COMP4434 Big Data Analytics

Individual Project

For internal use only, please do not distribute!

Individual Project

- Work independently on a real-world problem
- Deliverables and Grading (40%)
  - Mid report (5%)
  - Final report (8%)
  - Presentation (7%)
  - Code (20%)
- Submit deadline:
  - Mid report: 23:59 15 March 2021
  - Final report: 23:59 5 May 2021

Problem

- The project dataset contains information on user preference data on different teleplays.

- It is composed of two files
  - Teleplay.csv: It includes different features of teleplays.
  - rating.csv: It includes the rating of each user to every teleplay she/he has watched.

- You are required to implement using Python.
- Third-party packages are allowed.

Stage 1.1: Data Preprocessing

- In any big data analysis projects, data pre-processing is the first and an important step.
- Data preprocessing tasks:
  a) Detecting missing value: The rating dataset uses "-1" to represent missing ratings. You need to replace them with a null value to avoid the average to be distorted.
  b) Encoding categorical data: The teleplay dataset has a "type" column which includes "TV", "OVA", and "Movie". You need to encode them to 0, 1, and 2 because most algorithms require numerical inputs.
  c) Process Outliers. There are plenty of outliers inside the datasets. Deal with them according to the situation.

Stage 1.2: MapReduce

- Because there are huge amounts of data in the provided dataset, you are expected to design a MapReduce algorithm for data pre-processing in parallel.
- MapReduce task:
  a) The average rating of each teleplay can be computed by MapReduce based on the rating of each user to each teleplay in rating.csv.
  b) The average rating computed from rating.csv can be attached as a new feature to the corresponding teleplay in Teleplay.csv.

Stage 1.3: Linear Regression

- Task:
  a) Using python to design a linear regression model. Note that you are required to implement linear regression algorithms by yourself.
  b) Using the training dataset to train your model, which can predict the rating of teleplay.
  c) A grading dataset with hidden average rating column has been provided and you need to give predicted values through your trained model.
- Evaluation standard:

  Your predicted values will be compared with the true value and the grading is based on prediction accuracy.
- Only using the third-party related packages will lose the points for the coding.

Stage 2.1: Neural Network

- Task:
  a) Using python to design a neural network model, such as Multi-Layer Perception (MLP).
  b) Using the training dataset to train your model, which can predict the average rating of teleplay.
  c) A grading dataset with hidden average rating column has been provided and you need to give predicted values through your trained model.
- Evaluation standard:

  Your predicted values will be compared with the true value and the grading is based on prediction accuracy.

Stage 2.2: Contend-based Recommendation System

- Task: Build a teleplay recommendation system based on user's content preference.
  a) Calculate the similarity between teleplays.
  b) From existing dataset, find user 53698's favorite teleplay, and based on that, recommend similar teleplays for user 53698.

Stage 2.3: Collaborative Filtering-based Recommendation System

- Task: Build a teleplay recommendation system based on the preferences of other users, automatically recommend teleplay to users that they haven't watched yet.
  a) Design and train the prediction model via SGD from known ratings.
  b) Predict the ratings of the teleplays that are not rated by the user (user_id:53698).

Report & Presentation

- Mid Report (Stage 1.1, 1.2, 1.3)
  - Details of Data analytics
  - Details of model design and implementation
  - Primary Result
  - Summary of discoveries and future work
- Final Report (Stage 2.1, 2.2, 2.3)
  - Problem definition
  - Model design and analysis
  - Solutions and implementation details
  - Performance evaluation and discussion
  - Summary of discoveries and future work
- Presentation
  - Presentation slides in 15-20 pages
  - Record a presentation video in 10-15 minutes