# NLP TutorBot v9.1 — Design & Architecture Document

## 1. High-Level Architecture

PDF Documents –> Preprocessing –> Section Extraction –> Summarization | User Selects Topic –> Topic Summary + YouTube Video Links | User Asks Question –> Contextual Chunk Retrieval –> Answer Generation

## 2. Major Components

| Component | Description |
| --- | --- |
| PDF Preprocessing | Extracts headings and corresponding text sections from PDFs |
| Summarization | Uses facebook/bart-large-cnn to summarize each section |
| Embedding | Uses all-MiniLM-L6-v2 SentenceTransformer to generate embeddings |
| Question Answering | Uses: 1. deepset/roberta-base-squad2 for extractive QA 2. google/flan-t5-large for generative QA |
| YouTube Search | Uses youtubesearchpython or YouTube Data API |
| UI | Built using gradio |

## 3. Detailed Workflow

### A. PDF Parsing & Section Extraction

* PDF files are read from specified folder.
* Headings extracted using Regex pattern: r'^(\d+(\.\d+)\*)\s+.+'
* Sections extracted between headings.

### B. Summarization

* Section content summarized using: facebook/bart-large-cnn
* Summaries stored for each section.

### C. Embedding & Chunking

* Content split into ~1500 character chunks.
* Sentence embeddings generated via: SentenceTransformer('all-MiniLM-L6-v2')

### D. Question Answering Logic

1. **Chunk Retrieval**
   * Embed user question.
   * Calculate cosine similarity with content chunks.
   * Select most relevant chunk.
2. **Generative QA (Flan-T5)**
   * Prompt-based explanation using: google/flan-t5-large
3. **Extractive QA (Roberta fallback)**
   * Span extraction using: deepset/roberta-base-squad2

### E. YouTube Search

* Search YouTube for selected topic via: youtubesearchpython
* Displays top 3 video links.

### F. Gradio UI

* Interactive interface with:
  + Topic dropdown
  + Summary display
  + YouTube links display
  + Question answering

## 4. System Design Summary

| Module | Library |
| --- | --- |
| PDF | PyPDF2 |
| Summarizer | facebook/bart-large-cnn |
| Embedding | sentence-transformers |
| QA Extractive | deepset/roberta-base-squad2 |
| QA Generative | google/flan-t5-large |
| YouTube | youtubesearchpython |
| UI | gradio |

## 5. Key Features

* Automatic topic extraction
* Section summaries
* Semantic search for QA
* Explanation-style answers
* YouTube integration
* Fully local processing (except YouTube)

## 6. Future Improvements

* YouTube API v3 integration
* Multiple PDF knowledge base
* Better heading extraction with layout parser
* Chat history support
* Deploy as app or web service

NLPBot\_Refined.py is working code and C:\Users\sadas\PycharmProjects\NLP\_TutorBot\NLP\_TutorTrial.py is latest code tried with google api

C:\Users\sadas\PycharmProjects\NLP\_TutorBot\NLPBot\_ClaudeAI.py is old code compiling and working, but output is not precise. Should refine.