Développé par la Bibliothèque nationale de Finlande, Annif est un puissant outil d'indexation automatique de documents de tous types.

Annif, l'indexation automatique à la Bibliothèque nationale de Finlande

Les bibliothèques gèrent un grand nombre de métadonnées, pour différents types de documents. Le plus souvent, ces documents sont indexés à l'aide de mots sujets sélectionnés au sein d'un vocabulaire contrôlé, afin d'être recherchés plus tard. L'indexation manuelle de documents est un travail intellectuel très consommateur de temps. Par ailleurs, aujourd'hui, de nombreux documents sont disponibles sous forme numérique. Il devient dès lors possible d'automatiser ce travail d'indexation, à partir du texte intégral ou de certaines parties des documents, comme les titres ou les résumés. Pour ce faire, la Bibliothèque nationale de Finlande a développé Annif, un programme d'indexation automatique disponible en *open source*. Après l'avoir « alimenté » avec un vocabulaire SKOS[1] (en l'espèce la General Finnish Ontology, YSO) et les métadonnées disponibles sur le site finna.fi, Annif sait désormais comment assigner des termes d'indexation à de nouveaux documents, ce dans plusieurs langues. L'outil a été développé sur GitHub[2] et, grâce à la collaboration entre Zenodo et GitHub, il dispose d'un DOI[3].

DR

Bibliothèque nationale de Finlande.

LES PROCESSUS DES SYSTÈMES D'INDEXATION AUTOMATIQUE

Les systèmes d'indexation automatique suivent généralement un processus dans lequel les documents textuels sont, dans un premier temps, « préparés », par exemple en taguant les phrases et les mots à l'intérieur d'un texte, en convertissant les mots en minuscules, en enlevant les mots vides et en lemmatisant les déclinaisons, les pluriels (...) de telle façon que les variations grammaticales d'un mot soient réduites au lemme[4] qui identifie le sens du mot. Ensuite, les documents sont convertis dans une représentation vectorielle de leur fréquence d'apparition, ce qui, à l'aide d'un algorithme, permet d'établir des comparaisons entre les mots-sujets utilisés. On peut aussi comparer directement ces termes avec ceux présents dans un vocabulaire contrôlé. Dans les deux cas, on produit une liste de « mots-sujets candidats » pour le document. Pour déterminer quels mots seront finalement suggérés pour l'indexation du document, la liste des candidats est ensuite classée. Seuls les plus « prometteurs » sont retenus. Les algorithmes d'indexation automatique se répartissent selon deux types d'approche : lexicale et associative. Dans l'approche lexicale, les termes apparaissant fréquemment – ou estimés significatifs, sont comparés avec les termes du vocabulaire de référence. De tels programmes peuvent être relativement simples mais, souvent, tous les sujets n'étant pas forcément mentionnés dans le texte du document, certains ne seront par conséquent jamais suggérés par les algorithmes.

La technique associative, qui inclut les techniques d'apprentissage automatique[5], trouve au contraire des corrélations entre les mots présents dans les documents et les mots-sujets, en exploitant les masses importantes de données d' « entraînement ». Ces deux approches peuvent être considérées comme complémentaires, les meilleurs résultats étant généralement obtenus en combinant les résultats issus des deux approches.

NAISSANCE ET DÉVELOPPEMENT D'ANNIF

Créé début 2017, le premier prototype d'Annif fonctionnait suffisamment bien pour démontrer la pertinence de cette approche et pour que les professionnels soient intéressés par sa mise en production. Début 2018, le développement d'une nouvelle version d'Annif a été initié. Conçu en langage Pyton, le nouvel Annif utilise les *frameworks* Flask and Connexion pour un serveur web, et la fonctionnalité REST API[6]. Les fonctionnalités d'indexation sont gérées par différents algorithmes pouvant être utilisés séparé-

[1] Simple Knowledge Organization System, un des standards de base du web sémantique.

[2] github.com/NatLibFi/Annif

[3] doi.org/10.5281/zenodo.2578948

[4] Même s'il s'agit d'une approximation, on peut assimiler un lemme à un mot.

[5] Terme préféré à celui de « machine learning ».

[6] Dans une API de type REST, chaque requête doit contenir l'ensemble des informations nécessaires à son traitement.

ment, ou combinés dans ce qu'on appelle des « ensembles ». Chaque algorithme étant implémenté comme autant de modules séparés, de nouveaux algorithmes pourront être ajoutés ultérieurement.
Avant leur analyse par les algorithmes, les documents textuels doivent être préparés, ce qui est pris en charge par Annif grâce à des outils spécialisés[7] qui fragmentent le texte en phrases et en mots. Ensuite, les mots peuvent être normalisés, notamment avec des algorithmes de lemmatisation. Chaque projet Annif est lié au vocabulaire d'indexation qu'il utilise, sachant qu'un même vocabulaire peut être partagé entre plusieurs projets Annif. Un module spécifique gère le chargement et le stockage des données des différents vocabulaires, sous la forme de simples fichiers CSV ou de fichiers SKOS/RDF[8].

