DATS 6103 – Data Mining

Project Proposal

Group-F

Data Description:

- **Source:** https://www.kaggle.com/lava18/google-play-store-apps
- **Description:** This is a data set with about 9000 data points that show the statistics of various applications installed and used through the google play store.There are 2 files in the data set which show the statistics of each application and a set of sentiment evaluated user reviews (per application).

Problem Statement Proposals:

1. Determining and suggesting similar apps using KNN prediction or k-means clustering.
    a. *Sub Problem :* Finding 5 most similar unpaid applications to 50 top most paid applications in each category / Genre
    b. Recognizing mis-categorization of applications using above algorithms. If miscategorized, correcting the label or adding multiple genres to the database based on results.
2. Topic modeling for performing sentiment analysis of the user reviews in each application and determining the translated & common review.
    a. *Sub Problem :* Using NLP to analyse the user reviews and the language used by the users , detemine which sentiment it represents (Neutral,Positive,Negative)
3. Creating a decision tree to determine
    a. the Probable user rating (prediction for ranges [0-1,1.1-2, 2.1-3,….] etc)
    (OR)
    b. Will the user download it or not ?

    This would require entropy calculations to determine the features to be used for creating the decision tree and help calculate labels based on the attributes. The resultant label can be User Rating, Price,Download (Y/N) etc.