Text Summarization and Geo-Tagging Pipeline

1. Introduction

This project integrates automatic text summarization with geographic information extraction, providing a robust tool for analyzing long texts such as news articles, reports, or documents.The pipeline has two main objectives:

1. **Summarization**: Convert lengthy text into concise, informative summaries.
2. **Geo-Tagging**: Detect locations mentioned in the text and map them to geographic coordinates (latitude and longitude).

For example, given a news article, the system will produce a summarized version and identify all locations mentioned, such as "Paris," "New York," or "India," and convert these into precise coordinates.

The final outputs include:

- A clean CSV file with summaries and geolocation data.
- An interactive Gradio application for testing the pipeline live with any input text.

2.Workflow Overview

1. Data Preparation

- **Dataset Used**: CNN/DailyMail News Dataset (downloaded from Kaggle).
- **Cleaning Steps**:
    1. Remove rows with missing articles.
    2. Trim extra spaces in text.
    3. Optionally truncate extremely long articles to improve model efficiency.
- **Output**: A preprocessed CSV file (train_preprocessed_1000.csv) that is ready for summarization.

Libraries Used:

- **Pandas**: For data manipulation and CSV operations.

2. Summarization

- **Model Used**: facebook/bart-large-cnn from Hugging Face Transformers.
- **Reason for Choice**:
    - BART is a transformer-based encoder-decoder model, optimized for abstractive summarization.
    - Pre-trained on news datasets, it produces human-like summaries.

- **Pipeline Execution**:
    - Texts are processed in batches for efficiency.
    - Generated summaries are stored along with original articles.

Libraries Used:

- **Transformers**: For loading the pre-trained BART model and summarization pipeline.
- **TQDM**: For progress bars during batch processing.

3. Location Extraction and Geocoding

- **NER (Named Entity Recognition):**
    - **Library:** spaCy (en_core_web_sm)
    - **Purpose:** Detect geographical entities like cities, countries, and landmarks.
- **Geocoding:**
    - **Library:** GeoPy with Nominatim API.
    - **Purpose:** Convert detected place names into latitude and longitude coordinates**.**
    - Includes caching using shelve to avoid repeated API calls and improve speed.

Output:

- List of detected places.
- Latitude and longitude for each place.
- Primary location coordinates (first geocoded place) for simplicity.

4. Evaluation Metrics

To assess the quality of generated summaries, two metrics are used**:**

1. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**
    - Measures overlap of n-grams between generated and reference summaries.
    - Reports Precision, Recall, and F1-score for ROUGE-1, ROUGE-2, and ROUGE-L.
2. **BERTScore**
    - Uses deep contextual embeddings to evaluate semantic similarity.
    - Better captures meaning than surface-level word overlap.
    - Outputs Precision, Recall, and F1-score as percentages**.**

Libraries Used:

- rouge-score and evaluate for ROUGE.
- bert-score for semantic similarity evaluation.

Metric	Precision(%)	Recall(%)	F1 Score(%)	BERTScore
ROUGE - 1	98.88	16.93	28.72	81.24%
ROUGE - 2	89.33	15.14	25.73	80.84%
ROUGE - L	90.29	15.50	26.29	81.03%

.

5. Output Cleaning

- Remove duplicate rows.
- Drop rows with empty or incomplete summaries or references.
- Reset index for a neat CSV output.(output file-out_sample1000_final1.csv)

Libraries Used: Pandas

Final CSV Example Columns:

- id
- original_text
- summary
- places_found
- geocoded_all
- primary_lat
- primary_lon
- primary_place

6. Interactive Interface

- **Library:** Gradio
- **Purpose:** Provide a user-friendly web interface for testing.
- **Functionality:**
    1. Enter text in a textbox.
    2. **Outputs:**
        - Summarized text.
        - List of detected locations.
        - Corresponding latitude and longitude values**.**

3. Libraries & Tools Used

Library	Purpose
Transformers	Pre-trained BART summarization
spaCy	Named Entity Recognition (NER)
GeoPy	Geocoding detected locations
Pandas	Data preprocessing and CSV handling
TQDM	Progress visualization
Rouge-Score	Evaluate Summarization evaluation (ROUGE)
BERTScore	Semantic evaluation of summaries
Gradio	Interactive web interface

4. Future Improvements

- Fine-tune BART on domain-specific datasets to improve ROUGE and BERTScore.
- Integrate more advanced NER models for better location extraction.
- Allow multiple geocoding APIs and visualize locations on interactive maps.
- Add language support for non-English articles.

5. Conclusion

This pipeline provides a complete end-to-end solution for summarizing long texts while simultaneously identifying geographic entities. It is modular, allowing different summarization models or geocoding services to be integrated. The interactive interface makes it accessible even to non-technical users, while CSV output ensures the results are ready for further analysis.