Evan Dartez
CMPS 6730

Problem Overview

The goal of this project is to analyze the full archive of publicly released JFK assassination investigation documents to uncover latent themes, sentiment trends, and hidden relationships not obvious through traditional reading and historical methods. These documents originate from multiple government agencies and span decades of sensitive Cold War-era intelligence, policy discussions, and internal memos.

Due to the sheer volume of material—comprising tens of thousands of pages—manual review is infeasible. Furthermore, the documents exhibit considerable heterogeneity in format and quality, with most stored as scanned PDFs rather than text. As a result, direct application of NLP techniques is not immediately possible. To address this, the project incorporates an OCR preprocessing step to convert the non-textual PDFs into machine-readable form. The ultimate objective of this project is to build a scalable pipeline that enables semantic search, narrative discovery, and historical insight extraction over this highly dimensional, historically significant dataset.

Data

The primary dataset consists of the publicly released JFK archive, which includes a plethora of declassified documents pertaining to the assassination investigation. These documents are distributed in bulk as ZIP files containing scanned PDF images. The documents vary widely in structure, content, and quality—ranging from formal agency reports to handwritten memos and meeting transcripts.

For Milestone 1, a small subset of these files was selected to demonstrate the feasibility of OCR preprocessing and to begin exploring preliminary NLP techniques. The documents used for this phase were chosen to reflect a representative sample of formats and scanning quality. Subsequent stages of the project will apply topic modeling, entity extraction, and sentiment analysis to larger portions of the archive to reveal patterns and thematic structures embedded in the corpus.

Methods

For Milestone 1, a small subset of JFK investigation documents was used to prototype the preprocessing and NLP pipeline. These documents were packaged as scanned PDFs inside ZIP files and required several stages of processing to become usable for NLP tasks.

Evan Dartez
CMPS 6730

This pipeline consisted of the following components:

1. ZIP File Handling and PDF Extraction
2. PDF to Image Conversion
3. Image Preprocessing
4. OCR with Tesseract
5. Basic NLP and Exploratory Analysis

Rather than extracting all files to disk, PDFS were read in-memory from compressed ZIP archives using Python's zipfile module. This allowed for efficient processing of large archives without unnecessary file system overhead. Each page of each PDF was then converted to an image using the pdf2image library with a resolution of 200 dpi. Each image was converted to grayscale using Pillow to reduce noise. Text extraction was performed using python's wrapper for the Tesseract OCR engine. To improve speed, OCR was parallelized with ThreadPool allowing multiple pages to be processed concurrently. Progress was monitored using tqdm to provide real-time feedback. Once the OCR text was collected, a set of preliminary NLP experiments were performed:

- Word Frequency Analysis: Text was tokenized, lowercase, and filtered for English stopwords using nltk. A counter was then used to identify the top 20 most frequent words
- Bigram Collocation Analyis: The top 20 two-word phrases were computed with nltk.bigrams to identify common phrase structures and repeated name references.
- Document Length Statistics: Word counts per page were analyzed to compute average, minimum, maximum, and standard deviation of document lengths.
- Lexical Richness: The hapax legomena ratio was calculated to estimate lexical diversity in the dataset.

These methods provided a functional prototype pipeline for transforming raw scanned document archives into structured, analyzable text data suitable for downstream NLP tasks. Although limited in scope, this milestone demonstrated the feasibility of large-scale document analysis using OCR and basic NLP techniques, laying the groundwork for future expansion to the full corpus of JFK investigation files.

Because the OCR output was not manually validated, the NLP experiments should be interpreted as exploratory data analysis (EDA) rather than rigorous text modeling. Nonetheless, the insights gathered from this phase—particularly word usage patterns, phrase co-occurrence, and lexical structure—offer a meaningful first look at the corpus and help inform the design of subsequent modeling tasks such as topic clustering, sentiment tracking, and entity resolution.

Results

Following OCR and basic preprocessing, a subset of 201 pages from the JFK document archive was analyzed using a set of NLP techniques. These preliminary experiments were conducted to assess the structure, content, distribution, and linguistic features of the data, with the aim of informing future downstream modeling. The average page length was 563 words, with individual documents ranging from 65 to 1373 words. This variation reflects the diverse nature of the source material, which includes surveillance logs, memos, typed correspondence, and scanned forms. The standard deviation of document length was 267.8 words, indicating moderate variability in density across pages. The most frequent words included “david”, “halperin”, “page”, and “says”. Many frequent tokens appear to be surnames to persons of interest in the investigation such as Maurice Halperin. The Hapax Legomena ratio was 0.746, indicating a degree of lexical diversity. This would suggest that the corpus is semantically rich and well-suited for unsupervised topic modeling if the quality of the data is substantial. Bigram analysis revealed frequently occurring phrases such as “would like”, “maurice halperin”, “mexico city”, “reproduction issuing”, and “classified message”. Using spaCy’s pretrained model, over 800 named entities were extracted from the OCR text. The most common entity types were CARDINAL (n=180), ORG (n=170), PERSON (n=162), GPE (n=51). The most frequent named entities included Lo, Maurice Halperin, Mexico City, Prague, and John F. Kennedy. These point to intelligence targets, geographic nodes, and persons central to the investigation.

