Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) Ajinkya Morade

E-mail: ajinkya.morade1998@gmail.com

- Data sorting.
- Approach towards plan.
- Graphical representation.
- Bar plot and Word cloud.
- Implementation of various Models.
- Model selection and implementation.
- Data visualization.
- Sorting of values.
- Bar plot

2) Nitesh Gajakosh

E-mail: niteshgajakosh1998@gmail.com

- Various model implement.
- Project summery template.
- Model selection
- Data analysis.
- Frame work of project.
- Model presentation.
- Word cloud
- Analyzing results of model.
- Model validation

Problem definition:

In this project your task is to identify major themes/topics across a c ollection of BBC news article. You can use clustering algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) etc.

EDA on given Data set:

Digging into data we understand that

- There is no null value in the data set.
- News: Text document of news article of various types.
- Type: Type of news article such as Business, Politics, Tech, Entertainment and Sports.

Model selection and implementation:

After implementing various models on the given data such as Logistic Regression, Decision Tree Classifier, Naïve Bays Classifier, KNN Classifier, Random Forest Classifier. We get maximum accuracy with Logistic Regression and KNN Classifier train and test accuracy are 0.98, 0.95 and 096, 0.95 for Logistic Regression and KNN Classifier respectively. Both the models are of different types as Logistic Regression works on math base where KNN is a distance based it looks for its nearest neighbors.

Model	Train	Test
Logistic Regression	0.98	0.95
Decision Tree Classifier	1.00	0.86
Random Forest Classifier	1.00	0.91
Naïve Bays Classifier	0.95	0.92
KNN Classifier	0.96	0.95

Conclusion

In the past decades, text classification has been a hot topic and received attention from many scholars in areas such as natural language processing, information retrieval, and machine learning, etc. Online news services have emerged in past decades due to the pervasion of mobile devices and Internet access. Many readers, especially youngsters, relied on online news services rather than traditional sources such as newspapers, TV, and radio. In such case topic modeling of news came in front. It is important to classify news articles according to their types.

After implementing various models on the given data such as Logistic Regression, Decision Tree Classifier, Naïve Bays Classifier, KNN Classifier, Random Forest Classifier. We get maximum accuracy with Logistic Regression and KNN Classifier train and test accuracy are 0.98, 0.95 and 096, 0.95 for Logistic Regression and KNN Classifier respectively.

Please paste the GitHub Repo link.

Github Link:- https://github.com/ajinkyamorade/Topic-Modeling-on-News-Articles

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)