La fouille de textes en IST : les outils Istex-TDM

Pascal Cuxac

La fouille de textes en IST : les outils Istex-TDM

Pascal Cuxac [1]

1. INIST - CNRS
2 rue Jean Zay, 54500 Vandœuvre lès Nancy
pascal.cuxac@inist.fr

RÉSUMÉ. L'intelligence artificielle vient bousculer les habitudes des professionnels de tous domaines. À travers l'expérience récente de l'INIST, nous présenterons le développement et la mise en production de nouveaux services dans le domaine de l'Information Scientifique et Technique. Nous faisons le point sur la mise à disposition d'outils de fouille de textes, à destination de non spécialistes, aisément opérables sans connaissances préalables.

ABSTRACT. Artificial intelligence is changing the habits of professionals in all fields. Through the recent experience of INIST, we illustrate the development and production of new services in the field of Scientific and Technical Information. We take a look at the availability of text mining tools for non-specialists, which can be easily operated without any prior knowledge.

Mots-clés : Fouille de textes ; Intelligence artificielle ; IST ; Publication scientifique ; Offre de service ; Science ouverte

KEYWORDS: Text mining ; Artificial intelligence ; STI ; Scientific publication ; Open science

1. Introduction

Les données en libre accès se développent, que ce soit des collections issues de bibliothèques traditionnelles accessibles librement via Internet, mais également des entrepôts numériques regroupant des publications scientifiques nationales ou thématiques : Gallica, Europeana, HathiTrust's digital library, HAL, Isidore, Erudit... De récentes initiatives nationales ont également permis le développement d'archives scientifiques (ISTEX en France, SwissBib en Suisse, GBV en Allemagne, Scholars Portal en Ontario), et nous assistons à la montée en puissance de bases agrégeant des centaines de millions de publication comme CORE ou OpenAlex.

Ces réservoirs de données sont la matière première pour mettre en œuvre des méthodes de fouille de textes qui permettront d'analyser la production scientifique. Cependant, la qualité des données, leur richesse, leur format, sont les premiers écueils rencontrés. Bien entendu des outils existent, mais souvent difficiles à mettre en œuvre par un non spécialiste.

Dans cet article nous illustrons l'utilisation de l'IA dans le domaine de l'IST (Information Scientifique et Technique) à travers les récents développements réalisés à l'INIST[1] et l'offre de service Istex-TDM. Cela sera complété par une démonstration de l'offre, des outils proposés et de leur utilisation.

1. https://www.inist.fr/

2. IA et IST : les défis à relever

Les professionnels de l'information doivent répondre aux demandes croissantes de tableaux de bord pour mettre en évidence, entre autres, des taux d'accès ouverts en fonction des disciplines, des instituts ou tout autre indicateur.

Il existe certes des applications « presse bouton » mais leur utilisation dépend d'un abonnement. En plus du coût d'accès, les données ne sont pas toujours homogénéisées, et tous les domaines scientifiques ne sont pas toujours bien représentés, notamment les sciences humaines et sociales (Maddi et De La Laurencie, 2018).

On constate également la mise en ligne croissante, via GitHub ou GitBucket, de programmes permettant de traiter des données. Or leur mise en œuvre souvent complexe n'incite pas les non informaticiens à les utiliser. Des plate-formes comme Cortext, Gargantext et Weka sont aussi disponibles mais elles nécessitent souvent un niveau de connaissance des méthodes de TDM (Text and Data Mining) pour choisir parmi les algorithmes proposés et les paramétrer[2].

Si la fouille de textes a toujours été présente à l'INIST, ce n'est qu'avec le lancement du projet ISTEX[3] que des méthodes d'IA vont être développées pour être appliquées en grandeur nature sur de gros volumes de données, dans un processus industrialisé. Alors que l'IA n'était pas encore un mot-clé passé dans le langage commun, nous avons développé des méthodes d'enrichissement de données à partir notamment de techniques d'apprentissage automatique sous forme de modules intégrés à la chaîne de production (Cuxac et Thouvenin, 2017). 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.

Nous nous inscrivons dans le mouvement « Science Ouverte », en publiant tous nos codes, cependant nous voulons aller plus loin en faisant en sorte que qui que ce soit puisse les utiliser, quelque soit ses compétences. Cela doit répondre aux demandes d'utilisateurs, documentalistes ou chercheurs, qui souhaitent pouvoir utiliser ces programmes sur leurs propres données, et en pouvant choisir eux-mêmes les traitements dont ils ont besoin. Le public cible pour ces outils n'est pas le «data scientist», mais plutôt un utilisateur non expérimenté que ce soit en IA, en TDM ou en informatique. C'est un ingénieur ou un chercheur qui souhaite avoir des outils d'aide pour l'analyse de documents ou de corpus, sans avoir à maîtriser des processus complexes.

