Course ID: CSE-506
Total Points 100

# Assignment 2

Deadline: Nov 23, 2021

**Instructions**

1. The assignment is to be attempted in pairs.
2. Programming Language: Python
3. For Plagiarism, institute policy will be followed
4. You need to submit the readme.pdf, Code files and Model files.
5. Report, models and code in .py format should be submitted in the classroom in a zip folder with the name A2_RollNumber1_RollNumber2.zip.
6. You **can use any library** for pre-processing, training, doing experiments and post-processing in all questions.
7. One member should submit on google classroom while other member can mark turn in without the attachment.
8. In case of doubts, please comment on the classroom.

Dataset: Link; Evaluation metric: Precision@k
Reference Material: Paper

**Q1: (10 Points)** Perform exploratory data analysis (EDA) over the shared dataset. Analysis may include 1) finding frequently occurring values in categorical features, 2) Counting of Nan values per column, 3) Plotting correlations between features, etc.. Report at least three insights about the dataset.

**Q2: (60 points)** Using the association rule mining techniques, build a recommendation system for the dataset provided to you. Your system should be able to recommend movies given a customer profile. i.e. what set of movies a customer has watched in the past? The set can contain any non-zero number of movies. Let us say a customer has seen "avengers, iron man", the system can recommend the following movies "captain america, hulk, thor, doctor strange". The final scores will be computed using the metric "Precision@k". For a given customer profile, you will recommend four best movies, i.e. k=4.

**Q3: (30 points)** Visualize the maximal frequent pattern set.

**PS:** Q2 points will be given based on the performance on the test set. Refer to Table 1. Let $x$ denote the maximum "Precision@K" value reported by some student. The bucket will be formed as per Table 1, and each bucket will get the mentioned marks. (For ex. >90% of x will get 100% marks, >80% of x will get 90% and likewise, keep the inclusive and exclusive range in mind.).

| Precision | Points % |
|---|---|
| [.9x-x] | 100% |
| [.8x-.9x) | 90% |
| [.7x-.8x) | 80% |
| [.6x-.7x) | 70% |
| Remaining | 30% |

**Table 1:** Q2 performance points.

| Input | Inference |
|---|---|
| Titanic | poseidon<br>speed<br>apollo 13<br>... 1 more |
| speed, the rock | gladiator<br>taxi driver<br>... 2 more |

**Table 2:** Your output.csv should look like this.

Course ID: CSE-506
Total Points 100

# Assignment 2

Deadline: Nov 23, 2021

**Deliverables**

1. You need to make a inference code, which will take a input and save the predictions as output.csv (Example in Table 2).
2. Detailed explanation of assumptions made for solving the mentioned problems.
3. Report all visualizations and assumptions in the readme.pdf file.
4. Make a section "Learning's" in the readme.pdf and describe your learning's from this assignment.
5. Please provide references to all the sources used in completing this assignment.

PS: There will be an assignment deadline overlap with Assignment 3.