23. Building a RAG

(locally)

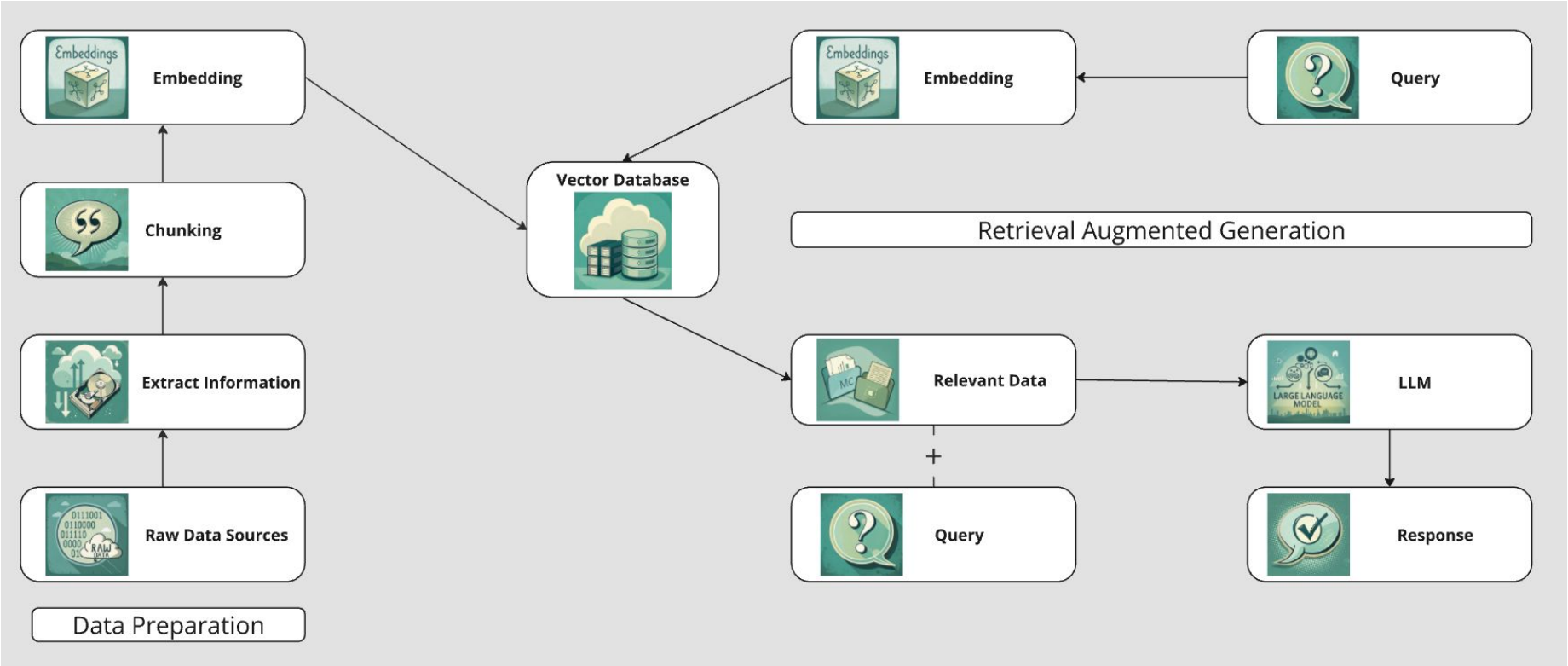
14 януари 2025

Контролно на 16.01.2025 (Четвъртък)

Какво за бога е RAG?

