Analytics Vidhya

Embeddings

Instructors

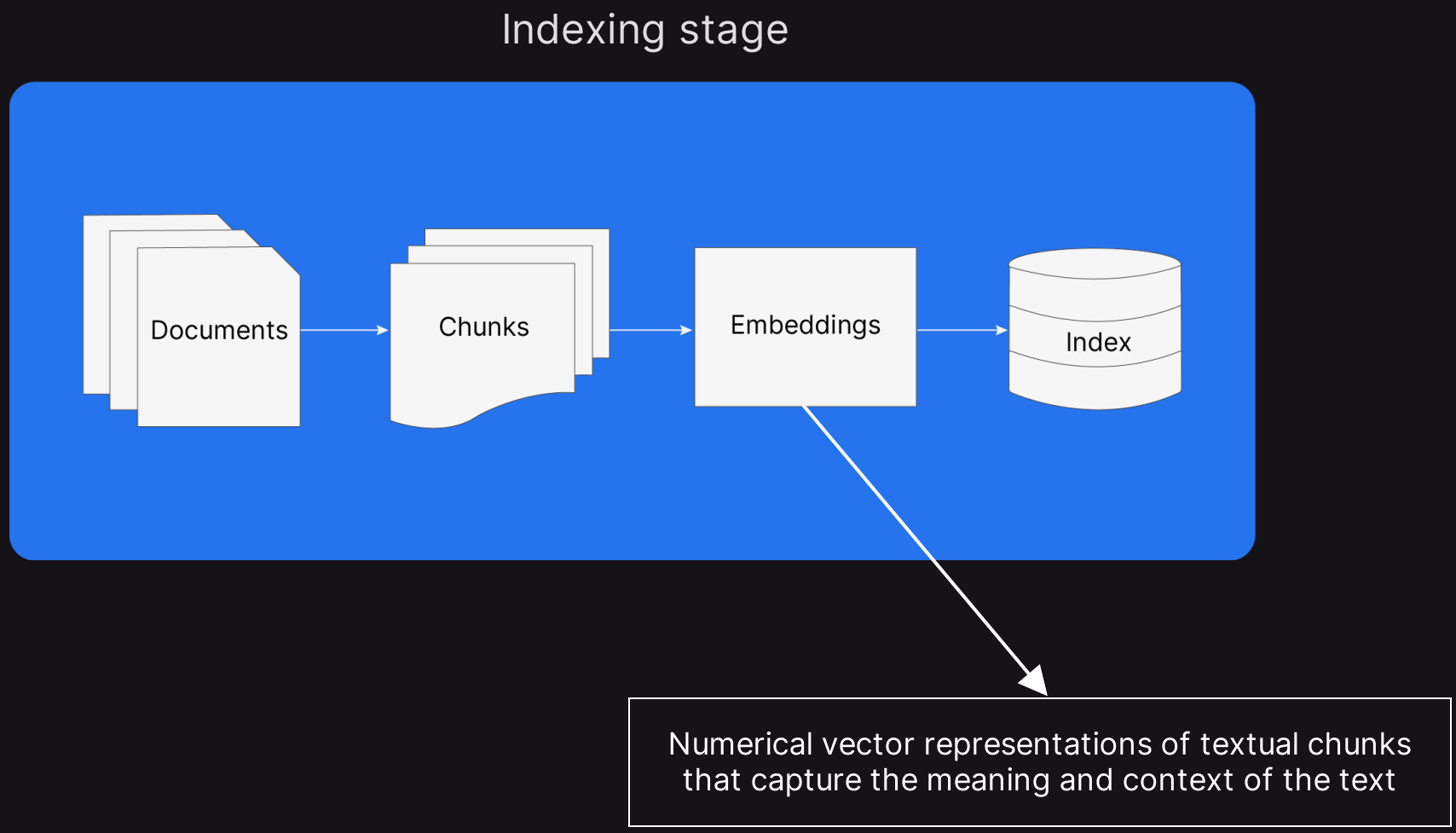
Prashant Sahu

