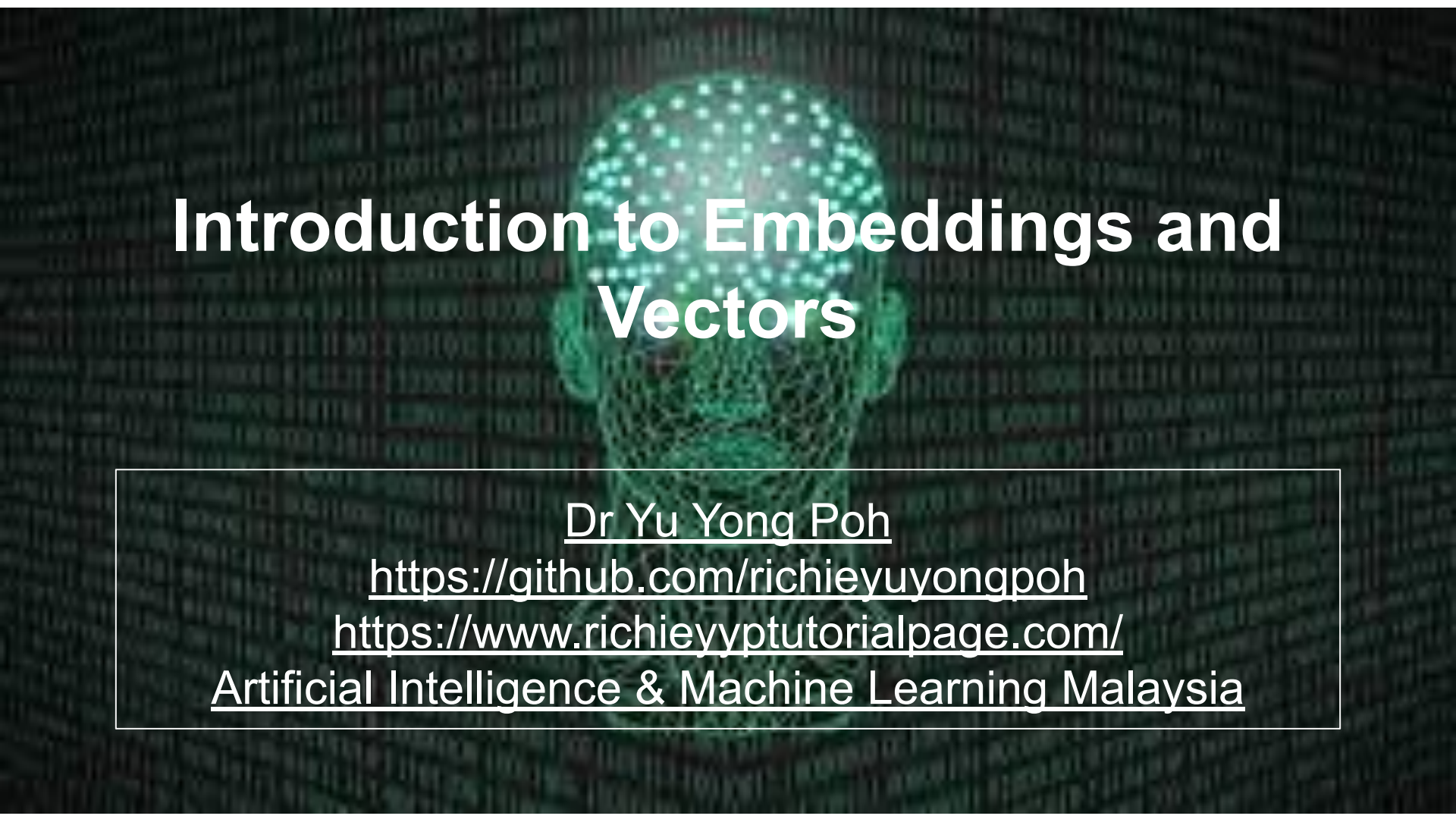


Introduction to Embeddings and Vectors

Dr Yu Yong Poh
https://github.com/richieyuyongpoh
https://www.richieyyptutorialpage.com/
Artificial Intelligence & Machine Learning Malaysia

Disclaimer:

The views and opinions expressed in this presentation are solely those of the presenter and do not necessarily reflect the views, opinions, or policies of any company, organization, or community group. This presentation is intended **for informational and educational purposes only**. Any references to specific companies, products, or services are for illustrative purposes and should not be considered as endorsements.

