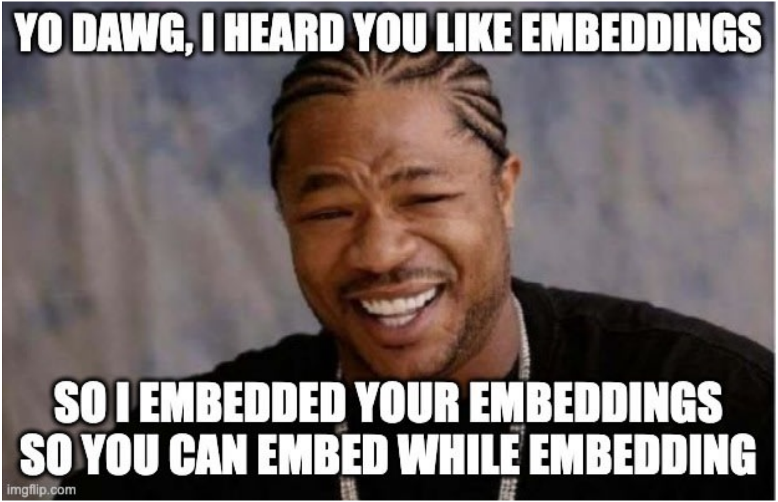
Document Retrieval with Embeddings

Joshua Cook, Chelle AI

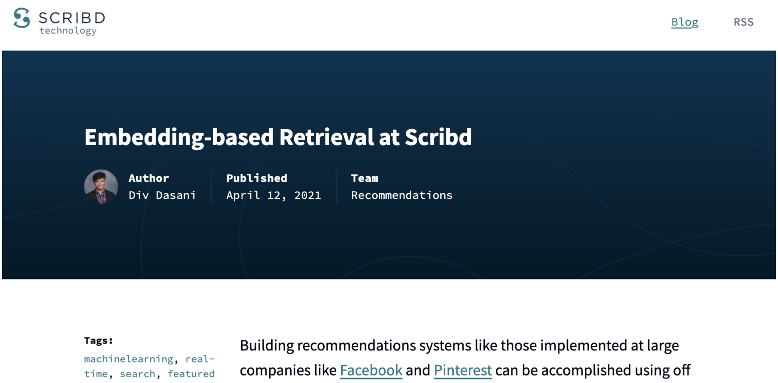
Introduction


YO DAWG, I HEARD YOU LIKE EMBEDDINGS
SO I EMBEDDED YOUR EMBEDDINGS SO YOU CAN EMBED WHILE EMBEDDING
imgflip.com

Importance of Document Retrieval

- Powerful, cheap strategy for building new product features
- Here a **document** is any textual record
  - e.g. a Hotel in a travel context
  - e.g. a Job in a job search context
  - e.g. an Influencer

Dirt-cheap recommender system. Use Document Retrieval to generate a list of candidates. Then use distance and metadata to rank them and then use.



SCRIBD technology

Blog RSS

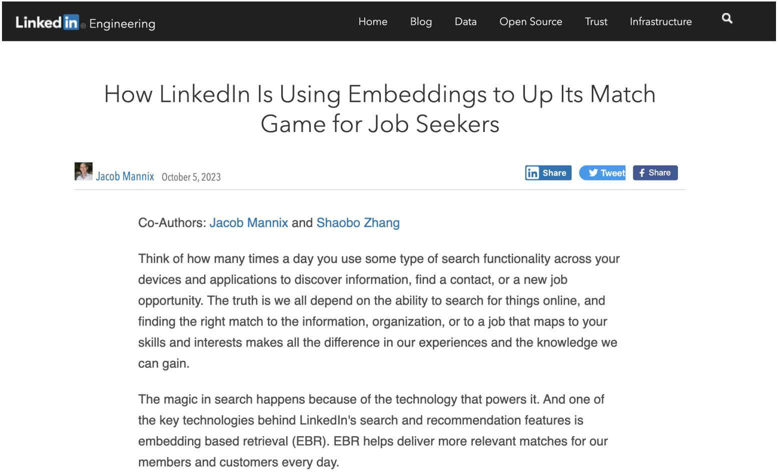
Embedding-based Retrieval at Scribd

Author Div Dasani | Published April 12, 2021 | Team Recommendations

Tags: machinelearning, real-time, search, featured

Building recommendations systems like those implemented at large companies like Facebook and Pinterest can be accomplished using off

https://tech.scribd.com/blog/2021/embedding-based-retrieval-scribd.html



How LinkedIn Is Using Embeddings to Up Its Match Game for Job Seekers

Jacob Mannix October 5, 2023

Co-Authors: Jacob Mannix and Shaobo Zhang

Think of how many times a day you use some type of search functionality across your devices and applications to discover information, find a contact, or a new job opportunity. The truth is we all depend on the ability to search for things online, and finding the right match to the information, organization, or to a job that maps to your skills and interests makes all the difference in our experiences and the knowledge we can gain.

