CSE 234
Data Systems for Machine Learning

Arun Kumar

Exercise 4
Exercise

Q1) [2pts] Which of these feature engineering system(s) optimize(s) materialization decisions for intermediate data?

A. Columbus

B. Vista

C. KeystoneML

D. All of the above

E. None of the above
Exercise

Q2) [2pts] Which of the following ML tools showed that it is possible to avoid some key-foreign key joins for ML over structured data without significantly hurting ML accuracy?

A. Othello
B. Hamlet
C. Macbeth
D. The Tempest
E. King John
Exercise

Q3) [2pts] Suppose you want to track how well your ad click prediction model performs on users from counties in CA and WA specifically and not just the whole US. Which class of functions can directly help you with this goal in Snorkel?

A. Labeling functions
B. Transformation functions
C. Slicing functions
D. All of the above
E. None of the above
Exercise

Q4) [6pts] Briefly explain two factors that make the “data acquisition” part of the sourcing stage challenging in many real-world ML applications.

Q5) [6pts] Briefly explain two approaches to labeling data that do not require domain experts to write labeling programs.

Q6) [6pts] Briefly explain two pros of building systems to support feature engineering specifically.

Q7) [6pts] Briefly explain one common data cleaning task and how you could automate it on large-scale data.
Q8) [6pts] Briefly explain two practical benefits of using TFDV in real-world ML application workflows.

Q9) [6pts] Briefly explain two ways in MLFlow and TFX are different in their design and functionalities.

Q10) [6pts] Briefly explain one technical debt explained in the paper we studied in class that is related to feature engineering. Also explain how you can pay off that debt.

Q11) [6pts] Briefly one way in which feature stores are similar to data warehouses and one way in which they differ.