The presenter **does not have any official affiliation or representation** with any company or community group mentioned during this talk. All information and content presented here are based on personal knowledge and experiences.

Please consult with appropriate professionals and organizations for specific advice and recommendations related to the topics discussed in this presentation.

Get the Deck

https://github.com/richieyuyongpoh/presentation/blob/main/Introduction%20to%20Embeddings%20and%20Vectors.pdf

Main References

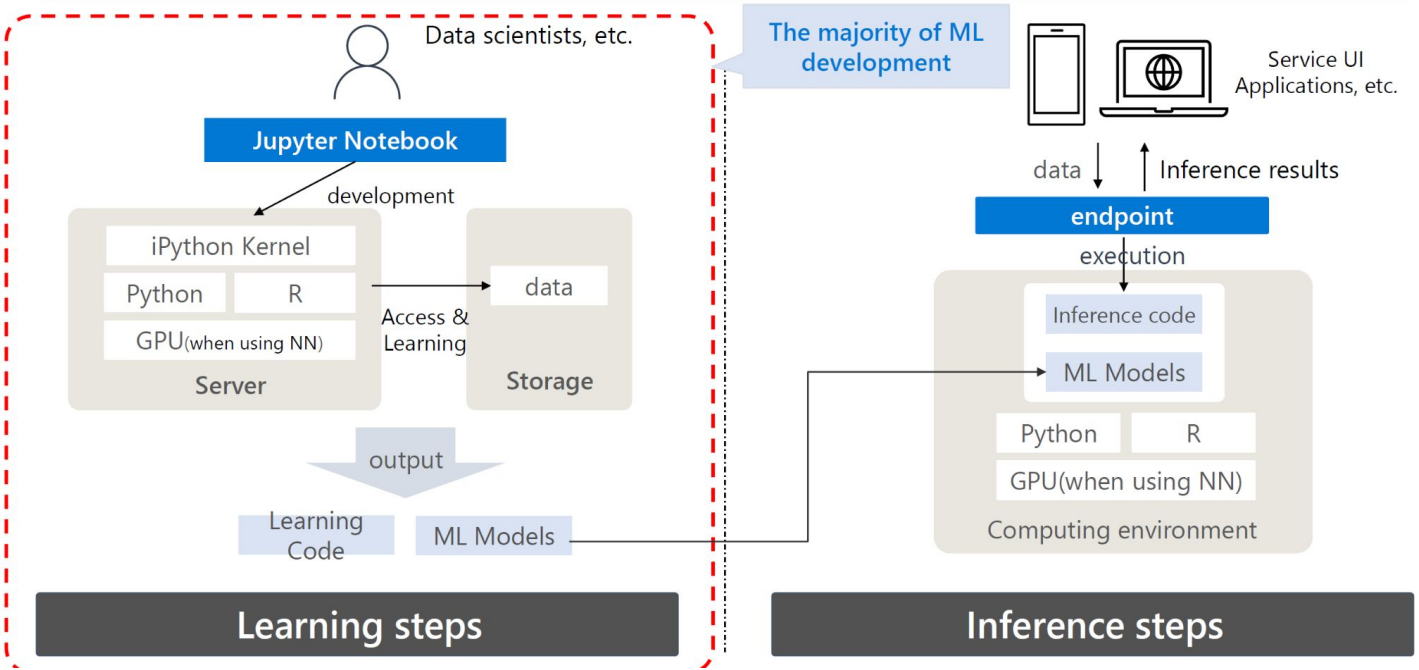
- Jimmy Liao, How will Development Change with LLMs
- Samuel Leonardo Gracio, Reinvent your recommender system using Vector Database and Opinion Mining
- LeewayHertz, What Role do Embeddings Play in ChatGPT-Like Model
- Bhaskar Mitra, Using Text Embeddings for Information Retrieval
- Prompt Engineering Guide, https://www.promptingguide.ai/
- Google Fundational Courses, Embeddings
- CloudFlare, What are Embeddings
- OpenAI, https://openai.com/

What will be covered?

