Building LLM Powered Solutions

Module 3: Introduction to Semantic Search

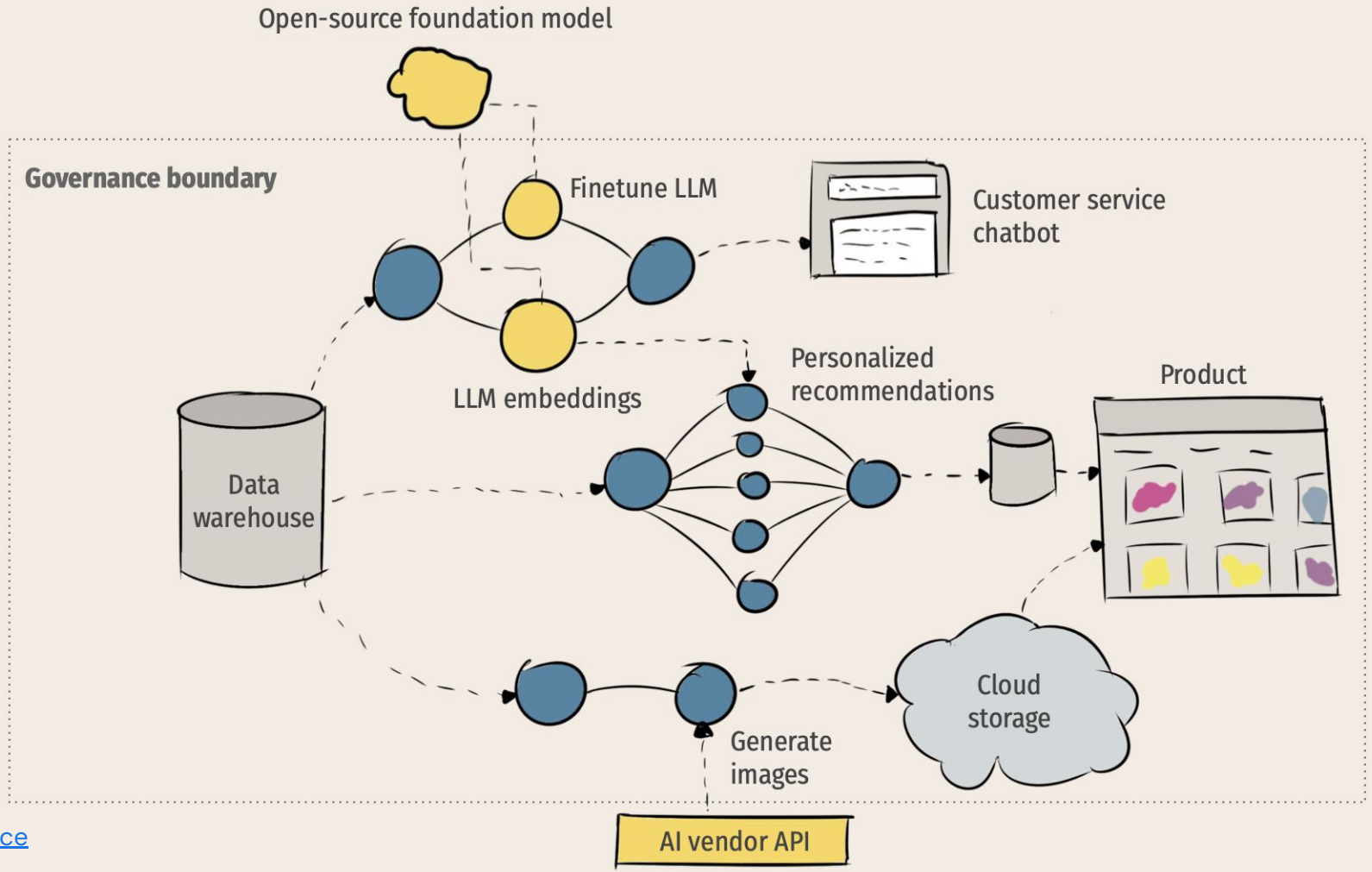
Hamza Farooq

Recap: What We Learned in Last Class

1. What is Self-Attention mechanism?
2. What is an Encoder?
3. What is Decoder?
4. What are Parameters?

Today Lecture Learning Outcomes

1. What is Semantic Search?
2. Sparse vs. Dense Vectors
3. What is Euclidean Distance?
4. Cosine Similarity
5. What is ANN?
6. Using FAISS and coding


Open-source foundation model
Governance boundary
Finetune LLM
Customer service chatbot
LLM embeddings
Personalized recommendations
Product
Data warehouse
Cloud storage
Generate images
AI vendor API

Source

BullsAI

Default

Unknown

Unknown

Unknown

Angel Midi Dress

Embrace sophistication with the Angel Midi Dress Black, crafted from 100% reclaimed polyester for eco-conscious elegance. Featuring a silky feel, a stunning plunging neckline, and a chic high leg slit, this dress combines sustainable fashion with a touch of allure. Perfect for any occasion that calls for a sleek and stylish statement.

