Big Data Application - Project progress report 1

Group information

ID	Name
21127170	Nguyễn Thế Thiện
21127326	Nguyễn Trần Trung Kiên
21127329	Châu Tấn Kiệt

Problem summary

Problem

Title: Anime recommendation system based on user ratings.

Description:

Analyze anime ratings data using Big Data and Machine Learning tools, in order to recommend animes a user has yet to watch, based on the user's rating history.

Input: User's rating history on watched animes.

Output: Ranking list for recommended animes fitting user's tastes.

Dataset

Anime Dataset 2023 [1] by username Sajid from Kaggle, which is a collection of user and anime ratings on one of the largest anime databases and communities - MyAnimeList (myanimelist.net).

The files contributing to the dataset:

users-score-2023.csv (1.16GB): **The main data** consisting of user ratings on anime titles, provided by 270K users on 16K anime titles, with a total of 24.3M samples, **99.48% sparsity rate** for only the observed users and anime titles. The file consists of 5 columns:

- Rating (0-10): User's rating on anime title.
- Anime ID, Anime Name, User ID, Username: details of user and anime in the ratings.

anime-dataset-2023.csv (15.92MB): Details of around 25K anime titles on MyAnimeList. The file consists of 16 columns:

- **Mal ID, Username**: User's ID and name on the platform.
- **Gender, Birthday, Location**: Personal information of the user, given voluntarily.
- **Joined, Days Watched, Mean Score, Watching, Completed, On Hold, Dropped, Plan to Watch, Total Entries, Rewatched, Episodes Watched**: User's implicit profile of their behaviors on the platform.

user-details-2023.csv (73.93MB): Details of around 730K users registered on MyAnimeList. The file consists of 24 columns:

- **Anime ID, Name, English name, Other name**: Anime ID, original name, translated English name and another (usually abbreviated) name it is known of.
- **Score, Rank, Popularity, Favorites, Scored by, Members**: Anime title's mean score, ranking and side information, on the platform.
- **Genres, Synopsis, Type, Episodes, Aired, Premiered, Status Producers, Licensors, Studios, Source, Duration, Rating, Image URL**: Extra information on the anime title.

Main tasks

1. **Data preprocessing**: Apache Spark preprocesses the datasets retrieved from static files, reformat, clean and prepare the raw data before feeding into the recommendation system.
2. **RS model training**: Train the dataset on ALS model from Apache MLlib.
3. **RS in use**: Input user's rating history to predict a ranking list for recommended animes which user has not watched.
4. **Dashboard**: Visualize predictions and trends with NetworkX-assisted Matplotlib.

Plan progress

1. **Preparation**: 100% done, 60% refining.

- Dataset basic preprocessing using Big Data techniques. [Done]
- Preprocessing refining: EDA logics, feature correlation. [Not done]
- Model preparation: choosing model, examine alternatives, small-scale test. [Done]
- Redis database setup and data ingestion. [Done, Removed]

2. **Model training & usage**: 25% done.

- RS model training. [Done]
- RS model training using Big Data techniques. [Done]
- RS visualization using Big Data techniques. [Not done]
- RS model in use for prediction. [Not done]

3. **Project conclusion**: 0% done.

- Research and experiment documentation and presentation with Canva. [Not done]
- Demonstration. [Not done]

Assignments

Note: The assignments here are summarized from the submitted work by each member of the group. For evidence, check out their corresponding directories.

Progress report no.	Who	Task(s)	Done	Not done	Expected results	Outcome

Progress report no.	Who	Task(s)	Done	Not done	Expected results	Outcome
1	Kiên	Set up real-time data streaming from database	Redis installation and configuration. Upload datasets to Redis in the correct formats.	Set up data streaming to Spark to perform data preprocessing.	Redis successfully streams by capped batches and as new data is ingested.	Redis correctly configured, but not any progress made for the streaming.
1	Thiện	Figure & apply data preprocessing strategies.	Load dataset in the correct format. Perform data transformation on categorical columns (anime dataset). Explanation on the strategies used.	Data cleaning: missing values, noisy data. Data transforming: standardization of numerical columns. Data reduction: dimensionality, support (one hot encoding).	Data reformatting, cleaning, filling missing values, transforming and dimension reduction.	Data reformatting, cleaning, transforming. Not yet for dimension reduction and filling missing values.
1	Kiệt	Choose and set up RS model, explanation on why choosing	Setting up the recommendation system models. Read datasets and train/test split. Explore the reason to use the ALS model.	None.	Choose a RS model and explain why choosing it. Training test on a small-scale unprocessed version of the dataset.	Choose a RS model and explain why choosing it, but have yet to consider alternatives. Not yet tested training.

Progress report no.	Who	Task(s)	Done	Not done	Expected results	Outcome
2	Thiện	Apply data preprocessing strategies (cont.). Set up basic UI for prediction program. Deploy RS into prediction program.	Preprocess on datasets. Deploy RS into prediction program: embed user inputs.	Set up basic UI for prediction program. Deploy RS into prediction program: deploy RS.	Data reformatting, cleaning, filling missing values, transforming and dimension reduction. Basic UI for prediction program.	Data reformatting, cleaning, filling missing values, transforming but without dimension reduction. Basic UI for only user input, not yet prediction since we did not have the RS model.
2	Kiên	Redis + Spark: set up data streaming to data preprocessing.	Set up data streaming from Redis to Spark.	Join data streaming to preprocessing.	Redis successfully streams by capped batches and as new data is ingested. Spark processes by batch as new data comes.	Redis successfully streams by capped batches and as new data is ingested. Not yet connected to processing phase.
2	Kiệt	Spark + ML: test model training with preprocessed datasets.	Setting up the UBCF/IBCF Recommendation System Linked the preprocessed dataset with Spark	Integrate with Redis Streaming Running and validating the model	Sucessfully test model training with preprocessed unified dataset to output a real-time model.	Sucessfully test model training with preprocessed unified dataset to output a model file. Unable to convert to a real-time RS model due to usage of ALS.

Progress report no.	Who	Task(s)	Done	Not done	Expected results	Outcome
3	Kiên	Report + Slide for project documentation.	??	??	Report and Slide documentation of the project progess and changes so far.	??
3	Kiệt	Model training on full dataset. Model visualization on Matplotlib	Model training on full dataset. Model visualization on Matplotlib	Optimizing the runtime of Spark for visualization Many visualization tools rely on pandas. DataFrame proved to be challenging Still looking for a solution	Model training on full dataset with the current output of other phases to a model file. Model visualization on Matplotlib	Completed ALS model on training with the full dataset, exported the model for visualization purpose Not yet completed with the visualization.
3	Thiện	Data preprocessing refining with EDA & feature correlation analysis.	Data preprocessing refining on user dataset.	Data preprocessing refining on anime & unified datasets.	Data preprocessing refining with full explanations.	All done on only the user dataset, but not the others.

References

[1] Sajid Uddin (2023). Anime Dataset 2023. *Kaggle: Your Machine Learning and Data Science Community*. https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset?resource=download

[2] Salvatore Sanfilippo (2009). Redis 7.4.2 (2025). https://redis.io

[3] Matei Zaharia (2014). Apache Spark 3.5.4 (2024). https://spark.apache.org/

[5] John D. Hunter (2003). Matplotlib 3.10.0 (2024). https://matplotlib.org/

[6] Aric Hagberg, Pieter Swart, Dan Schult (2005). NetworkX 3.4.2 (2024). https://networkx.org/

[7] Melanie Perkins, Cliff Obrecht, Cameron Adams (2013). Canva. https://canva.com