Technical Report: Email Matter Grouping System

Approach Summary:

In my solution for Qanooni's Email Matter Grouping System, I built a modular prototype designed to automate the organization and summarization of legal email communications. My work comprises three main stages:

- Synthetic Data Generation: I used the Faker library to generate a realistic yet controlled dataset of legal emails across multiple matters, complete with threaded replies and ambiguous subjects.
- Matter Grouping Model: I leveraged Sentence-BERT to extract semantic embeddings from the combined email text and then applied KMeans clustering to assign predicted matter IDs. I evaluated the performance using metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI).
- AI-Powered Matter Summarization: For each matter group, I created structured prompts that capture key elements (instructions, recipient details, scope of work, ongoing tasks, and chronology) and generated detailed 1–2 page summaries using Google GenAI's Gemini model via the google-genai package.

This approach allowed me to combine efficient data generation, robust clustering, and state-of-the-art LLM-based summarization to streamline legal matter management.

Features Used

- **Faker:** I utilized Faker to simulate realistic email data while ensuring that sensitive information was not exposed. This helped me generate a synthetic dataset that mimics real-world scenarios.
- **Pandas & Numpy:** I used these libraries for efficient data manipulation and numerical processing.
- **Scikit-learn:** I relied on this package for running the KMeans clustering algorithm and for computing evaluation metrics like ARI, NMI,
- **Sentence-Transformers (Sentence-BERT):** I employed a pre-trained model ("all-MiniLM-L6-v2") to convert email texts into dense embeddings. This allowed me to capture the nuanced semantic relationships critical in legal communications.
- **Google-genai:** I integrated the Gemini model (gemini-2.0-flash) through the google-genai package to generate comprehensive summaries from the grouped emails.

Methodology

1. **Synthetic Data Generation**

- **Technique:** I generated 20–25 synthetic emails covering 3–4 legal matters using the Faker library. I implemented legal templates with specific subjects and body content to ensure that each email accurately reflected a legal theme.
- **Key Components:** I randomized email IDs, senders, recipients, and timestamps to simulate real communication. Additionally, I generated threaded replies and forwarded messages to introduce context and realistic ambiguity, which is essential for testing the robustness of my grouping model.

2. **Matter Grouping**

- **Preprocessing:** I combined the subject and body of each email into a single text string and normalized it (e.g., converting to lowercase) for consistent processing.
- **Embedding Extraction:** I used the SentenceTransformer model ("all-MiniLM-L6-v2") to generate semantic embeddings for each email, capturing their full context and meaning.
- **Clustering:** I applied KMeans clustering with a fixed random state to group the emails into four clusters. I then compared the predicted matter IDs against the true matter IDs using ARI, NMI, and Silhouette Score to evaluate the clustering performance.

3. **AI-Powered Matter Summarization**

- **Prompt Engineering:** for each matter group, I sorted the emails chronologically and assembled a detailed prompt that includes a recap of instructions, recipient details, scope and fee information, ongoing tasks, and a chronology of events.
- **LLM Integration:** I integrated the Gemini model via google-genai to generate a 1–2 page summary for each matter group. This process automates the consolidation of key legal details, reducing manual effort and ensuring consistent output.

Challenges

- **Synthetic Data Complexity:** I had to ensure the synthetic emails included enough variability and realistic noise (e.g., ambiguous subjects and threaded replies) to truly mirror real-world legal communications while remaining manageable.
- **Clustering Tuning:** finding the right balance with KMeans was challenging. I needed to ensure the clustering model could accurately group emails by legal matter despite inherent language variability.

- **Summarization Quality:** creating prompts that reliably extract all essential details for each matter was challenging. Legal summaries require nuanced language, so fine-tuning the prompts and handling the model's output required careful attention.
- **Evaluation Limitations:** I discovered that using simple metrics does not fully capture the semantic quality of summaries. This made it necessary to complement automated evaluations with manual assessments.

Results

Grouping Evaluation Metrics

Adjusted Rand Index (ARI): 1.00
Normalized Mutual Information (NMI): 1.00

Summarization Evaluation

Jaccard Similarity Scores:

Matter 2 - Jaccard Similarity Score: 0.021

Matter 0 - Jaccard Similarity Score: 0.022

Matter 3 - Jaccard Similarity Score: 0.028

Matter 1 - Jaccard Similarity Score: 0.022

Scalability Assessment

To scale this system to support over 1,000 users managing 100,000 emails each, I would consider the following modifications:

- **Data Storage and Management:** integrate cloud-based databases (such as AWS DynamoDB or Google BigQuery) to manage the increased data volume. Distributed file systems and efficient data retrieval mechanisms would be necessary.
- **Distributed Processing:** I would leverage distributed processing frameworks like Apache Spark or Dask to parallelize embedding generation and clustering across large datasets.

- **Incremental Clustering:** instead of re-clustering the entire dataset on every update, I would explore incremental or online clustering algorithms to incorporate new emails in real-time.
- **Efficient LLM Integration:** deploy the summarization module as a microservice on GPU-enabled cloud instances or via serverless architectures to handle high summarization throughput with low latency.
- **Caching:** implement caching mechanisms for embeddings and summaries to reduce redundant computations, thereby improving overall efficiency.

Improvements

1. **Enhanced Evaluation Metrics:** in addition to Jaccard similarity, I would incorporate advanced metrics like ROUGE, BLEU to better capture the quality of summaries.
2. **Refined Prompt Engineering:** continuous refinement of the summarization prompt is necessary to ensure that all relevant legal details are captured. Fine-tuning the prompt based on feedback from legal experts can further enhance the output quality.
3. **Hybrid Clustering Techniques:** exploring a combination of heuristic methods (e.g., subject line analysis or sender-recipient matching) with embedding-based clustering may improve grouping accuracy.
4. **User Feedback Integration:** introducing a feedback loop where legal professionals can rate and suggest improvements to the summaries would allow for iterative refinement of both clustering and summarization components.
5. **Automation Pipeline Enhancements:** I would automate more parts of the data ingestion and processing pipeline to minimize manual interventions and errors.