Add to Cart

Olivia

New York

Summer dress (size XL), jeans (size XL)

34 years old

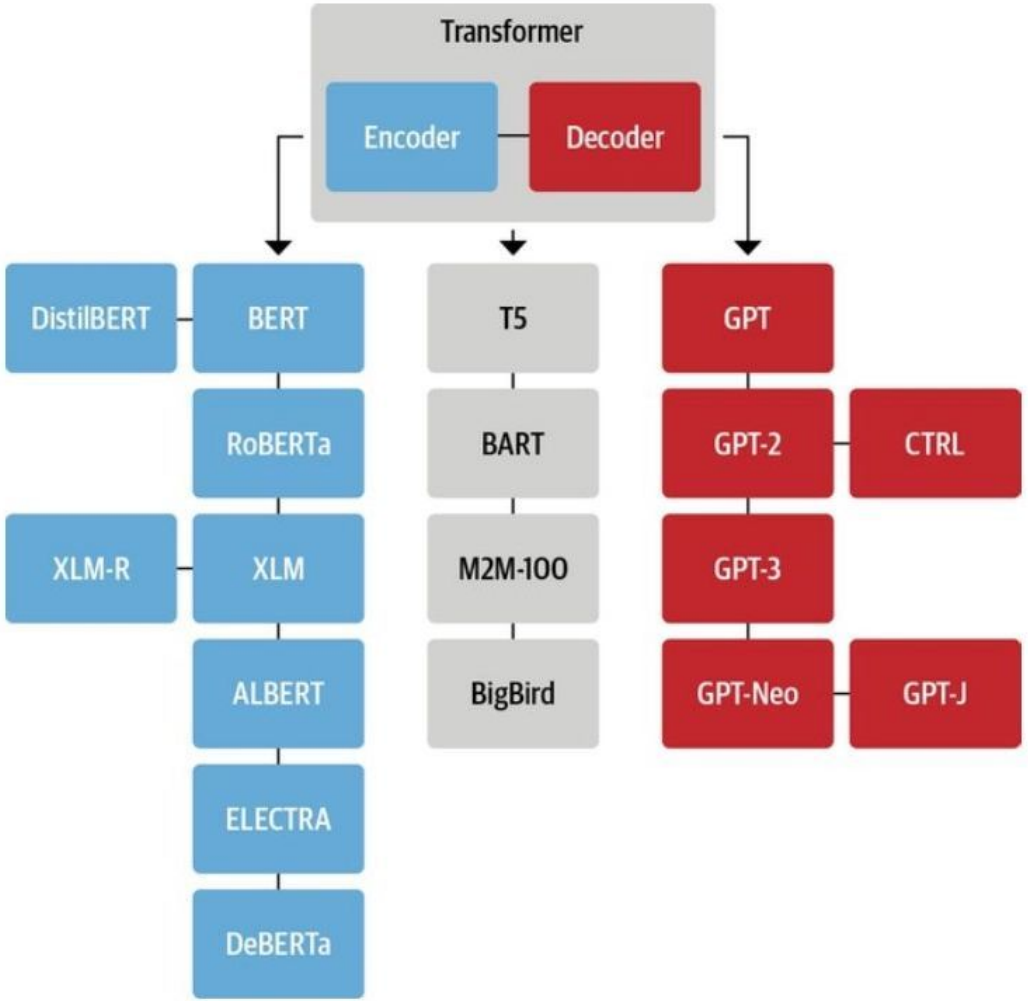
Angel Midi Broadway Dress

The Angel Midi Dress Black, combines high-end urban elegance with comfort. Its smooth fabric and graceful cut make it a versatile piece for any wardrobe, perfect for transitioning from work to social events.

