MISE EN ŒUVRE D'UN PROCESSUS DE DATA MINING

Analyse automatique des avis clients

Adrien Guille - SAE 5.03 - Université Lumière Lyon 2

Cette SAE a pour objectif de vous aider à comprendre, par la pratique, la méthodologie générale du Data Mining, ici appliquée à des données textuelles. Le livrable technique, à rendre par groupe (composition libre) de 4 ou 5 étudiants, ne comptera que pour un tiers de la note. Le reste de la note sera individuelle et basée sur une épreuve évaluant la compréhension des aspects méthodologiques.

Contexte et objectifs

Une bijouterie en ligne vous commande une étude pour l'aider à mieux comprendre ses clients, sur la base des avis textuels laissés par ces derniers. Elle vous fournit un échantillon d'avis, que le service commercial a manuellement classé en deux catégories : « négatif » et « positif ».

1. Mettre au point un système pour catégoriser automatiquement les futurs avis
2. Mettre au point un système pour synthétiser automatiquement les futurs avis selon les principales thématiques

Méthodologie

Vectorisation du texte : sac de mots et pondération tf-idf

Codage numérique du texte selon l'approche "sac de mots"

- Pré-traitements sur le texte : retrait des "mots vides", lemmatisation, retrait de la casse
- Construction du vocabulaire : seuils sur les fréquences, incorporation de n-grammes
- Transformation du codage "sac de mots" : pondération tf-idf, normalisation des vecteurs

Prédiction de la polarité : classification supervisée

Entraînement d'une régression logistique régularisée

- Régularisation de l'estimation des paramètres : L1, L2, Elastic-Net
- Sélection du meilleur modèle : meilleur combinaison codage du texte / hyperparamètres de l'entraînement

Annotation thématique : réduction de dimension

Factorisation de la matrice documents-termes en matrices non négatives

- Identification de bons hyperparamètres : fonction de perte Frobenius ou Kullback-Leibler,

nombre de thématiques

- Annotation manuelle des thématiques
- Mesure de l'importance des thématiques par rapport à l'avis : approche fondée sur l'analyse d'une forêt aléatoire

Contraintes techniques

Bibliothèques autorisées

- pandas, numpy, scipy, spacy, nltk, sklearn, matplotlib, seaborn

Évaluation

Rendu collectif (1/3 de la note)

Le rendu attendu est un notebook. Il ne doit pas présenter toutes les expériences menées, mais exclusivement le code permettant d'obtenir les deux systèmes demandés à partir des données initiales, et le code pour obtenir une prédiction et une synthèse thématique pour un nouvel avis fourni par l'utilisateur du notebook. L'évaluation du classifieur devra être présentée dans ce notebook. Il est attendu un effort sur la mise en forme du notebook, la restitution des résultats et l'interaction avec l'utilisateur. L'évaluation sera faite du point du vue du commanditaire.

Quizz individuel (2/3 de la note)

L'épreuve se tiendra à la fin de la SAE et portera sur les 3 parties de la méthodologie. Elle visera à évaluer ce que vous avez appris (au sens méthodologique, pas technique) à travers la réalisation de ce projet.