IS 603 GROUP PROJECT PROPOSAL

Team Members:

1. Asmita Deshpande (OF68333)
2. Parthiv Gandhi (FV59132)
3. Mostafa Cham (QT51576)

Background:

A semester-long group project to implement data science techniques with methods and algorithms such as clustering, data reduction, etc. The data set is taken from Kaggle.com. It consists of 1.45 million reviews of about 622 courses and is about 35 MB dataset file.

Dataset: https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera

Scope:

Coursera is the largest massive open online course (MOOC) provider in the world and hosts many courses from top universities in different subjects. We plan on exploring this dataset as it gives us information on how a Coursera course is reviewed, based on the course material, professor, the quality of course material, and what a peer learns at the end of the course. In the end, we will be able to find which Coursera course customer/student left a review and rating.

As an initial observation made by us, we can segregate the reviews into different clusters as some reviews are on the material, some as an overview and some values are not so useful. There are excellent examples " Material covered well, and easy to use application." or "Solid presentation all the way through. I really appreciated the intermittent questions that popped". Such reviews and data will help us more. We will be able to co-relate it with other multiple customers and the reviews for each course taken. We plan to find these relations with the help of different algorithms and methods such as classification, regression tasks, and clustering.

Dataset Brief:

This dataset contains mainly 5 columns and 622-course data. The detailed description:

1. review : Course review.
2. reviewer : The name of the reviewer who wrote the review.

3. date_review : It has the date when the review was posted.
4. course_rating : It has the rating score given by the reviewer to the course.
5. course_id : Contains the course title or id.

Preprocessing of Data:

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. This step is often required if the data is not presented in a clean format, if it is difficult to work with, or if it contains various data types.

Data Mining Techniques:

What methods plan to use - CLUSTERING
The objective is to apply and compare three different clustering algorithms to our dataset. The algorithms mentioned below are the ones we plan to use in this project. They have chosen as most commonly used algorithms in the field of clustering:

1) K-means Clustering
2) Dscan Clustering
3) Agglomerative clustering

Base on the fact that clustering is an unsupervised technique and there is no specific and pre-defined result, each algorithm should be evaluated individually. For comparison and evaluation of aforementioned algorithms, we are going to score each algorithm with two different metrics. The evaluation metrics are listed below:

1. Silhouette Coefficient
2. RMSE (Root Mean Squared Error)