Add to Cart

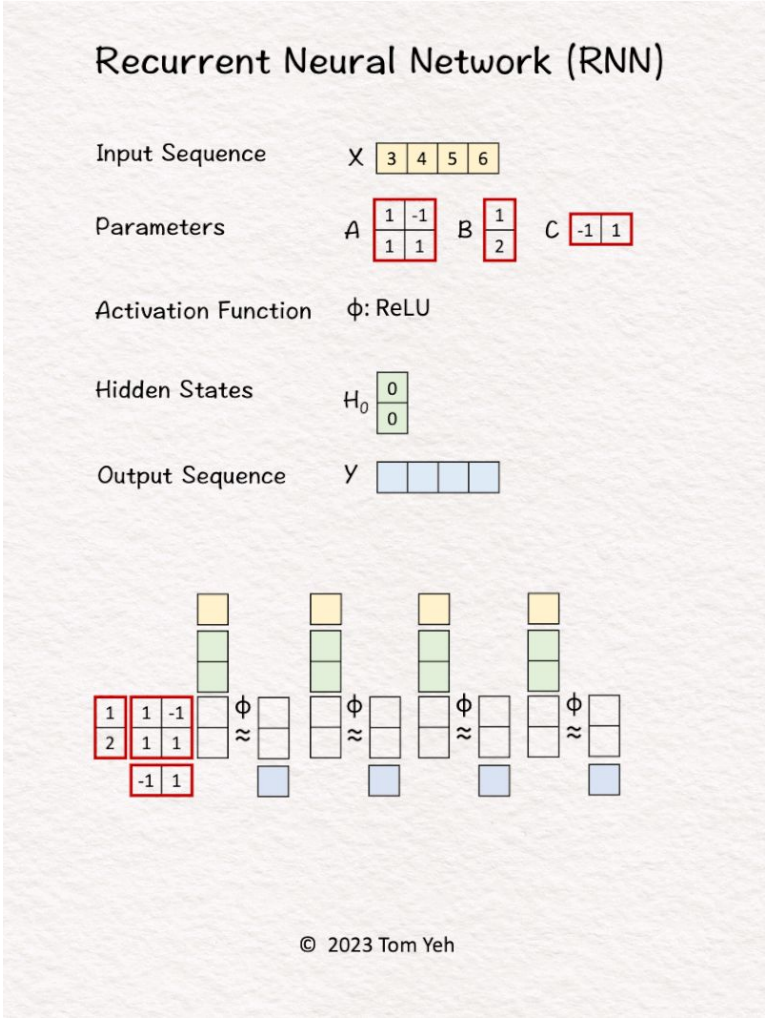
ash@getbullsai.com

Use case – Virtual Photoshoots


Transformer
Encoder
Decoder
DistilBERT
BERT
RoBERTa
XLM-R
XLM
ALBERT
ELECTRA
DeBERTa
T5
BART
M2M-100
BigBird
GPT
GPT-2
CTRL
GPT-3
GPT-Neo
GPT-J

Source

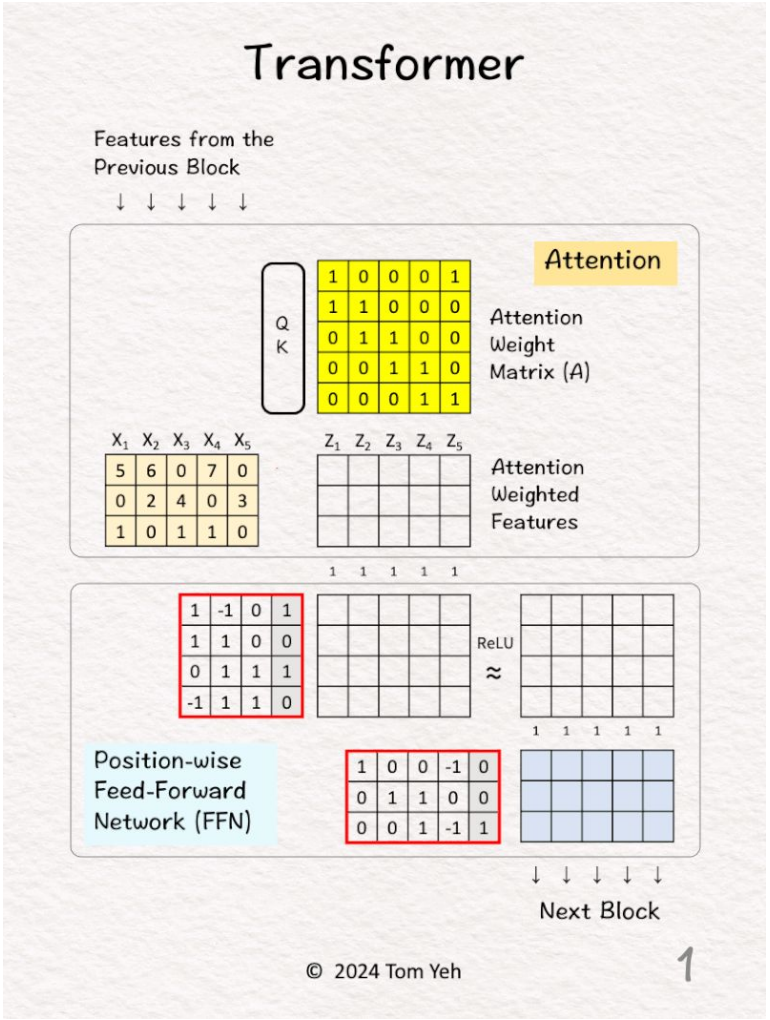
How RRN Works?



