Coursework Distributed Databases

Strijk Philippe, Wout Boeykens, Hannes Roegiers[1]

Samenvatting
This task discusses 3 Kaggle competitions

Sleutelwoorden
Apache — Spark — Maven — Big Data — Machine Learning — Kaggle — Distributed Databases

Co-promotor
[2] (Van Vreckem Bert)

Contact: [1] wout.boeykens@student.hogent.be; [2] hannes.roegiers@student.hogent.be; [3] philippe.strijk@student.hogent.be

Inhoudsopgave

1. Introduction

At the start of this task, our knowledge about Java Spark and the implementation of Machine Learning methods was quite limited, if not inexistant. Translating code from easy-to-use python scikit-learn to Java Apache Spark libraries proved to be a challenge that would send us down many knowledge-filled rabbit holes. The unique character of this assignment was evidently clear from the get-go. This coursework discusses 3 Kaggle competitions, namely the Quora Insincere Question Classification, <Wout Subject>, <Hannes Subject>.

2. Data Preprocessing

3. Model Selection

3.1 Logistic Regression

3.2 Random Forest Classifier

4. Conclusions