- The Big Picture with LLMs
- Embeddings and Vectors
- Applications and Use Cases

1: The Big Pictures with LLMs

Be more creative in the learning steps



1: The Big Pictures with LLMs

Prompt Engineering: A new paradigm for ML development

Prompt Processing※	If the information in the prompt itself is insufficient or difficult for AI to interpret Processing such as changing the way prompting is given
Few-shot Learning	Provide some example answers to the question at the prompt, A method of learning the answer format and behavior by prompting. Accuracy improvements may also be seen in several levels of examples.
Chain of Thought (CoT)	In large-scale language models, by giving a step-by-step thinking process, The ability to solve even difficult problems. ReAct and Self Consistency also carry on the CoT concept.
Recursively Criticizes and Improves (RCI)	let GPT itself examine and correct the output of GPT. The output is brushed up by repeated execution. In particular, it is often used to operate programming code.
ReAct	Not only linguistic generation from internal information, Dynamically recognize the required task from the prompt, The idea is to obtain information (grounding) using external APIs such as search and calculation, add that information, and return an answer.

1: The Big Pictures with LLMs

How to get a good prompt

Ex.	idea	Summary
1	Additional Questions	"If you don't have enough information, just ask," etc. Interactive from the user by looking at the status of the prompt Get additional information.
2	Text completion	Also done by Bing and others, make text completion and suggestions so that subjects and objects are not missing.
3	Non-English Utilization	Translate the context information behind the scenes into English, Convert input as English in the backend (in System Prompt). Conversion to a programming language called PAL (Program-Aided Language Model) is also effective.
4	Prompts using Templates	Make specific information easier to understand Embed input in a predefined template
5	GPT Calibration	Before letting GPT answer, first make the sentence easy for GPT to interpret. Incorporate steps to make GPT correct.
6	Non-text chat	Input by voice recognition Traditional UI (drop-down list, radio button, checkbox)

