Bank Review Analysis

Objective To analyse the customer reviews and predict the customer satisfication through the reviews.

Business Problem

- Data Processing
- Most frequent Pos and Neg words
- Classification of the reviews into Pos, Neg and Neutral
- Predict the rating
- Identifying theme of the reviews

Data Availability

The Customer review, rating and the bank name were given in the dataset.

Approach

1 Importing libraries and dataset
2 Data Audits

Check for duplicate rows
Check for missing values
Check for Language of reviews

3 Exploratory Data Analysis

Number of sentences per review
Number of word count per review
Number of stop words per review
Count of each parts of speech

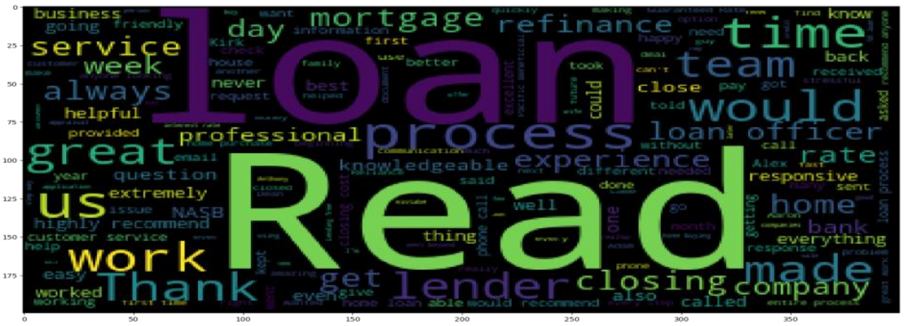
4 Data Preparation

- Train and Test split
- Creating Corpus

Cleaning
Case conversion (lower case)
Lemmatization / Stemming

5 Frequency distribution of words

- Frequency of words
  The frequency of each word in the corpus was found.



- Connotation of frequent words
  The connotation of each word was identified using the TextBlob library method.
  Where the polarity of the word object was used to classify the sentiment.
  It was found that there are,
  Positive words – 212
  Negative words – 104

6 Classifying the reviews as Positive, Negative or Nuetral

The TextBlob polarity was applied to each review and then classified based on the value derived.
Percentages of reviews:
Positive - 89.60 %
Negative - 7.92 %
Neutral - 2.47 %

7 Predicting the Rating of reviews

- Bag of words (Count Vectorizer)
- TF-IDF
- TF-IDF n_gram (Bigram)
  The above vectorizer were fit_transform () on the train corpus and transform () on the test corpus.

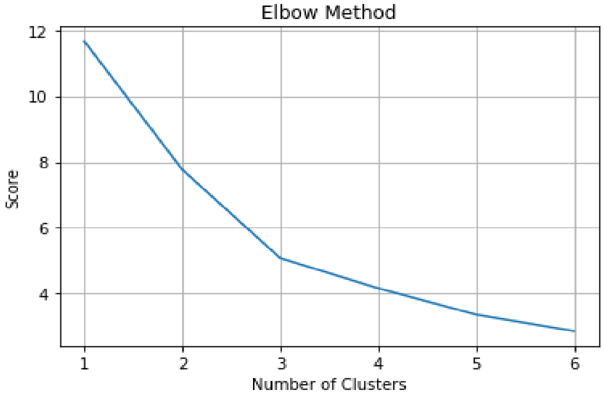
- Model implemetation + Evaluation

  - Naive Bayes
  - LinearSVC
    Both the algorithms were implemented on the Bag of words, TF-IDF and TF-IDF n_gram models and their respective Train and Test evaluation metrics such as confusion matrix, accuracy, AUC score, precision was noted.

    For both the algorithms, Bag of words model was giving a better performance comparatively.
    Accuracy train : 98.51 %
    Accuracy test : 95.05 %

8 Clustering Analysis

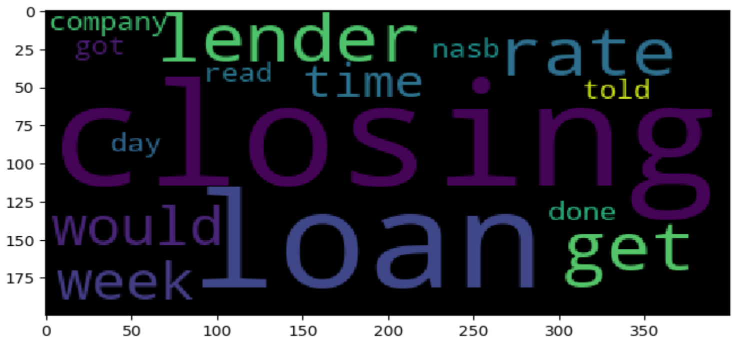
- Finding Optimum #Clusters
  The dimensions of the data were reduced using the PCA and the Elbow method was applied.
  It optimum #clusters were found to be 3.



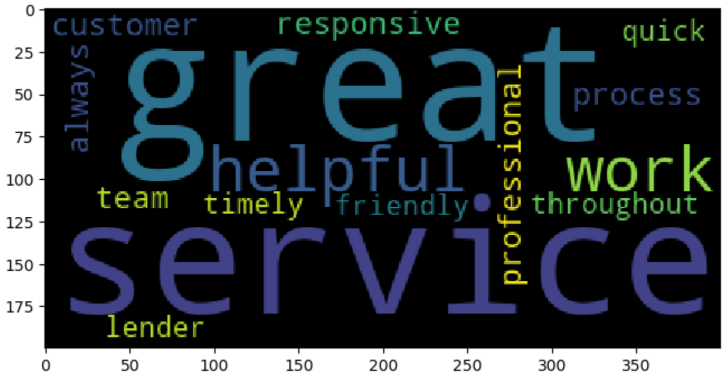
- K Means
  The K Means was implemented for #3 clusters.

