Assignment 3 - Hakuna Matata

2019101098
2019101102

=> To extract the association rules from the data set , we used apriori algorithm.First we preprocessed the data according to the given instructions. We have considered the movies rated above 2. We considered the user who has rated above 10 movies and divided the data set into 80% training and 20% testing. 20% of movies were removed from each user and formed the testing set.

Task-1:

=> We used apriori algorithm for getting the association rules. First we shuffled the data randomly as the last rows of the data set contains large values and the data set is not training properly.We applied an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets. For k=1 we created a table containing the support count of each item present in the dataset and then we compare candidate set item's support count with minimum support count. We will consider the item sets which are greater than support count. For k=2 we will generate candidate set C2 using L1 (this is called join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common.for this we will check if all subsets of an itemset are frequent or not and if not frequent remove that itemset.Now we will find support count of these itemsets by searching in dataset and those which are greater than minimum support. We will continue this algorithm until we find frequent item sets.We will stop if we don't find any frequent item sets.From these we generated association rules are generated using minimum confidence threshold.From the frequent itemsets we generate subsets and calculate the confidence and consider those itemsets which have greater confidence than minimum confidence threshold.

Task-2:
=>We considered the top 100 rules sorted based on support and 100 rules sorted based on confidence. And we selected the rules which are common in both lists.

Task-3:
=>We calculated average precision and recall for the association rules by calculating the Hitset(Intersection of R with test set) and we calculated the precision as the ratio of the Hitset and test set. Recall is calculated as the ratio of Hitset and recommendation set.

Task-4:
=> We Considered 10 users and for every user we plotted precision and recall.