8

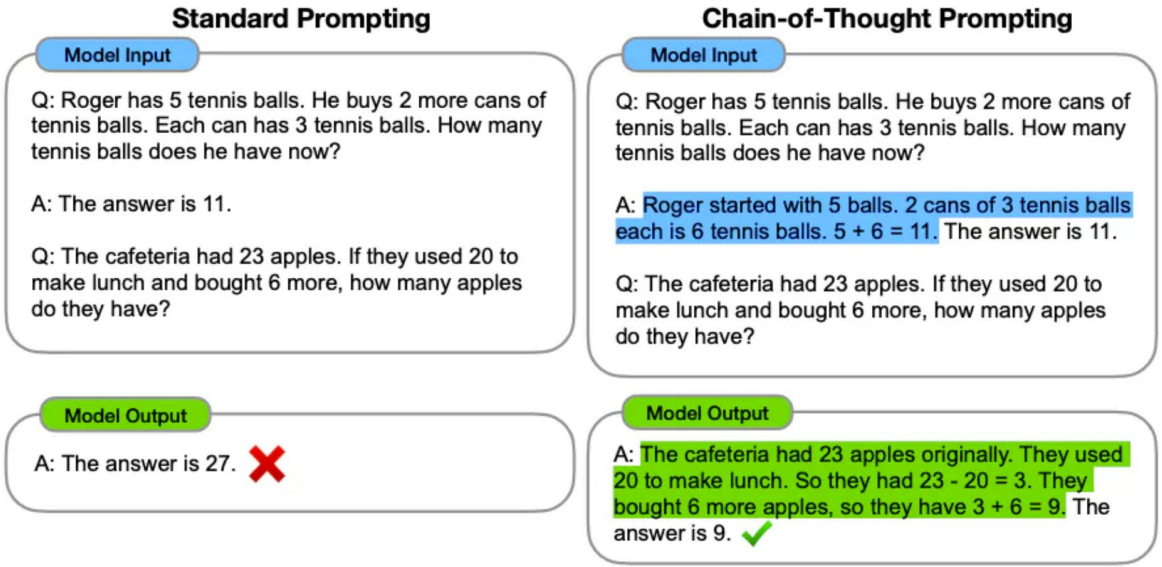
1: The Big Pictures with LLMs

Positioning between Fine tuning and Prompt Engineering

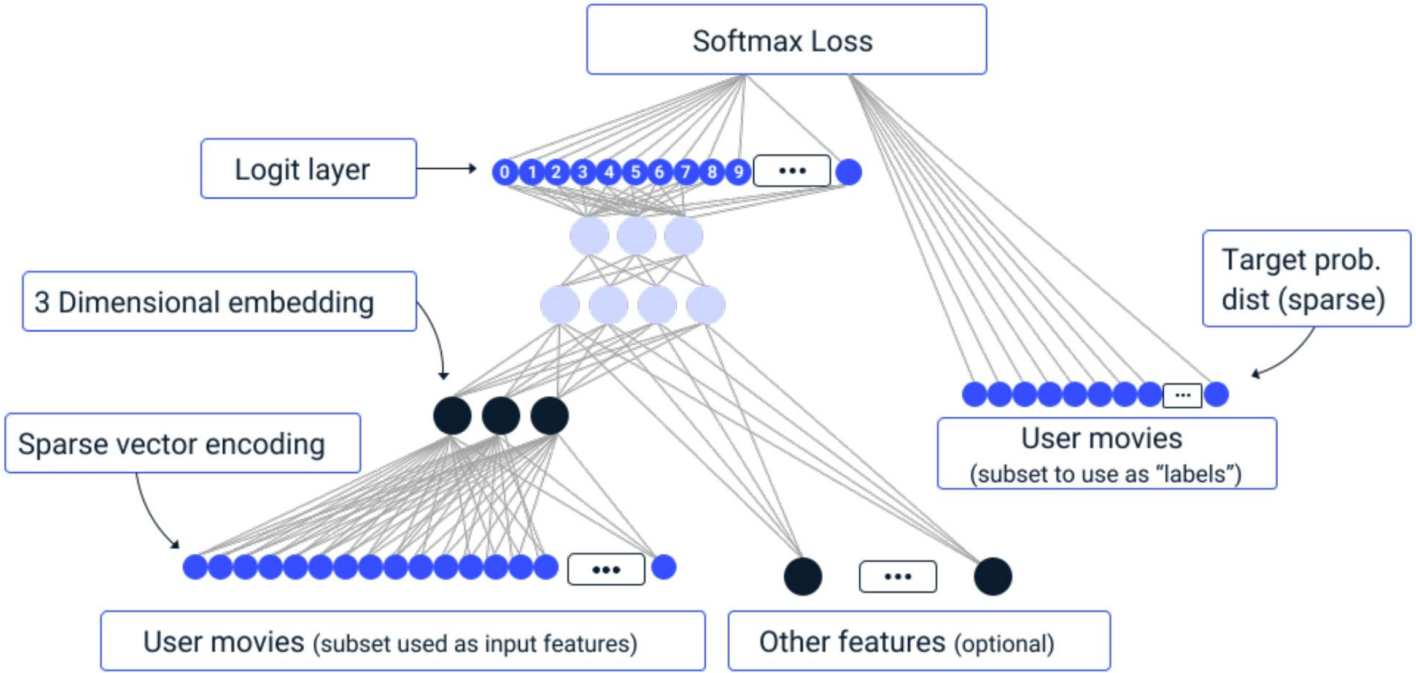
	Fine tuning Step	Prompt Step
What to get	Long term memory (remember general information, but vague)	inference (can handle the details by instruction, but has difficult for unknow area)
Limitations	Cost of resource and data processing	Token limit
Security / Quality	Confidential information while training Annotation Quality	Prompt Injection Be prepared for content filtering
When to use	Acquiring new task or Improve the task Add terminology, domain knowledge The reference info in Prompt too large	Improve accuracy of answers Task recognition Answer format rules Small amount of information

1: The Big Pictures with LLMs

Chain-of-Thought (CoT) Prompting



2: Embeddings and Vectors



LeewayHertz

2: Embeddings and Vectors

- A way of representing complex information, like texts and images, using a set of numbers
- It translates high-dimensional vectors into a lower-dimensional space
- Essentially, embeddings enable machine learning models to find similar objects. Given a photo or a document, a machine learning model that uses embeddings could find a similar photo or document.

2: Embeddings and Vectors

Organizing Movies by Similarity (1d)