Discussion

While the Milestone 1 pipeline successfully demonstrated the feasibility of OCR-based NLP analysis on a subset of the JFK archive, several limitations emerged—chief among them being the inconsistent and frequently poor accuracy of the OCR output.

Evan Dartez
CMPS 6730

Upon quick manual inspection of the raw OCR text, it became immediately clear that many pages were significantly degraded. Some documents—particularly those in typewritten memo form—yielded interpretable text, but many others were marred by fragmented words, garbled symbols, and boilerplate clutter. These distortions are likely caused by poor scan quality, non-standard formatting, and the presence of handwritten or degraded content.

Because no formal OCR accuracy validation was performed, all results in this phase should be considered exploratory. However, rather than relying on manual sampling in future iterations, the project will incorporate automated lexical validation. This approach will assess the interpretability of OCR outputs by comparing them against reference dictionaries or language models, using metrics such as word validity, vocabulary match rates, and overall text coherence to identify low-quality pages.

To address the issues observed in Milestone 1, several improvements are planned. OCR output will be graded for linguistic plausibility using dictionary match rates, heuristics based on symbol ratios, token lengths, and character entropy will help identify unreliable pages early in the pipeline, a second OCR pass using TrOCR, a transformer-based model optimized for very noisy or handwritten text, will be applied to low-confidence pages, a SQLite database will track page-level OCR attempts and scores, allowing the pipeline to resume efficiently and reproduce only failed or low-confidence results, and post-processing of the OCR text to cleanup irrelevant text such as boilerplate phrases, headers, and incoherent lines.

Despite the noisy OCR, the preliminary results yielded interesting findings. Named entity recognition and bigram analysis highlighted recurring references to Maurice Halperin, a historical figure linked to Soviet espionage, and Mexico City, a known Cold War intelligence hotspot. The frequency and contextual use of these terms suggest that even early-stage outputs from the pipeline can surface historically relevant themes and individuals. These insights help validate the overall direction of the project and underscore its potential to reveal narrative patterns hidden across a vast and fragmented corpus.

Milestone 1 revealed both the challenges and promise of large-scale historical document analysis. With improvements to OCR validation and fallback processing, future phases will be well-positioned to scale the pipeline and extract high-quality insights from the full JFK archive.

Evan Dartez
CMPS 6730

Related Works

1. **Blei, Ng, & Jordan (2003) – Latent Dirichlet Allocation**
   This foundational paper introduced LDA, a generative model for discovering topics in large corpora. LDA is directly relevant to this project's goal of identifying latent themes in the JFK archive. While the original work assumes clean, structured text, this project deals with noisy OCR text, requiring additional preprocessing before topic modeling can be applied.

2. **Jockers (2013) – Macroanalysis: Digital Methods and Literary History**
   Jockers explores how computational methods like topic modeling and text mining can uncover historical trends in literature. This project builds on a similar philosophy, applying NLP to reveal patterns and narratives in a large-scale historical document collection. However, the JFK corpus presents unique challenges due to its origin in scanned intelligence files, not born-digital literature.

3. **Liu, Zhang, & Wang (2014) – Text Mining in the Digital Age: Challenges and Opportunities for Historical Documents**
   This study discusses OCR errors, linguistic drift, and format inconsistency in historical corpora. The challenges described align closely with the issues encountered in Milestone 1. Like this paper, the project adopts preprocessing strategies and validation techniques to make historical text usable for modern NLP tools.

4. **Riedl & Pohl (2015) – Named Entity Recognition in Historical Texts**
   This work highlights the limitations of applying pretrained NER models to old or noisy text. The findings are relevant to this project's NER task, as preliminary results showed a mix of usable and distorted named entity outputs. In the future, this project may explore fine-tuning or customizing NER models for improved performance.
5. **Kumar et al. (2025) – Evolving Techniques in Sentiment Analysis: A Comprehensive Review**
   This recent review evaluates how modern sentiment analysis models can be adapted to non-modern or noisy texts. The techniques described in the paper will inform future

phases of the project, particularly once higher-quality OCR data is available. Unlike most prior work, this project deals with an unusually large and noisy real-world dataset.

Division of Labor

This is an individual project. All data preprocessing, OCR implementation, NLP experiments, evaluation design, and reporting were completed by me.

Timeline

Date	**Task**
April 17	Build prototype pipeline, run NLP prelims
April 17	Report and submit Milestone 1
April 26	Implement OCR fallback and per-page quality scoring
April 29	Process full JFK archive or at least a very large subset.
April 30 - May 1	Apply topic modeling, sentiment analysis, NER, etc
May 2	Final Report and submission

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.