**Resources:**

1. [**https://openai.com/index/new-and-improved-embedding-model/**](https://openai.com/index/new-and-improved-embedding-model/)

text-embedding-ada-002 text embedding model;

1. <https://huggingface.co/thenlper/gte-large>

gte-large text embedding model

1. LanceDB: <https://lancedb.com/>
2. Pinecone;

**--Algorithms:** [introduction of embedding](introduction%20of%20embedding)

1. [Latent semantic analysis](https://en.wikipedia.org/wiki/Latent_semantic_analysis) (LSA) --ingular value decomposition of the term-document matrix (fancy linear algebra, basically);
2. [Latent Dirichlet allocation](https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation) (LDA) --a statistical method called the Dirichlet process;
3. Word2vec; [Word2vec essentials](https://medium.com/@manansuri/a-dummys-guide-to-word2vec-456444f3c673)

**--Methodology/architecture:**

1. Neural methods:

* Well defined algorithms from the above to map bag-of-words to vector space of our embedding;
* input words🡪vector in embedding space;
* training data 🡪 which should be closer, which should be farther 🡪 adjust input vectors during training;

1. Deal with sequences:

* Sequential models:

--tokens (fully recurrent neural network);

--transformer; capture context and dependencies in a sequence; run in parallel with tensor operations; attention mechanism: token influence between each other

**Performance metrics: (metric choice)**

1. Evaluate LLM’s effectiveness and accuracy in handling queries;

**A close-up of a text

Description automatically generated**

**Updates:**

1. LanceDB installed;