1 POC

The Proof of Concept (POC) regarding the OCR pipeline establishes the technical feasibility of transforming a large corpus of scanned legal agreements into a structured dataset optimized for computational social science. The proposed system is anchored in a deep learning-based Optical Character Recognition (OCR) framework designed to digitize documents with high geometric fidelity, subsequently enabling complex algorithmic analysis to determine critical metrics including document length, the prevalence of boilerplate language, and the substantive scope of international cooperation.

1.1 Determining documents length

The core extraction layer is built upon the `docTR` (Document Text Recognition) library, utilizing a pre-trained predictor to process PDF documents ingested from the project's directory structure. To ensure the integrity of the analysis, the workflow incorporates a preprocessing step that programmatically discards the first page of every file, effectively stripping away the administrative cover sheets and metadata often appended during the download process. Unlike simple text-dumping tools, the model generates a hierarchical JSON output for each agreement, mapping the document's geometry into a nested structure of pages, blocks, lines, and words. This geometric preservation is crucial for the subsequent metadata extraction task, where a custom algorithm traverses the JSON hierarchy to calculate document volume. By iterating through every text block and summing the token counts at the line level, the system produces a highly accurate word count that ignores whitespace anomalies, resulting in the dataset which facilitates precise quantitative comparisons of agreements (Task 1).

1.2 Identifying areas of cooperation

Building upon this digitized foundation, the module implements a sophisticated NLP architecture to analyze the legal text. For the analysis of recurring clauses (Task 9), the system distinguishes between standardized "boilerplate" templates and bespoke diplomatic terms by leveraging the visual block structure detected by the OCR engine. To filter out noise such as page numbers, headers, and artifacts, the system discards any text blocks containing fewer than ten words. The remaining substantive clauses are encoded into dense vector embeddings using the `all-MiniLM-L6-v2` Sentence-Transformer, a model optimized for semantic similarity tasks. These embeddings are then subjected to HDBSCAN clustering with a Euclidean metric and a minimum cluster size of two, allowing the system to detect even minimal repetitions of legal phrasing. The classification logic is derived from these cluster assignments: clauses appearing in over 80% of the document corpus are categorized as standard diplomatic protocol, whereas clauses that fail to form clusters (labeled as -1) are identified as unique, custom-drafted terms specific to a particular negotiation.

1.3 Analyzing frequency of recurring clauses

In parallel with the structural analysis, the POC addresses the identification of cooperation areas (Task 1) through a Zero-Shot Classification approach powered by the `facebook/bart-large-mnli` model. This methodology allows the system to categorize agreements into specific sectors—such as "Green Energy," "Culture & Arts," or "Trade & Economic Development"—without the prohibitive requirement of a manually labeled training dataset. Recognizing the input constraints of Transformer-based models, the pipeline reconstructs the full text from the JSON output and applies a truncation strategy, limiting the input to the first 3,000 characters. This window is strategically selected to encompass the Preamble and the initial “Scope of Cooperation” articles, where the binding intent of the agreement is typically articulated. The model performs a multi-label classification on this truncated text, assigning tags only when the confidence score exceeds a 0.5 threshold. The resulting data is aggregated into a final CSV report, confirming that the infrastructure is fully operational and ready for the large-scale ingestion of the complete legal dataset.