[Source]

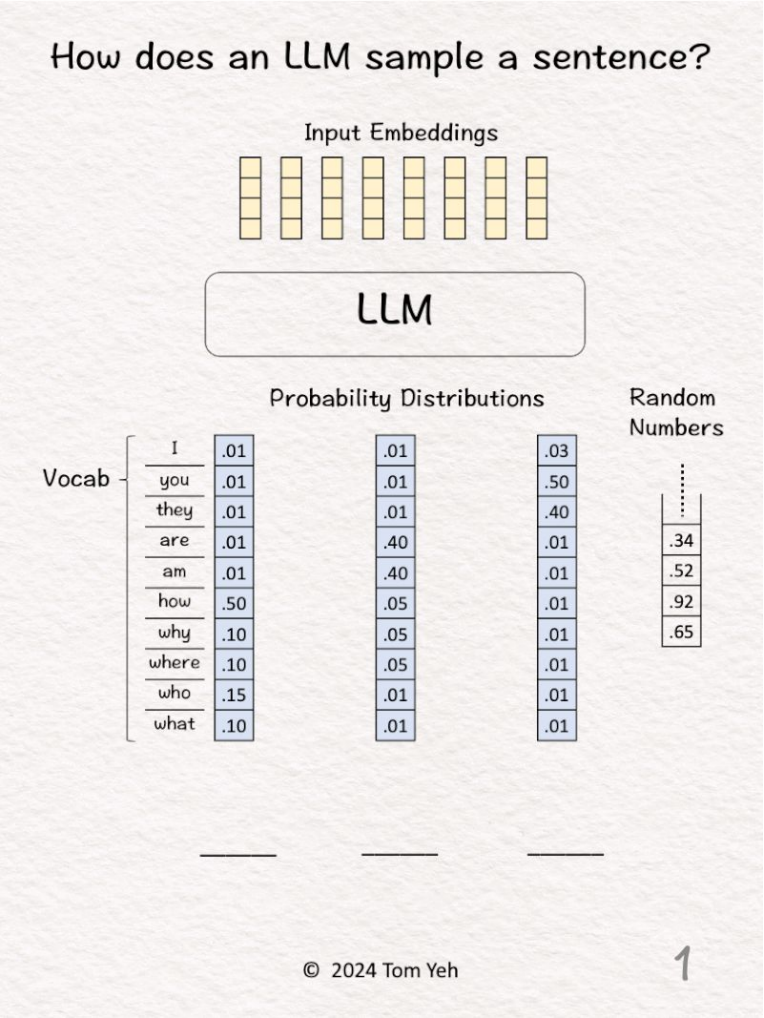
How Transformer Works?

Talk by Batool on Friday 12pm PT



[Source]

How Does an LLM sample a sentence?



[Source]

What is a Retrieval System?

Search and Retrieval Systems are essential tools for information retrieval, enabling fast and precise results with popular methods such as

- Euclidean Distance
- Cosine Similarity
- Approximate Near Neighbors,
- Locality Sensitive Hashing,
- Hierarchical Navigable Small World Graphs, and
- Quantization.

Why Semantic Search?

- Semantic search is a technique that understands the intent and context behind a user's query, rather than relying solely on keyword matching.
- It aims to deliver more accurate and meaningful results by understanding the user's query in a broader context.
- It helps overcome the limitations of traditional keyword-based search by focusing on the user's intent rather than the literal terms used in the query.