Manager - Data Science, Analytics Vidhya

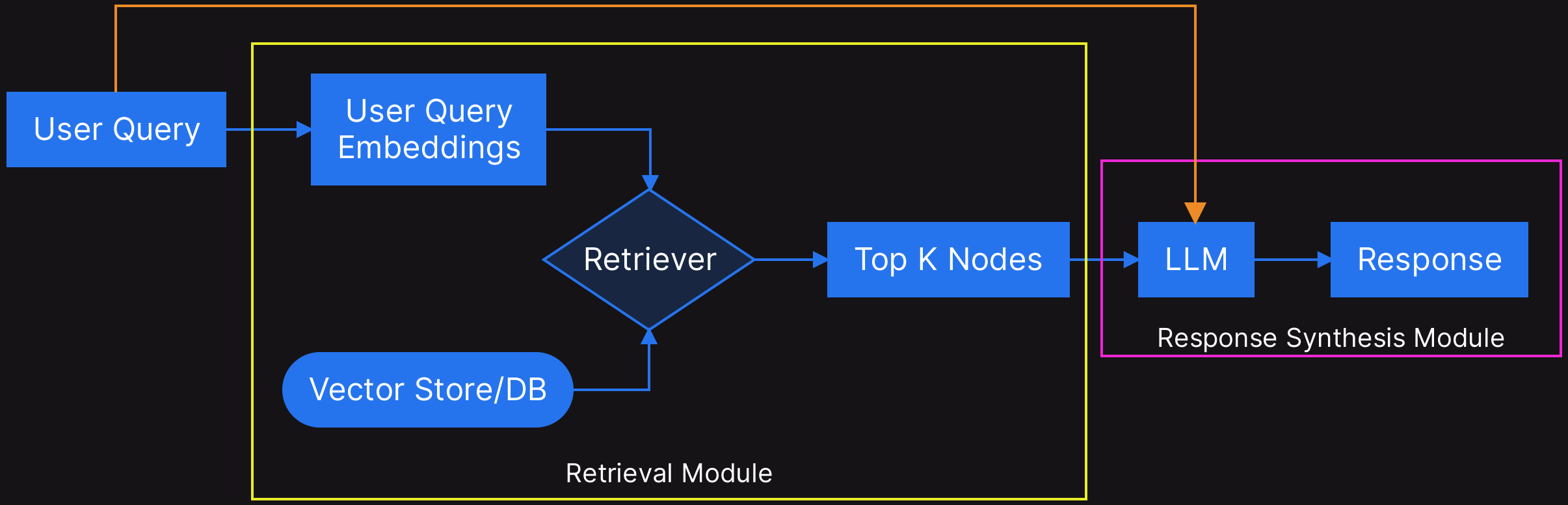
Ravi Theja

Developer Advocate Engineer, LlamaIndex

