Week 14: Vector Databases

Agenda

In this session, we will discuss:

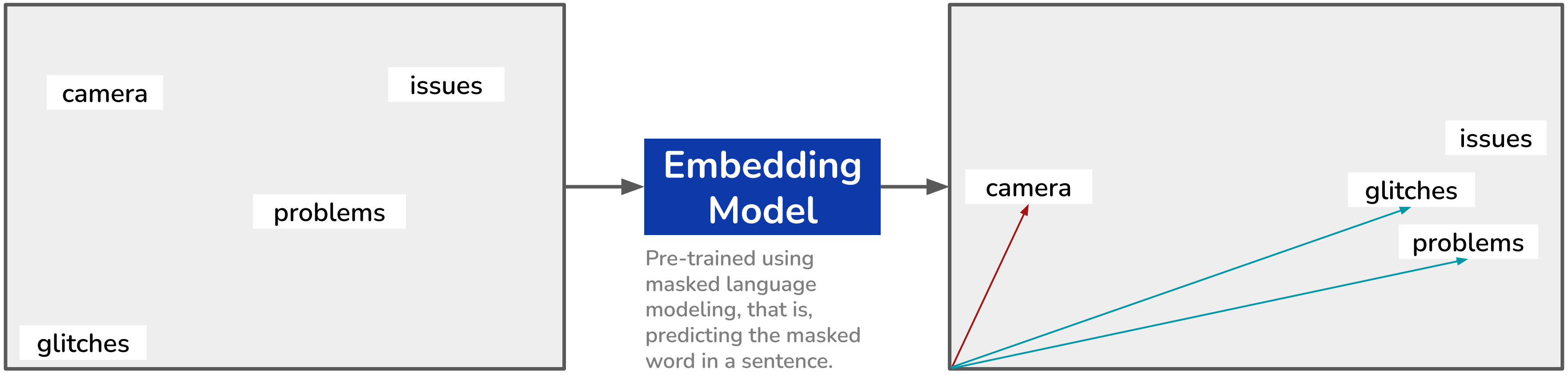
- Introduction to Embeddings
- Using Embeddings for Similarity
- Creating and Managing a Vector Database (Chroma)
- Using Vector Databases for Search

Representing Text

- The process of converting raw text data into a computer-readable format involves transforming it into numeric feature vectors.
- This conversion is known as text representation, which aims to capture both the linguistic information and the semantics of the text.
- Deep learning models are a popular method to create representations from input text referred to as *embeddings*.

Embeddings - An Introduction

- Embeddings are a type of word representation that allows words with similar meaning to have a similar representation.
- They capture semantic properties of words and relations with other words.



Pre-trained using masked language modeling, that is, predicting the masked word in a sentence.

Raw 2D Mathematical Space

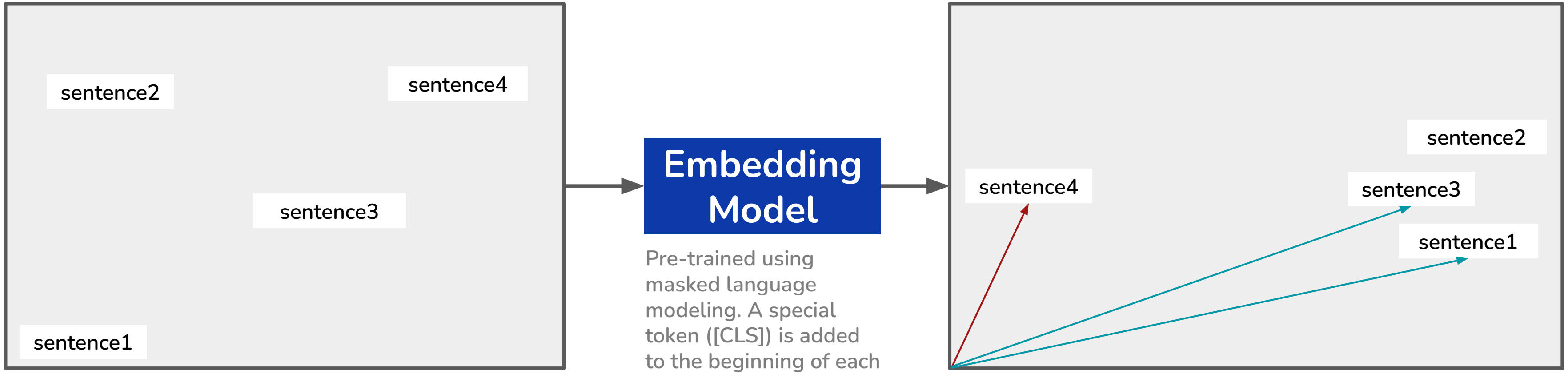
Each word is associated with a random list of two numbers. Words with similar meaning are scattered