FONCTIONNEMENT D'ANNIF

Actuellement, quatre algorithmes d'indexation ont été implémentés. Si « Maui » utilise une approche lexicale, «TF-IDF», « fastText » et «Vowpal Wabbit» utilisent, chacun à leur manière, une approche associative.
Tous les algorithmes d'indexation automatique ont leurs limites. L'attribution de mots sujets incorrects peut avoir de multiples causes, par exemple les homonymes (« son », qui peut renvoyer à la musique ou aux céréales), les noms mal interprétés (le général Boulanger pris pour une profession), les fausses corrélations, la fréquence d'un mot qui n'est pas pour autant un des sujets du document analysé, des données fausses utilisées pour l'entraînement, etc. Généralement, chaque algorithme produit ses propres erreurs. Pour améliorer la qualité des outils, favoriser leurs forces et diminuer l'impact de leurs faiblesses, une bonne stratégie consiste à les combiner entre eux. Pour l'indexation automatique, les méthodes de fusion sont une façon de combiner les résultats obtenus par différents algorithmes en construisant des ensembles et en choisissant les sujets retenus par le biais d'un arbre de décision appliqué aux « prédictions » de chacun des algorithmes. Ces arbres peuvent eux-mêmes être divisés en arbres de décision « invariants » ou « spécifiques ». Dans le premier cas, chaque sujet est traité de la même manière. Dans le second cas, le traitement varie pour chaque sujet.
Avec Annif, ces deux approches sont combinées. L'outil propose une interface de commande ligne à ligne utile pour le paramétrage initial, l'entraînement et l'évaluation des modèles. Il est également possible d'évaluer les algorithmes, en comparant les résultats avec ceux de corpus de documents indexés manuellement. Annif sert ainsi au calcul de nombreuses mesures de données d'évaluation.
Précisons qu'à partir de l'interface de commande ligne à ligne, les informations nécessaires doivent être chargées en mémoire séparément pour chaque opération et, à la longue, la procédure semble inefficace. Une fois terminé le paramétrage initial, il est donc préférable d'utiliser Annif comme un web service en phase de production, ou pour l'intégrer à d'autres systèmes. La mise en service de l'API d'Annif (REST) est relativement simple et s'intègre facilement aux outils standards de gestion de serveurs du type Apache Httpd.
L'API 'suggest' est au cœur de ce web service : en fournissant un document textuel en entrée, on obtient en sortie une liste (au format JSON) de suggestions de concepts pour l'indexer ainsi que l'appréciation quantifiée de leur niveau de pertinence. Autre fonctionnalité importante, «learn» est capable de mettre à jour les programmes en exploitant des correspondances vérifiées entre des documents décrits et les mots sujets choisis.

EXEMPLES D'UTILISATION

L'indexation par sujets peut fonctionner de façon semi-automatisée ou complètement automatisée. Dans le premier cas, l'algorithme fournit des suggestions que l'indexeur humain peut ou non accepter. La précision des suggestions peut être limitée, mais les suggestions doivent rester pertinentes.
Dans le second cas, les propositions sont automatiquement prises en compte dans l'indexation du document concerné.
Un outil de type semi-automatisé a été mis en place à l'Université de Jyväskylä, où Annif aide les étudiants de master et les doctorants à choisir les termes d'indexation les plus pertinents pour leur travail universitaire. Il est apparu que la moitié des termes suggérés par le système avait été accepté, soit par les étudiants, soit par les bibliothécaires ayant ensuite validé la sélection, prouvant l'efficacité de l'algorithme.
L'indexation entièrement automatisée devient indispensable pour des corpus importants, pour lesquels une intervention manuelle ou semi-automatisée n'est pas envisageable. Dans ce cas, les critères de sélection des termes sont forcément plus sévères, et seul un faible nombre de termes est retenu, du fait de leur fort coefficient de pertinence probable.
Un tel système a été utilisé pour deux corpus de documents de grande importance, la base finnoise de Wikipedia et celle du quotidien régional *Satakunnan Kansa*.
Après avoir procédé au déchargement d'un *dump* du Wikipedia finnois,tous les articles (plus de 450 000) ont été convertis en fichiers texte, en utilisant l'outil WikiExtractor. Chaque article a été ensuite analysé avec Annif, en appliquant des critères stricts de sélection des termes d'indexation, limités à trois sujets par article. En fait, la moyenne s'est établie à 1,56 sujet par article. Le système a permis de traiter 8 articles par seconde, soit un peu plus de 16 heures pour l'opération complète.
L'analyse de l'indexation attribuée a montré une prédominance des contributions consacrées au cinéma, aux sports d'équipe et aux groupes musicaux, mais aussi… aux navires de guerre et aux évêques, ce qui, après vérification, correspondait bien aux contenus des articles présents dans la base finnoise de Wikipedia. Compte tenu du caractère multilingue du vocabulaire YSO, cette opération pourrait être réalisée pour générer des termes d'indexation par exemple en suédois et en anglais.
Le même type d'opération a été réalisé sur l'ensemble des articles publiés entre 1987 et 2004 par le journal *Satakunnan Kansa,* soit plus de 110 000 articles. Ce traitement a duré 4 heures 30, soit un peu plus de 7 articles par seconde. Cette fois, les sujets privilégiés étaient liés à la politique locale, l'utilisation des devises et le festival local de jazz Pori. L'absence, dans le vocabulaire YSO, de termes pour désigner les Lieux a conduit à la mauvaise attribution de noms de bâtiments (par exemple d'églises) : dans la mesure où le concept « ville de Pori » n'est pas présent dans le vocabulaire, le système utilisait un terme « approchant », en l'espèce une certaine église de Pori – prouvant qu'il faut analyser les résultats avec attention pour éviter la répétition de ce type d'erreur.

