The objective is to determine if **customer age** and **customer feedback** can determine the category of a customer who buys products from a store. The category of a customer in the training data is “Promoter” and “Detractor” . Customer in the Promoter category always recommend the product to others, whereas the Detractor category always complains about the product and donot recommend the product.

Two datasets are provided, namely a training dataset(cust\_data\_train.csv) and a test dataset(cust\_data\_test.csv). You have to train a model on the training data and predict the probability on the test data that a particular customer is a promoter for that product. You also have to predict the customer category for the customer providing the feedback for the product

**Tasks to do**

1. Do an exploratory analysis to know whether age of the customer has any relation to the customer category in the training data( eg: create plots of distribution of age by customer category , check correlation between age and customer category)
2. Create a word cloud distribution of the customer feedback on the training data
3. Create tf-idf bigram token vectors from the customer feedback data(Do all the required pre-processing of the text data)
4. Use customer age and the tf-idf vectors as features
5. Train and cross-validate logistic regression, random forest, xgboost, neural network algorithm on the training data and use AUC metric to determine which algorithm gives better performance
6. Take the best algorithm and train it on training dataset and make predictions on the test dataset shared with you
7. Optional (Create a 100 dimensional doc2vec embedding for the customer feedback)

N.B. *Please use python for all the tasks*