PA4 Description

Design of the web service analysis system:

1. The code imports the essential packages for loading, processing, cleaning, clustering, and classifying data.
2. The load_input_data and preprocess_data functions load and sanitize the input data.
3. The perform_feature_selection function is used to extract features for clustering and classification using the preprocessed data.
4. Based on the three feature extraction models, classification models are trained in a for loop with the preprocessed data.
5. To execute clustering on the chosen features, the start_clustering function is called.
6. The main function is the driver code that calls the above functions in sequence to execute the pipeline. It also tracks the time taken for the pipeline to run.

Key Development aspects:

1. The description column was normalized by removing the stop words and punctuations from it.
2. The feature extraction models used were: tf-idf, topic modeling, word embedding modeling.
3. The classification algorithms used were: Decision tree, Naïve Bayes, KNN and Random Forest, out of which Random forest was found to be most accurate.
4. A pickling file was created so that predictions can be made on the future data.
5. Clustering Algorithms used were: K-means and DBSCAN

Evaluations:

1. The best feature selection model that amounted to most accuracy during classification was – tf-idf.
2. Tf-idf along with Random forest amounted to an accuracy of around: 74.8%
3. The results of evaluations for classification algorithms used is below:

```
Model Approach  1
Decision Tree accuracy: 0.6601289134438305
Naive Bayes accuracy: 0.18710865561694293
KNN accuracy: 0.49475138121546963
Random forest accuracy: 0.7481583793738491

Model Approach  2
Decision Tree accuracy: 0.038581952117863715
Naive Bayes accuracy: 0.07900552486187847
KNN accuracy: 0.0432780847145488
Random forest accuracy: 0.04152854511970534

Model Approach  3
Decision Tree accuracy: 0.11850828729281768
Naive Bayes accuracy: 0.12826887661141803
KNN accuracy: 0.14723756906077348
Random forest accuracy: 0.2176795580110497
```

Here Model Approach 1, 2 and 3 are tf-idf, topic modeling and word embedding respectively.

The predicted data looks like below:

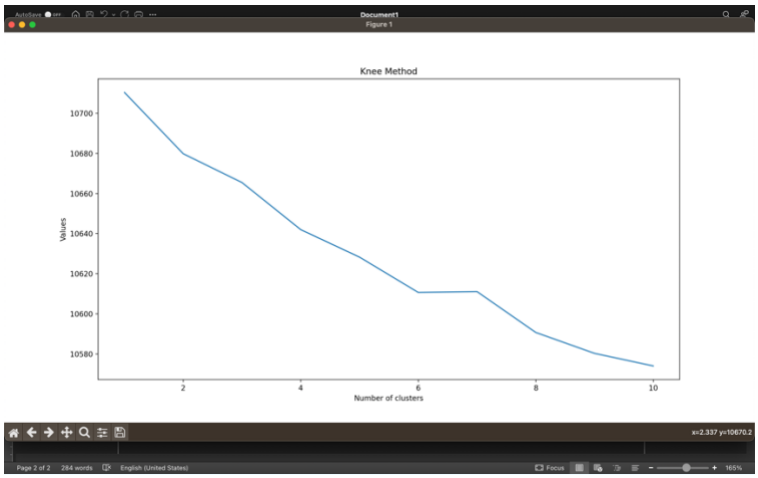
```
+----+-----------------+
|    | Random Forest   |
|----+-----------------|
|  0 | Science         |
|  1 | Games           |
|  2 | Messaging       |
|  3 | Social          |
|  4 | Internet        |
+----+-----------------+
```

