LlamaIndex and Index retrival from MongoDB:

The basic steps of this demo are:

Get data from a csvfile into a Mongo database

Using LLama3.2 offline model to generate the LLM Nosql query

Index the data using LlamaIndex. This will use HUggingFacemodel "all-MiniLM-L6-v2" under the hood and convert your text into vector embeddings.

Store the embedded data back into MongoDB. LlamaIndex will do this for you automatically.

Create a Vector Search Index in MongoDB. This is a manual step that needs to be performed in the MongoDB UI.

Enter the Python to answer questions about the data..

Choose how to display the results.

Prerequisites:

*Create cluster and Set up environment variables

Copy the connection string (make sure you include your password) and put it into a file for mongodb url

*Set up a python virtual environment and install dependencies

* Create appropriate mongoatlas index

```
{
  "mappings": {
    "dynamic": true,
    "fields": {
      "embedding": {
        "type": "knnVector",
        "dimensions": 384,  # all-MiniLM-L6-v2 has 384 dimensions
        "similarity": "cosine"  # or "euclidean"/"dotProduct"
      }
    }
  }
}
```

Model:

Centrally python is connected with LLM , mongoDB Atlas . The LLM model generates the query for mongoDB(Nosql query), which inturn is passed onto the collection find to get the matching documents.

Llama-index is also used to perform query search.

Challenges:

Additionally vector indices are added to the documents to perform llama_index search unfortunately the embedding model and limited resource outputs limited documents based on the vector scores. Which can further be enhanced by fine tuning the embedding model.