IS 603 MIDTERM PROGRESS REPORT

Team Members:

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

Dataset:

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

Problem:

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.

Implementation:

We started working on the data which we obtained from Kaggle. The data was itself cleansed but it needed further cleaning and preprocessing. The unzipped file was very large and had a lot of noise so we had to process it further before we run any clustering algorithm on the data.

Data Pre-processing:

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. As an initial step we omitted the rows which have null value because it makes the dataset noisy. We have removed unnecessary cells and columns from the csv file. For example the column “reviewer’s name” doesn’t have any impact on the data analysis pipeline.

We have also removed numbers and special characters for clustering because they do not carry any weightage in clustering. We have reduced the massive dataset to an acceptable size. The data was not cut at random, instead we have written a python program and also tried executing it in the Jupyter Notebook. We provided a list of stop words, which are the words used commonly in text and do not have a huge impact on the meaning of the text. Accordingly, they do not have a positive impact in the clustering process and we could remove them from the text.

Processing Textual Data:

1. **Tokenization:**

   Tokenizing is an important part for processing natural language. Tokenization basically means reducing a sentence to a small unit, i.e. A token to easily assign meaning to them. We use tokenization as a precursor to stemming.

2. **Stemming:**

   Stemming is used to reduce a word to its root form. For example: The words - Computer, Computing are reduced to the root word - compute. We employ this technique to maintain uniformity across our dataset.

Algorithms Used:

1. **Clustering:**

For clustering we have used two algorithms, K-means and Feature agglomeration. Also, we have used two metrics to evaluate and compare the algorithms, which are Davies-Bouldin Index and Silhouette Coefficient. The k-means algorithm works based on minimizing the in-cluster sum of squares. Feature agglomeration algorithm is a hierarchical clustering method which works by merging features using bottom top approach. Davies-Bouldin Index metric considers the distance between each data point and centroid of its assigned cluster and centroid of nearest cluster, where 0<= score and lower score is better. Silhouette coefficient metric takes account of mean distance of each data point with intra-cluster samples and nearest cluster samples, where -1<= score <=1 and 1 is the best score, -1 is the worst score and 0 indicates overlapping clusters.

2. **TF - IDF Vectorizing Method:**

We used the TF-IDF vectorizer method, which considers word frequency in all texts. The purpose of using this method is to deal with the most frequent words by penalizing them.

Challenges Faced:

The first challenge we faced was preprocessing of the data since the dataset CSV file had a lot of excess and unwanted columns. It consists of 1.45 million reviews and about 622 courses which is about 272 MB of data. The processing is very time-consuming, even using the GPUs in Google Colaboratory takes

The second obstacle we faced was we tried to run our CSV files in weka which was not successful. We further cleaned the data and again ran it in weka, but it didn't work. Therefore we continue to work in python and jupyter notebook for pre-processing and clustering.