Evan Dartez

CMPS 6730

Semantic Analysis of the JFK Assassination Records

Problem Overview

The objective of this project is to analyze the publicly released JFK assassination investigation archive using natural language processing (NLP) techniques. These documents, released by multiple government agencies, span decades of intelligence and policy decisions and include memos, surveillance logs, and declassified reports. The majority are stored as scanned PDF images, rendering manual review infeasible. The primary aim of the project is to construct a scalable pipeline capable of converting this heterogeneous, non-textual dataset into structured, analyzable text suitable for downstream semantic analysis. The overarching goal is to support narrative discovery, thematic modeling, and historical insight extraction from this historically significant corpus.

Data

The dataset comprises tens of thousands of scanned pages from the National Archives' publicly released JFK assassination records. These files are distributed in ZIP format and encapsulate a wide range of content formats and qualities. Due to processing constraints during this phase, the analysis was limited to approximately 63,000 pages. This subset, while substantial, represents only about 20% of the full archive. It was selected to capture a broad range of agencies, formats, and content types, including surveillance notes, internal memoranda, intelligence summaries, and policy documentation.

Methods

To enable text analysis on a corpus consisting primarily of scanned documents, the project first implemented a two-tiered OCR pipeline. Each page was rendered at 200 DPI and initially processed using Tesseract. For pages where Tesseract failed or yielded low-quality output, the pipeline invoked TrOCR, a transformer-based OCR model trained to recover content from handwritten or degraded text. This hybrid approach ensured a higher recall across heterogeneous scan qualities. After OCR, the output was normalized and evaluated for quality.

Evan Dartez

CMPS 6730

OCR quality was assessed using a combination of metrics. First, the valid word ratio was computed by comparing token outputs to a reference English dictionary. Second, a BERT-based cloze score was used to estimate the coherence of masked token predictions, providing insight into the contextual predictability of the output. Third, GPT-2-based perplexity was computed to assess overall fluency. These metrics were logged alongside engine-specific metadata in a SQLite database that allowed for efficient tracking and incremental updates.

Following successful text extraction, the documents underwent a standard NLP preprocessing sequence. Texts were tokenized and stripped of stopwords. Named Entity Recognition (NER) was performed using spaCy, and co-occurrence statistics were gathered via bigram analysis. To address the issue of inconsistent entity mentions (e.g., variations in spelling or formatting), fuzzy entity matching using RapidFuzz was applied with a 90% similarity threshold. Sentiment was also measured using TextBlob.

The final application of the pipeline was a Retrieval-Augmented Generation (RAG) system. All cleaned OCR outputs were chunked into 500-token segments with 50-token overlaps. These chunks were embedded using the MiniLM model and indexed using FAISS for efficient retrieval. Queries were enriched using named entity recognition to append relevant terms to improve retrieval precision. Retrieved contexts were passed to a quantized Falcon-7B model, which generated candidate answers. The system was evaluated using ROUGE (1, 2, L) and SQuAD-style metrics (Exact Match and Token F1).

Results

The entity extraction process across 193 processed document chunks surfaced frequent references to well-known organizations and geopolitical locations central to the JFK investigation. The most frequently occurring named entities included:

- [ORG] CIA: 6,238 mentions
- [GPE] Cuba: 3,400 mentions
- [ORG] FBI: 1,955 mentions
- [PERSON] JOHN F. KENNEDY: 1,789 mentions

Bigram analysis revealed several dominant collocations, including:

Evan Dartez

CMPS 6730

- "kennedy assassination"
- "united states"
- "john kennedy"
- "records act"

Token frequency analysis found the following terms to be most common:

- "secret"
- "committee"
- "subject"
- "agency"

These findings reaffirm the political, intelligence, and administrative focus of the corpus.

Sentiment analysis of the full corpus revealed a neutral tone overall, with a polarity of 0.03 and a subjectivity score of 0.41, suggesting that the language of the archive remains largely formal, factual, and emotionally restrained—as would be expected of government records.

To evaluate the efficacy of the RAG system, a sample question was submitted: "What does the Warren Commission conclude about Lee Harvey Oswald?" The enriched query successfully triggered relevant retrievals, and the generated response was: "The Warren Commission concluded that Lee Harvey Oswald was the sole assassin of President John F. Kennedy." While this answer was historically accurate, evaluation metrics revealed challenges in scoring semantically correct but syntactically divergent outputs. The ROUGE-1 score was 0.036, with similarly low ROUGE-2 and ROUGE-L scores. SQuAD metrics showed an Exact Match score of 0.00 and a Token F1 of 3.64. These outcomes reflect the sensitivity of automated metrics to minor phrasing differences, despite the underlying factual agreement.

Evan Dartez

CMPS 6730

Discussion

This project successfully demonstrated the viability of a scalable OCR-to-NLP pipeline capable of processing Cold War-era intelligence documents and extracting structured insights. The fallback to TrOCR significantly improved OCR coverage, especially for degraded or handwritten pages, and the implementation of post-OCR scoring enabled targeted quality control. Named entity analysis and bigram frequency distributions highlighted key individuals, organizations, and concepts central to the JFK assassination narrative. The sentiment profile of the documents reinforced the expectation that the language of intelligence and government reports is institutionally neutral.

The RAG-based semantic question answering system functioned as the central proof-of-concept for downstream utility. While it produced factually correct responses, traditional metrics such as ROUGE and SQuAD may not adequately capture the fidelity of the answers when they deviate from reference phrasing. This disconnect underscores the need for alternative evaluation frameworks that reward semantic congruence even in the presence of lexical variation.

A major constraint on the scope of this project was computational. Only 63,000 pages—roughly one-fifth of the full collection—were successfully OCR-processed during this phase. As such, any modeling results must be interpreted with the understanding that they reflect only a portion of the corpus. This limitation constrained the diversity of training data available for both semantic modeling and entity extraction, and limits the scope of any conclusions drawn from the current pipeline.

Future Work

Future phases of the project will extend OCR processing to the full archive in collaboration with the National Archives' Citizen Archivist Program, which offers a public platform for volunteer transcription and annotation. This partnership is expected to improve both the scale and the accuracy of the corpus.

Additionally, future technical improvements include the integration of full summarization results using BART to support narrative condensation; refinement of question-answering evaluation metrics to include fuzzy and paraphrastic matching; and fine-tuning of generative models on declassified intelligence-style documents to reduce hallucination and increase factual precision. Finally, topic modeling approaches such as

Evan Dartez

CMPS 6730

LDA and BERTopic will be explored to detect latent themes across the full corpus once OCR processing is expanded.

Evan Dartez

CMPS 6730

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research.

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.

Liu, Y., Zhang, Y., & Wang, H. (2014). Text Mining in the Digital Age: Challenges and Opportunities for Historical Documents. Historical Computing Journal.

Riedl, M., & Pohl, D. (2015). Named Entity Recognition in Historical Texts. Journal of Digital History.

Kumar, S., Patel, R., & Wang, Z. (2025). Evolving Techniques in Sentiment Analysis: A Comprehensive Review. International Journal of Computational Linguistics.