The magic in search happens because of the technology that powers it. And one of the key technologies behind LinkedIn's search and recommendation features is embedding based retrieval (EBR). EBR helps deliver more relevant matches for our members and customers every day.

https://engineering.linkedin.com/blog/2023/how-linkedin-is-using-embeddings-to-up-its-match-game-for-job-se

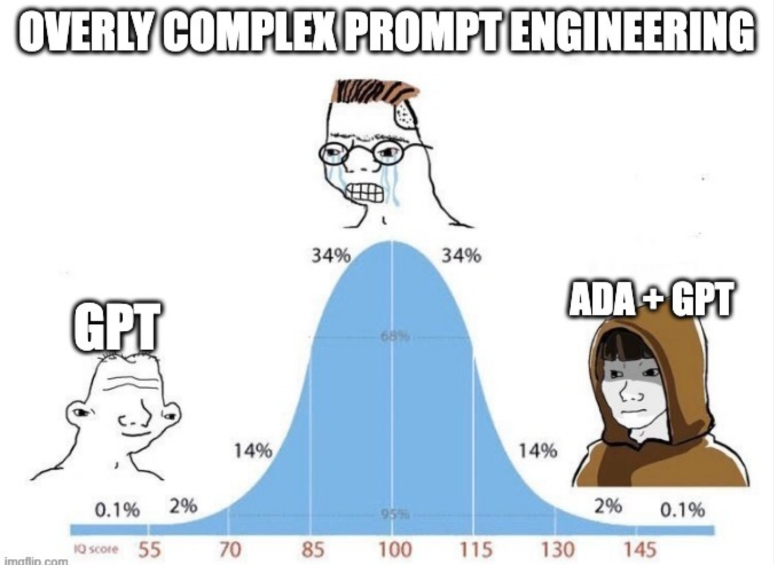
Traditional Methods and Their Limitations

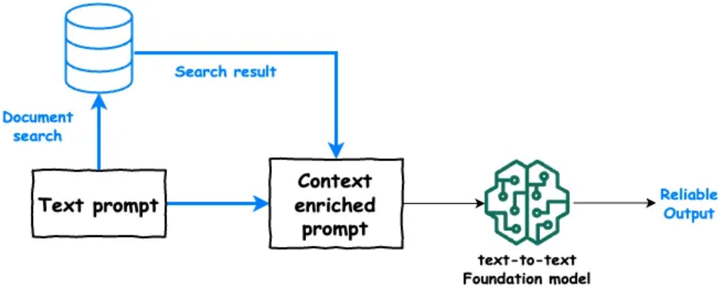
- **Keyword-based Searching**: Explicitly matching terms
- **Boolean Queries**: Using logical operators (AND, OR, NOT)

Limitations:

- Lack of semantic understanding
- Low relevance
- Manual optimization needed

Retrieval-Augmented Generation





https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview

Retrieval-Augmented Generation

- for open-domain question answering (QA) e.g. “Chat” with GPT
- the benefits of:
  - traditional information retrieval (IR)
  - with the power of large language models (LLMs)
- RAG uses a pre-trained LLM (GPT4) to generate answers to a given question
- but first uses an IR system to retrieve relevant documents for context

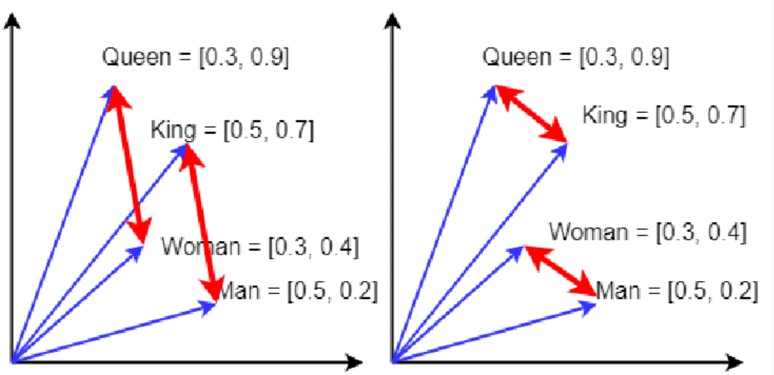
The Basics of Embeddings

What are Embeddings?

- Converting words to points
- Captures meaning in the math, especially distance and direction

Classic Example

Example: Man <-> Woman is similar to King <-> Queen

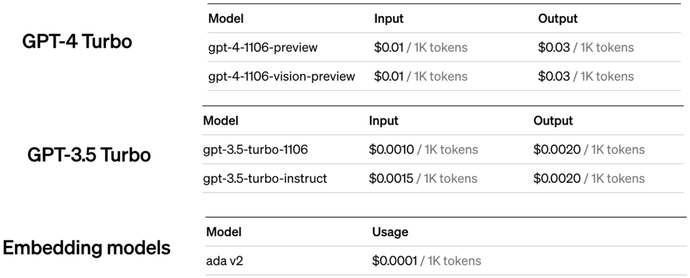


Difference from Traditional Methods

- Understands semantic context
- Example: Differentiates “Apple” the company from “apple” the fruit

Also

Pricing



GPT-4 Turbo

Model	Input	Output
gpt-4-1106-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
gpt-4-1106-vision-preview	$0.01 / 1K tokens	$0.03 / 1K tokens

GPT-3.5 Turbo

Model	Input	Output
gpt-3.5-turbo-1106	$0.0010 / 1K tokens	$0.0020 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

Embedding models

Model	Usage
ada v2	$0.0001 / 1K tokens

Demo

Brainstorming: Textual Datasets

(Three minutes, Groups of three)

Think of as many different textual datasets as you can three minutes.

Brainstorming: Building Context

(Five minutes, Groups of three)

Pick one set.

- Think of a context you can build with that data
    - What field or area or discipline or topic can this data be used to support GPT in answering questions?
- Name that context.
- Come up with five questions that you think GPT could answer give that context.

Brainstorming: Present

Share your context and questions with the cohort.