Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Kollana Harish, harishkollana@gmail.com :

1. Importing Libraries
2. Loading the dataset
3. Dataset Information
4. Data Cleaning
5. Data Analysis on Columns
6. Topic Modelling
    1. LSA - Latent Semantic Allocation
    2. LDA - Latent Dirichlet Allocation

7. Conclusion

Please paste the GitHub Repo link.

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

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

BBC is a broadcasting company and they publish news articles and users interact those articles from various channels. In this project we had identified major themes/topics across a collection of BBC news articles by clustering algorithms such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

We are provided with a dataset contains a set of news articles for each major segment consisting of business, entertainment, politics, sports and technology. You need to create an aggregate dataset of all the news articles and perform topic modeling on this dataset. Verify whether these topics correspond to the different tags available.

Initially we had combined all these articles of folders into one data frame by taking topics as another column and done text preprocessing using re and simple processing. We had removed /n lines, punctuation, stop words and moved forward with lemmatization of news text.

Then we had implemented Lsa and Lda models and we found out that tech news is more widespread than remaining column in Lsa topic modelling and biggest cluster is entertainment by using Lda topic modelling.