Shrek


Incredibles

The Triplets of Belleville


Harry Potter


Star Wars

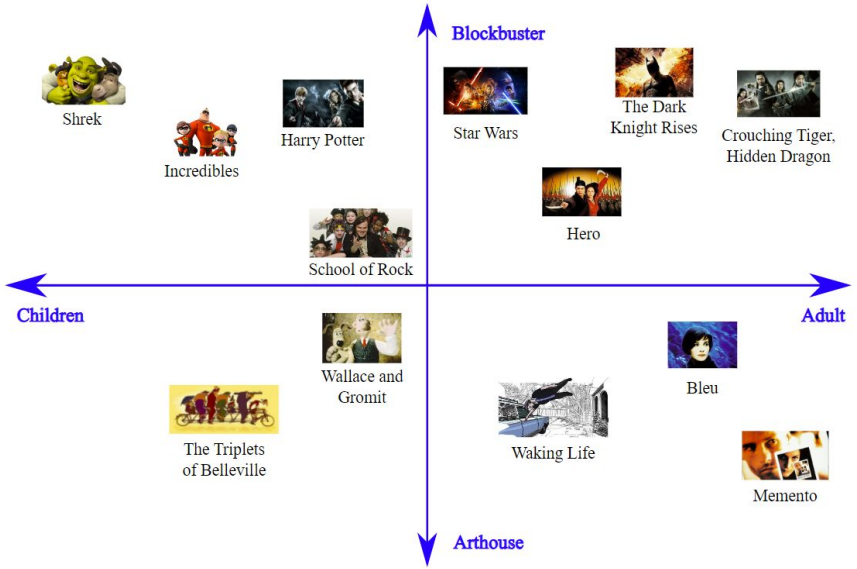

Bleu


The Dark Knight Rises


Memento

2: Embeddings and Vectors

Two-Dimensional Embedding



2: Embeddings and Vectors

d-Dimensional Embeddings

- Assumes user interest in movies can be roughly explained by *d* aspects
- Each movie becomes a *d*-dimensional point where the value in dimension *d* represents how much the movie fits that aspect
- Embeddings can be learned from data

2: Embeddings and Vectors

What is a vector?

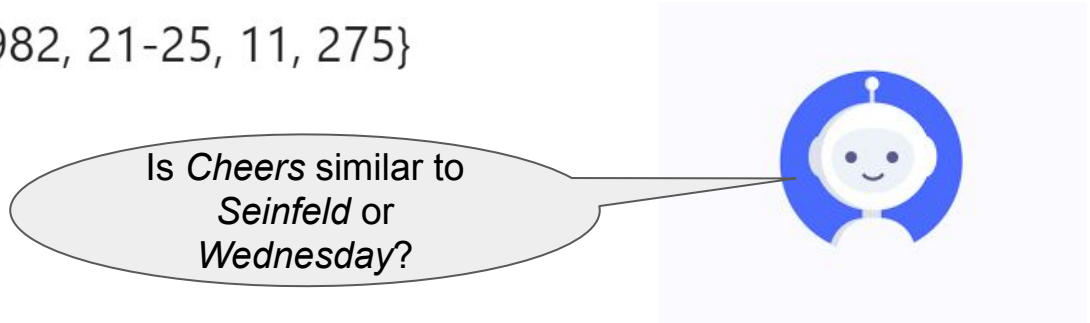
- An array of numbers that define a point in a dimensional space
- In short, a list of numbers
- Each number indicates where the object is along a specified dimension
- Vectors are used to search for similar objects
- A vector-searching algorithm simply has to find two vectors that are close together in a vector database.