3. L'approche par web-services : d'une IA intégrée dans un processus défini à une boîte à outil modulaire

Nous avons fait le choix de créer et déployer des applications d'IA sous forme de web-services[4] (WS), intégrables dans une chaîne de production comme ISTEX, mais

2. Cortext https://www.cortext.net ; Gargantex https://gargantext.org ; Weka https://waikato.github.io/weka-wiki/
3. https://www.istex.fr/
4. Un WS est une forme spécifique d'API (https://en.wikipedia.org/wiki/Web_service)

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 applicable sur ses propres données, 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 ;
- ils doivent être utilisables via 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 et ainsi applicable aux données bibliographiques, que ce soit sous forme de métadonnées ou de texte intégral (Bonvallot *et al.*, 2022). Pour aider l'utilisateur, le site internet ISTEX-TDM[6] recense les services en production : il permet à la fois d'identifier le service correspondant à ses besoins, connaître son url et avoir une aide sur son utilisation.

A partir de là, le service est utilisable via une interface graphique dans Lodex (outil open source de visualisation de données structurées (Gregorio *et al.*, 2019), (fig 1). 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 datamining.

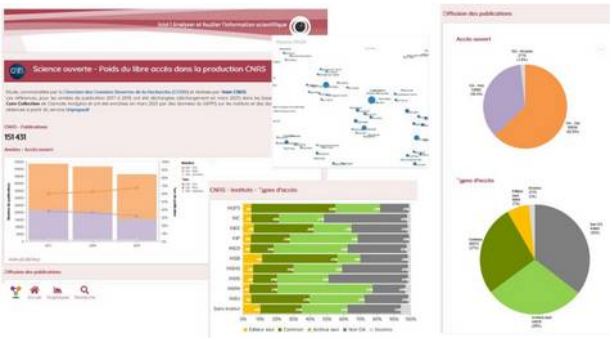

Figure 1 : Représentations graphiques sous Lodex (d'après Bonvallot et al. 2022).

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

5. https://www.inist.fr/projets/lodex/
6. https://services.istex.fr/

besoins exprimés, et ainsi d'évoluer continuellement pour répondre à de nouveaux usages.

4. Conclusions et perspectives

Nous avons mis en place un environnement approprié facilitant le déploiement de services de fouille de textes à partir d'algorithmes d'IA. Cela permet une grande souplesse quant à la modification, l'adaptation ou la création de nouveaux web services. Cette offre de service à destination de non spécialiste de fouille de textes (en priorité appartenant à un établissement de recherche publique), permet de façon extrêmement simplifiée d'exécuter des programmes complexes sans connaissances spécifiques a priori. Par rapport aux plateformes d'analyse de données cette solution est plus légère pour l'utilisateur et facilement interfaçable avec des outils de visualisation.

L'offre de service évolue rapidement, proposant de nouveaux web-services de façon régulière. Très prochainement nous allons mettre à disposition une interface simple permettant à l'utilisateur de charger ses données dans quelques formats simples (y compris csv), de choisir le traitement à faire, et d'être informé par mail avec un lien de téléchargement du résultat quand le traitement est terminé. Notre procédure de mise en production de ces web-services étant automatisé, nous proposons également de travailler avec des chercheurs afin de développer de nouveaux services performants et adaptés aux besoins.

Bibliographie

Bonvallot V., Parmentier F., Bourguignon L., Clauss I. et 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)

Cuxac P., (2022). L'IA et la fouille de textes à l'INIST : l'IA à portée de tous ?, *Arabesques* 107, 2022, : https://publications-prairial.fr/arabesques/index.php?id=3098)

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

Gregorio, S., Collignon A., Parmentier F. et Thouvenin N. (2019). LODEX : des données structurées au web sémantique (https://hal.science/hal-01990444). *Atelier Web des Données ,Conférence EGC, 2019*, Metz, France.

Maddi, A. et De La Laurencie A. (2018). La dynamique des SHS françaises dans le Web of Science : un manque de représentativité ou de visibilité internationale ? (https://hal.science/hal-01922266). working paper.