Scalable Semantic Search @ Course Hero

Kazem Jahanbakhsh, Harsh Seth, Reza Jamei, Saurabh Khanwalkar, Ishan Parmar, Jennifer Loui, Abbas Rizvi, Dayne Batsman, Vipul Gagrani, Isha Ghai

April 2022

Outline

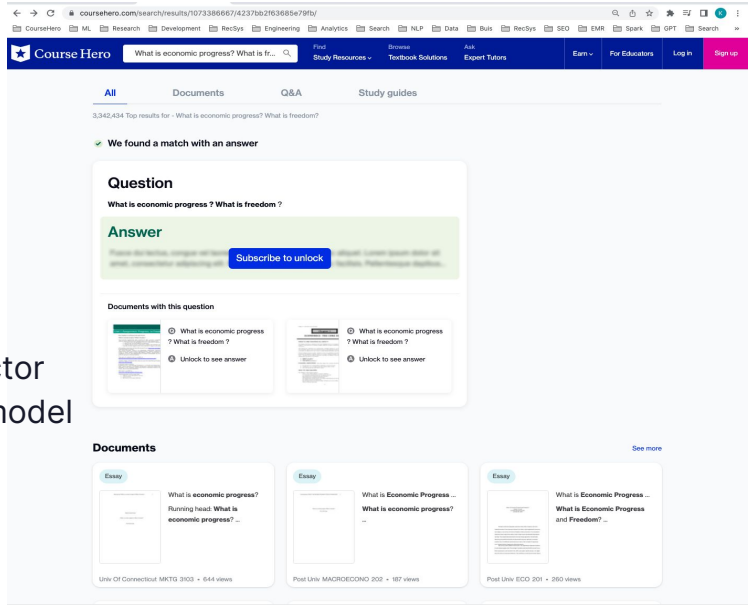
- **Demo**!
- **Semantic Search Problem**
  - Semantic Search Problem Definition
  - Semantic Search Architecture Overview
  - Sentence Transformers for Semantic Similarity Task
- **Semantic Search Evaluations**
  - Quora Dataset as Gold Judgment List
  - Universal Sentence Embedding vs Sent-BERT
  - USE vs Sent-BERT Embedding Observations
  - Performance Analysis: USE vs SBERT
  - A/B Test Results on Prod
- **Fine Tuning Sent-BERT**
  - Fine Tuning SBERT: Why, How, Challenges
  - Performance Analysis (I): SBERT-FT vs SBERT Precision
  - Performance Analysis (II): SBERT-FT vs SBERT Recall
  - SBERT-FT vs SBERT A/B Test Results
- **Wrap up**
  - What's next?
  - We are hiring Search/Platform/ML Engineers !

Semantic Search Problem

Semantic Search Use Cases @ Course Hero!

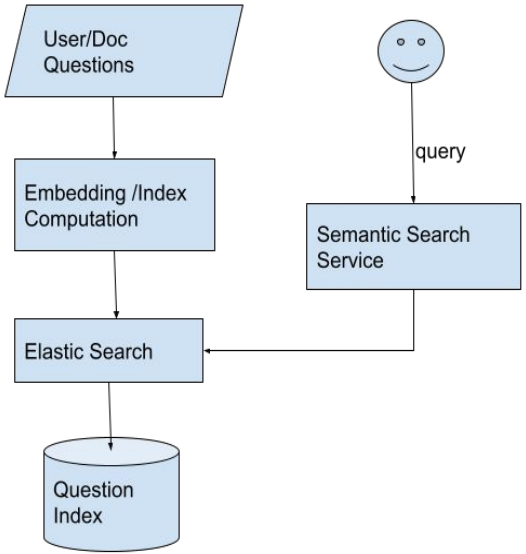
Semantic Search

- **Problem definition**:
  - Given a set of questions, we want to let users to search using a “**question query**” and find the most similar question to their query
- **ML metrics**:
  - Mean Average Precision@k
  - Recall@k
- **Prod metrics**:
  - Conversion rate (e.g. purchase rate)
  - Coverage
- **Embedding models**:
  - Embedding model takes a query text and generates a vector
  - We used **Universal Sentence Encoder v3** as the legacy model



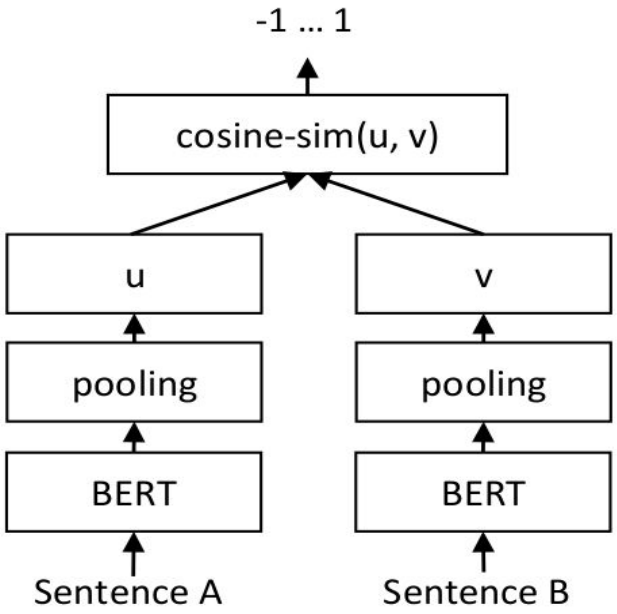
Course Hero Semantic Search Architecture

