Comment répondre à une question relative à la fouille de texte et de données ?

1. Définitions

La **fouille de texte et de données** ou ***text and data mining*** (TDM) est définie par l'article L. 122-5-3.-I du Code de la propriété intellectuelle comme « la mise en œuvre d'une technique d'analyse automatisée de textes et données sous forme numérique afin d'en dégager des informations, notamment des constantes, des tendances et des corrélations ».
En d'autres termes, il s'agit de l'extraction et l'analyse de données numériques massives via un algorithme sur des critères précis comme le nombre d'occurrences, les tendances…

La fouille de texte intervient dans différents domaines de l'intelligence artificielle, par exemple :

- **L'apprentissage automatique (*machine learning*)**, qui « vise à donner aux machines la capacité d'apprendre à partir de données, via des modèles mathématiques » (CNIL). **Exemple** : le Baromètre français de la Science Ouverte utilise des méthodes de *machine learning* pour détecter automatiquement au sein de publications scientifiques en texte intégral des mentions de jeux de données ou de logiciels.
- **L'apprentissage profond (*deep learning*)**, procédé d'apprentissage automatique qui imite le fonctionnement des neurones au sein d'un programme informatique, permet à ce programme d'apprendre en autonomie, notamment via l'analyse de textes et de données.

2. Cadre juridique

La fouille de textes et de données à des fins de recherche est « une pratique confirmée et opérationnelle en droit français grâce à la transposition de la directive européenne sur le droit d'auteur et les droits voisins dans le marché unique numérique » (Ouvrir la science).
En novembre 2021, la France a adopté et ajouté au Code de la Propriété Intellectuelle une ordonnance qui permet de reproduire des contenus protégés par des droits de propriété intellectuelle dans le but de conduire des activités de fouille à des fins de recherche scientifique, sans avoir à recueillir d'autorisation préalable des « titulaires de droits » ou à obtenir des licences de leur part, ce qui rendrait l'opération impossible à cette échelle. Aucune compensation n'est à verser aux titulaires de droits[1].

C'est dans cette dispense d'autorisation que réside « l'exception ». Cette exception ne doit, par ricochet, donner lieu à **aucune utilisation commerciale**, qu'elle soit directement exploitée par l'établissement de recherche en question, ou bien cédée à une entreprise, via une société d'accélération du transfert de technologies (SATT) par exemple.
Vous pouvez effectuer des activités de fouille à des fins de recherche scientifique, sur des textes et sur tous les types de contenus numériques : données, images fixes ou animées, sons, musiques, logiciels, etc.

L'essentiel à retenir

- **Un champ d'application étendu** : textes, mais également tous les types de contenus numériques : données, images fixes ou animées, sons, musiques, logiciels.
- **Des conditions à respecter**[2]
  - Le TDM peut être utilisé à des seules fins de recherche scientifique :
    * par un organisme de recherche ou une institution du patrimoine culturel ;
    * à condition que l'accès à la source ait lieu « de manière licite ».

1. https://www.ouvrirlascience.fr/je-publie-quels-sont-mes-droits/ - p. 18.
2. Stage URFIST « Droit des données de la recherche » - Lionel Loubet. Formation à distance à destination des membres des Ateliers de la donnée – 07 et 08 mars 2024 https://sygefor.reseau-urfist.fr/#/training/10518/

- Le TDM ne peut pas être utilisé dans deux cas :
  * dans le cadre d'un partenariat à but lucratif ;
  * lorsque les contenus à fouiller contiennent des données protégées (données personnelles - RGPD - ou sensibles).
- La conservation des copies des œuvres est possible sans limitation de durée, y compris pour la vérification des résultats de recherche,
- L'archivage et le stockage doit être réalisé « avec un niveau de sécurité approprié », c'est-à-dire respectant les mesures de sécurité et d'intégrité mises en place par le fournisseur des contenus,
- Il n'y a pas de restrictions spécifiques pour l'utilisation du *data mining* par Intelligence Artificielle - Communication de la Commission Européenne du 31 mars 2023.