Contextualized 2D Mathematical Space

Each word is associated with a specific list of two numbers. Words with similar meaning are clustered close to each other

Embeddings - An Introduction

- Sentence Embeddings are vector representations of whole sentences capturing their meaning.
- They are derived by averaging word embeddings or using specialized embedding models



Pre-trained using masked language modeling. A special token ([CLS]) is added to the beginning of each sentence during training.

Raw 2D Mathematical Space

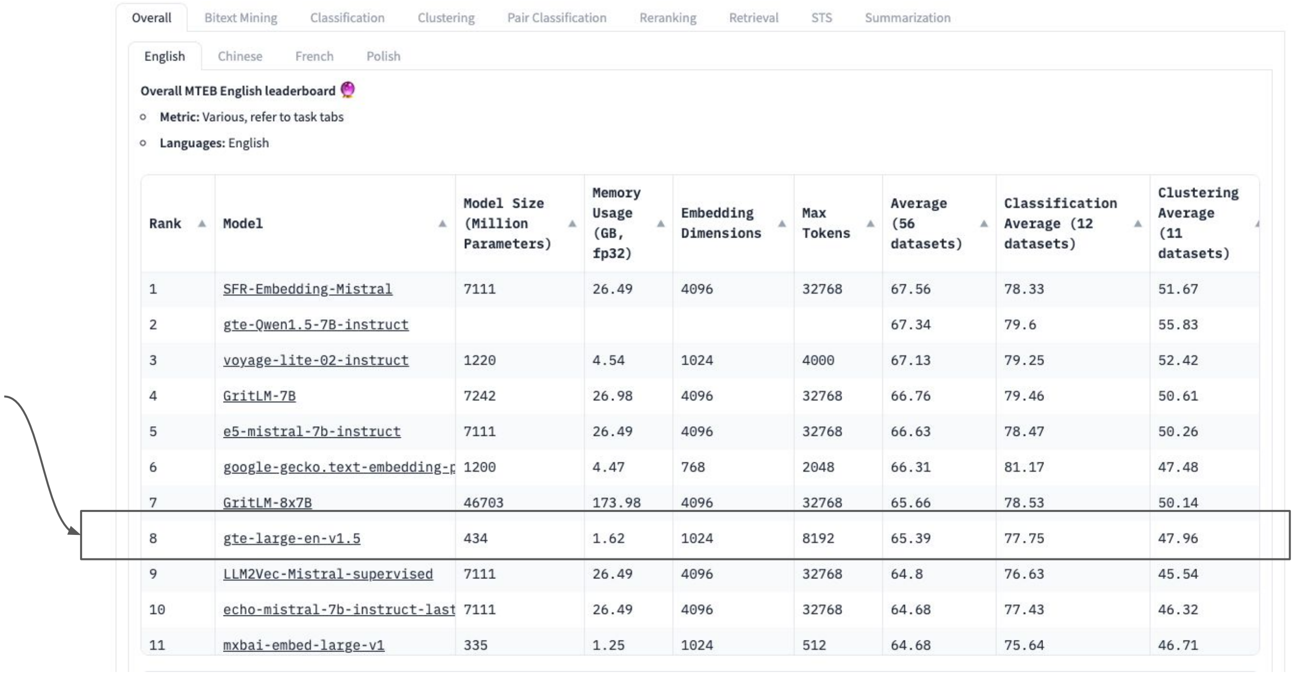
Each sentence is associated with a random list of two numbers. Sentences with similar meaning are scattered

Contextualized 2D Mathematical Space

Each sentence is associated with a specific list of two numbers. Sentences with similar meaning are clustered close to each other

Embeddings - An Introduction

Embedding models can be chosen depending on the downstream task (e.g., classification, summarization, ranking) using embedding leaderboards (for e.g., MTEB).