- **Source of questions:**
  - We have upstream NLP pipelines that produces document/user questions (100's of millions)
- **Embedding Computer / Index Insertion**
  - Embedding service ingests a question text and computes its embedding vector
  - The same service computes **LSH hash codes**
  - Last we update the ElasticSearch (i.e. ES) index
- **Semantic search service**
  - Computes query embedding & LSH hash code using query text
  - Runs NN to get top k most similar questions
  - The scoring part is implemented using an ES plugin



Sentence Transformers (SOA!)

- **sentence-BERT (SBERT)** was introduced in 2019
- SBERT outperforms previous models for Semantic Textual Similarity tasks
- SBERT is more scalable than BERT in terms of running time
- SBERT was trained using similar/dissimilar sentences
- SBERT supports a number of loss functions for training:
  - Cosine Similarity Loss
  - Triplet Loss
  - Margin MSE Loss and many more



Semantic Search Evaluations

Performance Analysis: USE v3 vs SBERT

Embedding & NN Algo	MAP@1	MAP@3	Recall@3	MAP@20	Recall@20	Recall@100	Recall@500
Semantic Search Prod (USE V3+LSH)	NA	0.628	0.647	0.646	0.826	0.901	0.927
USE V3+Faiss **(HNSW15)**	NA	0.629	0.65	0.65	0.84	0.933	0.975
SBERT (wo normalizing)+Faiss	0.7494	0.7462	0.7546	0.76711	0.9309	0.9806	0.9900
SBERT(L2 norm)+Faiss **(HNSW15)**	0.776	0.77	0.775	0.791	0.944	0.99	0.999
SBERT(wo normalizing)+LSH	0.7741	0.7653	0.7673	0.7821	0.92614	0.9643	0.9685

Benchmarking Semantic Search: USE v3 vs SBERT & LSH vs Faiss

A/B Test Results: USE vs SBERT

- We have hundreds of millions of questions so our index pipeline needs to scale
- We achieve this by leveraging **GPU**, **Amazon Batch Transform** and **Apache Spark** frameworks
- We ran an A/B test for **SBERT vs USE** for semantic search
- A/B results showed that while USE and SBERT are close in terms of CVR, with lower threshold for SBERT **we could increase the coverage by 32% :)**
- This was a clear win so we shipped SBERT model to prod
- Note that the two models have different score distribution so this requires some calibration before prod launch

Quora Dataset as Gold Judgment List

- Quora dataset contains a list of 400k question pairs
  - Goal was to detect duplicate questions on Quora website!
  - **Positive pairs**: A pair of questions that are semantically equivalent
  - **Negative pairs**: A pair of questions that are are NOT semantically equivalent
- Examples
  - Positive pair:
    - “How do you start a bakery?”
    - “How can one start a bakery business?”
  - Negative pair:
    - “What are natural numbers?”
    - “What is a least natural number?”
- Quora dataset could be used for benchmarking semantic search since we want to find similar questions
- **Quora Judgment List Methodology**
  - We sampled **~12k** questions from Quora questions
  - With quora question and semantic label we created a semantic graph
  - If q1 == q2 & q2 == q3 then we generate the following tuples as semantically equivalent questions:
    - q1 → (q2, q3)
    - q2 → (q1, q3)
    - q3 → (q1, q2)

Universal Sentence Embedding vs Sent-BERT

- Our goal is to improve MAP@k & Recall@k for semantic search
- Our legacy embedding model was **USE v3:**
  - USE v3 uses Deep Average Network (DAN)
  - USE vector dimensionality is **512**
- We tested pre-trained sent-BERT models:
  - For SBERT we picked “**paraphrase-mpnet-base-v2**”
  - SBERT vector dimensionality is **768** and max sequence length is **512**
  - The Base Model is “**microsoft/mpnet-base**”
  - This model was one of the best pre-trained models published in 2021 in terms of STS performance

USE vs Sent-BERT Embedding Observations

Sentence 1	Sentence 2	USE v3 cosine	USE v4 cosine	SBERT cosine	Observations
Kelly's organization corollary assumes which type of relationship among constructs?	Analytic Hierarchy Process Define Hierarchy model :	**0.835**	0.300	0.3932	Topics seem to be not related. One is about personal constructs and how people behave, the other is on the decision making process!
Kelly's organization corollary assumes which type of relationship among constructs?	To represent a relationship in the relational model , the primary key of one relation is placed into a second relation	**0.819**	0.3702	0.3812 4	Topics are different - one is psychology and people behavior and the other one is Database
Kelly's organization corollary assumes which type of relationship among constructs?	In object - relations theory , later relationships build upon :	**0.828**	0.4469	0.4755	Somewhat related but not quite!

