# News Group Classification

Team: 55

|  |  |  |
| --- | --- | --- |
| Seat Number | Section | Name |
| 2021170642 | 8 | يوسف محمد سيد محمد |
| 2021170480 | 6 | محمد عوض طلعت محمد |
| 2021170654 | 8 | يوسف وجيه وديع ناشد |
| 2021170649 | 8 | يوسف مصطفى محمد يوسف |
| 2021170483 | 6 | محمد مبروك فكري عبدالفتاح |
| 2021170641 | 8 | يوسف عمرو احمد زهران |

## PreProcessing:

Convert text to list of tokens (word tokenization)

Remove Stop Word from tokens

Remove Punctuation from tokens

Perform stemming on tokens using SnowballStemmer

Concatenate all tokens as single string

## Feature Extraction:

Apply TfidfVectorizer (TF-IDF) on train and test data in order to extract features from the preprocessed tokens

## Model Training and Testing:

Naive Bayes:

Training Accuracy = 0.9459961110129044

Testing Accuracy = 0.8069569835369091

Decision Tree (max-depth = 80):

Training Accuracy = 0.8463683441774

Testing Accuracy = 0.6197557089750398

Logistic Regression:

Training Accuracy = 0.9729783561280042

Testing Accuracy = 0.9121083377588953

## Visualizing Results:

Used Confusion Matrix to visualize the results as Precision, Recall and F1Score

## Saving the Model:

Save both Naive Bayes and Decision Tree models using pickle library