Project Report

The preprocessing pipeline focused on cleaning and preparing the dataset for efficient and accurate summarization. The steps include:

- **Text Cleaning**: Removal of extra spaces, special characters, and non-informative symbols.
- **Case Normalization**: Conversion of all text to lowercase for consistency.
- **Stopword Removal**: Filtering out common but non-informative words (e.g., "the", "and") using `nltk` stopword lists.
- **Tokenization**: Splitting sentences and words using `nltk` or `spacy` for structured processing.
- **Lemmatization**: Converting words to their base form to reduce vocabulary size while preserving meaning.

The extraction process aimed to identify the most relevant words from documents: The steps include:

- **Word Counter:** Counter was used to get the frequency of occurrence for the words in the sequence.
- **Visualization**: Plotted the most frequently used words.
- **Extraction**: I used Spacy (en_core_ web_sm) to extract entities and a function that uses regular expressions to capture important details like date, amounts and numbers from the text.

3. Summarization

A hybrid summarization approach was applied:

- **Extractive Summarization**: I used Sumy TextRank model to select the important parts of a sentence directly from the original text without altering the wording.
- **Abstractive Summarization**: I used transformer-based model - T5 to generate human-like summaries that paraphrase and compress the information.
- **Evaluation**: ROUGE metrics were used to measure the quality of generated summaries compared to reference summaries.