Capstone Project

SUMMARY

Member Name, Email and Project Details :

Name: Prasad Khedkar

Email : consistent_being@protonmail.com

Project Name : Topic Modelling on News Articles

Project Type : Individual

GitHub Repo link :

Github Link:- https://github.com/Prasad-Khedkar/Topic_Modelling_on_News_Articles

Project Description :

Topic modeling discovers the topics that occur in a collection of documents (corpus) using a probabilistic model. It is frequently used as a text mining tool to reveal semantic structures within a body of text.
Terms exhibiting similarity are grouped together and the topic is determined based on the statistical probability of occurrence of those words.

In any technical field it is common for general terms to take on specific, concrete meanings, and this can be a source of confusion.
In topic modeling the word “topic” takes on the specific meaning of a probability distribution over words.

The main objective of this project is to see how differerent algorithms correctly represent the news documents(BBC news collection) to the topics(which we already know) and evaluate the model performance.

Approach Taken :

1. **Raw Data Loading(into dataframe)**
2. **Data Cleaning**
3. **NLP Text Processing**
   - **Removing Tags/non-word characters**
   - **Removing Punctuations**
   - **Removing numbers**
   - **Removing Stop words**
   - **Lemmatization**
4. **Vectorizing the data**
5. **Training models on different algorithms.**
6. **Final Model Selection based on performance.**

Final Conclusions :

- **Tested on three different algoithms – Truncated SVD(also known as LSA), LDA(Latent Dirichlet Allocation) & NMF (Non-negative matrix factorization).**
- **Among three, NMF gave the best topic representation and is our final selection**