Big Data Project: Half-Page Summary

Ebrahim Golriz

This project aimed to answer: **How can we effectively predict user movie preferences based on past ratings?** using **Collaborative Filtering (CF) via the ALS algorithm in PySpark** on the Movielens dataset(ml-latest-small.zip), the approach focused on generating personalized recommendations. Findings aligned largely with expectations; the model achieved reasonable predictive accuracy (RMSE 0.87, Precision@10 0.65) and successfully produced user-specific suggestions. This means, on average, our model's prediction deviates from the actual user rating by less than one star. While state-of-the-art systems on optimized datasets might achieve lower errors, an average error below one star is often considered practically useful in real-world scenarios. A key challenge encountered was the inherent **"cold start" limitation** of CF, where predictions for new users are unreliable, as they haven't yet interacted with our data enough, so we cannot find similar users to them. To counter this problem, a **Content Based Filtering** system was implemented for the movies data, based on the year of release, user assigned tags and genres of the movies, this system let us have something to recommend to users even if they have only liked a few movies. In conclusion, this project successfully demonstrated the implementation of both Collaborative and Content-Based Filtering using PySpark, that shows how these complementary approaches can work hand-in-hand to address different aspects of movie recommendation. While future work focusing on hybrid systems promises further enhancements by combining their respective strengths, it's crucial to remember that the true effectiveness of any recommender system ultimately requires evaluation in real-world scenarios with actual users, as offline metrics like RMSE only capture part of the complex picture.