Different embedding similarity calculations

Fine Tuning Sent-Bert

Performance Analysis (I): SBERT-FT vs SBERT Precision

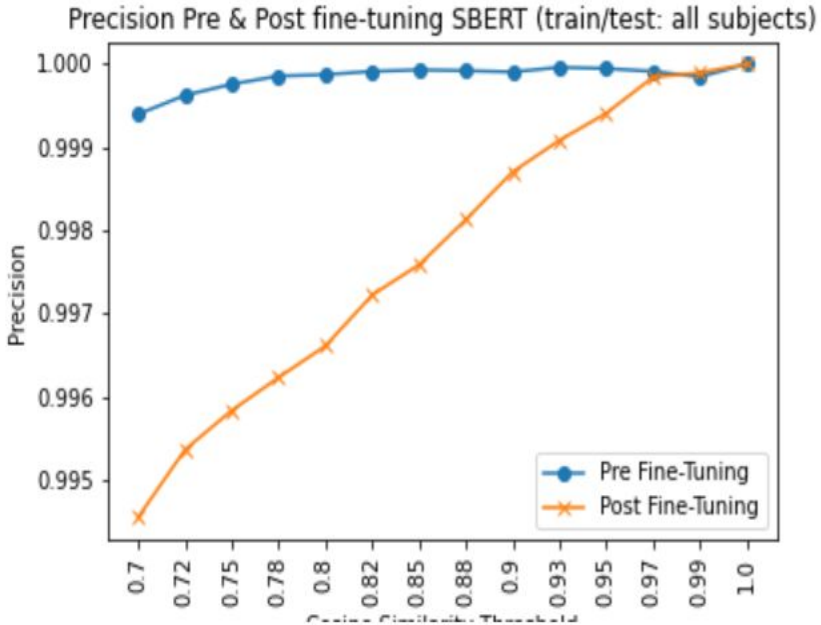


Figure 38: Precision of SBERT vs fine tuned SBERT trained and tested on all 40 subjects (x axes is cosine similarity - A/B test model)

Performance Analysis (II): SBERT-FT vs SBERT Recall

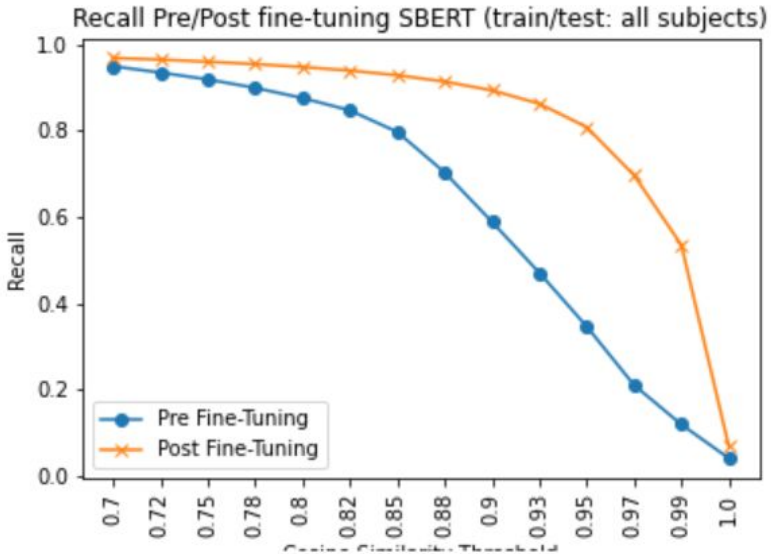


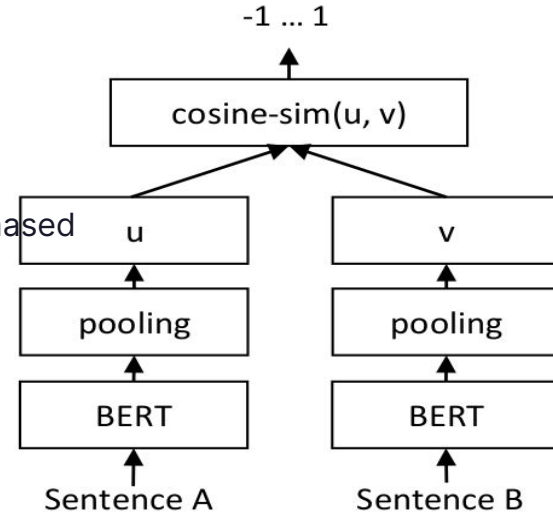
Figure 39: Recall of SBERT vs fine tuned SBERT trained and tested on all 40 subjects (x axes is cosine similarity - A/B test model)

A/B Test Results: SBERT-FT vs SBERT

- We generated SBERT-FT embedding & LSH hashes for one of our indices
- We ran an A/B test for SBERT-FT vs SBERT (“**paraphrase-mpnet-base-v2**”)
- SBERT-FT cosine similarity scores shifted after fine tuning
- We used **0.99** threshold for SBERT-FT and **0.92** for the Original SBERT
- CVR (purchase rate) was the same for the two models
- SBERT-FT model gave a **35% Lift in Coverage**
- Our plan is to ship SBERT-FT to production for all indices for semantic search use case

