L'IA et la fouille de textes à l'INIST : *l'IA à portée de tous ?*

L'Inist a pour objectif de développer des outils de fouille de textes intégrables dans des chaînes de production mais également utilisables par tous les acteurs travaillant avec des publications numériques.

Si la fouille de textes a toujours été présente à l'Inist[1], ce n'est qu'avec le lancement du projet Istex[2] que des méthodes d'intelligence artificielle vont être développées pour être appliquées en grandeur nature sur de gros volumes de données, dans un processus industrialisé.

Initié en 2012 par le MESR dans le cadre du Programme d'investissements d'avenir[3], Istex comprend dès le début un volet « enrichissement » *via* des méthodes de fouille de textes. Alors que l'IA n'était pas encore un mot-clé passé dans le langage commun, l'équipe Istex-RD d'alors commence à développer des méthodes d'enrichissement de données à partir notamment de techniques d'apprentissage automatique (*machine learning* ou ML), afin de déterminer automatiquement le domaine scientifique des articles. D'autres méthodes, comme l'extraction de mots clés ou d'entités nommées vont également voir le jour sous forme de modules intégrés à la chaîne de production, permettant le traitement des corpus au fur et à mesure de leur intégration[4].

À travers quatre axes de travail (structuration des documents, indexation automatique, reconnaissance d'entités nommées, catégorisation des documents) nous avons alors répondu aux trois principaux défis rencontrés :

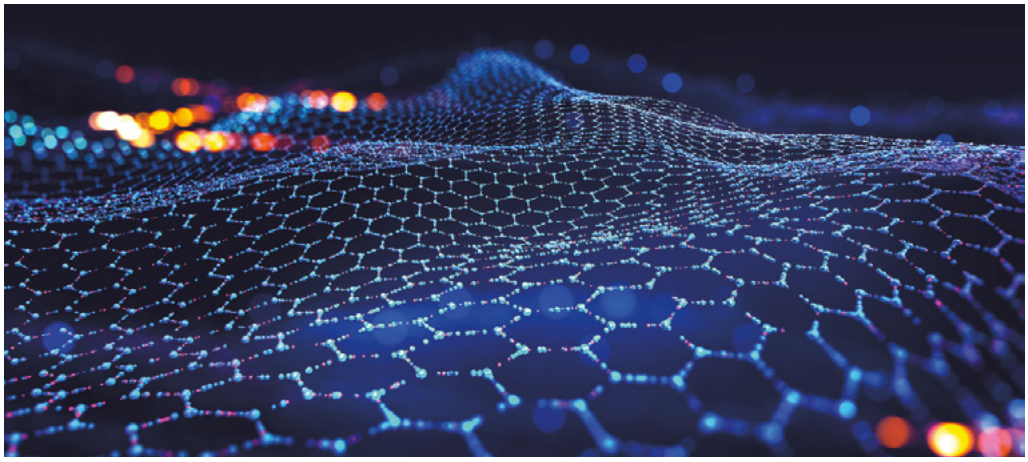
- Mise au point et intégration d'outils de TDM : entraînement, adaptation, mise en production.
- Passage à l'échelle : 25 millions de documents à traiter.
- Reversement des données : modélisation des données, réintégration, mise à disposition.

Si cette approche a donné de bons résultats, elle a montré un certain nombre de limites : développer et mettre en place un nouveau traitement est un processus complexe à mettre en œuvre, et surtout cela rend très difficile l'utilisation de ces programmes en dehors de la chaîne Istex.

D'UNE IA INTÉGRÉE DANS UN PROCESSUS DÉFINI À UNE IA INDÉPENDANTE DES DONNÉES

Dans une démarche « science ouverte », on constate la mise en ligne croissante, *via* GitHub ou GitBucket, de programmes permettant de traiter des données. Mais si tout un chacun a accès à ces algorithmes, leur mise en œuvre souvent complexe n'incite pas les non informaticiens à les utiliser. Nous nous inscrivons bien dans ce mouvement, en publiant tous nos codes, cependant nous voulons aller plus loin en faisant en sorte que qui que ce soit puisse les utiliser, quelles que soient ses compétences. Cela doit répondre aux demandes d'utilisateurs, documentalistes ou chercheurs, qui souhaitent pouvoir utiliser ces programmes sur leurs propres données de façon simple, et en pouvant choisir eux-mêmes les traitements dont ils ont besoin.