2: Embeddings and Vectors

vector similarity

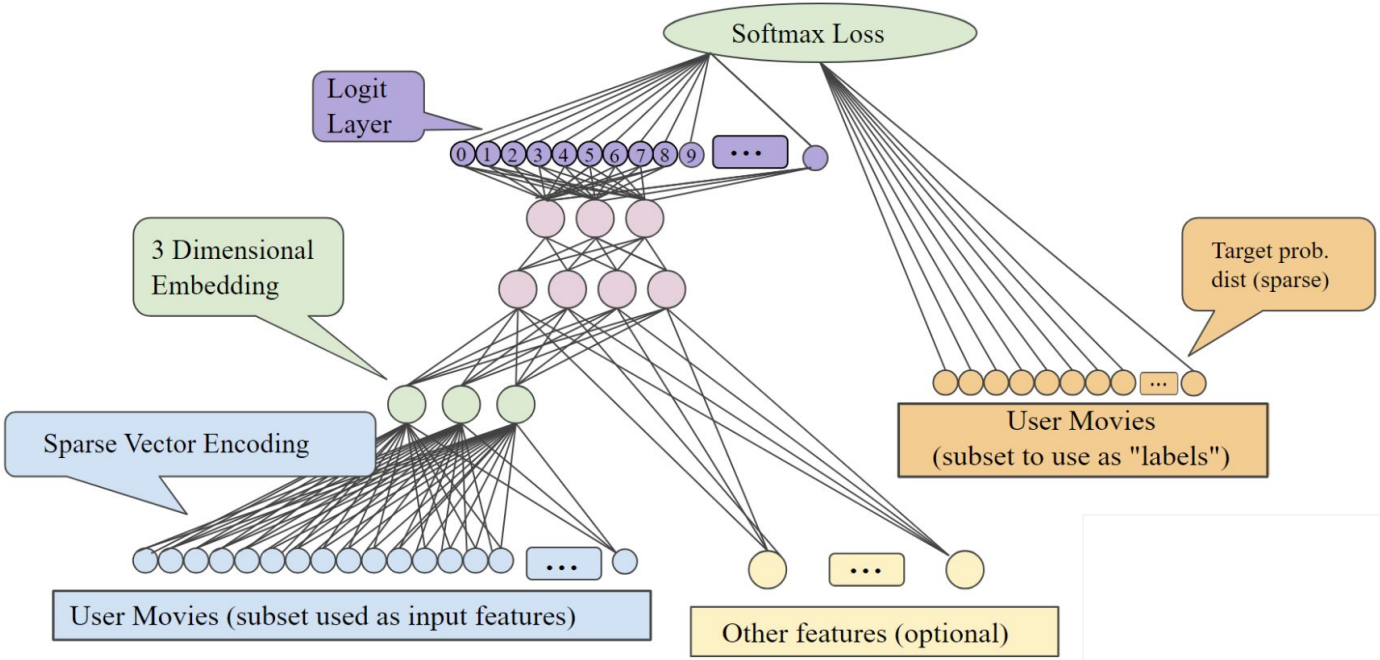
TV show	Genre	Year debuted	Episode length	Seasons (through 2023)	Episodes (through 2023)
Seinfeld	Sitcom	1989	22-24	9	180
Wednesday	Horror	2022	46-57	1	8

