Building a Search Engine

Jupyter Hub: https://hub.np.training/

Repo: npatta01/search-engine-workshop

About Me

Nidhin Pattaniyil
Linkedin
npatta01@gmail.com

ML Engineer on the Walmart Search team

Our Agenda

---

1. Token Based Retrieval
2. Embedding Based Retrieval
3. Approximate Nearest Neighbor (ANN)
5. Production Considerations

Who you are

- Beginner / Intermediate in the field of NLP ✅
- Beginner in the field of Search ✅

- If you already use embeddings & vector databases 😅

Our Agenda

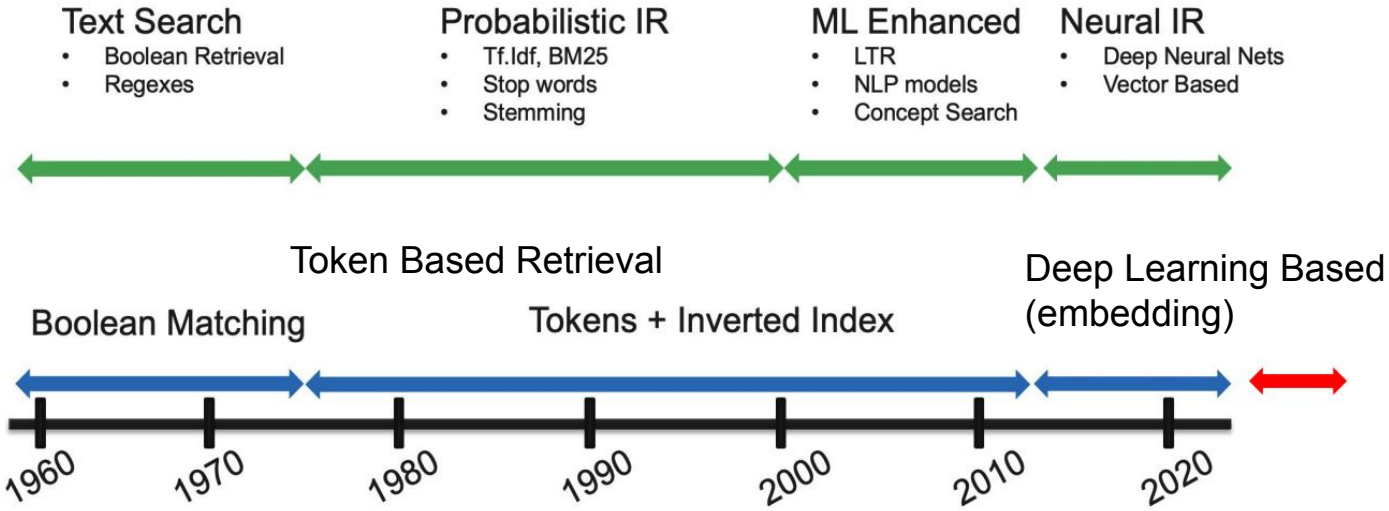
---

1. Token Based Retrieval
2. Embedding Based Retrieval
3. Approximate Nearest Neighbor (ANN)
4. Production Considerations

Learning Objectives

- Use ElasticSearch to implement a Token Based Retrieval
- Use SentenceTransformer Library to use pre-trained models for embedding retrieval
- Understand how to scale Embedding Retrieval using FAISS

History of Information Retrieval



Hughes, Simon. "Semantic Product Search – Vector Search for E-Commerce." Conference Presentation at Haystack 2021, https://haystackconf.com/files/slides/haystack2021/Hughes-Haystack_2021_Semantic_Product_Search.pdf, September 29, 2021.

Dataset

- Provided by https://unsplash.com
- Website to get high quality photography images
- 25K images and captions uploaded by photographers
-

Sample Queries

- person on top of mountain
- the boy and girl on a beach
- person in a desert
- Two dogs playing in the snow
- light at the end of the tunnel
- seven wonders of the world

Sample Queries (more than 4 words)

Detailed Intent

- water droplets on a leaf
- image of a man in a desert
- person on top of mountain

Location:

- ripley's aquarium of canada, toronto, canada
- the butterfly atrium at hershey gardens

Non English Queries

- salar de uyuni uyuni bolivia
- 沙漠青蛙 沙漠青蛙 (desert frog)
- por do sol no mar
- conhece te a ti mesmo (Greek for know thyself)

Metaphors / Slogan:

