Large Scale Data Analysis – Spring 2020

Exam Hand-in

This is the hand-in for the exam in Large Scale Data Analysis, Spring 2020. You will get your grade based on your written answers to the questions below.

This hand-in exam is composed of four questions, with equal weight. Each question contains several sub-questions. The first three questions concern the assignments. The fourth question concerns concepts and techniques introduced in class.

Question 1 (25%): Database-based ML

A. Define the open world and closed world assumptions. In this assignment, which of these assumptions do you operate under? Explain your answer.
B. Describe the model you have developed in Assignment 2: What was the insight you gained from the data exploration? What are the input and output of your model? How is the input obtained? How did you choose the model? How do you evaluate your model?

Question 2 (25%): Spark-based ML

A. Explain how you use SparkSQL to find the businesses that have been reviewed by more than 5 influencer users in the yelp! data set. Describe your query as well as the different steps required to be able to execute the query.
B. Is location a good predictor for the quality of a restaurant? Explain your answer.

Question 3 (25%): Automated ML Pipelines

A. Explain your design and implementation of the data cleaning processes. Explain how you automate these processes.
B. Describe in detail how the retrained models are evaluated in your solution. You should describe the evaluation method and its implementation.

Question 4 (25%): Topics from the class

A. What is data validation in machine learning pipelines?
B. What is the difference between correlation and causation?
C. What is the role of parallelism in Spark?
D. What is the role of OLAP in the data analytics lifecycle?