Cheers vector: {[Sitcom], 1982, 21-25, 11, 275}

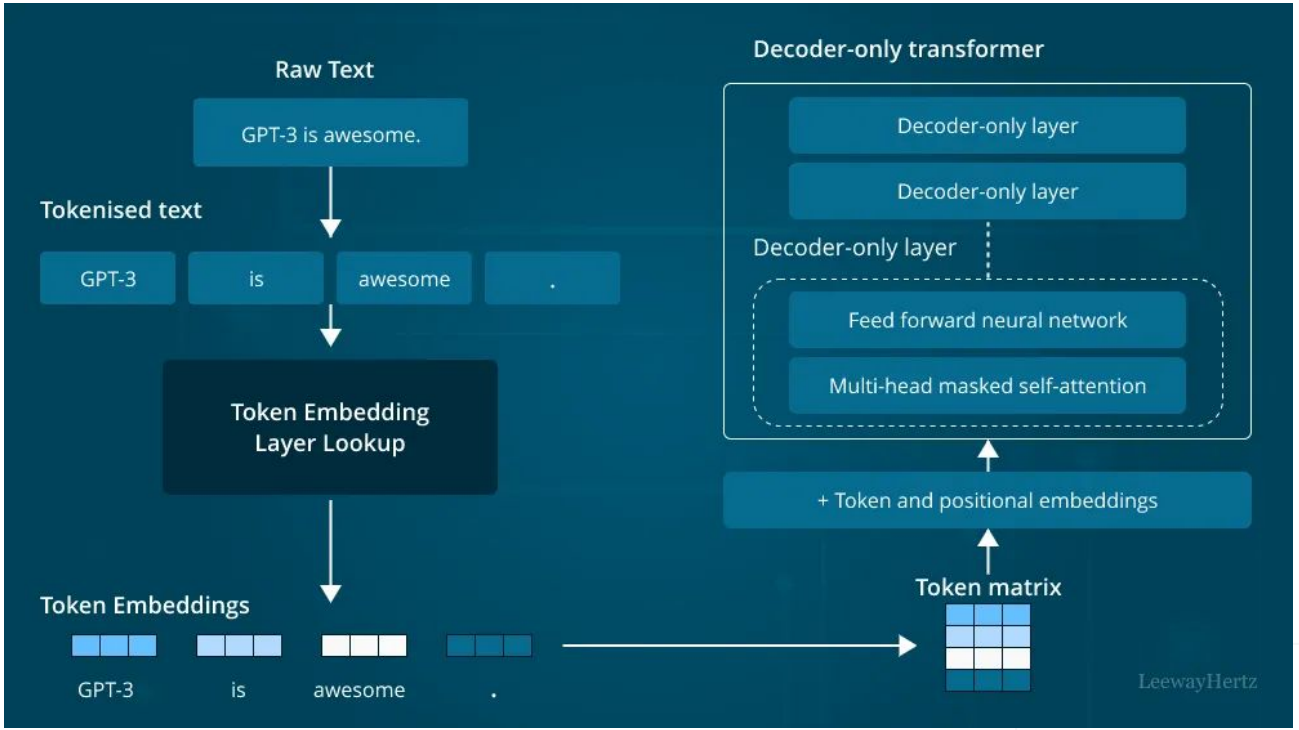


2: Embeddings and Vectors

Collaborative Filtering to predict movies to recommend:



2: Embeddings and Vectors

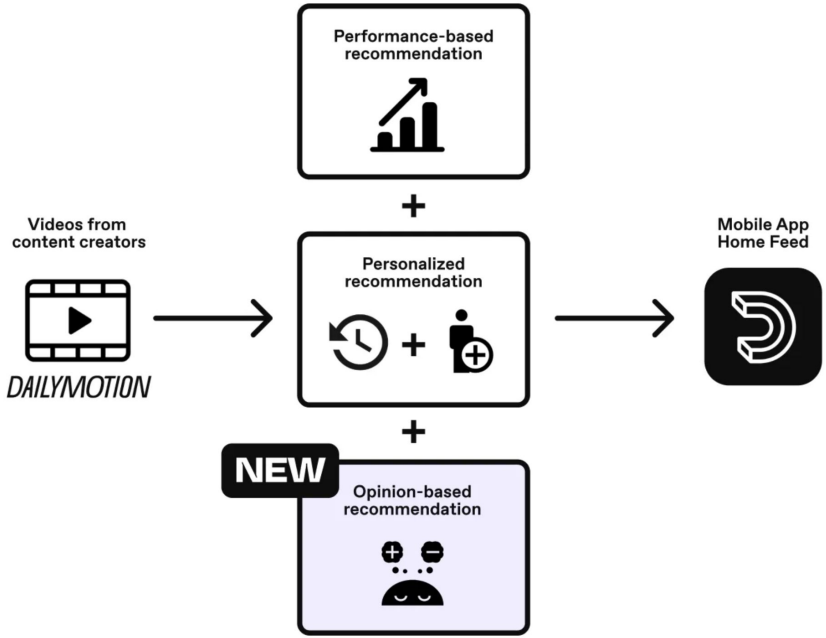


3: Applications and Use Cases

- FAQ Chatbot & NLP
- Customer Experience and Support
- Product Recommendation
- Financial Services
- Image and Video Recognition
- Medical Diagnosis
- Biometrics
- etc

