

Topic modeling neuronal non-paramétrique pour l'extraction d'insight client : application à l'industrie du pneumatique

Miguel Palencia-Olivar^{*,**}, Stéphane Bonnevey^{*,**}, Alexandre Aussem^{***}, Bruno Canitia^{*}

^{*} Lizeo IT, 42 quai Rambaud, 69002 Lyon, France

miguel.palencia-olivar, stephane.bonnevey, bruno.canitia@lizeo-group.com

^{**} Lab. ERIC, Université de Lyon, 5 Av. Pierre Mendès France, 69500 Bron, France

^{***} LIRIS, Université de Lyon, 25 Av. Pierre de Coubertin, 69100 Villeurbanne, France
alexandre.aussem@univ-lyon1.fr

Résumé. À l'ère des médias sociaux, les clients sont devenus des faiseurs d'opinion : toute personne intéressée par un produit peut rechercher des avis sur les plateformes d'internet. Le web-scraping reste souvent la seule voie d'accès, et malgré le recours à l'ETL, l'hétérogénéité des données rend la tâche d'extraction d'insight ardue et induit le besoin d'outils ad-hoc. Pour contourner ce problème, nous appliquons l'Embedded Dirichlet Process et l'Embedded Hierarchical Dirichlet Process dans un cadre industriel autour du cas du pneumatique. Ces topic models non-paramétriques apprennent les thèmes et leur nombre, des plongements de thèmes et des plongements de mots, de telle sorte à désambiguïser les mots et à offrir davantage de niveaux analytiques. Ils peuvent également servir à affiner des processus ETL. L'EDP et l'EHDP atteignent des niveaux de vraisemblance similaires voire plus élevés que ceux des techniques de l'état de l'art testées, sans ré-exécutions pour trouver le nombre de thèmes.

1 Introduction

Les topic models constituent un ensemble de techniques de référence en matière de text mining non supervisé. Leurs applications sont nombreuses, allant du résumé de rapports cliniques à l'extraction de tendances dans la littérature scientifique (Boyd-Graber et al. (2017)). Depuis l'introduction de l'Allocation de Dirichlet Latente (Blei et al. (2003)), le domaine a connu des développements dans de nombreuses directions. Les dernières en date consistent à mêler réseaux de neurones profonds et programmation probabiliste afin de tirer parti des propriétés et avancées des deux domaines (Miao et al. (2016); Srivastava et Sutton (2017)), en particulier en termes de plongements (Dieng et al. (2020)). L'inclusion de plongements semble prometteuse, puisqu'elle revient à ajouter des éléments contextuels. Ceux-ci semblent rendre plus aisée la capture de thèmes tout en permettant de visualiser les liens entre mots et entre thématiques. Ces capacités sont particulièrement intéressantes dans le cadre de l'analyse de réseaux sociaux. Toutefois, les applications de topic models neuronaux sur de tels jeux de données portent souvent sur des corpus provenant de sources uniques accessibles par API ; ce