PROPOSITION DE SUJET DE THESE

Campagne 2019

Laboratoire L3i

Sujet de la thèse :

Détection d'évènements sur la presse multilingue pour l'analyse de la propagation d'informations

Résumé du travail proposé :

Le sujet de thèse proposé rentre dans le contexte de l'exploitation des corpus massifs de presse ancienne en plusieurs langues. Alors que les bibliothèques nationales européennes investissent considérablement depuis des décennies pour la numérisation d'ouvrages, leur accès reste à ce jour extrêmement problématique en conséquence d'un processus de numérisation intrinsèquement imparfait, en particulier pour les documents anciens. On estime ainsi que sur Gallica, le portail numérique de la BnF, près de 10% des mots sont mal représentés, et qu'un grand nombre de ces erreurs concernent justement les entités nommées, qui font l'objet de 80% des requêtes (Chiron et al. 2017).

Cet état de fait crée un goulet d'étranglement nuisible au processus d'analyse des données historiques dans le cadre des humanités numériques, alors même que la presse historique est un outil essentiel pour les chercheurs en sciences humaines et sociales, que ce soit par exemple en histoire (Elart, 2017), mais aussi en littérature (Kalifa et al, 2011), en arts du spectacle (Amy de la Bretèque, 2017), en géographie (Eliot et al., 2012) ou encore en linguistique. C'est également un sujet d'intérêt pour le grand public (ANNO[1], le service dédié exclusivement à la presse ancienne de la bibliothèque nationale d'Autriche reçoit par exemple 2500 visiteurs uniques par jour).

Mots clés :

Fouille de données, analyse sémantique de contenus, humanités numériques

Informations complémentaires :

Encadrant(s) :

- Antoine Doucet (directeur de thèse)
- Cyril Faucher (encadrant scientifique)

Date de début du contrat : septembre ou octobre 2019
Durée du contrat : 3 ans

Contexte de l'étude et description du sujet

La presse historique numérisée est le résultat d'un investissement financier considérable des bibliothèques et des collectivités. Elles constituent des ressources patrimoniales exceptionnelles pour les chercheurs en sciences humaines et sociales. Ces masses de données numérisées restent actuellement hélas exploitées très en deçà de ce qui est possible. Ce projet de thèse veut contribuer à permettre une l'analyse quantitative des données historiques.

Dans ce cadre, l'objectif central du travail de thèse proposé est d'améliorer l'accès au patrimoine Européen constitué par la presse ancienne par les actions suivantes :

1. Extraire, de chaque article, les entités nommées (nom de personne, de lieu, organisations, pays) mentionnées, associées aux métadonnées de publication (a minima, date de publication et source) ;

[1] http://anno.onb.ac.at/

2. Détecter dans chaque article d'éventuels évènements émergents, sous une forme normalisée ;
3. Construire un modèle de visualisation de la diffusion spatio-temporelle des informations au sein de la presse historique européenne.

Références bibliographiques :

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, Jean-Philippe Moreux, *Impact of OCR errors on the use of digital libraries*, ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17), Toronto, June 19-23, 2017.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Jean-Philippe Moreux, *ICDAR 2017 Competition on Post-OCR Text Correction*, IEEE International Conference on Document Analysis and Recognition (ICDAR 2017), Kyoto, November 2017.

Élart Johann, « Boieldieu en France de la Révolution française à la Première Guerre mondiale : pour une étude des transferts culturels entre Paris et les départements », in Alexandre Dratwicki et Agnès Terrier (dir.), *Art lyrique et transferts culturels* [actes de colloque, Opéra-Comique, 19 et 20 janvier 2012], Bru Zane Mediabase, 2017,

Armel Fotsoh, « Recherche d'entités nommées complexes sur le Web - propositions pour l'extraction et pour le calcul de similarité », thèse de l'Université de Pau et des Pays de l'Adour, 2018.

Kalifa D., Régnier P., Thérenty M.-E., Vaillant A., *La Civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, nouveau monde éditions, 2011.

Gaël Lejeune, Romain Brixtel, Antoine Doucet, Nadine Lucas, *Multilingual event extraction for epidemic detection*, in the Artificial Intelligence in Medicine (AIIM) Journal, 65 (2), Elsevier, p. 131-143, 2015. (CORE=A, impact factor: 2.14)
Prix du meilleur article 2015 en"Public Health and Epidemiology Informatics" par l'International Medical Informatics Association [IMIA], parmi 1272 références.

Gaël Lejeune, « Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel », thèse de l'université de Caen Normandie, 2013.

Oskar Gross, Antoine Doucet and Hannu Toivonen, *Term Association Analysis for Named Entity Filtering* in Proceedings of the Text REtrieval Conference (TREC 2012), Gaithersburg, Maryland, USA, November 6-9, 10 pages, 2012 [CORE A]

Oskar Gross, Antoine Doucet and Hannu Toivonen, *Named Entity Filtering based on Concept Association Graphs*, in 14th International Conference in Computational Linguistics and Intelligent Text Processing (CICLing 2013), Samos, Greece, March 24-30, 12 pages, 2013 [CORE B]

Oskar Gross, Antoine Doucet, Hannu Toivonen, *Language-Independent Multi-Document Text Summarization with Document-Specific Word Associations*, in Proceedings of the ACM Symposium on Applied Computing (SAC 2016), Pisa, Italy, p. 853-860, 2016 [CORE B]

Nihel Kooli, « Rapprochement de données pour la reconnaissance d'entités dans les documents océrisés », thèse de l'université de Lorraine, 2016.

Candidatures

Emails : antoine.doucet@univ-lr.fr, cyril.faucher@univ-lr.fr, cyrille.suire@univ-lr.fr
Titre du mail : "Thèse 2019 laboratoire L3i"
Pièces jointes : CV et lettre de motivation en PDF