Information retrieval system

Goal

Input: a PDF or a **document** or corpus of information, and a **query**
Output: **snippet**(s) of relevant information and the formulated query based on the snippets.

Methodology

We divide the project into functional abstract classes

1. EmbeddingGenerator: embed([string]) -> [[float]]; responsible for converting text into an embedding
2. EmbeddingFactory: caches a list of embeddings for fast retrieval (e.g., uses MilvusDB, or elastic search).
3. EmbeddingComparator: compare(e1, e2); responsible for comparing two embedding
4. DocumentParser: responsible for parsing PDFs and returning list of strings that need to be embedded - chunking
5. RetrievalClass: uses retrieval and answer strategy to obtain the result
6. RetreivalStrategy: uses the strategy pattern and the previous classes to (e.g., comparator) with different ways to.
7. AnswerStrategy: given list of relevant documents and a query, formulate the answer

Timeline

Deadline: ~5th of April

Progress report: 21st of March

- Project structure & interfaces
- Experiments
- Responsibilities
- Literature review