Project Summary: (News Topic Modeling Project)

In this Project Text mining and Natural Language Processing (NLP) methods will be used as part of this project whereby it will be used to automatically uncover the hidden thematic themes of a set of world news stories that were scraped off The Daily Star site. The main aim will be to examine text information, which is not in any structure to indicate the existing trends and patterns in news coverage without prior labeling.

The basic developed algorithm is Latent Dirichlet Allocation (LDA), an unsupervised generative model to be applied in Topic Modeling. An end-to-end and full text analysis pipeline will be deployed. This will be initiated by automated data collection through web scraping followed by high degree of preprocessing of the text such as lemmatization, spell-checking and removal of stop words. An extraction of features will be made by building a Document-Term Matrix (DTM), in which textual information will be transformed into a numerical format that will allow algorithmic processing.

The effectiveness of unsupervised learning to find meaningful, latent themes in a corpus of documents shall be shown. An output given by the model will be a set of defined topics where each topic will be defined by its most probable words. Each article will also be assigned topic proportions that can show the combination of themes that will be addressed. To deliver these insights, visualizations, such as topic probability heatmaps and single-word clouds will be created.

This project will highlight the practical use of data science in media analysis as it offers journalists, researchers, and analysts a potent mechanism to rapidly comprehend the trends of narratives in the large amounts of news content. One of the basic NLP methods that are applicable to bigger data and more complicated processes will be presented.