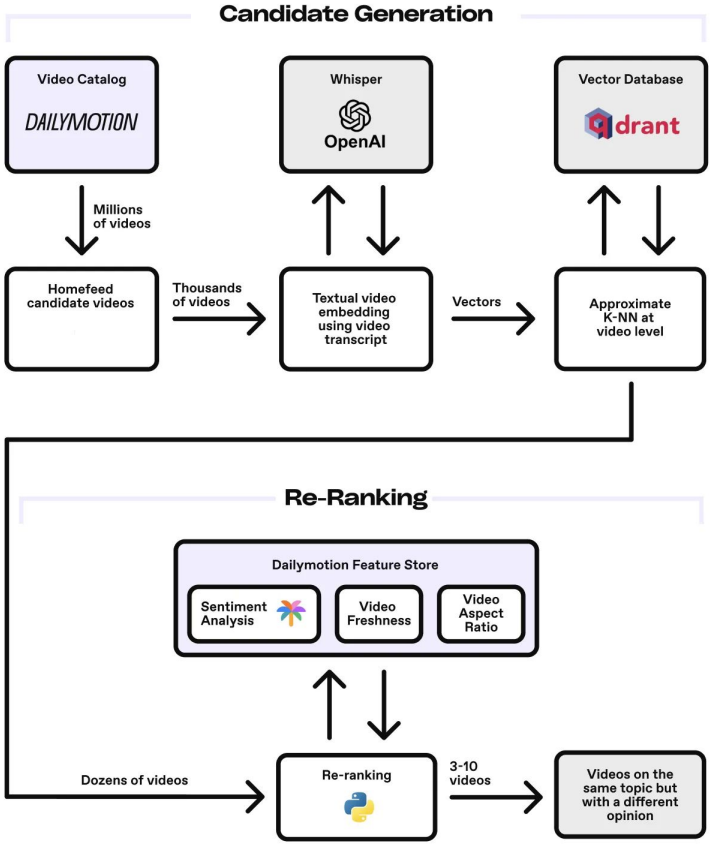
3: Applications and Use Cases

Dailymotion



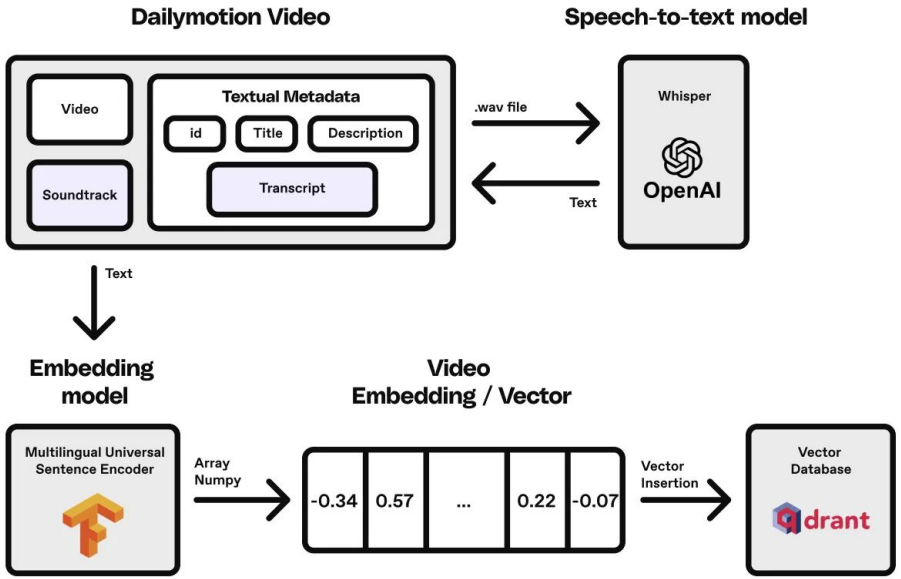
3: Applications and Use Cases

Dailymotion: The global architecture of the new opinion-based recommender system.



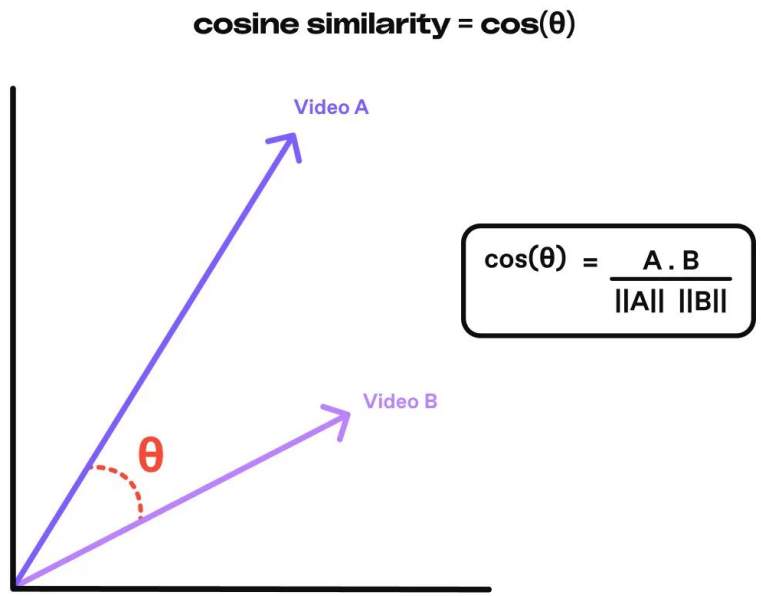
3: Applications and Use Cases

Dailymotion: Representation of transforming a video into a vector



3: Applications and Use Cases

Dailymotion:Cosine similarity between two embeddings



4: Simple Demo

https://www.richieyyptutorialpage.com/demo-python-series/simple-embedding-demo

4: FAQ Chatbot (Based on MySejahtera FAQ)

https://www.richieyyptutorialpage.com/demo-python-series/faq-chatbot


Q
A


THANK YOU...
You Rock!