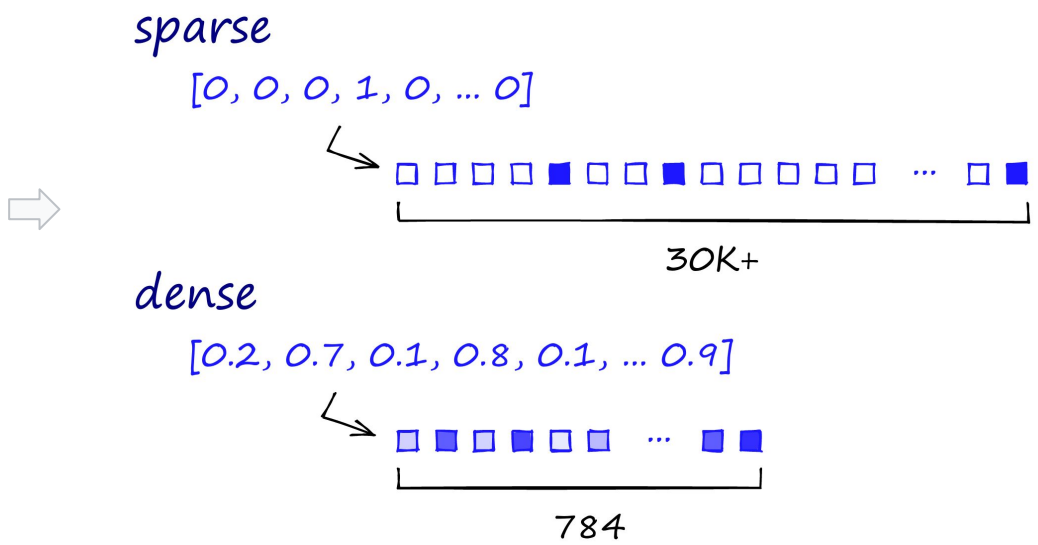
Find me a stylish Nike blue t-shirt specifically designed for golf, with a comfortable fit and moisture-wicking fabric, in size medium and at a reasonable price point

Search

Why should we represent text using vectors?

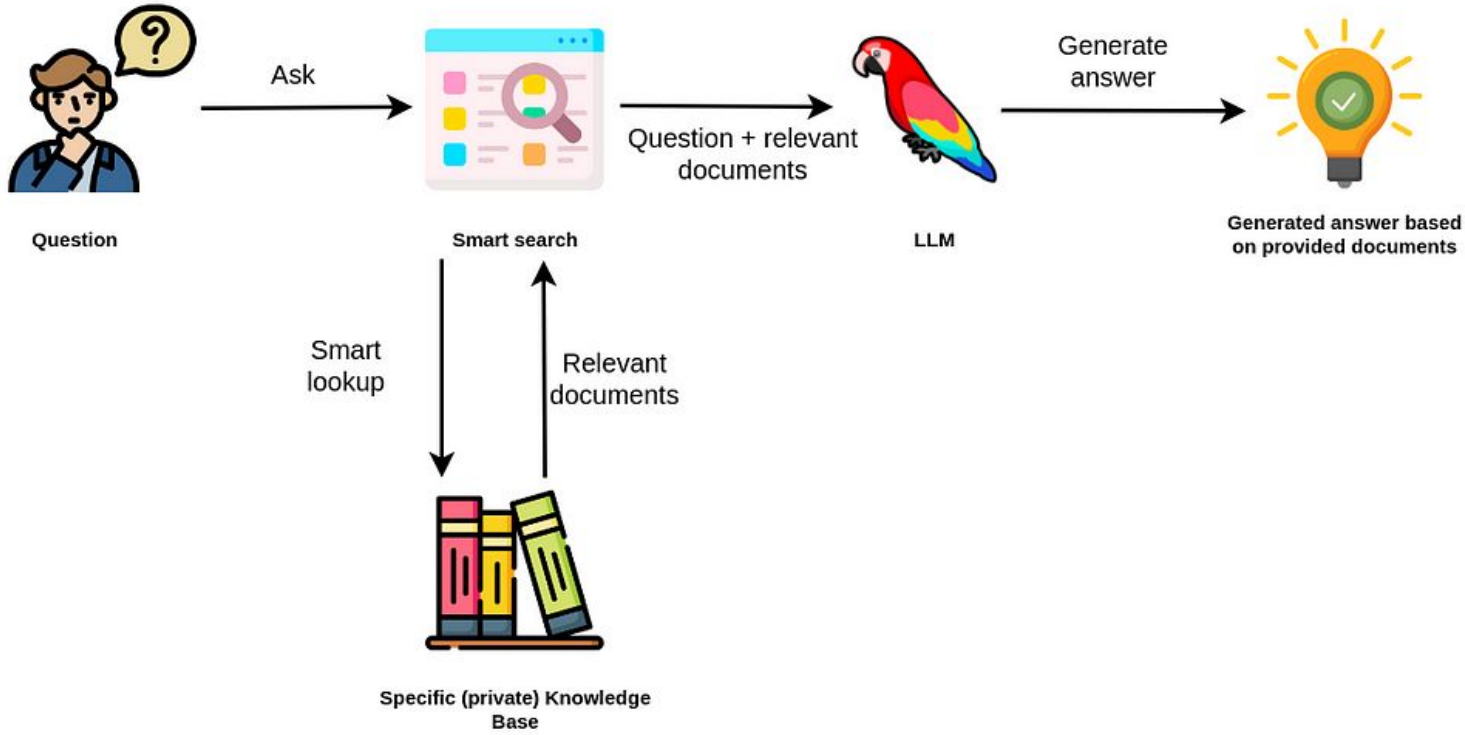
for a computer to understand human-readable text, we need to convert our text into a machine-readable format.

```
Bill ran from the
giraffe toward the
dolphin
```



