CSE299: Junior Design Project

Chatbot using Retrieval-Augmented Generation (RAG)

RAG Pipeline

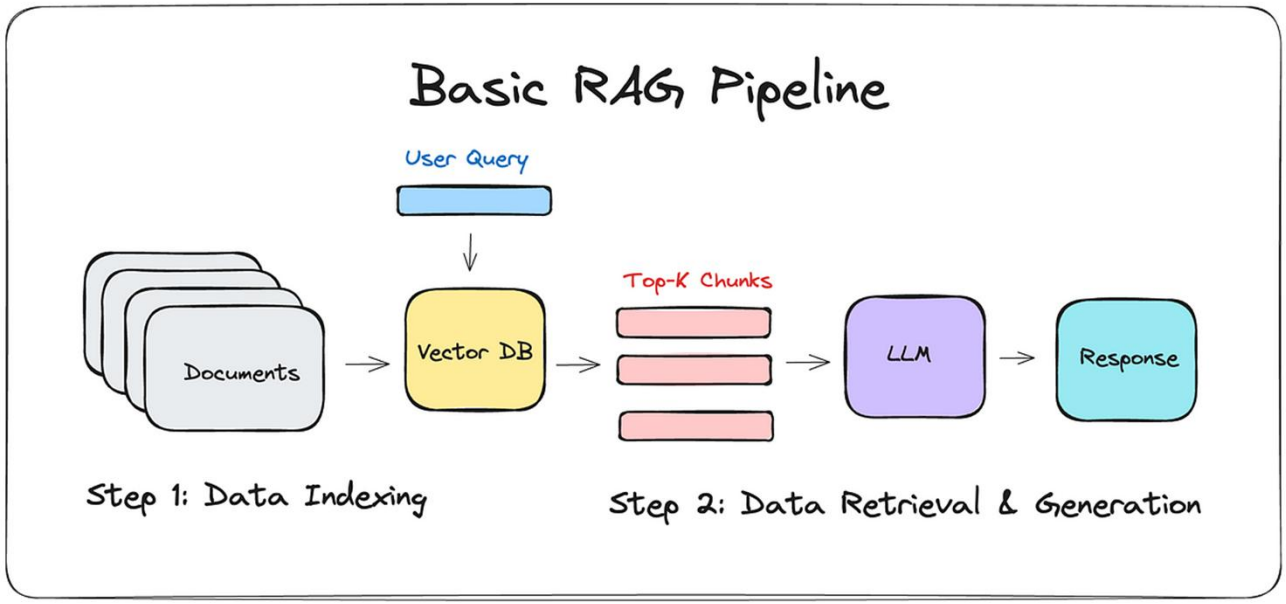

Figure: Overall RAG pipeline consisting of two main phases: **Data Indexing** and **Data Retrieval & Generation**.