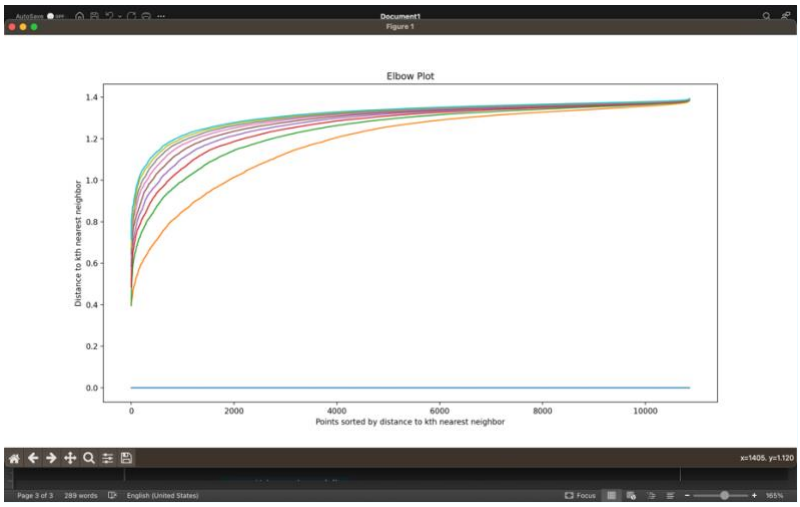
Here each row on the data frame corresponds to the row of the future/new data and predicted values are the categories.

4. Clustering algorithms used here were: K-means and DBSCAN.
5. DBSCAN with topic modelling was found to be most efficient.

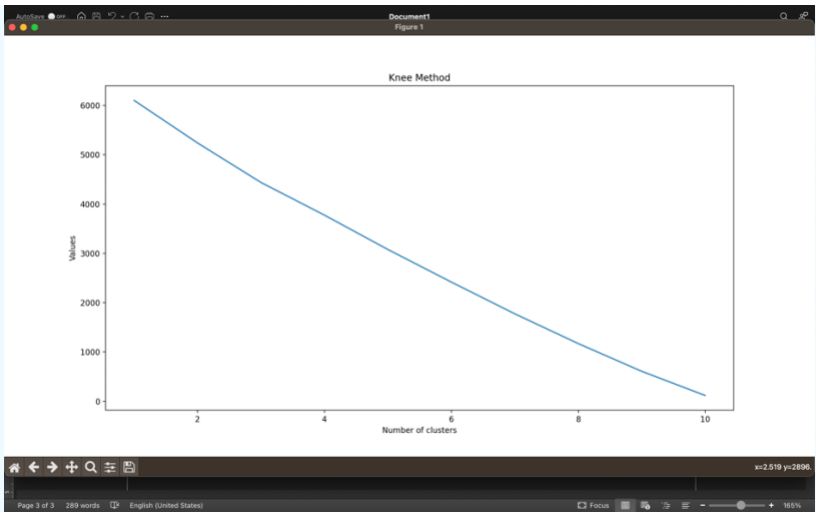
6. The hyperparameters for the clustering algorithms were obtained using the following graphs:

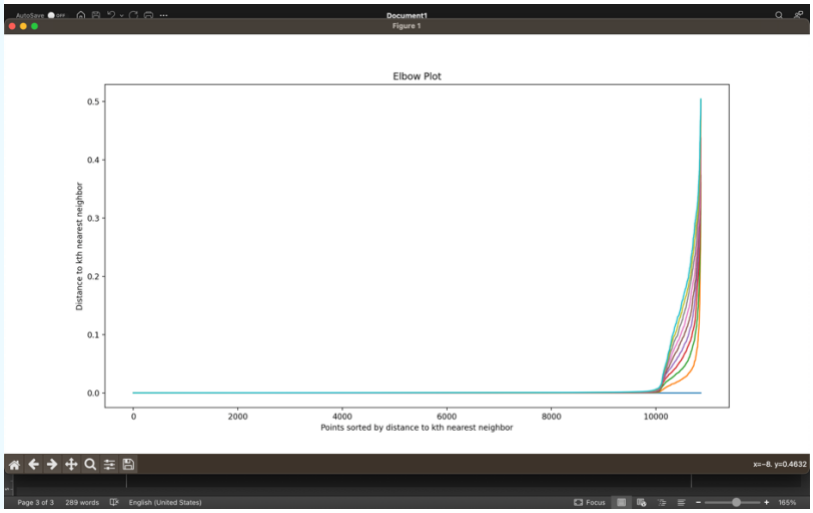
Using tf-idf:



