Etude approfondie des représentations de données textuelles dans l'apprentissage non supervisé

Mira Ait-Saada[*,**], Mohamed Nadif[*]

[*]Centre Borelli UMR9010, Université Paris Cité, 75006 Paris
[**]Caisse des Dépôts et Consignations, Datalab, 75013, Paris

Résumé. Les plongements de textes ont récemment suscité un grand intérêt dans plusieurs tâches telles que la classification de textes/documents et la réponse aux questions. Cependant, bien que de nombreux défis soient rencontrés dans le domaine de l'apprentissage non supervisé, on en sait beaucoup moins sur la pertinence de ces différents plongements lorsqu'on dispose d'un ensemble de documents non labellisés. Dans cet article, nous étudions l'utilisation de telles représentations sur des tâches non supervisées : le *clustering* de documents et la visualisation. Ainsi, pour répondre à l'objectif de *clustering*, nous proposons d'utiliser une *approche tandem* combinant des techniques de réduction de dimension et de *clustering*. Nous montrons d'abord l'avantage de s'appuyer sur le sous-espace obtenu par *Uniform Manifold Approximation and Projection* (UMAP) pour le *clustering* plutôt que d'utiliser la réduction de dimension basée sur l'Analyse en composantes principales (ACP), plus souvent utilisée. Ensuite, à travers des expériences réalisées sur des jeux de données réels, nous montrons l'efficacité de l'approche tandem proposée sur des modèles pré-entraînés par rapport aux stratégies de ré-entraînement proposées dans la littérature.

1 Introduction

Pour des besoins de labellisations de données de plus en plus massives, l'apprentissage non supervisé redevient un des principaux challenges de la science des données. De ce fait le *clustering* et la visualisation, par exemple, peuvent être très utiles pour créer de la valeur à partir des données non labellisées et ce notamment dans le contexte de données textuelles. Actuellement, un large éventail de représentations de données textuelles est proposé aux praticiens parmi lesquelles les sacs de mots sparses ou *Bag-Of-Words* (BOW) ainsi que les sacs de mots denses tels que Word2vec (Mikolov et al., 2013) et GloVe (Pennington et al., 2014) également appelés plongements de mots statiques. Plus récemment, Les représentations de textes/documents fournies par les Modèles de Langue Pré-entraînés basés sur les Transformeurs (MLPT) comme BERT (Devlin et al., 2019) et RoBERTa (Liu et al., 2019) qui produisent de différentes manières des représentations mot par mot pour représenter un document.

Malgré la multiplication des méthodes de plongement, il n'y a pas de réponse claire quant aux performances attendues dans un contexte non supervisé, où aucun label n'est disponible. En particulier, les plongements basés sur les Transformeurs suscitent de plus en plus d'intérêt,