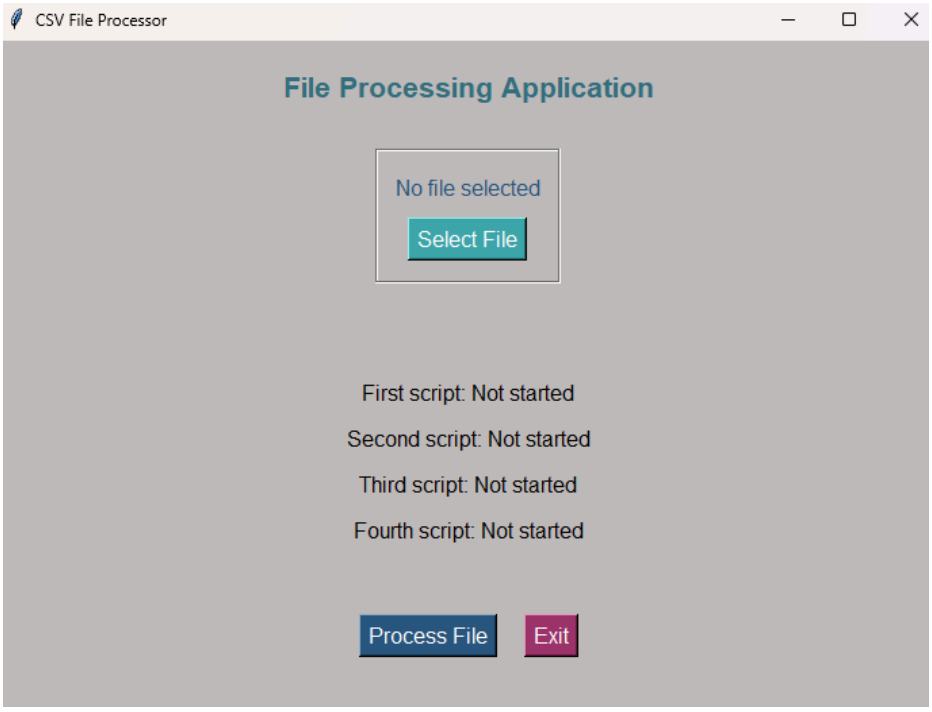
ARTIFICIAL INTELLIGENCE GROUP PROJECT

REFLECTIVE REPORT

SERGEJS KOPILS



Introduction

As the **Project Manager** for the **Adverse Media Monitoring and Client Risk Assessment System**, I was responsible for both **managing the team** and **developing Step 3** of the project. My management duties included organizing **weekly meetings**, maintaining the **GitHub repository**, assigning tasks, and ensuring deadlines were met. I documented key discussions, tracked progress, and provided feedback to keep the team aligned with project goals. Beyond management, I took a hands-on approach in **Step 3**, developing workflows for **data collection**, **web scraping**, **text processing**, and **risk evaluation**. These workflows played a key role in analyzing company-related risks and streamlining the project's technical implementation.

In **Step 3**, I implemented the **Google Custom Search JSON API** to retrieve URLs associated with specific company names. I managed **API rate limits** and optimized query structures to prevent service disruptions. For web scraping, I used **Requests** and **BeautifulSoup 4** to extract text from web pages, removing unnecessary HTML elements and ensuring clean data output. Text preprocessing was handled with **Regex (re)** to standardize and clean extracted content for further analysis. To handle large datasets efficiently, I incorporated **ThreadPoolExecutor**, allowing the simultaneous processing of multiple companies, which significantly reduced runtime. After processing, the data was structured using **Pandas** and exported as **CSV files** with proper formatting and encoding. The transition from the costly **GPT-4 API** to the more budget-friendly **Google Custom Search API** reduced the overall project costs to **€65**, while still maintaining data quality and meeting project requirements.

In addition to development, I conducted multiple **tests** to ensure the robustness, accuracy, and efficiency of the workflows. These tests were **not integrated into the main code** but provided valuable insights for refining the system:

- **Text Extraction Test:**
    - Verified that extracted text from various web pages was clean, free of noise, and structurally consistent.
- **Named Entity Recognition (NER) Test (using spaCy):**
    - Tested entity recognition to ensure company names and key terms (e.g., "ORG," "PRODUCT," "GPE") were accurately detected in the extracted text.
- **Sentiment Analysis Test (using TextBlob):**
    - Analyzed extracted text for sentiment classification (Positive, Negative, Neutral) to understand content tone.
- **Domain Frequency Analysis Test:**
    - Identified the most frequent domains in search results and cross-referenced them with high-risk indicators.
- **Duplicate Entry Test:**

    - Grouped duplicate company entries and analyzed inconsistencies in URLs, risk levels, and keywords.
- **Google Knowledge Graph API Test:**
    - Retrieved company-related metadata and validated the relevance and accuracy of contextual information.
- **NLP Analysis Test (using spaCy)::**
    - Applied **spaCy's NLP pipeline** to analyze extracted text for **Named Entity Recognition (NER)** and keyword presence.
    - Validated whether **specific company names appeared in the detected entities**, ensuring relevance.

These tests allowed me to validate the integrity of data extraction, assess the reliability of risk scoring mechanisms, and refine the overall workflow. They also highlighted areas where additional adjustments were needed, particularly in text preprocessing and keyword filtering.

Throughout the project, several challenges arose. Managing **API costs and quotas** required careful planning and continuous monitoring, especially during the transition away from **GPT-4 API**. Extracting text from diverse webpage structures often resulted in inconsistencies, requiring iterative adjustments to parsing logic. False positives in keyword matching also posed difficulties, as some terms appeared in misleading contexts. Balancing technical development with project management responsibilities demanded strong prioritization and time management skills, but regular progress reviews and clear communication helped address these challenges effectively.

Summary

In summary, my contributions to **Step 3** included developing workflows for **API integration**, **web scraping**, **text preprocessing**, and **risk evaluation**, supported by separate **code tests** to validate each component. I maintained clear documentation, ensured structured task management through **GitHub**, and coordinated team efforts via **weekly meetings**. The decision to transition from **GPT-4 API** to **Google API** reduced costs to **€65**, aligning with project constraints while delivering reliable results. This project strengthened my abilities in **project management**, **API integration**, **data processing**, and **technical troubleshooting**, equipping me for future challenges in data-driven projects.

Link to test codes:

API GTP-4

Testing libraries

Words 659