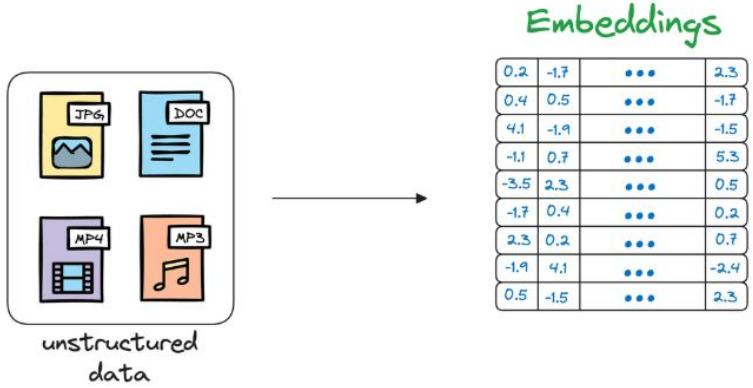
- **RAG -** Съкратено от **Retrieval Augmented Generator**
- Техника в Изкуствения Интелект(**AI**), която комбинира способностите на предварително тренираните езикови модели (като GPT-3, GPT-4, Claude Sonnet 3.5, Lamma, etc) с външни данни.
- Защо да използваме RAG-ове?
    - Подобряват точността на традиционните големи езикови модели- намаляват халюцинациите; моделът има само релевантна информация
    - Разрешават на моделите да отговарят на базата на нова информация или информация, до която не са имали достъп по време на тренирането
    - Разрешават ни да “нагласяме” моделите си спрямо бизнес нуждите ни или конкретна задача
- Слайдовете днес са към...

Как изглежда една базова RAG Архитектура:



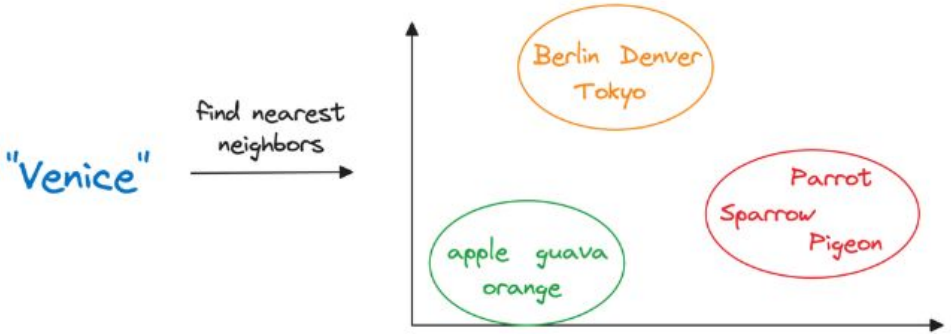
Векторна База Данни

- Специализиран тип база за съхранение, управление и търсене на неструктурирани данни
- Основния компонент е репрезентирането на неструктурираните ни данни, като ембединг(embedding- вектор улавящ семантиката на данните)



Векторна База Данни 2

- Ефективно индексиране
  - Flat index (ANN)
  - Inverted File Index
  - Product quantization
  - Hierarchical Navigable Small World
- Търсене
  - Semantic Search
  - Recommendation Systems
  - Image or Audio Retrieval



Пример - снимки









Пример - снимки

Март



Април

Май


Юни



Пример - снимки

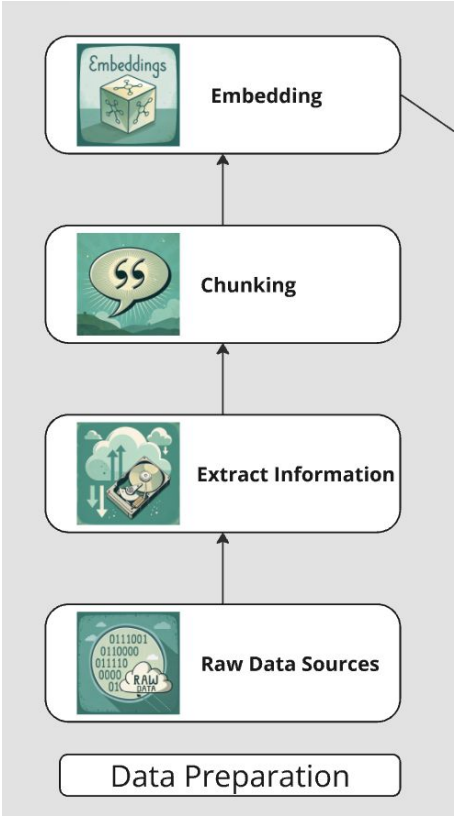
"Mountain images"

