2022

MIT 805 – ASSIGNMENT PART 2

U16098944

PANASHE MABWE

Contents

Brief Overview of Dataset 2

File Structure 2

Column Descriptions 2

MapReduce Algorithm 3

Intro 3

Implementation 4

Results 5

References 6

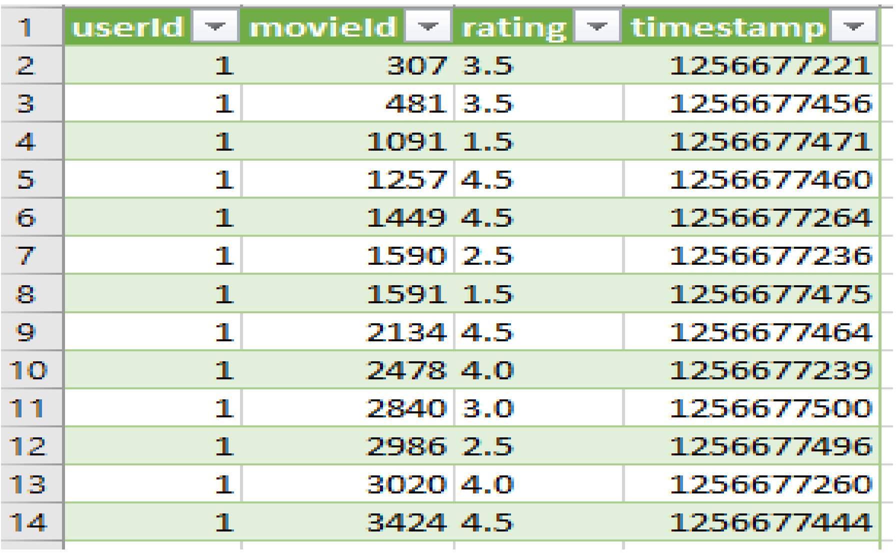
Table of Figures

Figure 1 - Map Reduce Algorithm 4

Figure 2- Power BI Visualisation 5

Brief Overview of Dataset

The dataset consists of movie related data provided by a movie recommendation service named MovieLens. The dataset consists of attributes in relation to a 5-star rating and tagging activity. The dataset was populated by approximately 162 541 users between the years of 1995 and 2019. Each user rated at least 20 movies and consequently the dataset has approximately 25 million rows in CSV format. Snippet of the dataset is shown below:



1	userId	movieId	rating	timestamp
2	1	307	3.5	1256677221
3	1	481	3.5	1256677456
4	1	1091	1.5	1256677471
5	1	1257	4.5	1256677460
6	1	1449	4.5	1256677264
7	1	1590	2.5	1256677236
8	1	1591	1.5	1256677475
9	1	2134	4.5	1256677464
10	1	2478	4.0	1256677239
11	1	2840	3.0	1256677500
12	1	2986	2.5	1256677496
13	1	3020	4.0	1256677260
14	1	3424	4.5	1256677444

File Structure

The dataset is in the form of a CSV (comma-separated values) file which is a text file in which the information is separated by commas. The CSV consists of 4 columns which are **userId, movieId**, **rating**, **timestamp.**

Column Descriptions

1. userId – unique identifier for each user
2. movieId – unique identifier for each movie watched by a user
3. rating – ratings are based on a 5-star scale. The rating from 0 to 5 which was provided by a user in terms of the users' opinion of the movie. The lowest score being 0 – indicating that the user did not enjoy the movie at all. The score 5 being of the most satisfactory.
4. timestamp – represent seconds since midnight of 1 January 1970 (UTC)

MapReduce Algorithm

Intro

Hadoop is a framework mainly utilized for big data that requires a substantial amount of processing and computing. Hadoop makes use of parallel processing and distributed storage to store and manage big data (Zhasa, 2022). Furthermore, Hadoop consists of 3 components including Hadoop Distributed File System (HDFS), Hadoop MapReduce and Hadoop Yet Another Resource Negotiator (YARN).

In this project, I used MapReduce to implement movie recommendations. The recommendation system is solely based on Item-Based Collaborative Filtering which in essence works by recommendation a particular movie based on the similarity of the movies watched prior (Qutbuddin, 2020). Furthermore, the similarity is calculated using the Cosine similarity based on the rating given by the customer per movie watched.

In simple terms, let's say User 1 watched movie A and movie B and rated the movies highly. User 1 really enjoyed the movies. User 2 watches both movies and similarly enjoys them as well. However, User 3 watches movie A only. The system will recommend movie A since most users who watched movie A and enjoyed it – watched movie B as well. Ultimately, in this project the map reduce algorithm finds the pair of movies that were watched by the same person, measures the similarity of the rating across all users who watched the pair and the outputs the movie and recommendation with similarity scores.

