RAG embeddings

Nico Morin

The Models

Openai: text-embedding-ada-002

Originally released in late 2022 this model is no longer openai’s premier embedding model but it is still on par with them without being massive

SBERT: all-MiniLM-L6-v2

An open source open licence embedding model easily accessible and usable through python

Initial testing

Results:

SBERT embedding:

Precision: 1.0000

MRR: 1.0000

openai embedding:

Precision: 1.0000

MRR: 1.0000

The dataset

Stanford Question and Answers Dataset (SQuAD):

- A question and answer dataset that contains over 100,000 different questions and answers

- Ideal for testing semantic search

- Has an extension that adds in unanswerable questions (not tested here)

SQuAD results

Results:

SBERT Embedding Model:

Precision@k:

P@1: 0.6400

P@3: 0.2844

P@5: 0.1853

P@10: 0.0953

Recall@k:

R@1: 0.6400

R@3: 0.8533

R@5: 0.9267

R@10: 0.9533

Mean Reciprocal Rank: 0.7634

OPENAI Embedding Model:

Precision@k:

P@1: 0.6467

P@3: 0.2800

P@5: 0.1787

P@10: 0.0920

Recall@k:

R@1: 0.6467

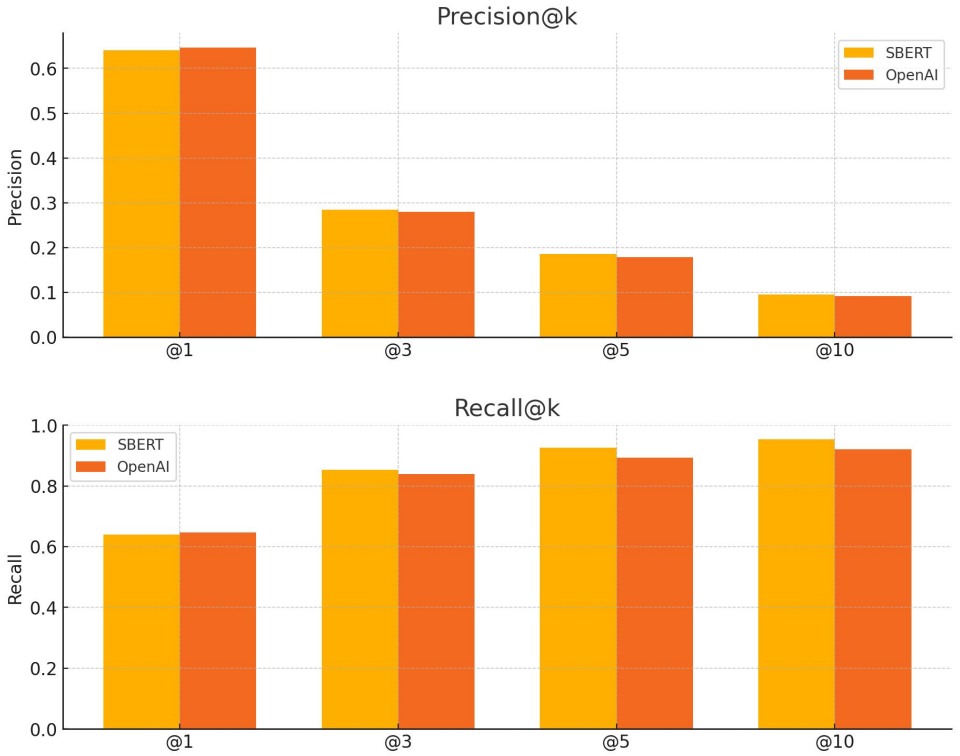
R@3: 0.8400

R@5: 0.8933

R@10: 0.9200

Mean Reciprocal Rank: 0.7537

Results cont.



Conclusions

SBERT outperformed openAI in almost all fields but not necessarily by much. However, SBERT has a leg up on the openAI models by being open source and easily adaptable to needs.

OpenAI’s model has the potential to be easily integratable with other openai products which may be useful.

Besides these small differences both models perform quite well.