Courriel: febrissy.75@gmail.com, lazhar.labiod@parisdescartes.fr *MLDS7*

TER | UNIVERSITÉ PARIS DESCARTES

Machine Learning for Data Science

Analyse non supervisée de sentiments

Mickael Febrissy et Lazhar Labiod

1 Contexte

La classification automatique ou *clustering* consiste à partionner un ensemble d'objets (instances) décrits par un ensemble de variables en groupes (classes) homogènes. Avec l'avènement du Big Data et de la science des données, le *clustering* est devenu une tâche encore plus importante.

De nos jours, on recense différentes plateformes (réseaux sociaux, revues, etc.) permettant le partage d'avis utilisateurs au sujet de divers contenus (films, séries, musique, politique, etc). Dans le contexte cinématographique, il est courant de pouvoir noter un film de façon relative (positive ou négative) ou sur une échelle de 1 à 5, et d'y associer un commentaire. Néanmoins pour ce dernier, il devient beaucoup plus compliqué d'extraire des groupes à travers la sémantique associée.

2 Travail à réaliser

1. Dans un premier temps, il s'agira de construire un dictionnaire de mots permettant d'encoder les avis utilisateurs dans un format vectoriel.

2. Deuxièmement, des méthodes de pré-processing basiques seront mises en oeuvre afin de palier aux contraintes sémantiques, par exemple les charactères spéciaux, la ponctuaction, les lettres majuscules, les mots vides (stop words), les pré/suf-fixes (racinisation), etc. Seuls les mots les plus pertinents seront retenus. Des pondérations et normalisations connues comme pertinente pour l'analyse de données textuelles seront envisagées.

3. Troisièmement, il conviendra d'effectuer la classification des données obtenues en faisant appel à des algorithmes de Nonnegative Matrix Factorization (NMF), de k-moyennes adaptés aux données directionnelles (Spherical K-means), d'autoencoder.

4. Enfin, une étude comparative des résultats des différents algorithmes sera réalisée en utilisant des critères d'évaluation appropriés. Il conviendra de trouver des outils appropriés pour décrire les caractéristiques des groupes obtenus.

Une description des données (taille, dimension etc.) doit être précisée, les méthodes utilisées doivent être maitrisées et enfin tous les résultats obtenus doivent être rigoureusement commentés.

3 Mots clés

Apprentissage non supervisé, Spherical k-means, NMF, R, Python, TensorFlow, Keras, Autoencoder, Deep Learning.

4 Références

https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/
http://www.cs.cornell.edu/people/pabo/movie-review-data/
https://keras.io/