Implementation

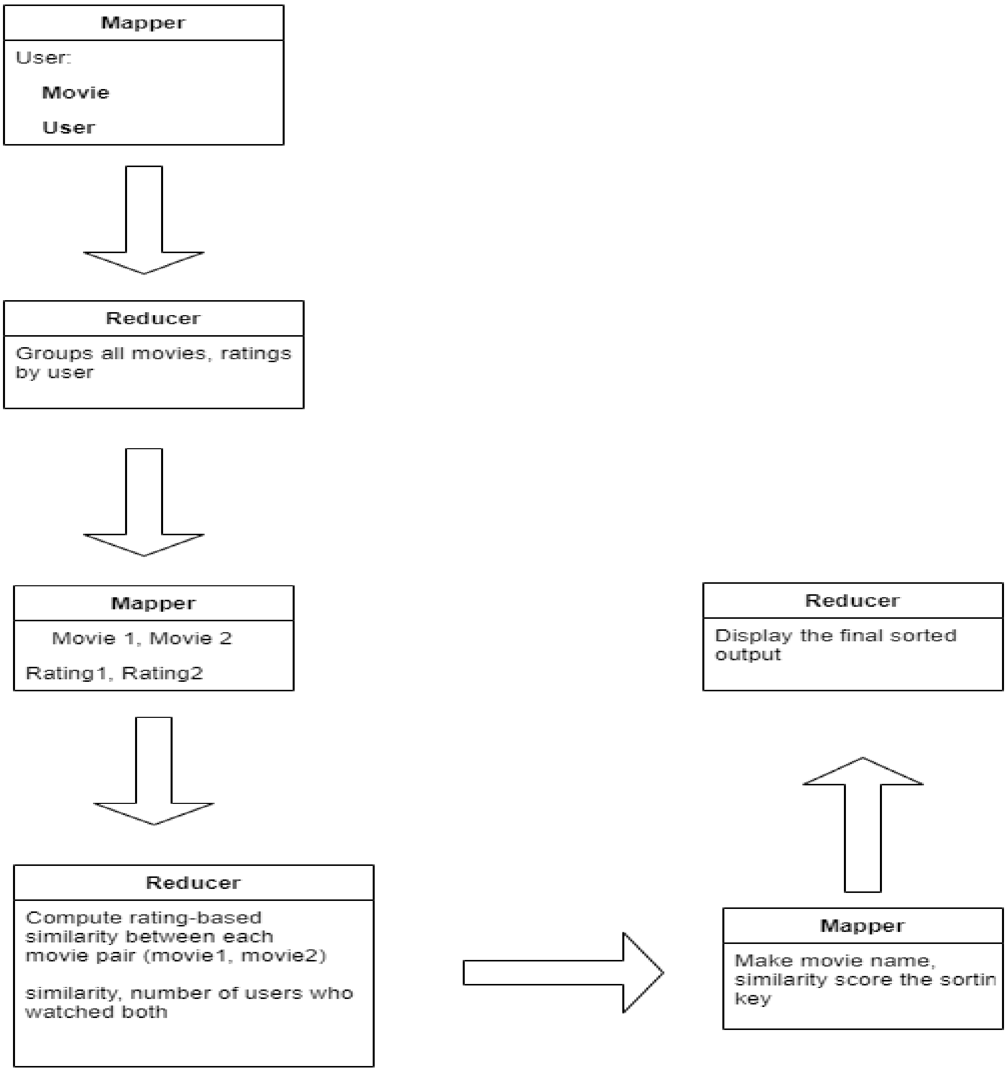

Mapper
User:
Movie
User
Reducer
Groups all movies, ratings by user
Mapper
Movie 1, Movie 2
Rating1, Rating2
Reducer
Compute rating-based similarity between each movie pair (movie1, movie2)
similarity, number of users who watched both
Mapper
Make movie name, similarity score the sortin key
Reducer
Display the final sorted output

Figure 1 - Map Reduce Algorithm

In terms of implementation – made use of a few mappers and reducers which are depicted above. In the initial stage, the mapper extracts the movie and rating from the source data as these are the only fields needed from the dataset. The reducer groups all the movies and rating by the user.

In the 2nd stage of the algorithm, the mapper outputs all permutations of every single pair of movies that the user watched. This stage is where most of the processing takes place. The list contains key value pairs where the key is the movie pairs then the value is the rating pairs associated with each movie. The reducer takes that movie pairs and calculates the similarities of the movies based on the ratings.

