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: Real-time anime recommendation system based on user ratings.

Description:

Analyze real-time 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.

anime-dataset-2023.csv (15.92MB): Details of around 25K anime titles on MyAnimeList.

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

Data storing and processing tool(s)

Redis [2]: NoSQL database supports real-time data streaming.

Apache Spark [3]: real-time data processing.

Recommender system (RS)

Apache Spark MLlib RS [3] as it comes with Apache Spark mentioned above, and is a well-known standard Machine Learning library. It contains Alternating Least Squares (ALS) matrix factorization to learn latent factors.

Data visualization

Type(s)

Network graph in order to present the interactions between the observed users and items.

Tool(s)

NumPy Matplotlib [5]: supports real-time data visualization for Python.

NetworkX [6]: a graph generating Python library.

Main tasks

1. **Data ingestion**: Set up Redis database with imported data from the dataset files, and set up data streaming connection.
2. **Data streaming & preprocessing**: Apache Spark Streaming simulates real-time data from Redis database, then clean and prepare the raw data before feeding into the recommendation system.
3. **Real-time RS model training**: Pre-built model from MLlib is trained by feeding real-time data, from Spark Streaming.
4. **RS in use**: Input user's rating history to predict a ranking list for recommended animes which user has not watched.
5. **Real-time dashboard**: Visualize analyzed data and predictions with NetworkX-assisted Matplotlib.

Plan progress

1. **Preparation**: 60% done.

- Set up Redis database and data streaming to Spark.
- Figure data preprocessing strategies, perform data preprocessing on dataset using Spark.
- Set up RS model from MLlib, learn its required input and output forms for training and testing.
- Figure out how to save RS model into a file for further training.

2. **Tool testing and systematic setups**: 0% done.

- Perform real-time data processing using Redis and Spark, with data visualization using NetworkX and Matplotlib.
- Set up a basic user interface to apply the use of RS.
- Test RS model training on small scale with multiple batches, with saving and loading RS model.

3. **Main events**: 0% done.

- Perform real-time RS model training on dataset.
- Research and experiment documentation.

- Application of RS model into the problem.

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

- Research and experiment documentation and presentation with Canva [7].
- Graphical demonstration.

Assignments

Note: Date used here is in form (YY/)MM/DD, time used here is in 24-hour format: HH:MM.

All works must be draft-documented in text files (md, txt, pdf, docs) upon finished working.

Sprint no.	Who	Job(s)	Tool(s)	Start	Due	Done	Not done
1-2	Kiên	Set up real-time data streaming from database	Redis, Spark	03/07 06:00	03/16 21:00	Set up Redis and configurations; Push dataset on live Redis DB	Streaming data to Spark
1-2	Thiện	Figure data preprocessing strategies	Spark	03/07 06:00	03/16 21:00	Data cleaning for categorical columns	Data cleaning for numerical columns; Data integration; Data transformation
1-2	Kiệt	Choose and set up RS model, learn its inputs from Spark and outputs	Spark, MLlib	03/07 06:00	03/16 21:00	Choose ALS model for batch processing; MLlib support for saving and loading models	(empty)

Self-assessment

Struggles

1. **Underestimated the difficulty and time of tasks**: Data preprocessing and data storing took more time than it should, especially when with little experience.
2. **Unexpected happenings from the outside**: Juggling multiple tasks and being overwhelmed.
3. **Post-vacation blues**: Some members have yet to recover from Lunar New Year vacation to keep high work efficiency, though relieved by task planning and active communication.

Problems:

1. **Problem re-definition**: Arised worries during data preprocessing about how to deal with empty values, such as what the problem really is and the steps to solving the problem. "The more I read, the less I know".

2. **Dealing with high dimensionality**: One-hot encoding on unordered labels gives too many columns, either dimension reduction (UBCF-IBCF) or multiple minimum support metrics could work.

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