# Lab Summary: Getting Started with Vector Search and Embeddings

## Lab Info

Title: Getting Started with Vector Search and Embeddings  
Learning Path: Advanced: Generative AI for Developers  
Platform: Google Cloud Skills Boost  
Duration: 1 hour 30 minutes

## Lab Objectives

- Create a Vertex AI Notebook instance and run the lab notebook.  
- Generate text embeddings using a pre-trained model.  
- Upload the embeddings to Google Cloud Storage.  
- Create and deploy a vector search index using Matching Engine.  
- Run semantic queries to retrieve the most relevant results.

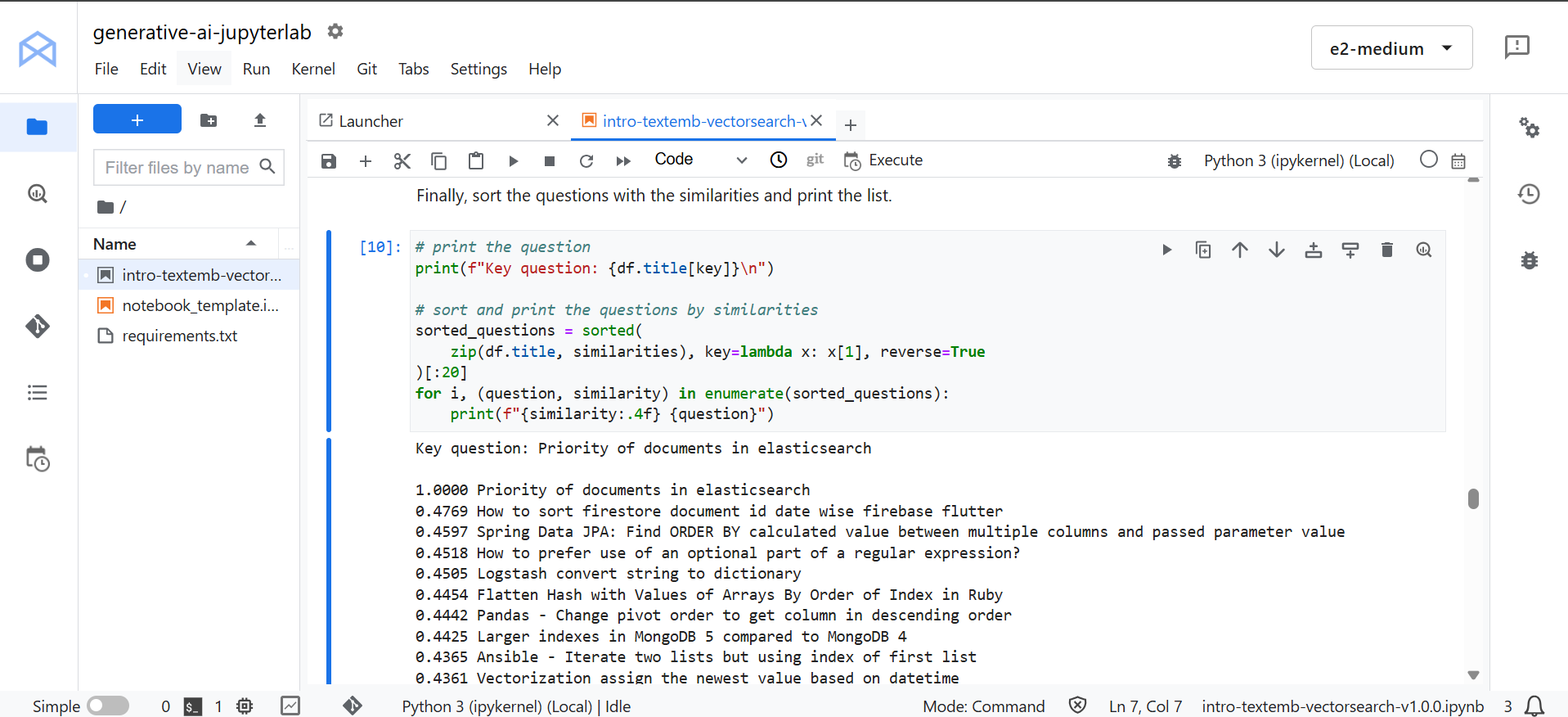
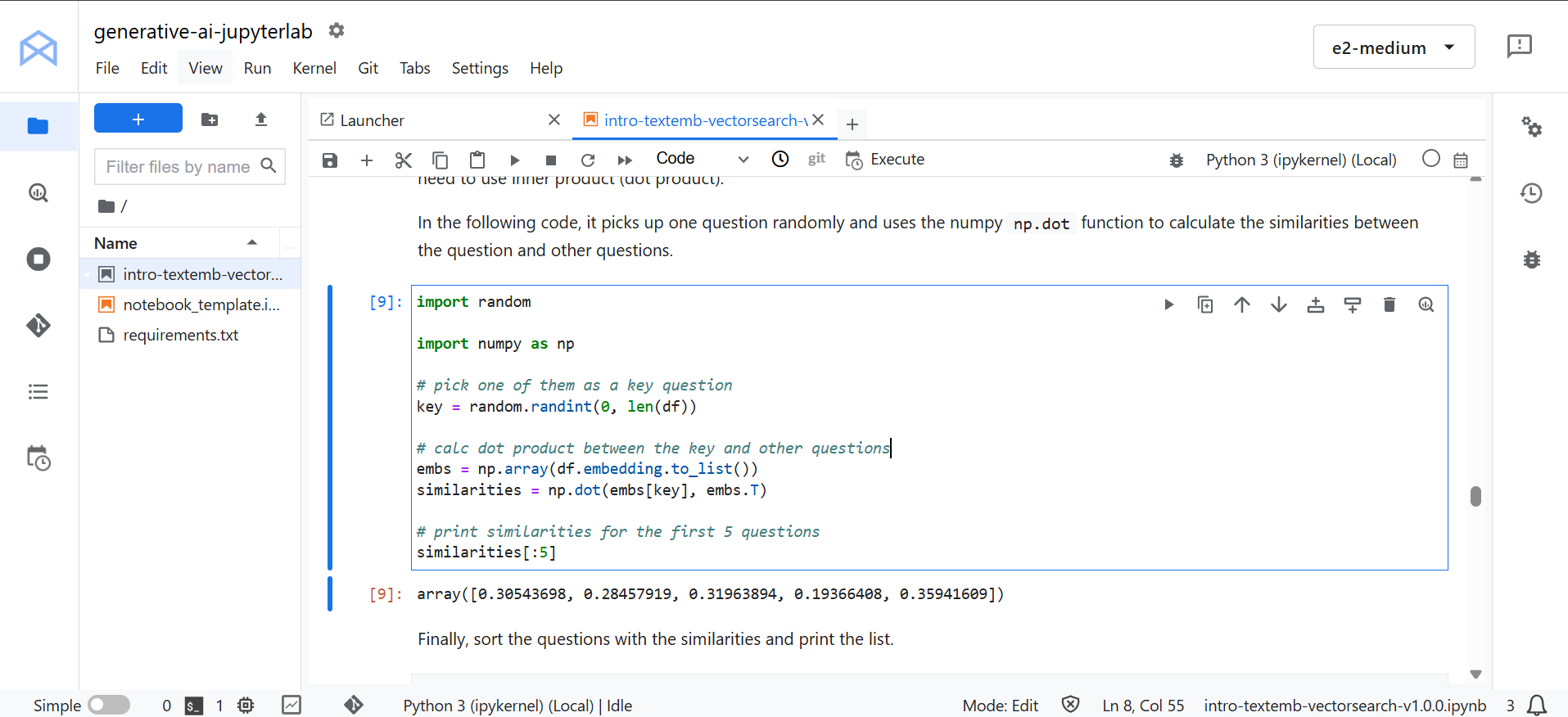
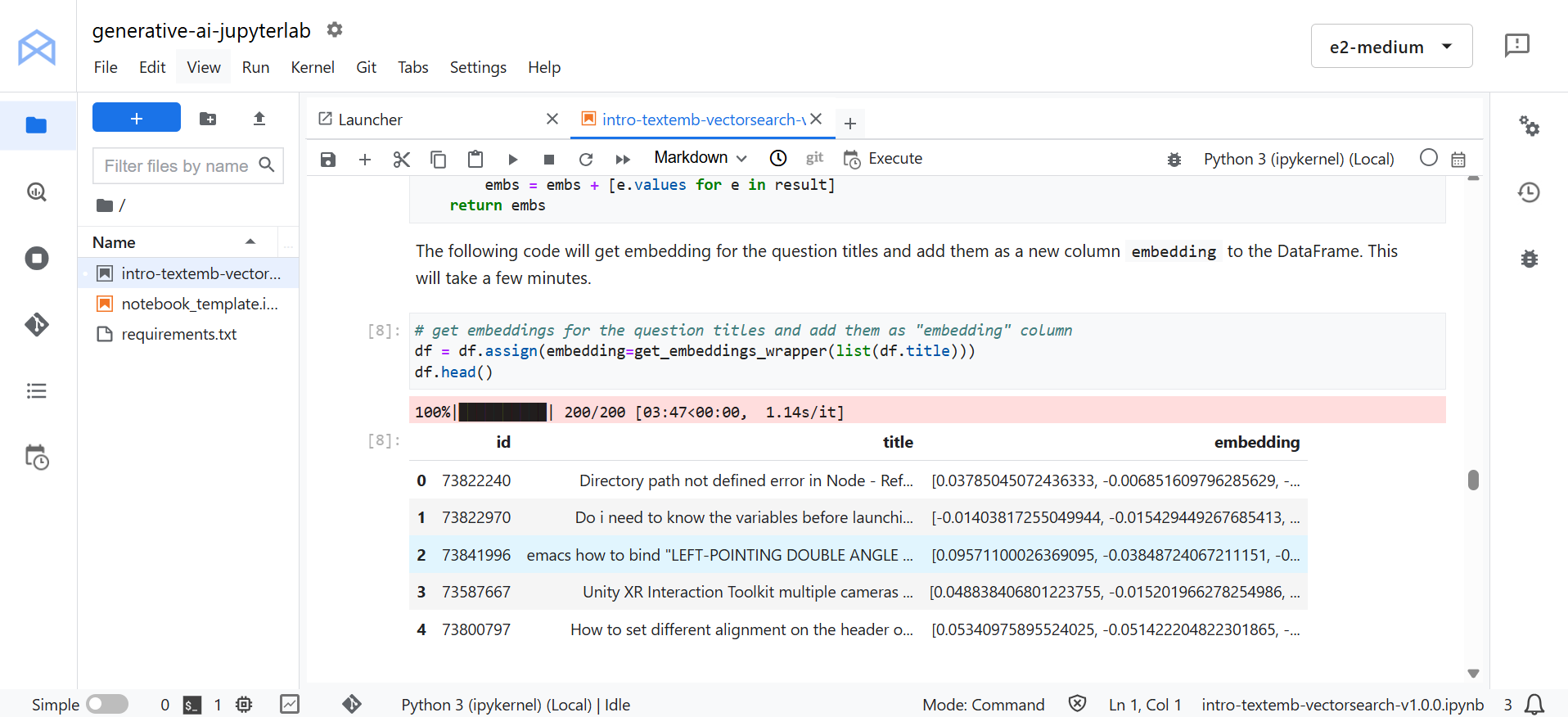
## Lab Content (Detailed)