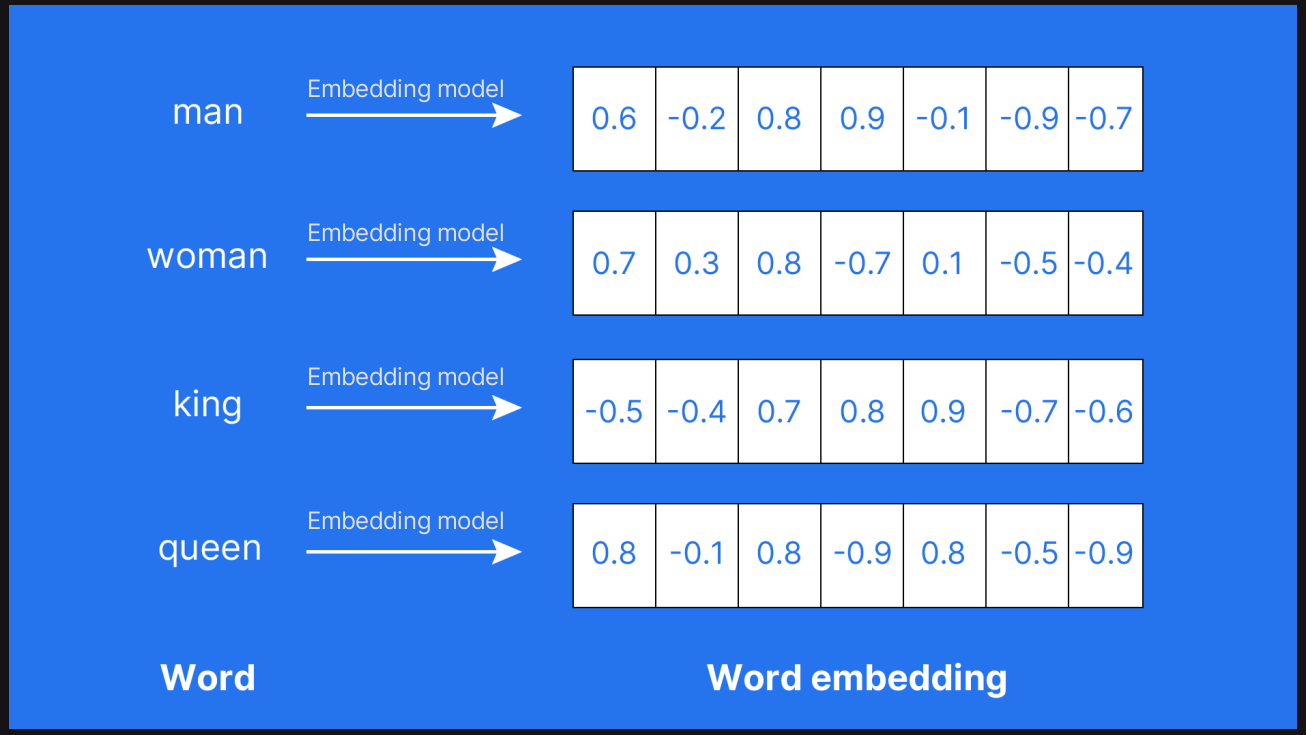
Why do we need Embeddings?



Why do we need Embeddings?

