DATA 612 Project 5 Report

Date: 2025-07-01

Implemented a PySpark ALS Recommender System inside Docker.

1. Setup & Implementation

- Python 3.9 with PySpark 3.4.1
- Docker container with Java 11 and Spark 3.4.1
- Ratings dataset loaded from CSV

2. Model Evaluation

- Spark ALS Model RMSE: 0.5904
- Baseline (mean rating) RMSE: 1.0770
- Improvement over baseline: 45.18%

3. Conclusion

The PySpark ALS model provides a significant improvement over the baseline. This Dockerized setup allows for easy deployment of the recommender system. While local single-node Spark is suitable for moderate data sizes, moving to a distributed Spark cluster becomes necessary for larger datasets and scalability.

4. Code Snippet from main.py

```
from pyspark.sql import SparkSession
from pyspark.ml.recommendation import ALS

spark = SparkSession.builder.appName("Data612Project5").getOrCreate()
ratings_df = spark.read.csv("/data/data.csv", header=True, inferSchema=True)
ratings_df = ratings_df.select("user_id", "hotel_id", "overall")
als = ALS(maxIter=10, regParam=0.1, userCol="user_id", itemCol="hotel_id", ratingCol="overall")
model = als.fit(ratings_df)
```