Fine Tuning SBERT: Why, How, Challenges

- Why?
  - Improve SBERT's recall while keeping precision high
  - Sent-BERT model has been trained on web datasets like Yahoo Q&A & stack exchange
  - These models have not seen course hero question language.
- How?
  - We used Siamese Networks to fine tune SBERT model
- Training dataset?
  - We sampled positive and negative pairs from 40 subjects
  - Positive pairs were sampled for a pair of query, question that has been purchased
  - Negative pairs were randomly sampled
- We used "Cosine Similarity" as the loss function
- Challenges:
  - We can't really use human to label data at scale!
  - Random negative pairs are not ideal since they are easy to be detected



Wrap up

What's next?

- We will be migrating from our internal ES semantic search pipeline to **Pinecone Vector DB technology** (www.pinecone.io)!
- Improve quality of training data for fine tuning SBERT:
  - Explore techniques for sampling hard negatives as well as random negatives
  - This will improve fine tuned model precision (CVR)
- We know that creating training data for supervised fine tuning is hard! What if we try a semi-supervised approaches like **unsupervised SimCSE** (Simple Contrastive Learning of Sentence Embeddings)
- Testing triplet loss, multiple negative ranking and other loss functions for fine-tuning

We are hiring Search/Platform/ML Engineers!

- Course Hero is growing fast! We just did a series C 380ml fundraising in December 2021
- We have a number of positions for search and recommendations
- We are hiring for **Staff ML platform Engineer** and **Search Engineer**
- We hire in Redwood City (California) and Vancouver/Toronto in Canada
- Come and talk to us if you are interested!
- Visit www.coursehero.com/jobs to see the list of positions

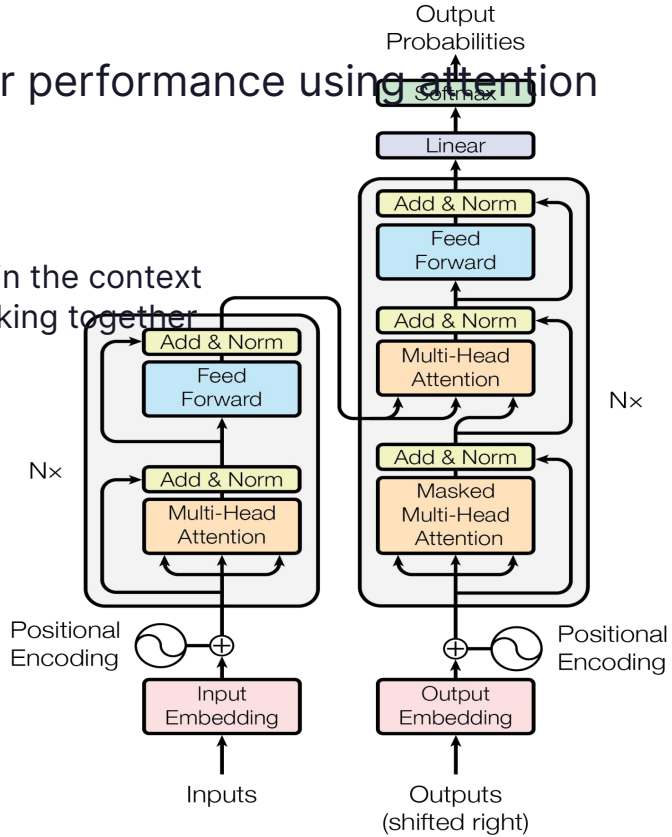
References

- Universal Sentence Encoder paper
- USE models: https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder
- Sentence-BERT website: https://www.sbert.net/
- Pre-trained sentence-transformers models: https://huggingface.co/sentence-transformers
- SBERT Loss functions: https://www.sbert.net/docs/package_reference/losses.html
- Quora duplicate dataset: https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs
- Sentence Transformers: Meanings in Disguise by Pinecone
- Attention Is All You Need
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- SimCSE: Simple Contrastive Learning of Sentence Embeddings
- Pinecone Vector Similarity Search DB

Extra slides

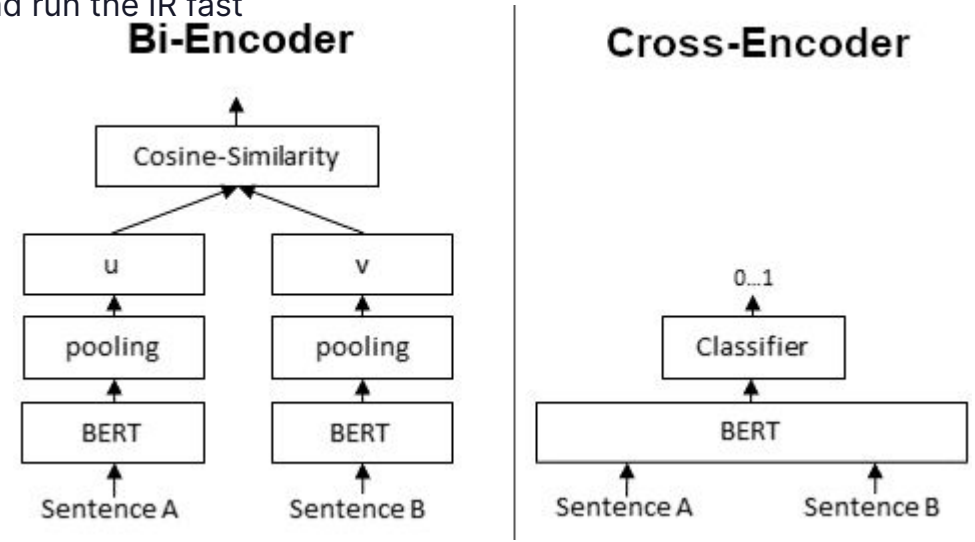
NLP Transformers: Attention is All You Need!

