DocScribe: Intelligent Document Analysis Agent

Brief Project Report

1. Project Overview & Methodologies

DocScribe is an AI-powered agent designed to extract, summarize, and synthesize insights from unstructured text. Acting as both a *Research Agent* and *News Aggregator Agent*, it integrates multiple **Natural Language Processing (NLP)** techniques to deliver concise and relevant intelligence.

Data Preparation

The system uses the *Kaggle News Category Dataset* accessed via the Kaggle API. Headlines and short descriptions are merged into a `full_text` field, then processed through:

- Tokenization, lowercasing, and stop-word removal.
- Lemmatization for consistent word forms.

Information Extraction

For information extraction, DocScribe employs a layered approach. Custom regular expressions identify structured patterns such as dates, monetary values, and percentages. SpaCy's Named Entity Recognition (NER) captures entities like people, organizations, and locations, while a transformer-based model from Hugging Face provides deeper, context-aware recognition, including less common categories such as events and products.

2. Results, Challenges, and Model Performance

Performance Highlights

- **Regex**: Highly precise for fixed patterns but brittle with variations.
- **SpaCy NER**: Robust and adaptable to general text.
- **Transformer NER**: Superior contextual accuracy, handling complex entities effectively.
- **Summarization**: Extractive maintained factual accuracy but lacked brevity. Abstractive (*T5-small*) consistently produced concise, coherent summaries, even from short texts.

Evaluation Method

Due to the lack of human-written reference summaries, quantitative evaluation (e.g., ROUGE scores) was not possible. Instead, qualitative review assessed:

- Coherence of generated text.
- Relevance to the query.
- Conciseness and readability.

Challenges

- Short, ambiguous news snippets reduced context for accurate extraction.
- Transformer models required high computational resources, necessitating GPU acceleration.
- Varied data quality affected summarization consistency, favoring abstractive methods for shorter inputs.

3. Use Case & Architecture

Primary Use Case: *Geopolitical Insight Monitor* — assisting analysts in rapidly identifying events, entities, and trends from large news corpora.

The system then retrieves documents using semantic search techniques, ranking them by relevance and recency. Depending on the task, DocScribe orchestrates a sequence of tools. For example, in topic summarization, individual article summaries are generated and then synthesized into a single unified narrative

A two-tiered memory system supports these operations. Short-term contextual memory maintains conversational flow within a session, allowing for natural follow-up queries. A long-term semantic cache stores processed data, document embeddings, extracted entities, and summaries, enabling trend analysis and faster responses to recurring queries.

Workflow

1. **Query Parsing & Intent Recognition** – Interprets user requests (summarize, extract entities, extract financials) and identifies relevant parameters (topic, entities, timeframe).
2. **Intelligent Document Retrieval** – Uses semantic search with vector embeddings, prioritizing relevance and recency.
3. **Dynamic Tool Orchestration** –
    - **Topic Summarization**: Generates article summaries → synthesizes them into a unified narrative.
    - **Entity & Timeline Extraction**: Identifies people, organizations, and dates → aggregates and ranks them.
    - **Financial Analysis**: Extracts monetary values, percentages, and associated dates → presents results with context.

Memory System

- **Short-Term Memory**: Maintains session context for natural follow-up queries.
- **Long-Term Semantic Cache**: Stores embeddings, extracted entities, and summaries to support trend analysis and improve efficiency over time.