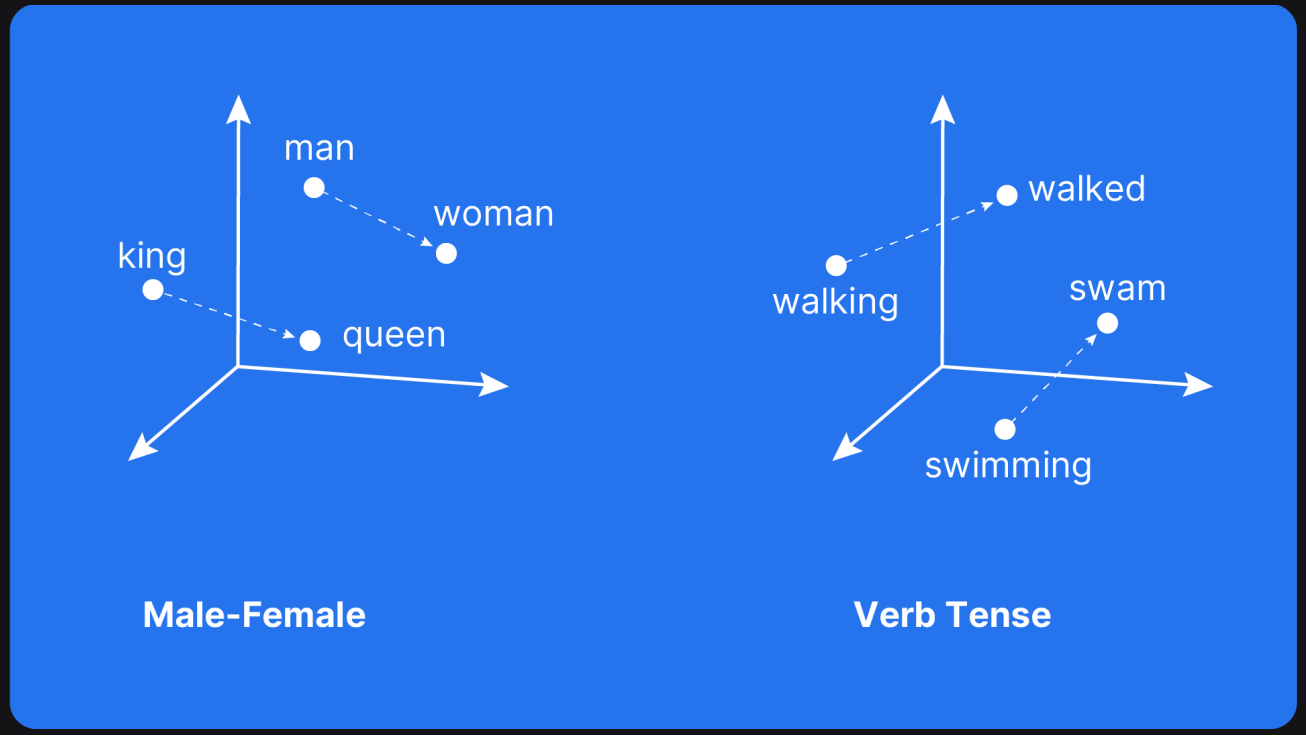


What are Embeddings?



Embeddings represent the text data in the numerical format

What are Embeddings?



They capture the semantic relationships in the language

Interpreting Embeddings

- Cosine similarity is used to determine how similar two vectors are, regardless of their magnitude.
- The value of cosine similarity ranges from -1 to 1, where 1 indicates that the vectors are identical, 0 indicates no similarity, and -1 indicates complete dissimilarity.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

where A & B are the text embedding vectors of 2 different pieces of text (words or phrases or document chunks)

Applications of Embeddings

1. Finding Most Similar Words

Word: "king"

Most Similar Words: ["queen", "monarch", "prince", "ruler", "emperor"]

Applications of Embeddings

1. Finding Most Similar Words

Word: "king"

Most Similar Words: ["queen", "monarch", "prince", "ruler", "emperor"]

2. Finding the Odd one out

Word: ["breakfast", "lunch", "dinner", "car"]

Odd One Out: "car"

cosine_similarity(breakfast, avg_vector_embed) = 0.954

cosine_similarity(lunch, avg_vector_embed) = 0.965

cosine_similarity(dinner, avg_vector_embed) = 0.963

cosine_similarity(car, avg_vector_embed) = 0.891

Applications Embeddings

3. Sentence Similarity

Sentence 1: "The cat sits on the mat."

Sentence 2: "A feline is sitting on a rug."

```
[51]: vec1 = embed_model.get_text_embedding("The cat sits on the mat.")
      vec2 =  embed_model.get_text_embedding("A feline is sitting on a rug.")

      cosine_similarity(vec1, vec2).round(3)

[51]: 0.925
```

Analytics Vidhya

Applications Embeddings

3. Sentence Similarity

Sentence 1: "The cat sits on the mat."

Sentence 2: "A feline is sitting on a rug."

```
[51]: vec1 = embed_model.get_text_embedding("The cat sits on the mat.")
      vec2 =  embed_model.get_text_embedding("A feline is sitting on a rug.")

      cosine_similarity(vec1, vec2).round(3)

[51]: 0.925
```

4. Document Clustering

Cluster 1

"AI is transforming the tech industry."

"The new AI model is impressive."

Cluster 2

"Climate change impacts the environment."

"Renewable energy is the future."

Closed source Embeddings

- OpenAI Embeddings
- CohereAI Embeddings
- Google Gemini Embeddings
- JinaAI Embeddings

