Projet conception d'une application Big Data

Ehouarn SIMON – ehouarn.simon@toulouse-inp.fr

Certificat Big Data

Novembre 2019

L'objectif de ce projet est de concevoir une application Big Data et de la déployer dans une infrastructure de calcul. Ce projet utilise les notions vues en cours et mises en pratique en TP. Le projet se compose de deux parties, une relative à la conception d'une application de traitement de données, l'autre à l'étude de performances sur une architecture de calcul. Seule la première partie est décrite dans ce fichier.

Conception d'une application Big Data

Le but de cette partie est l'implantation et l'évaluation d'une méthode de descente de gradient pour la résolution d'un problème de filtrage collaboratif. Nous avons à notre disposition un fichier contenant les évaluations d'un certain nombre de films par différents utilisateurs d'une plateforme. Ces utilisateurs n'ayant pas noté, ni même vu, l'ensemble des films à disposition, l'objectif est de pouvoir estimer ces notes « manquantes » depuis l'ensemble de notes, films et utilisateurs à notre disposition. L'objectif est de pouvoir proposer de futures recommandations de films pertinentes pour chacun des utilisateurs.

Ayant stocké ces données sous forme d'un matrice R=[r_{ij}] (ligne → utilisateur, colonne → film, entrée de la matrice → note), ce problème peut se modéliser comme la recherche d'une factorisation de rang faible de R. Ceci conduit au problème d'optimisation suivant :

$$\min_{P,Q} \sum_{(i,j) \text{ t.q. } \exists R_{ij}} \left(r_{ij} - q_j^\top p_i\right)^2 + \lambda(||p_i||^2 + ||q_j||^2)$$

avec (p_i) et (q_j) les lignes de P et Q.

On cherche ainsi le couple de matrices P et Q, de rang fixé, tels que le produit entre P et la transposée de Q minimise l'écart aux données R.

Pour ce faire, vous avez à votre disposition un notebook jupyter *Projet-Optimisation.ipynb* à compléter, ainsi que le fichier de données *ratings.dat*. Ce notebook est en python et fait intervenir la librairie pyspark, vous permettant de manipuler le modèle de programmation Spark. Le fichier de données *ratings.dat* doit se trouver dans un répertoire *data/* par défaut. Il vous est demandé de répondre aux différentes questions du notebook, afin de mettre en place la résolution du problème de recommandation dans un environnement Spark.

Une fois la version notebook du TP réalisé, il vous est demandé d'en extraire les sources python afin de remplir le fichier *Projet-Optimisation.py* afin de pouvoir exécuter votre programme de filtrage collaboratif dans l'infrastructure de calcul, que vous utiliserez dans la seconde partie du projet. Néanmoins, vous pouvez directement développer l'application dans ce fichier, plutôt que passer par le notebook, si vous le souhaitez.

Condition de travail

- Le projet est à faire en binôme. Vous utilisez vos ordinateurs personnels.
- Vous devez notifier par email la composition de votre binôme avant le 15 novembre 2019 (midi).
- Vos réalisations (code source incluant le notebook jupyter, script et un rapport) devront être envoyées par email avant le 31 janvier 2020 (midi).
- Les emails – groupes et rendus - sont à envoyer à destination d'Ehouarn Simon et de Daniel Hagimont
- Le suivi du projet se fera via les discussions Slack créées pour le Certificat.