### Task 1 – Open the notebook in Vertex AI Workbench

I began by launching a Vertex AI Workbench environment and cloning the provided lab notebook.  
This step was essential to execute the code and interact with the cloud-based embedding and vector tools.

### Task 2 – Generate embeddings

I used Vertex AI’s Embeddings model to convert question titles into numerical representations (embeddings).  
Then I saved the embeddings in a JSON file and uploaded it to a newly created Google Cloud Storage bucket.  
This process involved running predefined cells in the notebook and using Python to assign the embeddings column to the dataframe.



### Task 3 – Create and deploy an index

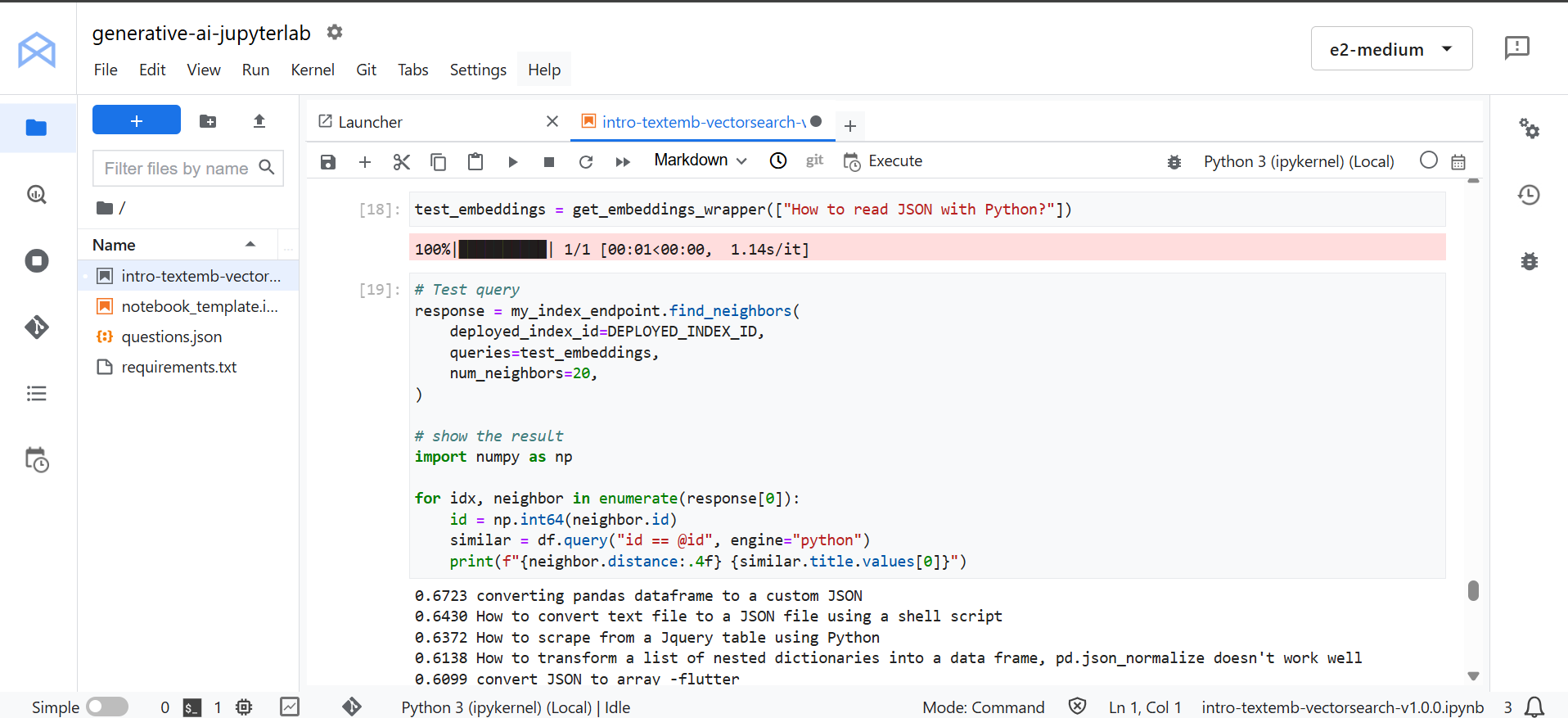
Using the `aiplatform` library, I created a Matching Engine Index by calling the `create\_tree\_ah\_index` function.  
Key parameters like dimensions (768), distance measure (DOT\_PRODUCT\_DISTANCE), and approximate neighbors count (20) were configured.  
The deployment process took about 20–30 minutes to complete.

A screenshot of a computer

AI-generated content may be incorrect.

### Task 4 – Run a query

I randomly selected a question from the dataset to act as a key query, then computed the similarity between it and other questions using NumPy’s dot product.  
Finally, I sorted and printed the questions based on similarity scores, showing the closest semantic matches.  
This highlighted how vector search can go beyond keyword matching by understanding the meaning behind text.



## What I Learned

- I learned how to generate embeddings from text and represent data semantically.  
- I used Google Cloud Storage to manage and access my JSON files in the notebook.  
- I successfully created and deployed a vector search index with Google’s Matching Engine.  
- I practiced querying the index using similarity scores to retrieve semantically relevant results.  
- I gained practical experience in building intelligent search systems without writing APIs or backend code.  
- This lab enhanced my understanding of how vector search can be applied in real-world AI applications such as chatbots, recommendation engines, and semantic retrieval.