RAG Pipeline

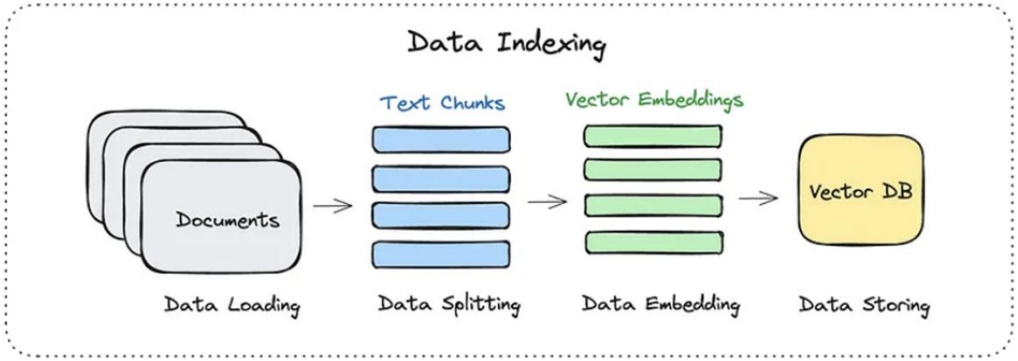

Figure: Data indexing

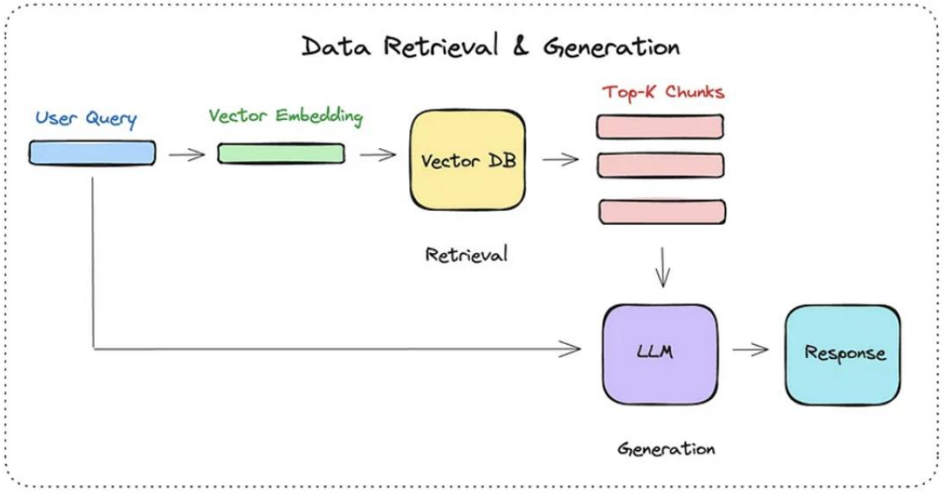

Figure: Data retrieval & generation

RAG Components & their Usage

- **Document Loader**: To load a document (PDF, doc, txt, etc.)
- **Text Splitter**: Split the document into smaller text chunks
- **Embedding Model**: Compute embeddings of the text chunks
- **Vector Database**: Required to store the embeddings
- **Query**: Question provided by the user
- **Retrieval**: Retrieves the "*k*" most similar text chunks based on the user query
- **LLM**: Generates final response based on the retrieved text chunks

Some Applications of RAG Chatbots

- Question Answering
- Summarization
- Human-Machine Conversation
- Traditional tasks (Recommendations, classification, search)
- Generation (based upon retrieved content)

Other Remarks

- Nice Graphical User Interface (GUI)
- Easy to use
- Quantitative/Qualitative evaluation of the product
- Fastness – Use of GPU
- Hosting your model - https://huggingface.co/
- Make use of version controlling – github or bitbucket

Key Challenges & Failure Points

- System Complexity
- Noise (good/bad)
- Gaps between retrievers and generators
- Increased context size (opportunity?)
- Performance (latency, suboptimal, etc.)
- Still reliant on data (missing, formatting, cleaning)
- Testing is hard (and on that note…)

Overall Marking

Grading Tools Used for Grade Assessment	Weeekly reports						Proposal [CO4]	Final			X	Y
	W1	W2	W3	W4	W5	Weekly progress [CO1]		Final Project Demo [CO2]	Final Presentation + Viva [CO5]	Final Report [CO3]	Total	Letter
Respective Percentages (Out of Total 100%)						40%	15%	10%	20%	15%	100%	Grade
Marks for Respective Grading Tools	10	10	10	10	10	50	15	10	20	15		

Week	Date	Tasks
1		Group formation
2		Informal Project Proposal Presentation
3		Weekly Update 1 to Canvas + Progress presentation
4		Progress presentation
5		Weekly Update 2 to Canvas + Progress presentation
6		Proposal presentation + Proposal report submission to Canvas
7		Weekly Update 3 to Canvas + Progress presentation
8		Progress presentation
9		Weekly Update 4 to Canvas + Progress presentation
10		Progress presentation
11		Weekly Update 5 to Canvas + Progress presentation
12		Final presentation + Demonstration + Viva
13		Project Submission to Canvas (PPT, Code, Final report)

Useful Resources

- https://medium.com/@drjulija/what-is-retrieval-augmented-generation-rag-938e4f6e03d1
- https://www.openxcell.com/blog/rag-pipeline/
- https://www.analyticsvidhya.com/blog/2023/10/rag-pipeline-with-the-llama-index/
- https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2

THANK YOU