Source

RAG Systems… coming back

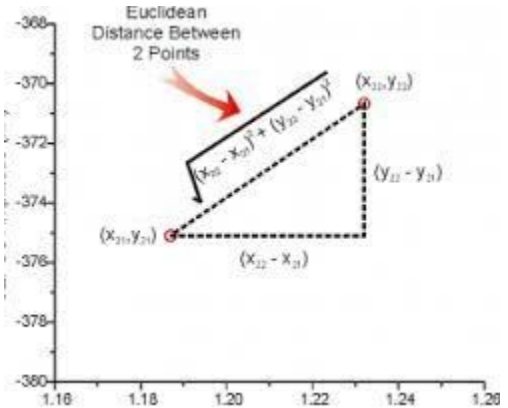


Source:
Knowledge Graphs & LLMs: Multi-Hop Question Answering

How do we measure distance?

IndexFlatL2 measures the L2 (or Euclidean) distance between all given points between our query vector, and the vectors loaded into the index. It's simple, very accurate, but not too fast.

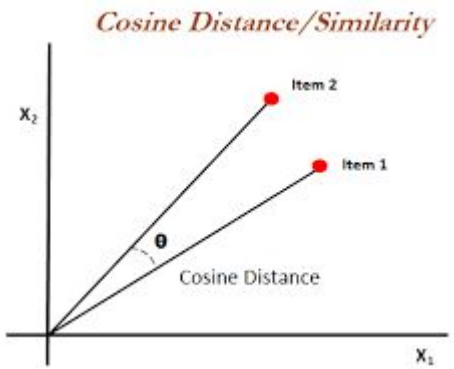
Euclidean distance is a way of measuring the distance between two points in a Cartesian plane. It is calculated by taking the square root of the sum of the squares of the differences between the corresponding coordinates of the two points.