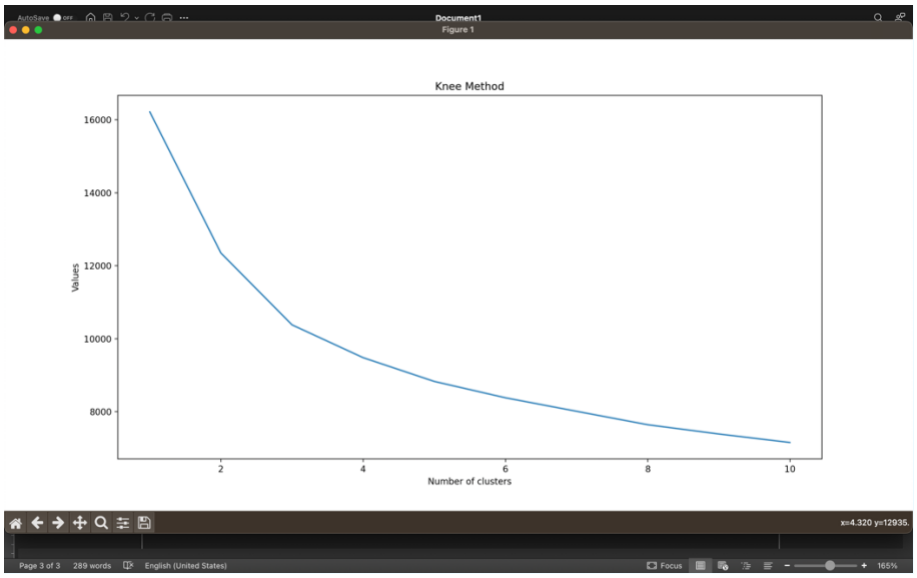


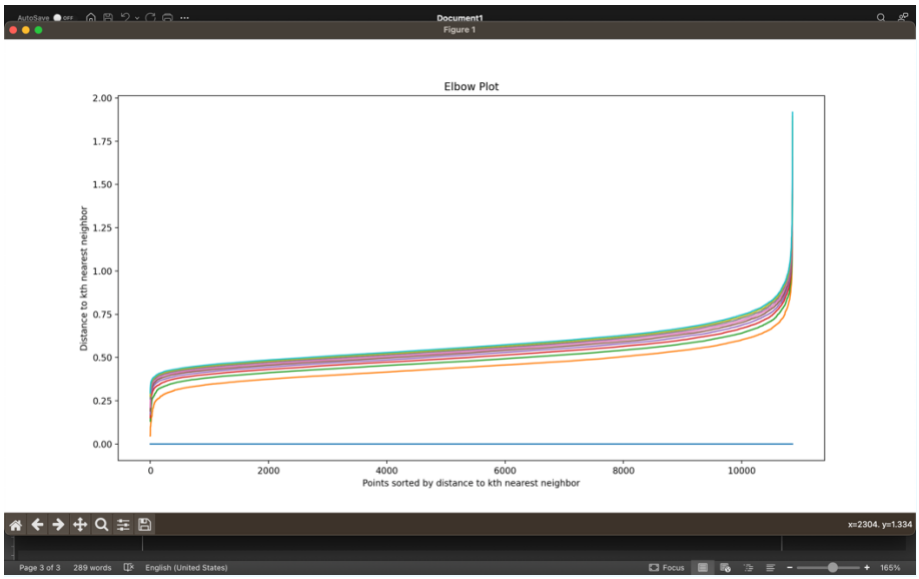
Using topic modeling:





Using word embedding:





7. The clustering results suggest that topic modelling combined with DBSCAN produced the best clustering:

```
Silhouette score for k-means clustering using feature 1 :  0.001444202905902985
DBSCAN clustering resulted in only one cluster. Cannot compute silhouette score.
Something was not right with feature extraction model 1
Silhouette score for k-means clustering using feature 2 :  0.5055255720177435
Silhouette score for DBSCAN clustering using feature 2 :  0.8569733989011805
Number of clusters:  10
Silhouette score for k-means clustering using feature 3 :  0.14671053
DBSCAN clustering resulted in only one cluster. Cannot compute silhouette score.
Something was not right with feature extraction model 3
Total time taken by the code to run is: 452.9378409385681
```