Overall MTEB English leaderboard 🔮

- **Metric:** Various, refer to task tabs
- **Languages:** English

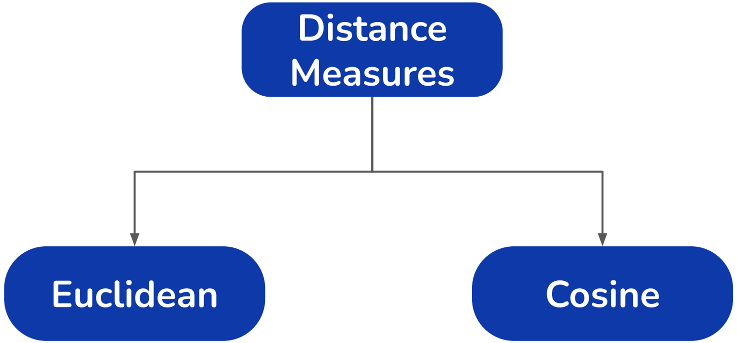
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)
1	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67
2	gte-Qwen1.5-7B-instruct					67.34	79.6	55.83
3	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42
4	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61
5	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26
6	google-gecko.text-embedding-p	1200	4.47	768	2048	66.31	81.17	47.48
7	GritLM-8x7B	46703	173.98	4096	32768	65.66	78.53	50.14
8	gte-large-en-v1.5	434	1.62	1024	8192	65.39	77.75	47.96
9	LLM2Vec-Mistral-supervised	7111	26.49	4096	32768	64.8	76.63	45.54
10	echo-mistral-7b-instruct-last	7111	26.49	4096	32768	64.68	77.43	46.32
11	mxbai-embed-large-v1	335	1.25	1024	512	64.68	75.64	46.71

Embedding model used in this course

https://huggingface.co/spaces/mteb/leaderboard

Using Embeddings for Similarity

A pair of texts is deemed to be similar if they are close to each other (i.e., less distant) in the embedding space.



