Assignment:

When constructing an NLP pipeline for news scraped from various sources, it is crucial to first eliminate unnecessary and redundant sentences from the news content. This step is important as it ensures that the output generated by the subsequent summarization and named entity recognition models is accurate and relevant. By removing irrelevant and repetitive sentences, the NLP pipeline can focus on the important information and provide a clearer and more concise summary of the news.

A quick solution that can be deployed to achieve this is using TF-IDF which is a numerical statistic that is commonly used in natural language processing and information retrieval. At the sentence level, it can be used to identify the most important sentences within a document by measuring the relevance of each sentence based on the frequency of specific words or phrases. By assigning a weight to each sentence based on its TF-IDF score, it is possible to rank the sentences in order of importance and use this information to generate a summary of the document or perform other NLP tasks.

Other techniques include TextRank, Latent Semantic Analysis, Latent Dirichlet Allocation, and the use of sentence embeddings for similarity computation and clustering, such as using cosine similarity to compare the vector representations of sentences.

These vector representations can be obtained through techniques such as bag-of-words or more advanced methods like BERT embeddings.

dataset of news: https://drive.google.com/file/d/1KsWWrhGaK0_IMXPaYBOEXXE8IExqogmJ

Leveraging any of the above methods or anything else you can come up with, do a test-train split of 10-90 and generate the cleaned responses for the test set.

Submission-
The code must be submitted to a Github repository that includes a Readme file describing the methodology used and a table with at-least these specified columns for the test set.

Original Content	New Content	Removed Lines	Further Metrics.
.....	....	..	