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

Siddhardha Reddy Naredla

September 26, 2025

1 Objective

The goal of this project is to build a system that:

- Summarizes news articles from CNN-DailyMail dataset.
- Enables question-answering (QA) on generated summaries.
- Logs geolocation of news articles in a structured table.

2 Dataset

- Dataset: CNN-DailyMail News Text Summarization (train.csv)
- Source: `https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarizatio`
- Sample: 100 preprocessed summaries with geotags stored in `summaries_geotagged.csv`

3 Pipeline Design

The project pipeline is divided into the following stages:

3.1 Data Preprocessing

- Loaded dataset in Pandas.
- Cleaned and preprocessed text.
- Extracted necessary columns for summarization and geolocation.

3.2 Text Summarization

- Model: **T5-small** (transformers library)
- Input: Article text
- Output: Summarized text stored in CSV
- Evaluation: ROUGE metrics

3.3 Geolocation Extraction

- Named Entity Recognition using `spaCy` (`en_core_web_sm`)
- Extracted GPE (Geo-Political Entities) from summaries
- Converted to latitude and longitude using `Geopy`
- Logged results in CSV along with summary

3.4 QA Pipeline (RAG-style)

- TF-IDF vectorizer to embed summaries
- Cosine similarity to retrieve relevant summary context
- T5-small model used to generate answers
- Interactive QA service implemented for continuous questions

4 Implementation

4.1 Flask Deployment

- Simple UI: Enter question → get answer → see context
- Connects frontend with QA backend
- Steps to run locally:
  1. Install dependencies: `pip install -r requirements.txt`
  2. Run Flask app: `python app.py`
  3. Access: `http://127.0.0.1:5000/`

5 Results & Evaluation

5.1 Summarization Metrics

Metric	Score
ROUGE-1	0.38
ROUGE-2	0.22
ROUGE-L	0.28

Table 1: ROUGE Scores for T5-small Summarization

5.2 QA Evaluation

- Accuracy checked manually on 100 samples
- Answers retrieved using RAG-style retrieval + T5-small generation

6 Discussion

- Model choice: T5-small for efficiency
- Retrieval using TF-IDF for lightweight RAG-style QA
- Flask deployment allows interactive demonstration

7 Future Work

- Use FAISS vector database for faster retrieval
- Replace T5-small with larger LLMs for higher QA accuracy
- Add visualization dashboard (Streamlit/Gradio)
- Analyze and mitigate biases in summarization and QA

8 References

- CNN-DailyMail Dataset: `https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail`
- HuggingFace Transformers: `https://huggingface.co/docs/transformers/index`
- SpaCy Documentation: `https://spacy.io/usage`