Vous avez dit text-mining ? Zéro règle et pourtant

Patrice Lopez est caractérisé par un profil particulier, il est informaticien indépendant mais collabore avec l'Inria comme collaborateur extérieur depuis 2010 Après un parcours à l'Office Européen des Brevets notamment en tant qu'adminstrateur de services d'information scientifique et technique, il a cherché comment promouvoir la science ouverte et naturellement, par son expertise informatique, s'est orienté vers le développement de logiciels Open Source. Il défend le Machine Learning (apprentissage automatique) comme soutien aux métiers et fonctions des professionnels de l'information scientifique et technique, avec comme ambition une appropriation de nouveaux outils et nouvelles pratiques par la communauté de l'enseignement supérieur et de la recherche.

Communication : Présentation des projets que vous menez actuellement dans vos fonctions et activités

Patrice Lopez : Il y a trois ans, j'ai créé mon entreprise dont la raison sociale est de développer et faire vivre des logiciels libres. De nouvelles technologies sont aujourd'hui disponibles pour soutenir les activités d'information scientifique et technique et améliorer l'accès aux données associées.
Pour illustrer mon travail, je peux citer deux logiciels Open Source :

- Grobid pour extraire automatiquement des informations contenues dans les documents plein textes en PDF. C'est un outil très utilisé par exemple par ResearchGate, ou des organismes de recherche comme le CERN, la NASA, ou encore dans HAL.
- entity-fishing pour l'extraction et la désambiguïsation d'entités qui constitue un outil d'enrichissement sémantique adapté à de gros volumes de données, et proposant des hypothèses d'interprétation en lien à Wikidata et Wikipedia.

Ces outils reposent sur l'apprentissage automatique. Le but est d'améliorer les outils des chercheurs, en particulier les moteurs de recherche et les processus d'accès à la littérature scientifique, mais également à plus long terme de les aider à mieux gérer et valoriser les connaissances et données de la recherche.

Avec l'Inria, je participe à des soumissions de projets et accompagne ceux-ci en encadrant techniquement les personnes impliqués. La majorité de ces projets sont orientés vers les humanités numériques avec le Consortium pour une infrastructure de recherche européenne, l'ERIC DARIAH.
Cependant, avec le projet Opaline par exemple, en partenariat avec BrailleNet, nous touchons plutôt une problématique sociétale. Le but est d'utiliser le text mining pour rendre automatiquement les ebook accessibles aux personnes non-voyantes et mal-voyantes. En structurant automatiquement les ebook, il est possible d'effectuer une transcription braille de qualité, de créer des systèmes de lecture et de navigation vocale, d'exploiter des marqueurs de présentation (tels les tours de parole). D'autre part, la reconnaissance d'entités nommées aide la transcription des noms propres qui suivent une codification braille différente

C. : Dans le cadre de vos activités, pouvez-vous définir les compétences nécessaires aujourd'hui pour les professionnels de l'information avec lesquels vous collaborez. Comment vos compétences s'articulent dans le cadre des projets que vous portez ?

P. L. : Le cœur de mes activités repose sur une maîtrise des techniques de Machine Learning. Le concept existe depuis les années 50, mais grâce à une certaine maturité technologique, cette approche prend réellement son ampleur depuis une dizaine d'années. Ceci ne s'apparente pas au Semantic Web qui s'intéresse à la représentation et l'échange de données structurées, mais à du Text and Data Mining (TDM) où l'objectif est de créer des données structurées et des connaissances à partir de données brutes.

Pour souligner le contraste avec les approches de type systèmes experts, grammaires ou automates, toujours très présentes, je dis que ma démarche implique « zéro règle », sous-entendu zéro règle ou connaissance créée manuellement. Mes outils utilisent des exemples annotés pour apprendre une tâche donnée, ou des données structurées existantes comme Wikipedia, ou encore capture des interactions manuelles d'utilisateurs.

L'autre caractéristique de mes outils est de viser autant que possible la généricité, à l'inverse d'une tendance établie à se restreindre à des domaines particuliers. Par exemple GROBID va traiter tous les PDF d'articles scientifiques, indépendamment du domaine, de l'éditeur et de la langue. entity-fishing va extraire des entités sur tous les domaines couverts par Wikidata et Wikipedia, donc en pratique sur tous les domaines scientifiques, et ne fait pas appel à des connaissances additionnels d'experts du domaine.