- Paper titled *Attention Is All You Need* showed superior performance using attention mechanism compared to RNN
- Transformers compose of:
  - **Positional Encoding**: ability to consider order of a sequence
  - **Self-attention**: attention applied between a word & other words in the context
  - **Multi-head attention**: several parallel attention mechanisms working together
- Pretrained transformer models:
  - Google has released BERT
  - New transformer models generalized much better than RNNs
  - Swap the last few layers for different use cases:
    - Classification
    - Q&A
    - POS-tagging
- BERT spawned a number of new models:
  - distilBERT
  - RoBERTa
  - ALBERT

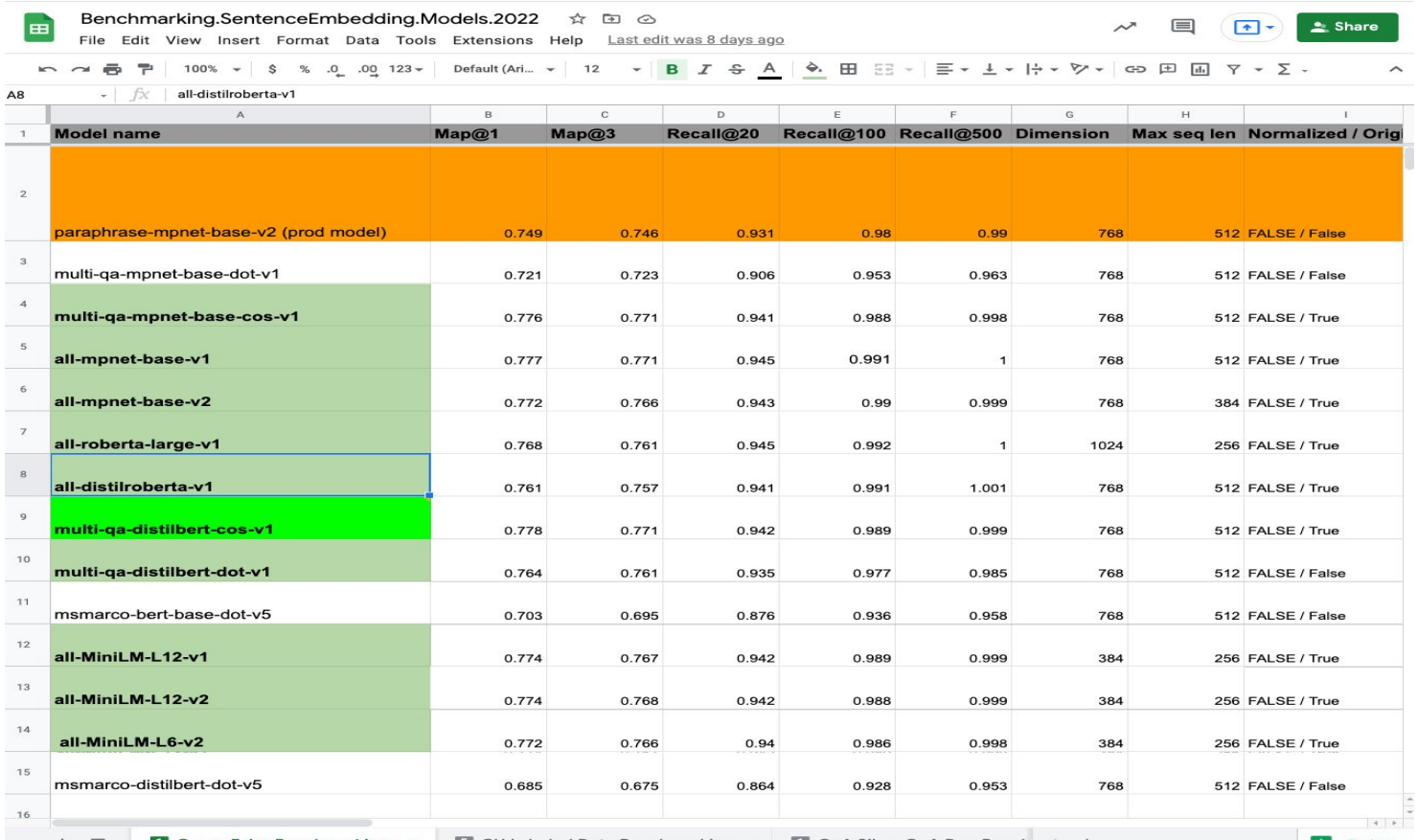


Sentence Similarity

- Cross-encoder:
  - We can use a cross-encoder & pass two questions and get their similarity
  - Cross-encoder outperform Bi-Encoder performance but doesn't scale for IR!
- BERT/GloVe:
  - We can computing a sentence embedding by averaging token embeddings output by BERT
  - Advantage: we can generate embeddings and run the IR fast
  - Disadvantage: Low accuracy!



