Data Challenge - Detailed Report

CS F415 - Data Mining

Abhay Koushik --- 2015A7PS0056P

Gautham V --- 2014B4A70637P

Problem 2:

1. **Preprocessing** : The given data was not nice, so it was processed. Please refer `prep2.py` .
    - Split student ID to obtain category
    - Split the DateTime value to obtain Date and Time separately
    - Split Time to get hours alone (because 16:00:00 to 16:59:59)
    - Changed hours (00->24, 01->25, 02->26) to get ordinal hour values *Note: this takes a long time, rowwise operation*
    - Mapped ItemID to name using `monthlyPriceList.csv` so that viewing combos would be easier.
    - The items were split as per their (nonzero) ratings: items below a rating of 2 and above a rating of 4 were chosen to be put in combos.
    - Average rating for each item over each month was found.
2. **Data Mining**: Refer `arule2.py` .
    - `arule2.py` :
        - Used the Association Rule Inducer method from the Orange package.
        - Using base parameters in the Association Rule Inducer method `(support = 3.5e-3)`
        - Obtained association rules and frequent itemsets.
        - Used these rules (along with their support and count) as the various combos.
    - **Note:** the output from `arule2.py` was piped to `rules2.csv` via command line.
3. **Post-Processing**: Refer `post2.py` .
    - `rules2.csv` is obtained from the AR Mining, it is processed to get `preArules.csv` .
    - Each rule in this file is then split into its constituent items, and saved in `PriceProcessing.csv`

- The proper formats was selected to make `ComboMeals.csv`.
- Finally, the proper combos were selected to make `Meals2.csv`.