Compte tenu de cela, les profils de personnes avec lesquelles j'interagi sont au nombre de quatre :

1. Les informaticiens qui développent des plateformes comme Hyper article en ligne (HAL) ou INSPIRE du CERN. Ces personnes disposent de compétences très techniques en développement, ingénierie logiciel et architecture informatique. Ils intègrent par exemple mes outils d'extraction automatique de métadonnées ou de références bibliographiques, et parfois y contribuent.
2. Les spécialistes d'une discipline, cela peut être des astrophysiciens, des cancérologues, ou des chercheurs en sciences humaines et sociales (SHS). Dans le projet CENDARI par exemple, des historiens interviennent dans la numérisation d'archives sur la première guerre mondiale. Ces spécialistes apportent des données et des problèmes intéressants à résoudre.
3. Les professionnels de l'information scientifique et technique (IST) qui, par exemple, imaginent comment exploiter mes outils. Leur expertise couvre typiquement les systèmes de bibliothèques (résolveurs de liens), la bibliothéconomique, l'indexation, la connaissance des métadonnées et des formats, la maîtrise des vocabulaires des domaines scientifiques concernés.
4. Ceux que l'on désigne maintenant comme les « data scientists ». Il s'agit de personnes faisant le lien entre les techniques d'analyse de données (statistiques, machine learning, modélisation) et les processus métiers et économiques. Sur la problématique de Text mining, ces profils ont deux origines :

a. des informaticiens spécialisés soit dans l'apprentissage automatique appliqué aux documents, soit dans l'ingénierie et le traitement des données à grande échelle (Big Data) – ce sont des profils rares, très recherchés sur le marché de l'emploi,
b. des spécialistes de l'information de type « ingénieur documentalistes » capables de constituer des jeux de données d'entrainement. La valeur ajoutée primordiale dans le monde des données sont les annotations, provenant de la « curation de données »s ou provenant d'activités manuelles capturées par des moyens automatiques.

C. : Aujourd'hui, comment projetez-vous l'évolution des projets, des besoins liés à ces derniers ? Quelles compétences vous semblent primordiales à investir ou renforcer pour les professionnels de l'information ? Quel rôle à jouer pour ces derniers ?

P. L. : Aujourd'hui les compétences que j'identifie sont de plusieurs natures :
Juridique. Nous sommes au cœur de nombreux débats concernant la propriété intellectuelle et la protection des données personnelles. Par exemple, même pour des documents disponibles en accès libre, il faut tenir compte de la notion de réutilisabilité des données (cf Le nouveau régime juridique de la réutilisation des informations ...)

-) car par défaut aucun TDM n'est possible. Au-delà de la réutilisation des données se pose la question des œuvres dérivées que constituent les modèles d'apprentissage. Qui est propriétaire des droits d'exploitation de tels modèles? Ceci fait intervenir des métiers comme juristes en propriété intellectuelle, professionnels de l'information, services de valorisation des EPST…. et dépasse le champ du chercheur, des informaticiens peu sensibilisés par ailleurs.
- Communication. Les acteurs se côtoient, s'ignorent parfois. C'est là qu'intervient la notion de médiation, de vulgarisation, de pédagogie, de valorisation des outils, des expériences et le besoin d'intermédiaires entre les acteurs. Présenter, faire savoir, diffuser, convaincre, ce qui existe « dans le monde », toute la diversité des logiciels open source, bases de connaissances collaboratives, des ressources crées par la communauté. Développer une science ouverte qui ne soit pas guidé par le profit mais par le partage et qui s'affranchisse des monopoles des grands éditeurs scientifiques, implique apprendre à mieux communiquer et coopérer
- Evaluation. Il existe en IST un fort besoin d'évaluer de façon indépendante les outils existants et d'étudier l'impact auprès des usagers des nouveaux logiciels, processus et organisations. Cela est essentiel pour savoir comment améliorer les outils, comment mieux les interfacer dans les systèmes personne/machine, comment imaginer de nouvelles façons de travailler avec les données numériques.
- Gestion de projet. Les nouveaux projets en information scientifique supposent le travail en équipe d'une pluralité de profils différents. Nous avons besoins de chefs de projets capables de conjuguer toutes ces compétences et visions. Cela demande une compréhension technique, des métiers, des processus,

mais également la maitrise de méthode de travail collaborative, réactive, sans processus hiérarchiques contre-productifs et chronophages.

C. : Quelles sont les interactions avec l'INIST que vous entrevoyez ?

P. L. : Mon expérience dans le cadre d'ISTEX et un travail régulier depuis 2014 avec la constellation des équipes RD et Data m'amène à suggérer ces quelques pistes :

- Aller plus loin dans les activités d'enrichissement des ressources, c'est la grande singularité d'ISTEX au niveau mondial : par exemple un effort de structuration et d'uniformisation des plein textes ou connecter les ressources ISTEX à de grandes bases de connaissances scientifiques. Une telle valeur ajoutée ouvre des possibilités de recherche uniques aux ayant droits français.
- Pousser encore un peu plus loin l'intégration des ressources ISTEX dans les environnements de travail des chercheurs, ce qui passe par exemple par une meilleure description des données, plus d'activations des ressources de type Google Scholar, ou l'ajout de nouvelles fonctionnalités dans le bouton ISTEX.

Le défi à relever pour l'INIST est bien sûr s'adapter à une période de grands changements en IST où on assiste à une nouvelle organisation des métiers et des services mais aussi des relations entre les individus. Mais je pense qu'il s'agit également d'avoir la possibilité d'influencer ces changements et ces nouvelles organisations de façon pro-active pour défendre une certaine vision de la science et de la recherche publique qui valorise le partage et l'ouverture des connaissances.