3. Enjeux

Les activités de *text and data mining* sont porteuses de nombreux potentiels pour la découverte scientifique et le développement de nouvelles connaissances. Elles doivent permettre au monde de la recherche de bénéficier des progrès rendus possibles par l'analyse des *big data*, en autorisant les chercheurs à opérer des fouilles automatisées dans l'immensité des documents scientifiques disponibles. Pour la recherche scientifique, l'exploration de données permet notamment :

- d'exploiter, de croiser et de réutiliser massivement des produits de la recherche (ressources sémantiques, publications…), opération impossible manuellement ;
- de sélectionner plus rapidement la littérature scientifique utile pour un sujet ;
- de gagner du temps et de l'argent en contribuant à l'accélération de l'innovation ;
- de développer des enrichissements adaptés à ces textes par ajout de métadonnées ;
- d'automatiser certaines tâches pour se concentrer sur d'autres missions ;
- de faciliter l'accès aux documents, l'exploration de corpus et les analyses thématiques ou bibliométriques ;
- de favoriser la découverte de nouvelles tendances et la recherche transdisciplinaire.

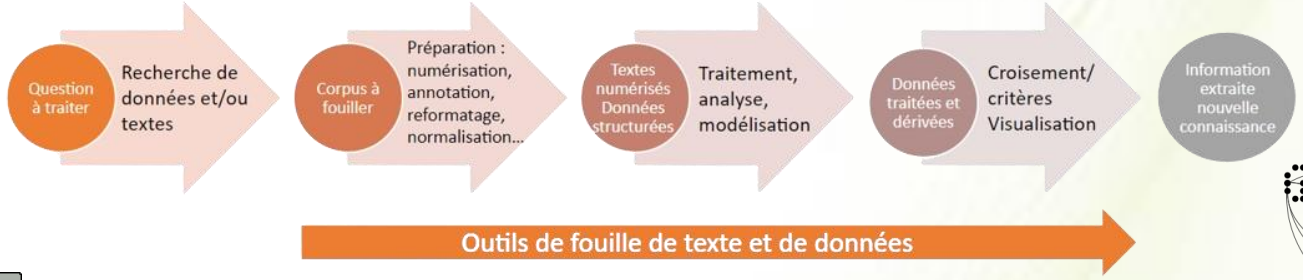
Exemple : dans le domaine de la santé, les techniques de *text mining* sont de plus en plus utilisées par les chercheurs. Le clustering d'informations permet par exemple d'extraire des informations à partir des livres de médecine de manière automatisée. Ceci permet de gagner du temps et de réaliser des économies.

Le TDM fait appel à de nouvelles compétences et de nouveaux métiers (développeurs spécialisés, ingénieurs de la connaissance…) et nécessite de nouveaux parcours de formation, à compétences multiples.

4. Outils et logiciels

La fouille de texte et de données fait appel à plusieurs méthodes pour extraire, explorer des données et les rendre exploitables. Cette pratique implique de disposer :

- de données d'entrée (ou *input*) : publications et données au sens large (tableaux, images, statistiques, textes…) devant être préparées avant la fouille :
  - pour les textes : numérisation puis création de métadonnées enrichies pour faciliter l'extraction de connaissance ;
  - pour les données, après extraction : la curation comprend des étapes de vérification, normalisation, annotation, reformatage, enrichissement, et structuration.
- d'outils de traitement permettant plusieurs actions telles que la sélection, le tri, l'organisation, l'analyse, la visualisation…



Inspiré de : https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

Les logiciels de TDM offrent des fonctionnalités pouvant couvrir toute la chaîne, de la préparation des données à la visualisation des résultats, en passant par l'application de calculs algorithmiques et l'exploitation de données de sortie et de modèles.
Les logiciels spécifiques à la fouille de texte traitent le langage naturel par des techniques linguistiques (étiquetage grammatical, ontologies, règles syntaxiques). Puis, sur les corpus de documents ainsi structurés, ils appliquent des algorithmes d'analyse sémantique permettant la classification automatique, l'analyse de tendances,
des associations de concepts.
Enfin, sur la base de critères choisis, l'utilisateur final pourra croiser et rapprocher plus facilement des données et ainsi les interpréter pour en sortir du sens.

