CHAPTER 3
RESEARCH METHODOLOGY

The methodology of this project is as follows

1. Identification of a problem and Literature Study

The very first stages of research are to seek for a case studies to be solved by a certain or specific algorithm with the assumption that the problem will be solved completely or at the very least more effectively. The literature used in this project are mostly related to recommender system such as the one utilized in Netflix and Spotify. Most journal collected discusses about issues relating to recommender algorithm such as collaborative filtering, or some ranking algorithm and many data science famous classification method such as Pearson correlation, distance similarity measures, and many familiar algorithms such as KNN or Decision Tree.

The recommender technique of each algorithm or method will likely produce slight difference if not similar metrics based on the resulting formula of each method performed by the code and the complexity of the formula. The real practical definition of recommender system basically a recommender system is the program in which the algorithm or method is to compute similarity between two entities to give a targeted output. This engine usually trying to find the level of similarity between two entities. Then, the computed result can be used to calculate various results of any kinds. This recommendation technique is commonly found in YouTube queue based on our favorite videos, or Spotify's Radio which plays random music based on the previous or our favorite musical taste like 'Heavy Drumbeat', 'Rock', etc.

2. Data Gathering

The data that is collected for the purpose of this project is the movie lens dataset obtained from grouplens.org. It consists of multiple csv files such as user data, movie data and

ratings. The data will be analyzed in python using .ipynb or jupyter notebook for easier reading of graph plot or similar approach and per cell debugging. Debugging one cell at a time is easier for this kind of project as there are a lot of plot and data frame tabulation and using only .py file for compilation will take a lot of time, even though the expected output of this project may be compiled to a single .py file.

3. Method & Completion

The Program is utilized to manage data pre-processing and little to medium analysis of the data provided by group lens(the CSV Files), then some comparison and grouping of userId column and movieId is executed. After the pre processing is completed the data can be analyzed by correlation, Manhattan distance, euclidean distance or many more method such as Pearson correlation whcih might be handy for comparison as well if needed. The final result will be a list of recommended movies to a particular user and the most similar user based on the algorithm specified in the code.

4. Software & Tools

Software that utilized in this project are all open-source software or tools that can be modified according to the manual of the respective owners. Python 3 is the programming language utilized for this research and with the help of Jupyter Notebook for easier per cell debugging. All the library needed for the completion of this work will be demonstrated and explained

5. Expected implementation and metric between two algorithms

The most important step of every recommender system is scoring. Different kinds of big tech company working on movie subscription service will offer the next recommender movie using percentage point to see the accuracy of their algorithm.

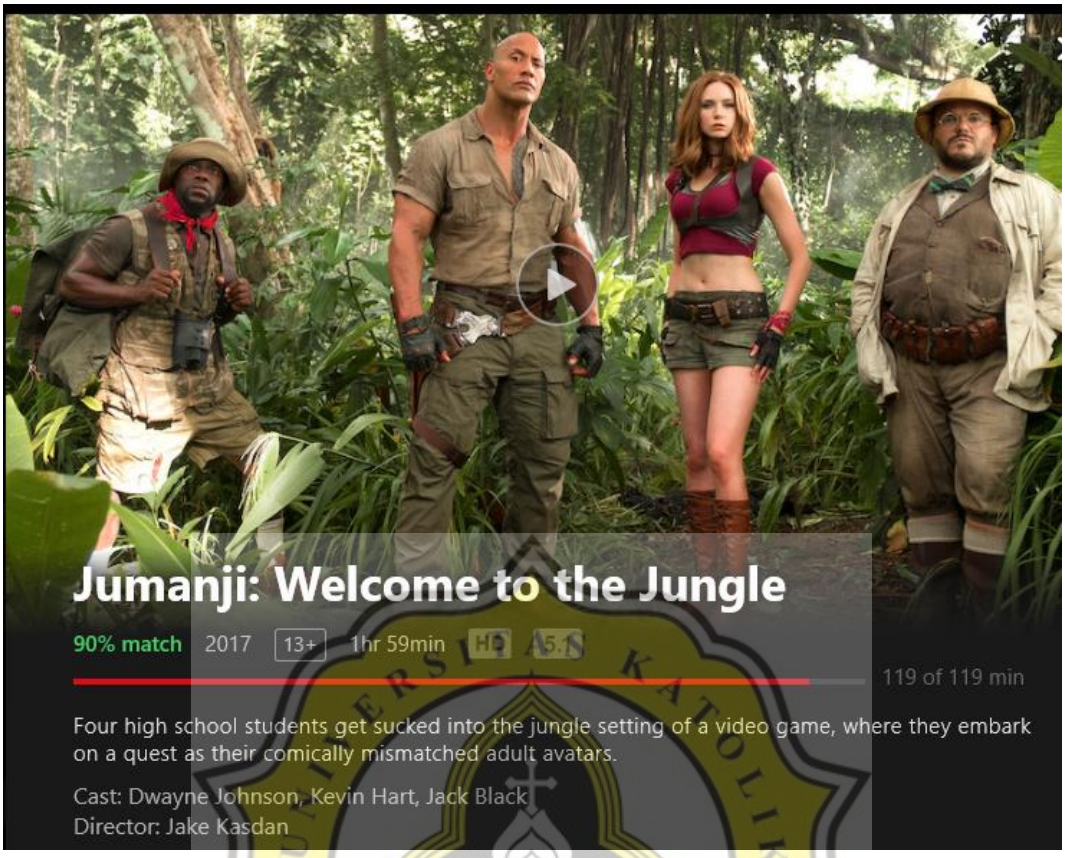

Figure 3.1 Netflix Recommender

The range of percentage point is considered as activation function. It ranges from 0 to 1 and usually called sigmoid function. In this project it is called a heuristic function to map (0,1) so the result will only show a score of between 0 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Figure 3.2 Sigmoid Activation Function

This is an example of sigmoid activation function. This project utilizes a heuristics approach of this function by replacing “e to the power of negative x” to my own implementation of the algorithm that is $\mathbf{Return}\ \frac{\mathbf{1}}{(\mathbf{1}+\boldsymbol{sum\_of\_squares})}$