Benchmarking Pre-Trained Models I



Benchmarking.SentenceEmbedding.Models.2022

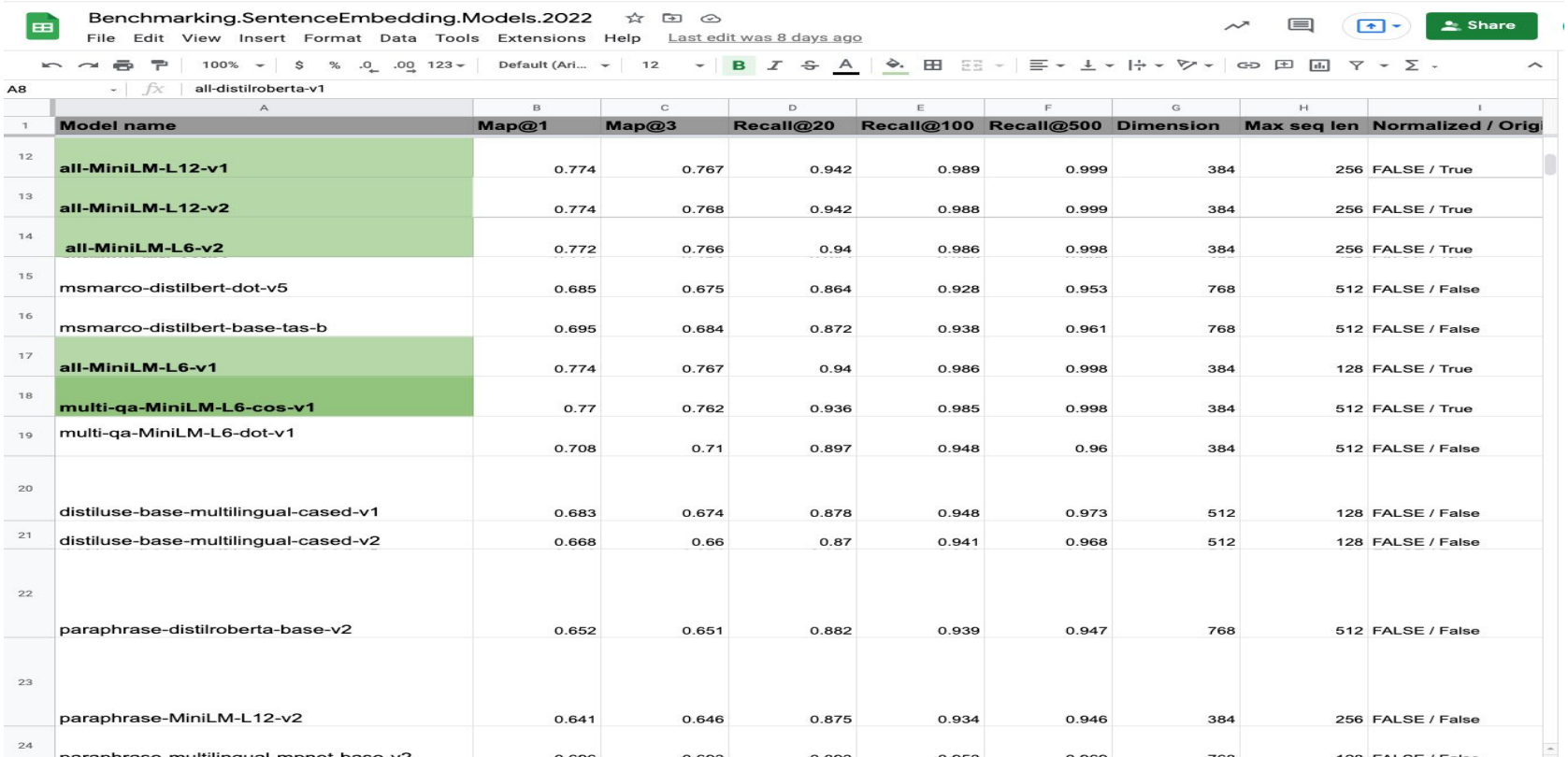
File Edit View Insert Format Data Tools Extensions Help Last edit was 8 days ago

