**1. Introduction**

Implementation of RAG (Retrieval-Augmented Generation)[1] for QA NLP chatbot which can master science QA.

**2. Dataset**

context, question, answer and its description in the form of text.

**3. Database**

Due to memory contraints and high computation power needed to calculate vector and index those vectors for similarity search. I went for vector db which does the same job but in less memory needed and we can calculate in distributed manner also.

I used weaviate(https://github.com/weaviate/weaviate) Vector DB. With 768 embedding.

**4. Data Prepartion:**

Dataset used is KonstantyM/science\_qa (<https://huggingface.co/datasets/KonstantyM/science_qa>). Title and text were used in the form of 100 words to be vectorized and putted into the vector db.

**5. RAG Context Encoder**

Context encoder as "facebook/dpr-ctx\_encoder-multiset-base"which is pretrained encoder developed and trained by meta.

**6. RAG Retriver**

Vector similarity search between context vectors and question vector with cosine similarity in weaviate db with help for custom similar knowledge retriver.[2]

**7. RAG Generator**

QA RAG generator uses "valhalla/bart-large-finetuned-squadv1" from hugging face which is finetuned on squadv1 dataset which is famous for its open domain QA data quality and used for benchmarks.

**References**

[1] https://arxiv.org/pdf/2005.11401.pdf

[2] https://github.com/huggingface/transformers/blob/main/examples/research\_projects/rag/use\_own\_knowledge\_dataset.py