Пример - снимки

- Векторната база не пази само векторите, а и самите неструктурирани данни
- Това ни разрешава да връщаме точните данни, когато бъдат търсени

Data Preparation

- Raw Data Source -> например някой пдф
- Extract Information -> Обработваме и почистваме текстовите данни
- Chunking -> разделяме на по-малки части
  - Защо това е важно?
- Embedding -> прекарваме всеки chunk през ембединг модела



Ollama

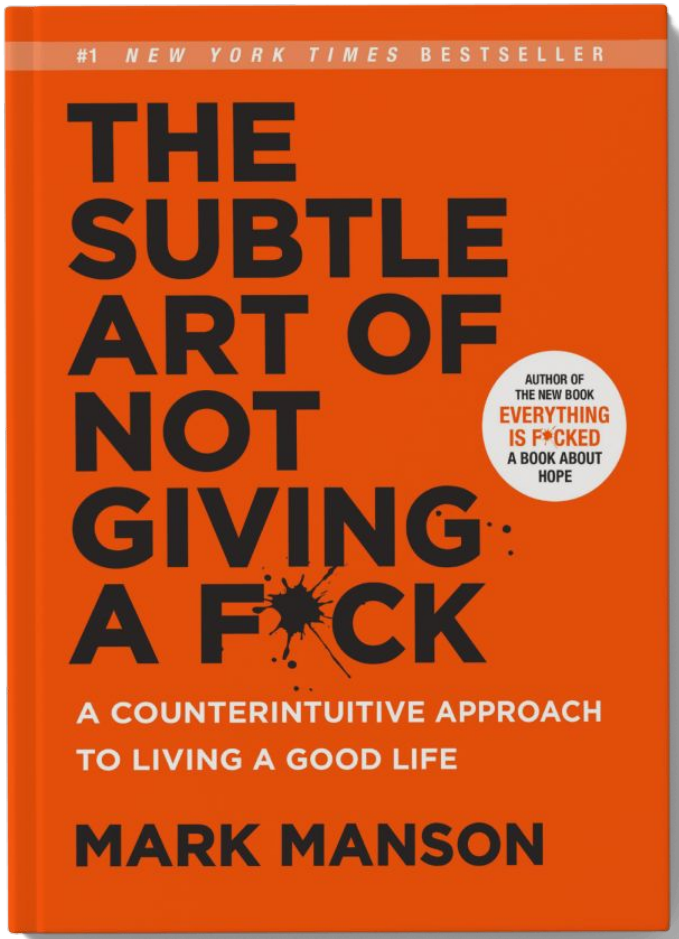
- https://ollama.com/ <- Инсталираме си я
- Защо точно Ollama?

Векторна база данни - again...

- Ще използваме Qdrant
- Необходимо е да инсталираме и Docker

```
1. Install qdrant-client- `pip install qdrant-client`
2. Make sure you have docker installed
3. Open a terminal and run `docker pull qdrant/qdrant`
4. Run `docker run -p 6333:6333 -p 6334:6334 -v $(pwd)/qdrant_storage:/qdrant/storage:z
qdrant/qdrant`
```

Книжка


#1 NEW YORK TIMES BESTSELLER
THE SUBTLE ART OF NOT GIVING A F*CK
AUTHOR OF THE NEW BOOK EVERYTHING IS F*CKED A BOOK ABOUT HOPE
A COUNTERINTUITIVE APPROACH TO LIVING A GOOD LIFE
MARK MANSON

Процес

- Обработваме пдф-а
- Почистваме текста
- Разделяме на парчета (Chunking)
- Минаваме през модела за векторизиране (Embedding)
- Съхраняваме във векторната ни база данни
- Задаваме въпроси! 🎉
- Лимитации?

Въпроси?