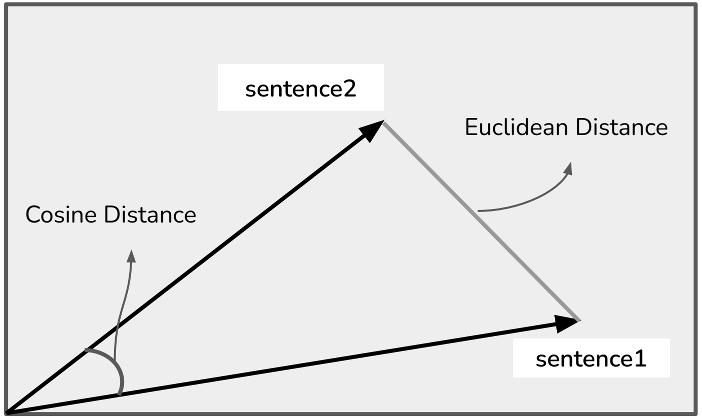
Distance Measures

Euclidean

Cosine

Straight line distance between two vectors

Difference between the angles of two vectors



Embedding Space

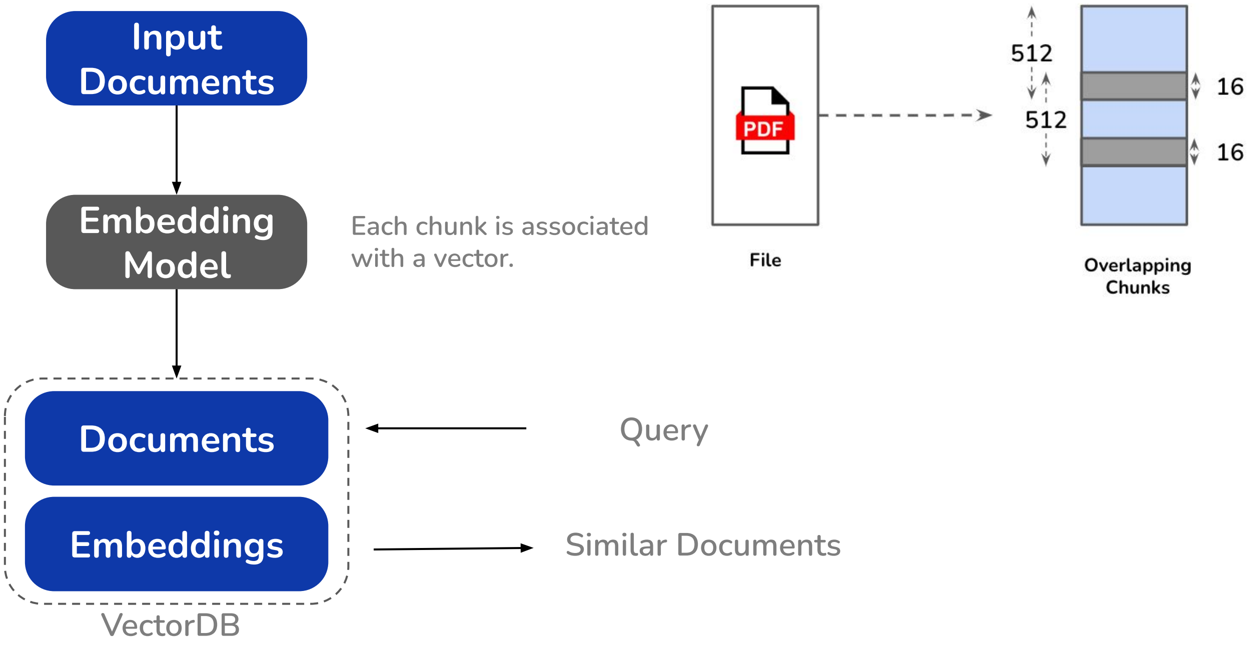
[Notebook]
embeddings_similarity.ipynb

Vector Databases

Vector databases are specialized in storing and retrieving vectors associated with unstructured data. Given input queries, the database can retrieve relevant documents using similarity search.

Input documents are split into chunks of a certain size.

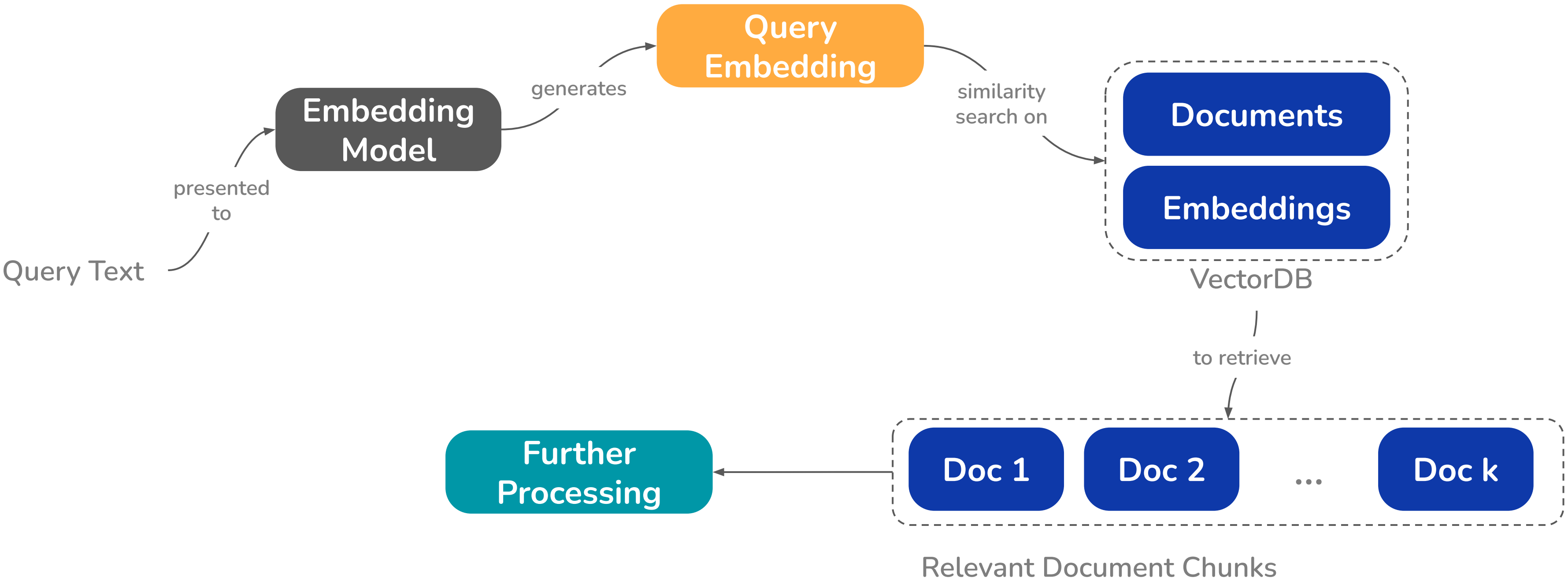
The vector database is pre-populated by indexing all the document chunks and the vectors created using the embedding. Indexes are organized into collections.



