# Project Overview: AI-Powered Document Summarization and Q&A System

## Week 1-2: Setting Up the Environment and Collecting Data

## 1. Setting Up The Environment

To ensure a seamless development experience, we start by configuring the necessary tools and installing only the required dependencies.

### 1.1 Essential Libraries

The following libraries are used:

• datasets – Enables the loading and processing of huge datasets.

• nltk – Enables text processing, including tokenization and cleaning.

• pandas – Effectively maintains structured data.

## 2. Data Collection

To train the summarization and Q&A models, we use well-established datasets.

### 2.1 Dataset Overview

• CNN/DailyMail – A collection of news stories and their summaries, perfect for training summarization models.

• SQuAD 2.0 – A dataset that contains context paragraphs with related questions and answers, which is used to construct Q&A models.

### 2.2 Dataset Sources

These datasets can be accessed through:

* [CNN/DailyMail on Hugging Face](https://huggingface.co/datasets/abisee/cnn_dailymail)
* [SQuAD 2.0 on Hugging Face](https://huggingface.co/datasets/rajpurkar/squad_v2)
* [SQuAD 2.0 Official Explorer](https://rajpurkar.github.io/SQuAD-explorer/)

## 3. Data Preprocessing

Preprocessing ensures that the text data is well-structured and optimized for model training.

### 3.1 Preparing Summarization Data

• Removing unnecessary spaces and line breaks.

• Standardizing text formatting for consistency.

• Splitting text into sentences to enhance processing efficiency.

### 3.2 Preparing Q&A Data

• Cleaning text by eliminating unwanted characters.

• Structuring context, questions, and answers correctly.

• Addressing missing or incomplete entries in the dataset.

## 4. Storing Processed Data

The cleaned datasets are stored in JSON format to maintain structure and facilitate easy access during model training.