Nous avons fait le choix de créer et déployer des applications d'IA sous forme de web services (WS), intégrables dans une chaîne de production comme Istex, mais également directement utilisables par tout utilisateur désirant traiter ses propres corpus. Ainsi nous passons d'une IA intégrée dans un processus défini à une IA indépendante des données,



Crédit Adobe stock

[1] www.inist.fr

[2] www.istex.fr

[3] https://franceuniversites.fr/wp-content/uploads/2012/04/Projet_Istex.pdf

[4] Cuxac P., Thouvenin N. (2017) : Archives numériques et fouille de textes : le projet Istex. *Atelier TextMine, conférence EGC,* 24 janvier 2017, Grenoble, France. https://textmine.sciencesconf.org/data/pages/TextMine17.pdf

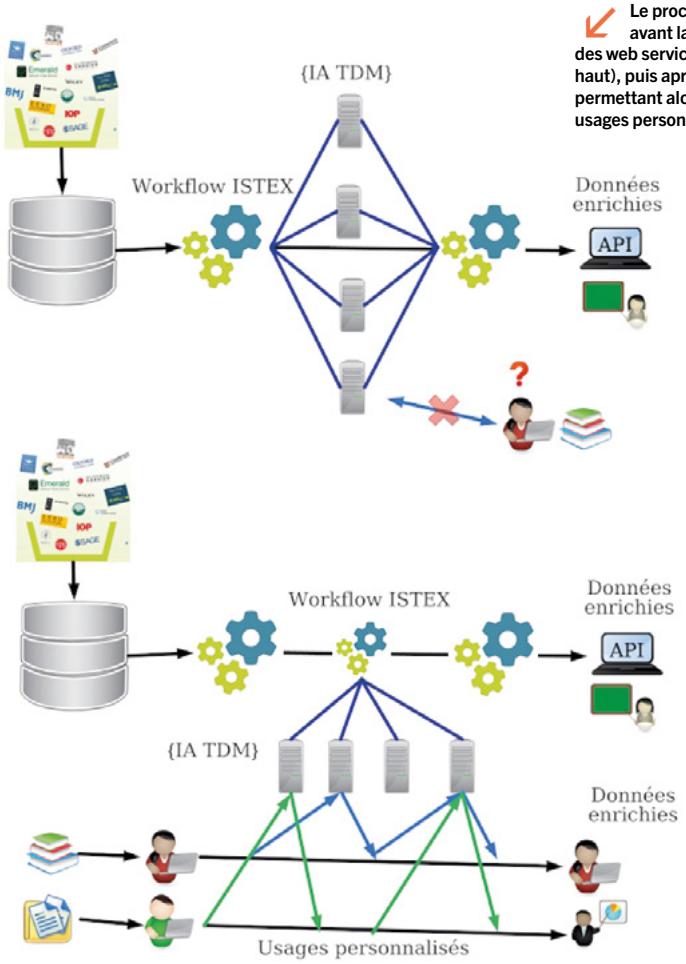
avec des contraintes minimales, utilisable par des non spécialistes, et largement extensible pour répondre à de nouveaux besoins.