Share

A8 all-distilroberta-v1

	Model name	Map@1	Map@3	Recall@20	Recall@100	Recall@500	Dimension	Max seq len	Normalized / Orig
2	paraphrase-mpnet-base-v2 (prod model)	0.749	0.746	0.931	0.98	0.99	768	512	FALSE / False
3	multi-qa-mpnet-base-dot-v1	0.721	0.723	0.906	0.953	0.963	768	512	FALSE / False
4	multi-qa-mpnet-base-cos-v1	0.776	0.771	0.941	0.988	0.998	768	512	FALSE / True
5	all-mpnet-base-v1	0.777	0.771	0.945	0.991	1	768	512	FALSE / True
6	all-mpnet-base-v2	0.772	0.766	0.943	0.99	0.999	768	384	FALSE / True
7	all-roberta-large-v1	0.768	0.761	0.945	0.992	1	1024	256	FALSE / True
8	all-distilroberta-v1	0.761	0.757	0.941	0.991	1.001	768	512	FALSE / True
9	multi-qa-distilbert-cos-v1	0.778	0.771	0.942	0.989	0.999	768	512	FALSE / True
10	multi-qa-distilbert-dot-v1	0.764	0.761	0.935	0.977	0.985	768	512	FALSE / False
11	msmarco-bert-base-dot-v5	0.703	0.695	0.876	0.936	0.958	768	512	FALSE / False
12	all-MiniLM-L12-v1	0.774	0.767	0.942	0.989	0.999	384	256	FALSE / True
13	all-MiniLM-L12-v2	0.774	0.768	0.942	0.988	0.999	384	256	FALSE / True
14	all-MiniLM-L6-v2	0.772	0.766	0.94	0.986	0.998	384	256	FALSE / True
15	msmarco-distilbert-dot-v5	0.685	0.675	0.864	0.928	0.953	768	512	FALSE / False
16									

Benchmarking Pre-Trained Models II



Benchmarking.SentenceEmbedding.Models.2022

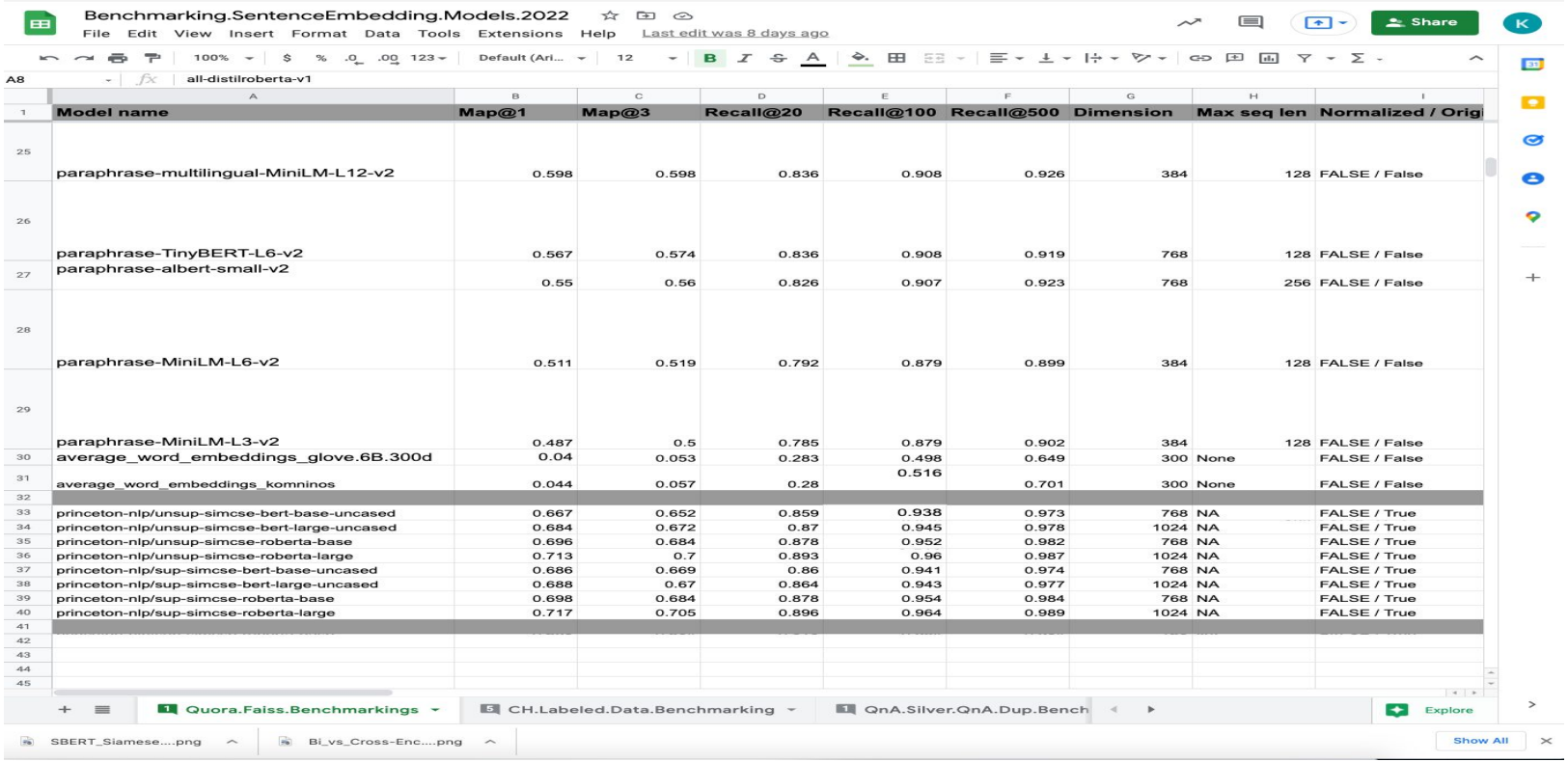
File Edit View Insert Format Data Tools Extensions Help Last edit was 8 days ago