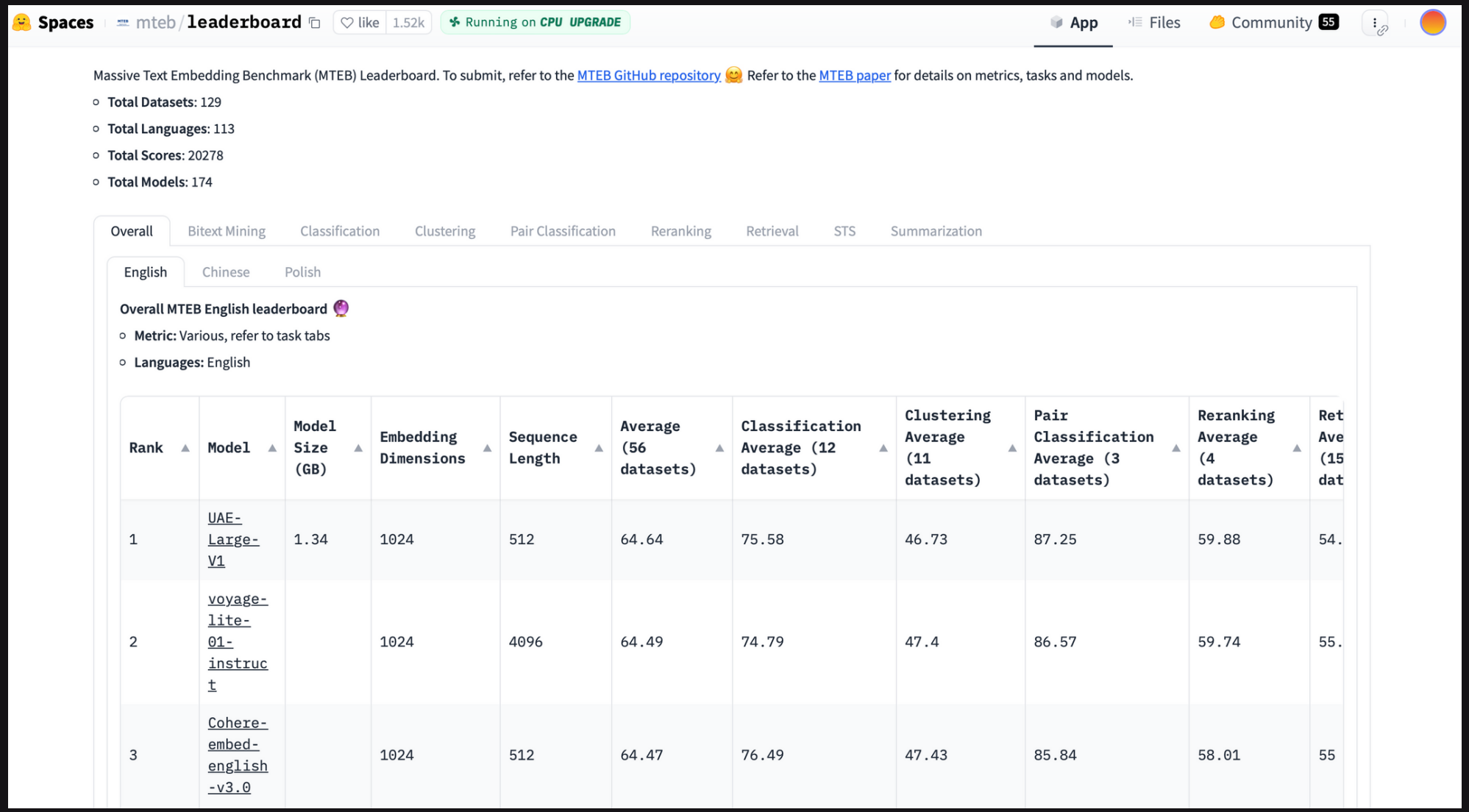
Open source Embeddings

- BERT / DistilBERT
- BGE
- mpnet
- e5

How to select the right embeddings?

1. Look for domain specific embeddings
2. State-of-the-art embeddings

Massive Text Embedding Benchmark (MTEB)



Rank	Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Ret Ave (15 dat
1	UAE-Large-V1	1.34	1024	512	64.64	75.58	46.73	87.25	59.88	54.
2	voyage-lite-01-instruct		1024	4096	64.49	74.79	47.4	86.57	59.74	55.
3	Cohere-embed-english-v3.0		1024	512	64.47	76.49	47.43	85.84	58.01	55

How to select the right embeddings?

1. Look for domain specific embeddings
2. State-of-the-art embeddings
3. Finetune embeddings

Thank You

Analytics Vidhya