Les méthodes implémentées peuvent être complexes, mettant en œuvre des réseaux neuronaux élaborés, avec un nombre élevé de paramètres à optimiser. Afin de faciliter au maximum leur usage, les web services doivent répondre à un certain nombre d'exigences :

- Chaque service ne doit répondre qu'à un seul besoin.
- Il n'y a pas de paramétrage par l'utilisateur.
- Il doit y avoir un seul format d'entrée/sortie très simple (ex : un JSON identifiant/valeur).
- Ils doivent être utilisables *via* une interface graphique (dans l'outil de visualisation Lodex[5])

Les modèles de ML sont construits par des spécialistes TDM, avec l'aide d'experts pour la constitution des corpus d'apprentissage et la validation des algorithmes, puis utilisés par les WS mis à disposition : l'IA devient alors accessible à tous et facilement applicable aux données bibliographiques, que ce soit sous forme de métadonnées ou de texte intégral[6]. Pour aider l'utilisateur, le site Internet objectif-TDM[7] recense les services en production : il permet à la fois d'identifier le service correspondant à ses besoins, de connaître son URL et d'avoir une aide sur son utilisation. À partir de là, le service est utilisable suivant les compétences de chacun : *via* une interface graphique dans Lodex (outil *open source* de visualisation de données structurées[8]), en ligne de commande, ou intégré dans un programme informatique.

Les nouveaux services proposés permettent l'utilisation de méthodes apportant une forte valeur ajoutée aux données traitées sans qu'il soit nécessaire de mobiliser des compétences en informatique, ou en *data mining*. Le système est suffisamment souple pour rapidement mettre en production de nouveaux services répondants à de nouveaux besoins ou à des données ayant des caractéristiques propres. En effet, les nouvelles approches font appel à des modèles de langage, modèles statistiques capables par exemple, à partir de la distribution de séquences de mots, de deviner une succession de mots. Avec le développement des approches neuronales (et le *deep learning*), on trouve d'importantes ressources pré-entrainées sur de gros volumes de données et dans des langues différentes. Ces gros modèles « génériques » peuvent avoir un intérêt mais peuvent aussi entraîner des biais importants[9] ; nous l'avons constaté, par exemple, dans le cas d'analyses de publications de la fin XIXe siècle au début XXe en français, où le style est fondamentalement différent de celui de nos jours. C'est le cas également dans le cas de corpus scientifiques spécialisés où le vocabulaire scientifique est très mal appréhendé *via* des modèles génériques. Pour faire face à ces cas-là, nous devons pouvoir nous adapter en créant nos propres modèles, adaptés aux données à traiter.



Le processus IA avant la création des web services (en haut), puis après (en bas), permettant alors des usages personnalisés

À l'Inist-CNRS, nous avons mis en place un environnement approprié facilitant le déploiement de ces services à partir d'algorithmes d'IA (plus spécifiquement de TDM). Cela permet une grande souplesse quant à la modification, l'adaptation ou la création de web services.

Cette nouvelle offre de service est donc là pour répondre à de multiples finalités et s'adresse à tous les professionnels de l'IST qui ont besoin, par exemple, de détecter des thématiques scientifiques, de classer des documents, ou encore de les enrichir pour faire de la bibliométrie. Elle propose des services assez génériques pour être utiles au plus grand nombre, mais est également capable de s'adapter aux besoins exprimés, et ainsi d'évoluer continuellement pour répondre à de nouveaux usages.

Pascal Cuxac
Responsable du Service Text & Data Mining, Inist-CNRS
pascal.cuxac@cnrs.fr

[5] www.inist.fr/projets/lodex

[6] Bonvallot V., Parmentier F., Bourguignon L., Clauss I. and Gregorio S. (2022) : Le TDM pour tous grâce à des web services au sein de LODEX, outil libre de visualisation, *Revue des Nouvelles Technologies de l'Information, Extraction et Gestion des Connaissances,* RNTI-E-38, 2022, 445-452 https://editions-rnti.fr/render_pdf.php?p=1002758

[7] https://objectif-tdm.inist.fr

[8] Gregorio, S., A. Collignon, F. Parmentier, and Thouvenin N. (2019) : LODEX : des données structurées au web sémantique https://hal.archives-ouvertes.fr/hal-01990444. *Atelier Web des Données Conférence EGC,* 2019, Metz, France.

[9] Bender E.M., Gebru T., McMillan-Major A., and Shmitchell S. (2021) : On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the *2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 610–623.