- Cluster exploration
  The top features of each clusters were identified.

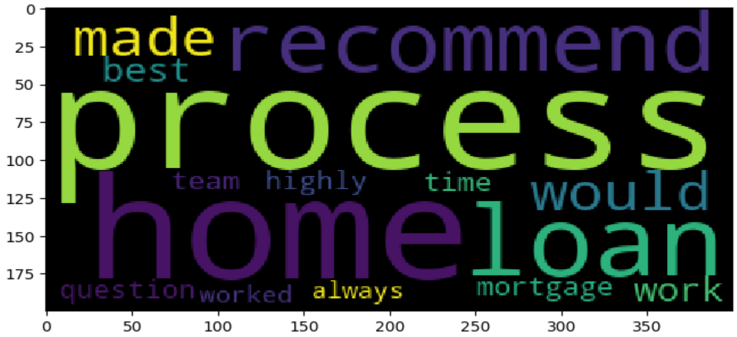
Cluster 1:



Cluster 2:



Cluster 3:

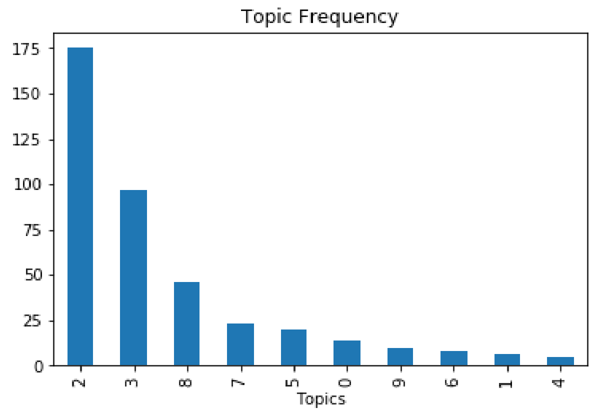


9 Topic Mining

- LDA
  The LDA was applied to the train data by filtering the extremes no_below=3 and no_above=0.9.
  The number of topics was set as 10, chunksize was set as 10, passes=20.
  Below are topics got from LDA,

```
（0,
  '0.100*"would" + 0.073*"company" + 0.036*"bank" + 0.029*"mortgage" + 0.025*"never" + 0.025*"thing" + 0.024*"quick"
+ 0.022*"called" + 0.020*"respond" + 0.019*"received"'),
 (1,
  '0.091*"every" + 0.057*"way" + 0.053*"deal" + 0.051*"financing" + 0.049*"done" + 0.047*"well" + 0.039*"interest" +
0.038*"step" + 0.038*"response" + 0.034*"want"'),
 (2,
  '0.054*"loan" + 0.049*"great" + 0.036*"mortgage" + 0.028*"service" + 0.027*"best" + 0.025*"team" + 0.025*"worked" +
0.022*"lender" + 0.022*"alex" + 0.021*"refinance"'),
 (3,
  '0.082*"process" + 0.059*"home" + 0.052*"time" + 0.045*"question" + 0.044*"recommend" + 0.037*"always" +
0.035*"made" + 0.032*"experience" + 0.030*"work" + 0.030*"easy"'),
 (4,
  '0.070*"house" + 0.065*"friend" + 0.059*"know" + 0.059*"jon" + 0.056*"barrett" + 0.051*"phone" + 0.034*"job" +
0.033*"mr" + 0.033*"recommended" + 0.033*"another"'),
 (5,
  '0.070*"closing" + 0.043*"option" + 0.041*"email" + 0.034*"helped" + 0.034*"day" + 0.029*"refinancing" + 0.027*"good"
+ 0.026*"back" + 0.022*"without" + 0.022*"peter"'),
 (6,
  '0.065*"got" + 0.062*"go" + 0.054*"adam" + 0.042*"even" + 0.038*"find" + 0.038*"fantastic" + 0.032*"people" +
0.030*"better" + 0.030*"document" + 0.026*"online"'),
 (7,
  '0.132*"u" + 0.053*"work" + 0.033*"answered" + 0.027*"thank" + 0.026*"certainly" + 0.025*"hard" + 0.025*"care" +
0.024*"able" + 0.023*"kept" + 0.023*"issue"'),
 (8,
  '0.063*"rate" + 0.047*"went" + 0.037*"week" + 0.034*"get" + 0.033*"customer" + 0.024*"time" + 0.024*"read" +
0.023*"took" + 0.020*"call" + 0.019*"provided"'),
 (9,
  '0.076*"need" + 0.046*"nasb" + 0.028*"loan" + 0.026*"anything" + 0.022*"fee" + 0.019*"selling" + 0.019*"various" +
0.018*"explaining" + 0.017*"letter" + 0.016*"market"')
```

- Topic Frequency
  The frequency of each topic in the data was found and plotted.



- Topic inspection

  The Frequent identified from the plot

Topic 7 (Home loan related)
0.044*"process" + 0.037*"u" + 0.032*"loan" + 0.031*"home" + 0.030*"time" + 0.030*"work" + 0.028*"read" + 0.026*"rate" + 0.023*"recommend" + 0.022*"great"

Topic 9 (Mortgage, NASB (North American Savings Bank))
0.067*"mortgage" + 0.052*"would" + 0.044*"closing" + 0.038*"week" + 0.034*"loan" + 0.031*"alex" + 0.028*"email" + 0.028*"bank" + 0.020*"nasb" + 0.018*"day"

Topic 3 (Credit card, Information enquery)
0.083*"company" + 0.065*"went" + 0.032*"kirk" + 0.028*"one" + 0.026*"call" + 0.025*"much" + 0.025*"next" + 0.024*"credit" + 0.022*"say" + 0.020*"information"