Final Report

NLP/LLM: Text Summarization & Geo-Tagged QA Pipeline

Author: Rajendra Dayma

1. Introduction

This project leverages large language models (LLMs) for summarizing news articles and identifying geo-referenced content from structured and unstructured sources. The goal is to extract useful insights and locations from long text inputs in an efficient and scalable manner.

2. Dataset

Dataset used: CNN-DailyMail News Summarization (from Kaggle).

Fields: article, highlights (human summary)

3. Model Architecture

Model Used: facebook/bart-large-cnn via HuggingFace Transformers.

Pipeline:

- Load the article and truncate (max 1024 tokens)
- Run summarization using BART
- Extract location mentions using SpaCy (NER: GPE, LOC)
- Use Geopy to convert locations into latitude and longitude
- Store output in a CSV file

4. Tools & Libraries

- transformers (BART)
- spacy (NER)
- geopy (geolocation)
- pandas, tqdm, numpy

5. Results

Final output includes:

- Summarized article

- Location extracted from article

- Latitude, Longitude via geocoding

A CSV of 100 processed entries was generated with <X>% success rate.

6. Challenges

- Managing token limits for long articles

- Geocoding ambiguous place names

- Speed optimization for batch processing

7. Future Work

- Add question-answering layer using GPT or LangChain

- Use vector database for semantic search

- Visualize locations on an interactive map