Nota bene : les outils seuls ne suffisent pas, un contrôle qualité par l'équipe est nécessaire. Les formats et standards (souvent disciplinaires) à intégrer doivent être définis et les procédures de traitement doivent être formalisées. Cela participe aux bonnes pratiques de gestion FAIR des données produites.

Quelques exemples de logiciels

Parti pris : ne seront cités ici que des logiciels et outils open source ou libres.

Le bien-nommé Grobid (*Generation of Bibliographic Data*), développé par l'INRIA, extrait des contenus, de l'information bibliographique puis propose une analyse statistique de termes récurrents. Par exemple, Grobid pourra extraire automatiquement du PDF d'une publication scientifique son titre, son résumé ou encore les auteurs qui y sont mentionnés. Softcite et DataStet fonctionnent en complémentarité avec Grobid ; le premier pour détecter les mentions de logiciels, le second pour les mentions de jeux de données.

CorTexT, développé par le LISIS, est « une plateforme de développement méthodologique, d'ingénierie logiciel et d'appui à l'analyse de corpus textuels pour les Sciences Humaines et Sociales. [...] L'objectif est de doter les utilisateurs d'informations et d'outils innovants aux fins de traiter, caractériser, analyser et quantifier des données textuelles peu ou pas calibrées » (source).

GarganText est un outil de visualisation terminologique d'un corpus textuel développé par l'ISC-PIF. Il produit des cartes interactives et évolutives. Il permet de construire une carte thématique de mots, de nourrir un article de type « état de l'art » et évite de passer à côté d'une thématique incontournable sur une problématique donnée. L'outil a permis par exemple d'établir une cartographie interactive de la recherche sur le coronavirus et de ses liens avec les autres maladies, d'analyser des milliers d'articles, de faire ressortir les thématiques abordées et leur organisation puis de dégager les termes les plus représentatifs du corpus, ici « vaccins efficaces ».

De nombreux outils de TDM sont également proposés par l'Inist-CNRS pour faciliter l'exploration ou l'exploitation de corpus de textes. Il s'appuient sur le réservoir d'archives scientifiques ISTEX, parmi eux :
- ISTEX-DL : pour le téléchargement massif des résultats
- LODEX : pour la visualisation et l'exploration d'un corpus
- DATA.ISTEX : des exemples de corpus prêts à l'emploi

Le médialab de SciencePo recense et décrit des outils utiles dans le cadre d'opérations de TDM, tels que :
- Minet (*webmining*)
- OpenRefine (nettoyage)
- Dicto (annotation)
- Gazouilloire (collecte massive de données sur X)
- Hyphe (réseau d'acteurs web)

5. Un cas concret rencontré en bibliothèque : le Baromètre de la science ouverte

L'extension du Baromètre français de la science ouverte aux données de la recherche et aux codes logiciels a été menée grâce à des outils de *machine learning*. Concrètement, les mentions de jeux de données et de logiciels ont été repérées automatiquement sur les publications françaises préalablement téléchargées en PDF. Si une majorité de documents étaient en accès libre, l'accès aux publications sous droit pour réaliser des opérations de TDM a nécessité un échange avec les éditeurs concernés, notamment Elsevier.

Quelles étapes pour mener à bien ce projet ?

1. Délimitation du corpus des publications à télécharger ;
2. Création d'un jeton de téléchargement sur l'API d'Elsevier ;
3. Échanges avec le service commercial pour obtenir la possibilité de télécharger les PDF (seul le format XML était disponible) ;
4. Paramétrage du proxy de l'Université de Lorraine pour autoriser ces téléchargements massifs sans risque de coupure ;
5. Lancement des téléchargements puis application des algorithmes sur les PDF.