Each chunk is associated with a vector.

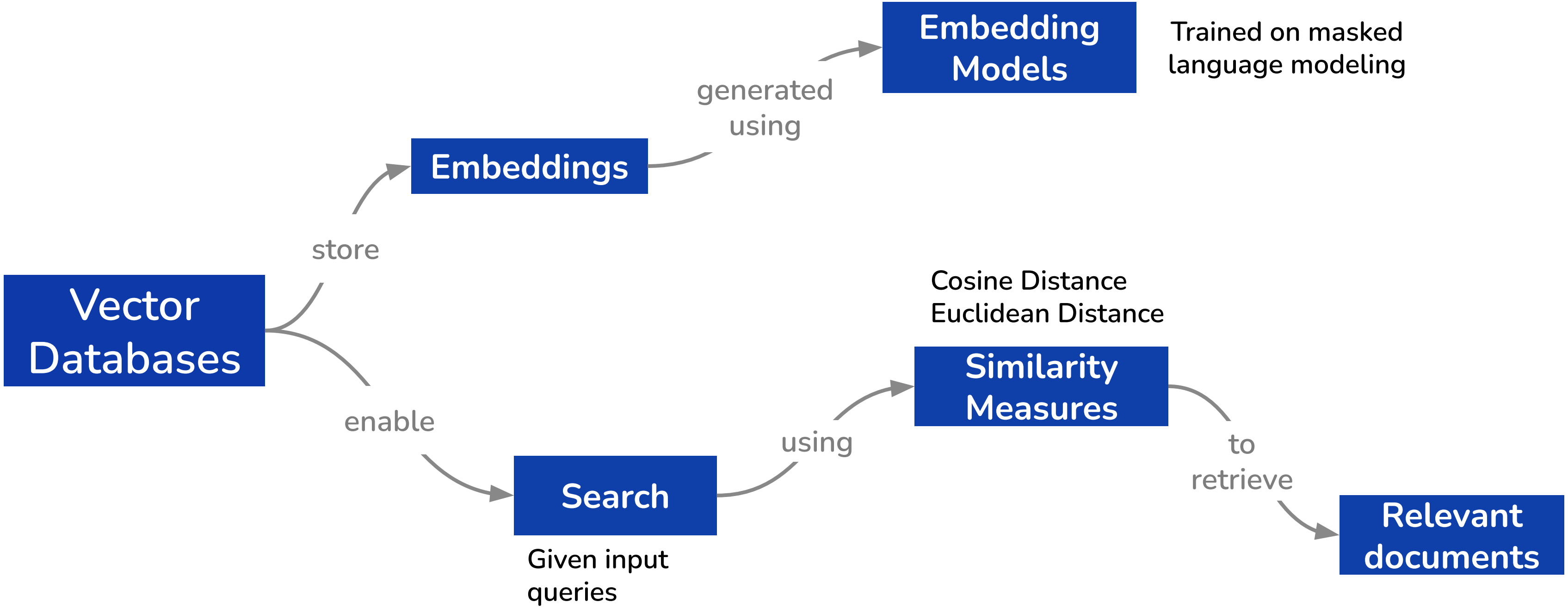
Vector Databases

Vector databases are specialized in storing and retrieving vectors associated with unstructured data. Given input queries, the database can retrieve relevant documents using similarity search.



[Notebook]
vectordb_search.ipynb

Summary


Vector Databases
store
Embeddings
generated using
Embedding Models
Trained on masked language modeling
enable
Search
Given input queries
using
Similarity Measures
Cosine Distance
Euclidean Distance
to retrieve
Relevant documents