USAGES « NON CONVENTIONNELS »

Même si Annif est destiné essentiellement à l'indexation semi ou entièrement automatisée, le fait qu'il soit disponible sous forme d'API autorise son intégration dans des outils variés pour des usages allant au-delà du domaine de l'indexation sujets.

DR

Un premier prototype, entièrement mobile, est basé sur une application web qui utilise un service d'OCR[10] disponible sur le *cloud*. Le processus est relativement lent, du fait du chargement du fichier image, mais fonctionne sur n'importe quel smartphone disposant d'un navigateur web. Un second prototype, utilisant Android et le service Google ML Kit, génère une OCRisation en temps réel. Dans les deux cas, le système propose à l'utilisateur une liste de sujets dans un temps beaucoup plus court que celui que prendrait la lecture exhaustive du document. Pour l'instant, ces prototypes n'ont pas été testés dans des conditions réelles de production.

Une autre application « non conventionnelle » a été développée lors d'un hackathon organisé par la Bibliothèque nationale de Finlande : il s'agit de *Finna Recommends*. une extension du navigateur Chrome. Ajoutée à la barre de commande, l'extension propose à l'utilisateur, à partir de n'importe quelle page web préalablement sélectionnée, une liste d'ouvrages présents à la bibliothèque sur des sujets comparables. En fait, le texte sélectionné est envoyé à l'API Annif, les trois sujets principaux sont extraits, puis le catalogue de la bibliothèque est interrogé avec les mêmes mots, pour obtenir, en un clic, une liste de livres disponibles.

Dernier en date des prototypes développés à partir d'Annif, Annifbot, à la manière des «chatbot», interroge l'utilisateur sur ses centres d'intérêt qu'il convertit en mots sujets à l'aide du vocabulaire YSO, avant de lancer des requêtes dans les différentes bases de données à l'aide de Finna API pour proposer livres et images pertinents. La fonctionnalité est la même que si l'utilisateur recherchait lui-même dans le catalogue, sauf que le dialogue automatique se substitue au formulaire de recherche. Dans le futur, ce type d'outil pourrait rendre plus interactive l'utilisation des interfaces de recherche des bibliothèques.

L'AVENIR DES SYSTÈMES D'INDEXATION AUTOMATIQUE

Pour les bibliothèques et les organismes documentaires, l'automatisation de l'indexation constitue un besoin clairement identifié. Cela nécessite des outils pratiques, bien intégrés aux systèmes d'information et proposant une indexation de bonne qualité. Des systèmes commerciaux sont disponibles, mais ils sont souvent onéreux, disposent d'une base de vocabulaire limitée et posent des problèmes de langue et de systèmes propriétaire peu paramétrables. Par ailleurs, la plupart des outils disponibles en *open source* ne peuvent pas être intégrés dans d'autres systèmes. Dans ce domaine, Annif constitue donc une alternative innovante, souple et évolutive.

Les prochaines étapes sont d'ores et déjà planifiées : ajouter des algorithmes, incorporer de nouveaux types de vocabulaires (noms de lieux, termes de classification comme la CDU ou la Dewey). À l'avenir, Annif sera utilisé pour améliorer l'indexation de corpus de documents électroniques, mais aussi pour d'autres types de documents, encore dépourvus d'indexation.

Osma Suominen
Spécialiste du Système d'Information
Bibliothèque Nationale de Finlande
Osma.suominen@helsinki.fi

[7] Nommément NLTK, le langage naturel de la boîte à outils Python.

[8] Respectivement des fichiers de type tabulé et utilisant le cadre de modélisation RDF, Resource description format.

[9] help.ubuntu.com/lts/serverguide/httpd.html

[10] Reconnaissance optique de caractères