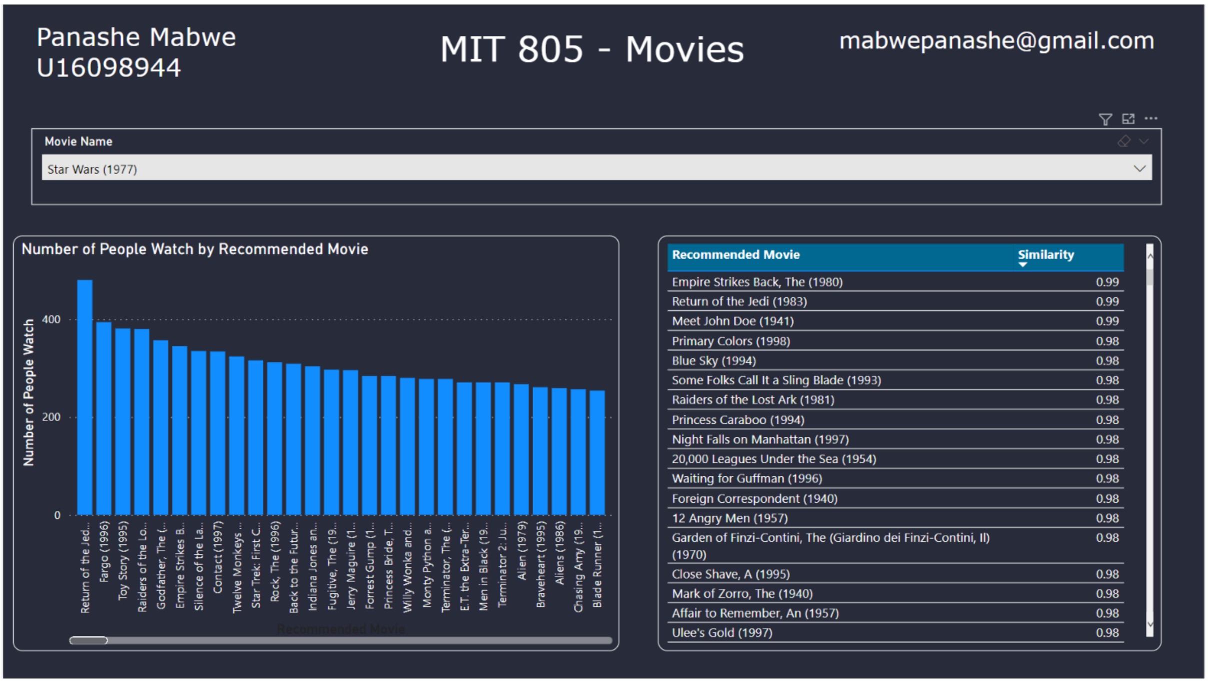
In the final stage, the mapper sorts the movie by name and for every movie, there is another movie associated to that movie sorted by the similarity score. The reducer is responsible for the final output.

Results

I decided to utilize all the attributes such as userID, movieID and Rating as the attributes allowed to answer questions such as:

- What is the similarity between movie A and movie B?
- What movie recommendation should I watch if I for example watched Star Wars?
- How many people viewed that movie pairing?

I made use of Microsoft Power BI to visualize my results as depicted below:



Panashe Mabwe
U16098944

MIT 805 - Movies

mabwepanashe@gmail.com

Movie Name: Star Wars (1977)

Recommended Movie	Similarity
Empire Strikes Back, The (1980)	0.99
Return of the Jedi (1983)	0.99
Meet John Doe (1941)	0.99
Primary Colors (1998)	0.98
Blue Sky (1994)	0.98
Some Folks Call It a Sling Blade (1993)	0.98
Raiders of the Lost Ark (1981)	0.98
Princess Caraboo (1994)	0.98
Night Falls on Manhattan (1997)	0.98
20,000 Leagues Under the Sea (1954)	0.98
Waiting for Guffman (1996)	0.98
Foreign Correspondent (1940)	0.98
12 Angry Men (1957)	0.98
Garden of Finzi-Contini, The (Giardino dei Finzi-Contini, Il) (1970)	0.98
Close Shave, A (1995)	0.98
Mark of Zorro, The (1940)	0.98
Affair to Remember, An (1957)	0.98
Ulee's Gold (1997)	0.98

Figure 2- Power BI Visualisation

I made use of 3 visualizations which are a filter, bar chart and table to visualize the results. A user can search for any movie via the filter with the title Movie Name.

In this example, the user selected the movie Star Wars and below the filter on the left, we have a bar chart depicting the movies that other users watched combined with Star Wars. Most people watched Return of the Jedi which makes sense as they are both Star War movies. Furthermore, the bar chart shows the number of people who watched the movie pairing which in this case is approximately 400 people. People also watched Fargo and Toy Story according to the bar chart.

On the right, we have a table depicting the similarities calculated per movie pairing. Empire Strikes Back which is another Star Wars movie has a similarity of 99% indicating a high correlation between Star Wars, Return of the Jedi and Empire Strikes Back.

References

Chancellor, S., Konstan, J., Terveen, L., & Yarosh, L. (2019). *MovieLens Datasets*. Retrieved September 30, 2022, from https://grouplens.org/datasets/movielens/.

Qutbuddin, M. (2020, March 7). Comprehensive guide on Item Based Recommendation Systems. Retrieved October 30, 2022, from https://towardsdatascience.com/comprehensive-guide-on-item-based-recommendation-systems-d67e40e2b75d

Zhasa, M. (2022) What is Hadoop? Components of Hadoop and how does it work [updated], Simplilearn.com. Simplilearn. Available at: https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop (Accessed: October 30, 2022).