Share

A8 all-distilroberta-v1

	Model name	Map@1	Map@3	Recall@20	Recall@100	Recall@500	Dimension	Max seq len	Normalized / Orig
12	all-MiniLM-L12-v1	0.774	0.767	0.942	0.989	0.999	384	256	FALSE / True
13	all-MiniLM-L12-v2	0.774	0.768	0.942	0.988	0.999	384	256	FALSE / True
14	all-MiniLM-L6-v2	0.772	0.766	0.94	0.986	0.998	384	256	FALSE / True
15	msmarco-distilbert-dot-v5	0.685	0.675	0.864	0.928	0.953	768	512	FALSE / False
16	msmarco-distilbert-base-tas-b	0.695	0.684	0.872	0.938	0.961	768	512	FALSE / False
17	all-MiniLM-L6-v1	0.774	0.767	0.94	0.986	0.998	384	128	FALSE / True
18	multi-qa-MiniLM-L6-cos-v1	0.77	0.762	0.936	0.985	0.998	384	512	FALSE / True
19	multi-qa-MiniLM-L6-dot-v1	0.708	0.71	0.897	0.948	0.96	384	512	FALSE / False
20	distiluse-base-multilingual-cased-v1	0.683	0.674	0.878	0.948	0.973	512	128	FALSE / False
21	distiluse-base-multilingual-cased-v2	0.668	0.66	0.87	0.941	0.968	512	128	FALSE / False
22	paraphrase-distilroberta-base-v2	0.652	0.651	0.882	0.939	0.947	768	512	FALSE / False
23	paraphrase-MiniLM-L12-v2	0.641	0.646	0.875	0.934	0.946	384	256	FALSE / False
24	paraphrase-multilingual-mpnet-base-v2								

Benchmarking Pre-Trained Models III



Benchmarking.SentenceEmbedding.Models.2022

File Edit View Insert Format Data Tools Extensions Help Last edit was 8 days ago

Share

A8 all-distilroberta-v1

	Model name	Map@1	Map@3	Recall@20	Recall@100	Recall@500	Dimension	Max seq len	Normalized / Orig
25	paraphrase-multilingual-MiniLM-L12-v2	0.598	0.598	0.836	0.908	0.926	384	128	FALSE / False
26	paraphrase-TinyBERT-L6-v2	0.567	0.574	0.836	0.908	0.919	768	128	FALSE / False
27	paraphrase-albert-small-v2	0.55	0.56	0.826	0.907	0.923	768	256	FALSE / False
28	paraphrase-MiniLM-L6-v2	0.511	0.519	0.792	0.879	0.899	384	128	FALSE / False
29	paraphrase-MiniLM-L3-v2	0.487	0.5	0.785	0.879	0.902	384	128	FALSE / False
30	average_word_embeddings_glove.6B.300d	0.04	0.053	0.283	0.498	0.649	300	None	FALSE / False
31	average_word_embeddings_komninos	0.044	0.057	0.28	0.516	0.701	300	None	FALSE / False
32									
33	princeton-nlp/unsup-simcse-bert-base-uncased	0.667	0.652	0.859	0.938	0.973	768	NA	FALSE / True
34	princeton-nlp/unsup-simcse-bert-large-uncased	0.684	0.672	0.87	0.945	0.978	1024	NA	FALSE / True
35	princeton-nlp/unsup-simcse-roberta-base	0.696	0.684	0.878	0.952	0.982	768	NA	FALSE / True
36	princeton-nlp/unsup-simcse-roberta-large	0.713	0.7	0.893	0.96	0.987	1024	NA	FALSE / True
37	princeton-nlp/sup-simcse-bert-base-uncased	0.686	0.669	0.86	0.941	0.974	768	NA	FALSE / True
38	princeton-nlp/sup-simcse-bert-large-uncased	0.688	0.67	0.864	0.943	0.977	1024	NA	FALSE / True
39	princeton-nlp/sup-simcse-roberta-base	0.698	0.684	0.878	0.954	0.984	768	NA	FALSE / True
40	princeton-nlp/sup-simcse-roberta-large	0.717	0.705	0.896	0.964	0.989	1024	NA	FALSE / True
41									
42									
43									
44									
45									

Quora.Faiss.Benchmarkings | CH.Labeled.Data.Benchmarking | QnA.Silver.QnA.Dup.Bench | Explore

SBERT_Siamese....png | Bi_vs_Cross-Enc....png | Show All