1**. Dataset Collected**:

NEWS SUMMARY (kaggle.com)

2. **Ipython Notebook (done)**

3. **Dataset Brief Description**:

The first dataset consists of 4515 examples and contains the Author's name, Headlines, URL of the Article, Short text, and Complete Article. I gathered the summarized news in shorts and only scraped the news articles from the Hindu, Indian Times, and Guardian. The period ranges from February to August 2017

The second dataset consists of 98402 rows with 2 columns labeled as headlines and text.

To increase the intake of possible text values to build a reliable model as we are working on text summarization on news articles, we have merged these datasets before preprocessing and cleaning. Now the dataset contains 102915 rows and 2 columns labeled as text and summary while the text column has some null values.

4. **Data Dictionary:**

Raw Datasets:

Dataset 1: Name: news_summary.csv

Column Name	Data Type	Description
author	String/object	Contain the news author-name
date	String/object	Date of publication
headlines	String/object	Headline Of the news
read_more	String/link	News link
text	String/object	Short or summary of the article
ctext	String/object	The main content of the news.

Dataset 2: Name: news_summary_more.csv

Column Name	Data Type	Description
headlines	String/object	Contains the headline of the news.
text	String/object	Contains the content of the news.

Merged Dataset:

From the raw dataset, we found, we merged it to create our preferred dataset the so-called “Tuned News Summary”.

Column Name	Data Type	Description
text	String/object	Contains the long text or description of the article. Which will be used to train the model.
summary	String/object	Contains the summary of the particular article. The model will predict this part.

5. In iPython Notebook

6. For an average length of 70 words of text the dataset has a summary of 10 words. The ratio is 7:1