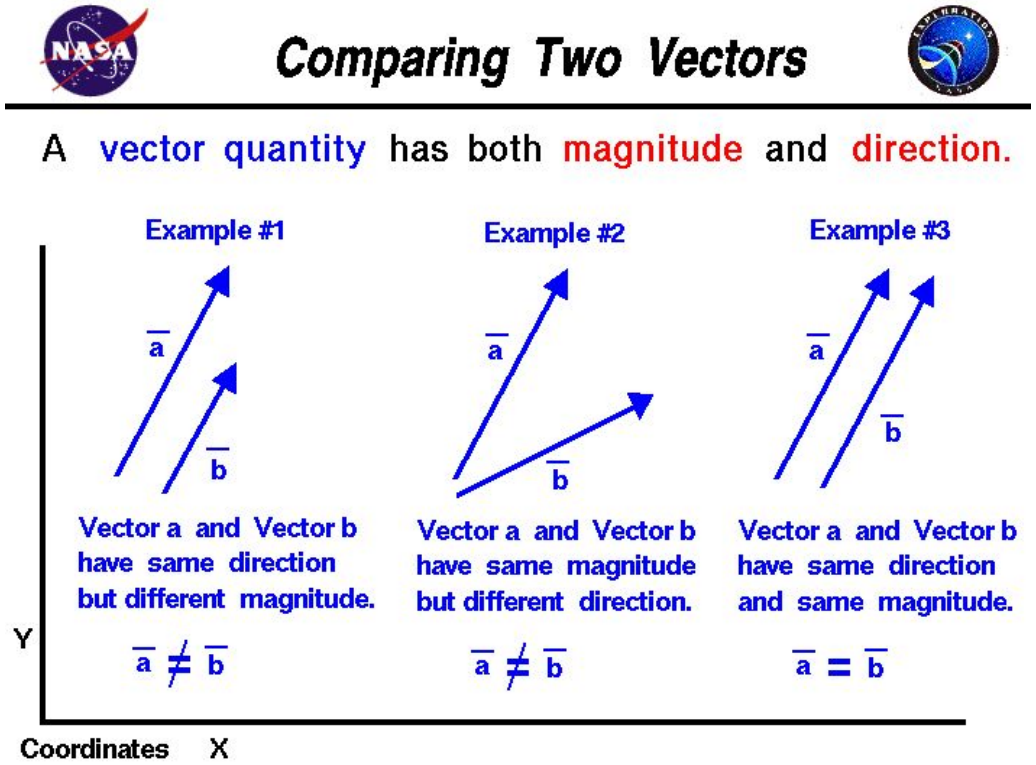
Cosine Similarity

Cosine similarity is a measure of similarity between two vectors. It is calculated by taking the dot product of the two vectors and then dividing by the product of their magnitudes.

In simpler terms, it is a measure of how similar the directions of two vectors are, regardless of their magnitudes.



Magnitude and Direction in a vector



Drawback of Cosine and Euclidean

Cons of Euclidean distance when dealing with large search space:

- *Sensitivity to dimensionality*: Euclidean distance becomes less effective as the number of dimensions increases, known as the "curse of dimensionality."
- *Computational complexity*: Calculating Euclidean distance requires computing the square root, which can be computationally expensive, especially with large search spaces.
- *Lack of normalization*: Euclidean distance is not inherently normalized, meaning features with larger magnitudes can dominate the distance calculation.

Drawback of Cosine and Euclidean

Cons of Cosine similarity when dealing with large search space:

- One of the main drawbacks of cosine similarity is that it is not as effective as Euclidean distance at measuring similarity between vectors with different magnitudes. This is because cosine similarity only considers the direction of the vectors, not their magnitudes. This can make it difficult to find similar vectors in a dataset where the magnitudes of the vectors vary widely.
- Another drawback of cosine similarity is that it is not as robust to noise as Euclidean distance. This is because cosine similarity is only affected by the direction of the vectors, not their magnitudes. This means that noise in the vectors can have a significant impact on the cosine similarity between the vectors.

This shirt costs $55.

```
<[CLS]> <this> <shirt> <costs>
<$> <55> <.> <[SEP]>
```

This shirt costs $559.

```
<[CLS]> <this> <shirt> <costs>
<$> <55> <##9> <.> <[SEP]>
```

Similarity: **0.9530616**

Source

Introducing FAISS

- Faiss is a library — developed by Facebook AI — that enables efficient similarity search.

- So, given a set of vectors, we can index them using Faiss — then using another vector (the query vector), we search for the most similar vectors within the index.

- In vector similarity search, we use an index to store vector representations of the data we intend to search.

Hence...

- Using either of these approaches means that we are no longer performing an exhaustive nearest-neighbors search but an approximate nearest-neighbors (ANN) search — as we no longer search the entire, full-resolution dataset.
  - LSH
  - HNSW
  - Quantization

Homework

Explore all the techniques mentioned here:

- Comprehensive Guide To Approximate Nearest Neighbors Algorithms | by Eyal Trabelsi | Towards Data Science
- Record a one minute video of each member in the group to record and talk about each individual technique, use loom.com to submit link

Check this Link:

https://www.linkedin.com/posts/tom-yeh_deeplearning-gan-artificialintelligence-activity-7152664087968055296-1AJb?utm_source=share&utm_medium=member_desktop

This Prof has very handful gifts/1 pager to explain difficult concepts. You might want to find some relevant information and add it here to explains some concepts.

Thank you.

Appendix