Stage de Master Recherche 2016-2017 :
Titrage automatique des thématiques identifiées dans les corpus

Responsables de stage locaux (TETIS & LIRMM) : **Mathieu Roche**, **Pascal Poncelet**
Autres encadrants (ERIC & Hubert Curien) : **Julien Velcin**, **Christophe Gravier**

Localisation :
UMR TETIS (AgroParisTech, Cirad, Cnrs, Irstea)
500, rue J.F. Breton, 34093 Montpellier Cedex 5, France

Contact :
mathieu.roche@cirad.fr
pascal.poncelet@lirmm.fr
julien.velcin@univ-lyon2.fr
christophe.gravier@univ-st-etienne.fr

1 Contexte

De nombreux travaux de fouille de textes permettent (i) de faire émerger les descripteurs linguistiques les plus significatifs (mots, syntagmes) à partir d'un corpus puis (ii) de les regrouper. Ceci permet de mettre en exergue, de manière automatique, les thématiques abordées dans les textes facilitant l'organisation et l'indexation des documents, la recherche d'information, la compréhension et l'analyse des textes, ou même les résumer.

La réalisation du premier point s'appuie, en grande partie, sur l'utilisation de méthodes d'extraction de la terminologie à partir de textes (Hasan & Ng, 2014). Les approches de la littérature combinent des méthodes linguistiques et statistiques (Frantzi *et al.*, 2000; Pazienza *et al.*, 2005). De tels travaux ont récemment été proposés dans le cadre d'une collaboration de quatre laboratoires : ERIC (Lyon), Laboratoire Hubert Curien (Saint-Etienne), LIRMM (Montpellier) et TETIS (Montpellier) (Velcin *et al.*, 2016).

La deuxième étape du processus consiste à regrouper les descripteurs linguistiques permettant de mettre en relief les différentes thématiques abordées dans les textes. Pour découvrir des structures thématiques "cachées" dans les corpus, les méthodes appelées "topic models" sont largement utilisées comme le modèle probabiliste génératif LDA, i.e. Latent Dirichlet Allocation (Blei *et al.*, 2003).

Une fois les thématiques identifiées, une des problématiques aujourd'hui réputée difficile consiste à leur attribuer un titre à partir de l'ensemble des descripteurs linguistiques identifiés. Une telle tâche a des similitudes avec les travaux sur le titrage automatique de textes qui s'appuie sur des méthodes d'extraction de la terminologie et de génération de textes (Lopez *et al.*, 2014).

2 Travail à réaliser

Le travail de stage qui sera effectué dans le cadre du projet *Songes* [1] (Science des Données Hétérogènes) s'articulera autour des tâches suivantes :

1. Compléter l'état de l'art des approches les plus récentes ayant adopté une démarche similaire.
2. Proposer et mettre en œuvre une approche qui se déclinera selon les 4 étapes suivantes :

1. http ://textmining.biz/Projects/Songes/

- Identifier les descripteurs linguistiques (mots, syntagmes) propres à chaque topic obtenus avec différentes approches de l'état de l'art ;
- Sélectionner les descripteurs les plus pertinents par filtrage statistique et/ou sémantique ;
- Identifier les phrases les plus pertinentes au regard des descripteurs sélectionnés à l'étape précédente (approche de Recherche d'Information) ;
- Extraire les syntagmes les plus pertinents à partir des phrases identifiées à l'étape précédente.

3. Expérimenter les propositions sur des données réelles issues de divers domaines (actualités, agriculture, etc.). Dans ce contexte, un protocole d'évaluation devra être défini et mis en œuvre.

Notons que la méthodologie proposée pourrait avoir des applications directes pour d'autres tâches comme le titrage de clusters ou le titrage de nuages de mots.

Références

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

FRANTZI K. T., ANANIADOU S. & MIMA H. (2000). Automatic recognition of multi-word terms : the c-value/nc-value method. *Int. J. on Digital Libraries*, **3**(2), 115–130.

HASAN K. S. & NG V. (2014). Automatic keyphrase extraction : A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1262–1273, Baltimore, Maryland : Association for Computational Linguistics.

LOPEZ C., PRINCE V. & ROCHE M. (2014). How can catchy titles be generated without loss of informativeness ? *Expert Syst. Appl.*, **41**(4), 1051–1062.

PAZIENZA M. T., PENNACCHIOTTI M. & ZANZOTTO F. M. (2005). *Terminology Extraction : An Analysis of Linguistic and Statistical Approaches*, In S. SIRMAKESSIS, Ed., *Knowledge Mining : Proceedings of the NEMIS 2004 Final Conference*, p. 255–279. Springer Berlin Heidelberg : Berlin, Heidelberg.

VELCIN J., ROCHE M. & PONCELET P. (2016). Shallow text clustering does not mean weak topics : How topic identification can leverage bigram features. In *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing, DMNLP 2016, co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD 2016, Riva del Garda, Italy, September 23, 2016.*, p. 25–32.