Certificat Data Science: Projet

Nous avons un fichier CSV (tableur) contenant des informations sur les avis (*reviews*) donnés par les clients sur les produits qu'ils achetés sur Amazon ainsi que les notes qu'ils ont donné (*rating*). Le but de projet est d'explorer et de faire une analyse de sentiments des avis. Le projet consiste donc de deux tâches d'exploration et de classification des avis. Chaque ligne du fichier CSV contient trois champs: (i) le nom du produit acheté, (ii) l'avis d'un client qui a acheté le produit en question sous forme de texte, et (iii) un champ sentiment indiquant si l'avis est "positif" ou "négatif". Vous pouvez télécharger le fichier CVS compressé qui vous sera communiqué avec cet énoncé. Dans la tâche 1, vous aller utiliser ce que vous avec vu avec Jamal. Pour la tâche 2, vous allez utiliser ce que vous avez vu avec Dario.

Tâche 1: Analyse des sentiments
Dans cette tâche, vous êtes invités à développer une solution pour prédire si un avis est positif ou négatif. le résultat de cette tâche a de l'intérêt, par exemple, pour classifier les futurs avis pour lesquel on aura pas de sentiment associés.

1. En utilisant les méthodes de text mining vues en TP, construire l'ensemble des caractéristiques pour conduire votre classification (par ex. matrice termes-document, tfidf, etc.). A noter que les labels correspondent aux valeurs de la colonne "sentiment" dans le document CSV.
2. Proposer une méthodologie pour séparer votre ensemble de données en ensemble d'entraînement et en ensemble de tests.
3. Proposer et utiliser un ensemble de trois méthodes d'apprentissage supervisé (ex. arbres de décision, random forests, svm, bayesien naïf) pour construire un modèle d'analyse de sentiment.
4. Proposer un ensemble de mesures pour comparer les performances de ces méthodes (accuracy, precision, recall, f-mesure, ROC, AUC).

Tâche 2: Exploration des données
Dans cette tâche, vous êtes invité à explorer les données pour répondre aux questions suivantes en utilisant Spark.

1. Reprendre l'exemple WordCount vu en cours et le faire exécuter sur le fichier CSV.
2. Créer à partir du fichier CSV une data frame avec les champs (nom, avis, sentiment).
3. Calculez le nombre d'avis positifs par produit
4. Calculer le nombre d'avis positifs et le nombre négatifs par produits
5. Y'a t'il des enregistrements avec des avis vides ?

Pour les questions 2 à 5, vous pourrez les faire sur sur Spark RDD/Dataframes ou HIVE.