Proposition de PPD pour la formation
Master 2 MLDS - Université de Paris

Information concernant les encadrants

Encadrant(s) : Séverine Affeldt et Lazhar Labiod, MCF, Université de Paris
Email : severine.affeldt@u-paris.fr, lazhar.labiod@u-paris.fr

---

Description générale du projet

Intitulé du projet :

Interface web : Analyse avancée des textes courts

Application aux Tweets

Contexte:

Les Tweets sont devenus aujourd'hui une forme très répandue de communication sociale sur Internet, des milliards de tweets sont générés chaque jour. Découvrir des connaissances à partir de ces données suscite beaucoup d'intérêt de la part de l'industrie et de la recherche. La découverte de connaissances à partir de ce type données devient une tâche de recherche intéressante et stimulante, qui a beaucoup retenu l'attention des chercheurs. Les Tweets sont des textes courts ne comportant que quelques mots, ils peuvent être arbitraires, bruyants et ambigus. Tous ces facteurs rendent difficile la représentation efficace Tweets et la découverte de connaissances. Traditionnellement, la modélisation par thématique a été largement utilisée pour découvrir automatiquement les informations thématiques cachées dans les documents à contenu riche.

De manière générale, il existe deux groupes de modèles thématiques, à savoir les modèles probabilistes génératifs, tels que l'allocation de Dirichlet latente (LDA) et la factorisation matricielle non négative (NMF). Les modèles basés sur NMF apprennent les thématiques en décomposant directement la matrice de documents-termes, qui est une représentation matricielle de corpus de Tweets en deux ou trois matrices de facteurs de rang inférieur. Les modèles basés sur le NMF ont montré des performances exceptionnelles en matière de réduction de dimensions et de clustering pour les données de grande dimension.

Objectifs:

Les Tweets sont des textes courts, ils ont une information contextuelle limitée et sont clairsemés, bruyants et ambigus. Par conséquent, l'apprentissage automatique de thématiques (topics) à partir de ces textes reste un défi important. L'objectif de ce projet est d'aborder ce problème afin de répondre de manière efficace aux différents challenges posés par ce type de données textuelles aux méthodes basées sur des modèles probabilistes ou sur la factorisation matricielle.

Réalisations attendues:

1. Constitution d'un benchmark de données Tweets à partir de Twitter
2. Création d'une Interface web avec Python Streamlit, Dash, ou R shiny pour l'analyse des Tweets

3. Implémentation des méthodes de nettoyage de tweets et intégration à l'interface web
4. Utilisation des approches de "Topics modeling" (LDA, NMF) et intégration à l'interface web

Références

[1] Kais Allab , Lazhar Labiod, Mohamed Nadif : A Semi-NMF-PCA Unified Framework for Data Clustering. IEEE Trans. Knowl. Data Eng. 29(1) : 2-16 (2017)
[2] Kais Allab , Lazhar Labiod, Mohamed Nadif : SemiNMF-PCA framework for Sparse Data Co-clustering. CIKM 2016 : 347-356
[3] Aghiles Salah , Melissa Ailem , Mohamed Nadif: Word Co-Occurrence Regularized Non-Negative Matrix Tri-Factorization for Text Data Co-Clustering. AAAI 2018 :