Utilité d'un couplage entre Word2Vec et une analyse sémantique latente : expérimentation en catégorisation de données textuelles.

Oussama Ahmia*, Nicolas Béchet*, Pierre-François Marteau*, Alexandre Garel**

* IRISA, Université Bretagne Sud,
Rue Yves mainguy BP 573 56000 VANNES cedex
nom.prénom@irisa.fr, http ://www-expression.irisa.fr
** 2 Place Saint-Pierre, 44000 Nantes
a.garel@octopusmind.info, http ://www.octopusmind.info

Résumé. Nous réexaminons dans cet article les méthodes de vectorisation de textes dans le cadre d'une étude de classification de documents. Nous étudions les méthodes basées sur des plongements de mots (word2vec) ou de documents (analyse sémantique latente, ou sac de mots associées à diverses pondérations) ainsi que certaines combinaisons de ces méthodes. A cette fin, nous évaluons ces méthodes de vectorisation en utilisant trois modèles de classification (un perceptron multicouches, une machine linéaire à vecteurs supports optimisée par descente de gradient stochastique et un classifieur multinomial naïf de Bayes). Nos résultats montrent que le modèle proposé pour associer les méthodes word2vec et LSA, qui conjugue les deux caractérisations complémentaires du contexte d'occurrence des mots (local pour word2vec et global pour LSA), permet de produire une vectorisation robuste, en général plus discriminante que les autres approches testées.

1 Introduction

Avec la croissance rapide de l'information en ligne, la nécessité de développer des méthodes pour trouver, filtrer et gérer ces ressources de manière rapide et efficace devient d'autant plus prégnante. La classification de données textuelles consiste à classer de façon automatique des textes dans une ou plusieurs catégories. Cette dernière a déjà été appliquée à plusieurs problématiques notamment l'extraction d'information (Kushmerick et al., 2001), l'analyse de sentiments (Dey et Haque, 2009), la détection de SPAM (Jindal et Liu, 2007), etc.

Afin d'exploiter des algorithmes d'apprentissage automatique sur des données textuelles, il est souvent nécessaire de représenter le texte sous la forme d'un vecteur de taille fixe, ceci afin de plonger la donnée dans un espace métrique.

De nombreuse méthodes de "vectorisation" ont été développées au fil des années. La plus utilisée étant la méthode dite *sac de mots* (bag of words) (Harris, 1954) qui consiste à décrire un texte par les occurrences (fréquences) des mots qui le composent. Cette méthode considère que tous les mots dans un document donné ont le même poids ce qui est problématique pour