Project Title: Text Summarization Using NLP and Generative AI

Team Members and Roles

- **Neeraj Ghate**
Student ID: 110960079
Email: neeraj.ghate@ucdenver.edu
Role: Data Collection, Preprocessing, and Model Development
Responsibilities: Collect and preprocess data from sources. Implement web scraping techniques. Assist in model development, training, and fine-tuning.

- **Prasad Belsare**
Student ID: 111438300
Email: prasadarun.belsare@ucdenver.edu
Role: Model Implementation, Fine-tuning, and Evaluation
Responsibilities: Implement and fine-tune NLP models for text summarization. Evaluate the model's performance. Collaborate in data collection and preprocessing tasks.

- **Sakshi Kangane**
Student ID: 111030609
Email: sakshisopan.kangane@ucdenver.edu
Role: Web Interface Development, Integration, and Testing
Responsibilities: Develop a user-friendly web interface where users can upload documents for summarization. Handle backend integration with the NLP models to ensure seamless functionality. Collaborate in data collection, model development, and fine-tuning processes.

Problem Statement : The rapid proliferation of information has made it challenging for individuals to digest large volumes of textual data. Automatic **text summarization** can address this by condensing documents into manageable, high-quality summaries. This project aims to develop models for summarization using **NLP** and **Generative AI** techniques.

The Data Source(s) You Intend to Use:

For our project on text summarization, we will be using CNN/Daily Mail dataset, ensuring that it provides sufficient volume and quality for our needs:

News Summarization Dataset: CNN/Daily Mail

- Source: CNN/Daily Mail Dataset
- This dataset contains over 300,000 news articles and their corresponding summaries, making it an ideal source for training and evaluating models on summarizing news articles.
- Access & Quality: We have confirmed access to this dataset via Hugging Face's datasets library, ensuring that we can obtain it without throttling or restrictions. The dataset is widely used in NLP tasks, so we are confident in its quality and completeness.

Data Acquisition and Preprocessing Plan:

- Data Collection: We will download or scrape the datasets in batches to ensure smooth data collection without overwhelming any API or source restrictions.

- **Preprocessing:**
  - **Cleaning:** Removing unnecessary metadata (like authorship, dates, etc.) and irrelevant text (e.g., boilerplate text from websites).
  - **Text Normalization:** Lowercasing, removing special characters, and performing tokenization using libraries like NLTK or SpaCy.
  - **Data Augmentation:** For diversity in summarization tasks, we may perform minor alterations on some texts, such as shuffling sentence order to simulate noisy input.
  - **Joining & Splitting:** Where necessary, we will split longer documents into meaningful sections to ensure the summarization model can process them effectively, while also ensuring that key context is preserved.

Project Goals

1. **Accuracy**: Generate coherent and contextually accurate summaries.
2. **Efficiency**: The solution should generate summaries in real time, with minimal lag for large datasets.
3. **Scalability**: The model should support summarization of text ranging from a few paragraphs to multiple pages.
4. **Human Interaction**: Users should have the option to adjust or correct the generated summaries through the web interface.

Tools and Technologies to be Used

- **Programming Language**: *Python*
- **Frameworks**: *Hugging Face Transformers*. *TensorFlow* or *PyTorch* , *NLTK*, *SpaCy*. *BeautifulSoup/Scrapy*.
- **Compute Resources**: *Google Colab* or *AWS EC2 with GPU*: For training and running models on large datasets.
- **Evaluation Tools**: ROUGE score, BERTscore, and human evaluators for quality checks.

Products Produced

1. **Summarization Models**: Capable of generating extractive and abstractive summaries.
2. **Data Preprocessing Pipelines**: Scripts to clean and process the input data for training and testing.
3. **Web Interface**: A platform where users can upload documents and get summaries.
4. **Final Report**: Comprehensive documentation of methodology, data, results, and future scope.