre d'un apprentissage par modèle : dans le stock de documents détenus par une bibliothèque, certains ont été empruntés au moins une fois et d'autres jamais. Peut-on dégager des variables pertinentes qui déterminent le fait pour un livre d'être emprunté ou pas, et ainsi, prévoir pour des nouveaux documents leur probabilité d'être empruntés en fonction de leurs métadonnées ?
Réponse : oui à 75 %. Plus précisément, après avoir entraîné plusieurs algorithmes de classification binaire[4] sur un fichier csv contenant les métadonnées de tous les exemplaires de monographies créés entre 2015 et 2021 au SCD, on obtient pour le plus performant d'entre eux un taux de précision de 75 % (l'algorithme a raison 3 fois sur 4 lorsqu'il prédit que les exemplaires du jeu de test seront empruntés) et arrive à détecter 75 % des exemplaires réellement prêtés (mesure du rappel)[5]. Au-delà de la résolution algorithmique du problème posé, l'intérêt d'une telle démarche est aussi d'être amené, durant l'étape de préparation des données[6], à explorer les collections d'un nouveau point de vue en déployant une analyse centrée sur l'usage qui va potentiellement au-delà des analyses traditionnelles de politique documentaire, à l'image de ce prototype de tableau de bord[7] qui reprend la plupart des visualisations créées pour la génération du modèle.

L'IA AU SERVICE DU BAROMÈTRE *OPEN ACCESS* DE L'UNIVERSITÉ

Enfin, le troisième cas d'usage concerne le set de métadonnées de publications collectées puis enrichies avec les données d'Unpaywall qui sert de base au baromètre *open access* de l'établissement. En plus des indicateurs d'ouverture par date de publication, date d'observation et structures de recherche, nous souhaitions également produire des vues et données chiffrées par disciplines, à l'instar du baromètre national. Afin d'indexer le millier de publications UCA qui n'étaient pas

Crédit Adobe stock

présentes dans le jeu de données du ministère, et pour lesquelles nous ne disposions donc pas de variable discipline, la solution la plus évidente à mettre en œuvre a consisté à entraîner plusieurs algorithmes de multi-classification sur le jeu de données national déjà étiqueté (lui aussi grâce à un algorithme de *machine learning,* d'ailleurs) puis à « prédire », grâce au meilleur modèle obtenu, la discipline associée à chaque publication du *dataset* local.
Alors, certes le ticket d'entrée pour se familiariser avec ce type d'usages de la donnée est un peu élevé quand la plupart des métiers de la data (*data scientist, data engineer, data miner…*) sont des profils historiquement absents de nos organigrammes. Mais le retour sur investissement, même au niveau d'un établissement, peut se révéler très positif, en termes de connaissances métier supplémentaires qui ne sont pas produites par le travail courant sur les collections ou les reporting classiques, ou encore en produisant des algorithmes remplaçant avantageusement des règles métiers difficiles à maintenir et souvent appliquées manuellement (pour des opérations de monitoring sur la qualité des données par exemple).

Géraldine Geoffroy
SCD Université Côte d'Azur
geraldine.geoffroy@univ-cotedazur.fr

[3] http://azur-scd.com/apps/poc-recommandations-engine

[4] https://github.com/azur-scd/poc-loan-predict/blob/main/notebooks/ml_algorithmes.ipynb

[5] L'algorithme est entraîné sur 80 % du *dataset* de départ, et évalué sur les 20 % restants d'observations qui lui sont inconnues en mesurant l'écart entre valeurs réelles et valeurs prédites. La précision mesure la qualité des prédictions positives tandis que le rappel prend en compte les faux négatifs pour évaluer le taux de couverture des prédictions positives.

[6] https://github.com/azur-scd/poc-loan-predict/blob/main/notebooks/data_processing.ipynb

[7] https://azurscd-poc-loan-predict.herokuapp.com