PROPOSITION DE STAGE

Année 2018

Laboratoire L3i

Sujet de stage :

Extraction d'entités nommées et analyse d'opinion sur la presse ancienne

Résumé du travail proposé :

Au cœur du domaine émergent des humanités numériques et dans le cadre de la fédération Numeric de l'Université de La Rochelle, le sujet de stage proposé vise à s'appuyer sur un corpus de presse quotidienne fourni par la Bibliothèque nationale de France (BnF) afin d'y détecter dans un premier temps des entités nommées (noms de personne, de lieu, et d'organisation), et dans un second temps des indicateurs quant à l'évolution de l'opinion liée à ces entités nommées. On pourra ainsi mesurer l'évolution au fil du temps de l'opinion vis à vis d'une personne, d'un pays, etc. Si le temps le permet, nous pourrons ensuite utiliser des approches d'analyse de données pour découvrir des associations voire des corrélations entre différentes évolutions. A terme, ces éléments viseront à être exploités par des historiens, afin de leur permettre de d'appuyer (ou de suggérer) des théories qualitatives sur la base de l'analyse de données quantitatives issues de corpus de presse à grande échelle.

Les principales tâches du stagiaire seront tout d'abord les suivantes :

- Organiser les articles (déjà extraits) en une structure adéquate ;
- Réaliser l'extraction d'Entités Nommées (EN) ;
- Réaliser l'extraction de mots clés ;
- Réaliser l'extraction des opinions associées aux entités nommées.

Puis si possible :

- Proposer une méthode de visualisation des évolutions d'opinion pour un ou plusieurs EN ou mots-clés (par exemple, un graphe avec une courbe par EN) ;
- Proposer une méthode d'association des évolutions (par exemple sur la base de l'extraction de règles d'associations) ;
- Fournir un outil de visualisation des résultats correspondants aux besoins des historiens.

Toutes les techniques à mettre en œuvre s'appuient sur des outils génériques de recherche de motifs, qui ne requièrent quasiment aucune connaissance en traitement automatique des langues.

L'objectif premier du stage est de réaliser rapidement un « *proof-of-concept* » sur la réalisation d'un système plus large. Il faudrait impérativement arriver à un outil ou ensemble d'outils réalisant toutes les opérations ci-dessus, ce qui ne posera pas de difficultés particulières car, pour chaque étape, des outils sur étagère sont disponibles, en interne ou distribués en ligne.

L'objectif second est d'améliorer certaines des sous-étapes dans un contexte multilingue, car la finalité (qui dépasse largement le cadre de ce stage) est de réaliser un outil permettant de traiter des corpus de presse quotidienne simultanément dans toutes les langues dans lesquelles ils sont fournis.

Nous pourrons par exemple découvrir des motifs séquentiels tels que :

[émergence de sentiment positif sur « Catalogne » ➔ émergence de sentiment négatif sur « Espagne » avec un taux de confiance de 70%].

Mots clés : **fouille données, analyse sémantique de contenus, humanités numériques**

CPER NUMERIC:

- **Thème : Patrimoine numérique et industries du tourisme (e-patrimoine / e-tourisme)**
- **Actions concernées : e-Patrimoine**

Informations complémentaires :

Encadrant(s) : Antoine Doucet, Mickaël Coustaty

Equipe :

- ☒ **Images et Contenus**
- ☐ **Dynamique des systèmes et adaptativité**
- ☐ **Modèle et Connaissance**

Domaine d'application stratégique :

- ☐ **E-éducation**
- ☐ **Environnement et développement durable**
- ☐ **E-culture**
- ☒ **Valorisation de contenus numériques**

Cadre de coopération : Humanités numériques

Date de début du stage : janvier 2018

Durée du stage : 6 mois

Financement : CPER/FEDER

LOCALISATION DU STAGIAIRE (dans quel bureau) : à déterminer

BESOIN MATERIEL (indiquer qui fournit ce(s) matériel(s)) : à déterminer

Contexte de l'étude :

Dans le cadre d'une collaboration avec la Bibliothèque nationale de France et notamment les universités d'Innsbruck (Autriche) et de Helsinki (Finlande), nous visons l'exploitation temporelle des nombreuses archives numérisées, avec de nombreuses finalités dans le cadre des humanités numériques.

En lien avec l'ANR MRSEI Digistory, ce stage pourra faire office de Proof of Concept en vue de projets plus larges en cours de construction et d'évaluation.

Prérequis et contraintes particulières :

Master 1 informatique ;
Bon niveau anglais ;
Un intérêt pour une poursuite du travail à l'issue du stage serait un plus.

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.

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.

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.

Contacts – liens :

Email : {antoine.doucet,mickael.coustaty}@univ-lr.fr