6. Quels services d'appui au TDM développer ?

Les bibliothèques universitaires et services de documentation en général, de par leur expertise en matière d'information scientifique et technique, et par leur habitude de gestion de la documentation numérique, ont toute légitimité pour proposer des services de TDM. En fonction des profils et des compétences présents dans les équipes d'appui à la recherche, plusieurs types de services peuvent être proposés, du plus simple au plus complexe en termes de mise en œuvre.

Communication et promotion d'outils

Le service le plus simple à mettre en place est la création, sur le site des bibliothèques par exemple, d'une page de présentation sur le TDM, le contexte juridique, les possibilités et les outils mis à disposition de la communauté scientifique.

Exemples

- Université des Antilles : promotion des services ISTEX accessibles suite à une adhésion ;
- Université de Poitiers : production d'une fiche pratique sur l'exception TDM ;
- Université de Berne (Suisse) : fourniture d'une page web avec toutes les ressources éligibles au TDM, les modes d'emploi et liens API pour chaque ressource.

Formation

Au-delà de l'information mise à disposition sur un site web, les services IST peuvent également proposer des formations, notamment au niveau Doctorat. Ces services sont généralement assurés en partenariat avec d'autres services universitaires.

Exemples

- Le Centre des Humanités Numériques des Grands Moulins, dans lequel des personnels des bibliothèques sont impliqués, propose notamment des formations à la fouille de données et à l'apprentissage automatique ;
- Le médialab de Sciences Po propose un accompagnement complet à l'usage des méthodes numériques ;
- Les bibliothèques de l'Université de Cambrige (Royaume-Uni) proposent un libguide complet pour se former en ligne.

Accompagnement au TDM

Pour les équipes les plus avancées, l'accompagnement au TDM représente la dernière étape. Cet accompagnement peut aborder, selon des formes variées, des questions juridiques comme techniques. Ce dernier niveau suppose de solides compétences.

Exemples

- Échanges avec le service commercial d'un éditeur réticent au TDM ou ayant coupé l'accès suite à un usage de TDM inattendu ;
- Aide au paramétrage d'une API de TDM, comme celles proposées par Elsevier ou Wiley ;
- Échanges avec les services informatiques pour paramétrer les proxys de l'établissement et assurer la faisabilité de cet usage.

Mise en place d'un partenariat avec un interlocuteur incontournable en France

Un partenariat avec l'Inist-CNRS, les plateformes universitaires de données (IR* Progedo), l'IR* Huma-Num ou encore des laboratoires tels que le médialab de Sciences-Po est recommandé pour se doter des compétences nécessaires pour développer des services avancés. C'est le cas par exemple du pôle Données et Humanités numériques (DHUNE) de la MSH Mondes, qui, « en tant que relai des IR* Huma-Num et Progedo, facilite l'accès aux ressources et services de ces infrastructures ».

7. Liens utiles

- Webinaire GTSO Couperin - Données « Text and data mining : explorons les cités enfouies ! » du 21/06/2023

  Présentation de Maxime Crépel *Analyse des récits médiatiques sur l'IA.*

  Présentation de Lionel Villard *Logiciel CorText Manager - Extraction d'information et analyse socio-sémantique pour les sciences humaines et sociales*
- Webinaire CNRS Inist - Découverte du TDM, Fabienne Kettani-Schmittheisler, 2023
- Rapport de mission : Transposition des exceptions de fouille de textes et de données : enjeux et propositions. Conseil supérieur de la propriété littéraire et artistique, 2020
- DIST – CNRS, Qu'est-ce que le text and data mining ? In : *Une Science ouverte dans une République numérique — Guide stratégique*. OpenEdition Press, 2017
- Supports de formation : ANF TDM 2023 | Exploration documentaire et extraction d'informations (CNRS) 12-13 oct. 2023
- Le réseau MATE-SHS organise régulièrement des webinaires pour présenter des outils et des méthodes de traitement des données, les Tuto@Mate