- light at the end of the tunnel
- there is no planet b

Multiple Candidates

- seven wonders of the world

Long Query / Single Intent

- nova scotia duck tolling retriever (dog breed)

Sample Images

Woden pole on the sand beach

From a sunny afternoon spent wandering around the Glasshouse at RHS Wisley. I loved the texture and colours of these 🌿

Our Agenda

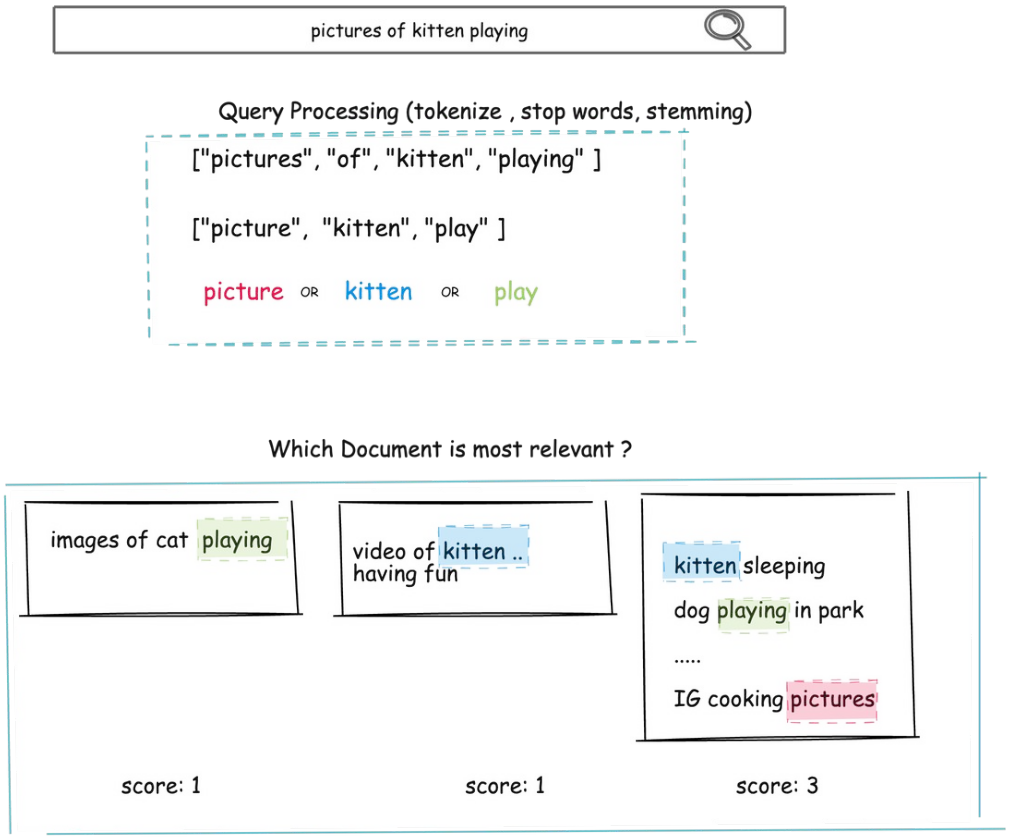
---

1. **Token Based Retrieval**
2. Embedding Based Retrieval
3. Approximate Nearest Neighbor (ANN)
4. Production Considerations

Token based retrieval

Boolean Retrieval

- Queries and documents are represented as bag of words
- Query terms are connected with boolean operators
- Disadvantages:
  - Filtering more than retrieval
  - Terms have same weights
  - No ranking



Better token retrieval algorithm

If a term appears often in a document, then the document is relevant ✅

If a lot of documents contain the term, then the term doesn't contribute much ❌

Better token retrieval algorithm

If a term appears often in a document, then the document is relevant ✅

Token Frequency (t, d)

Number of times term (t) appears in a document (d)

If a lot of documents contain the term, then the term doesn't contribute much ❌

Inverse Document frequency

Better token retrieval algorithm

If a term appears often in a document, then the document is relevant ✅

Token Frequency (t, d)

Number of times term (t) appears in a document (d)

- limit the influence of token frequency on score
- consider document length

If a lot of documents contain the term, then the term doesn't contribute much ❌

Inverse Document frequency

Better token retrieval algorithm: BM25

If a term appears often in a document, then the document is relevant ✅

Token Frequency (t, d)

Number of times term (t) appears in a document (d)

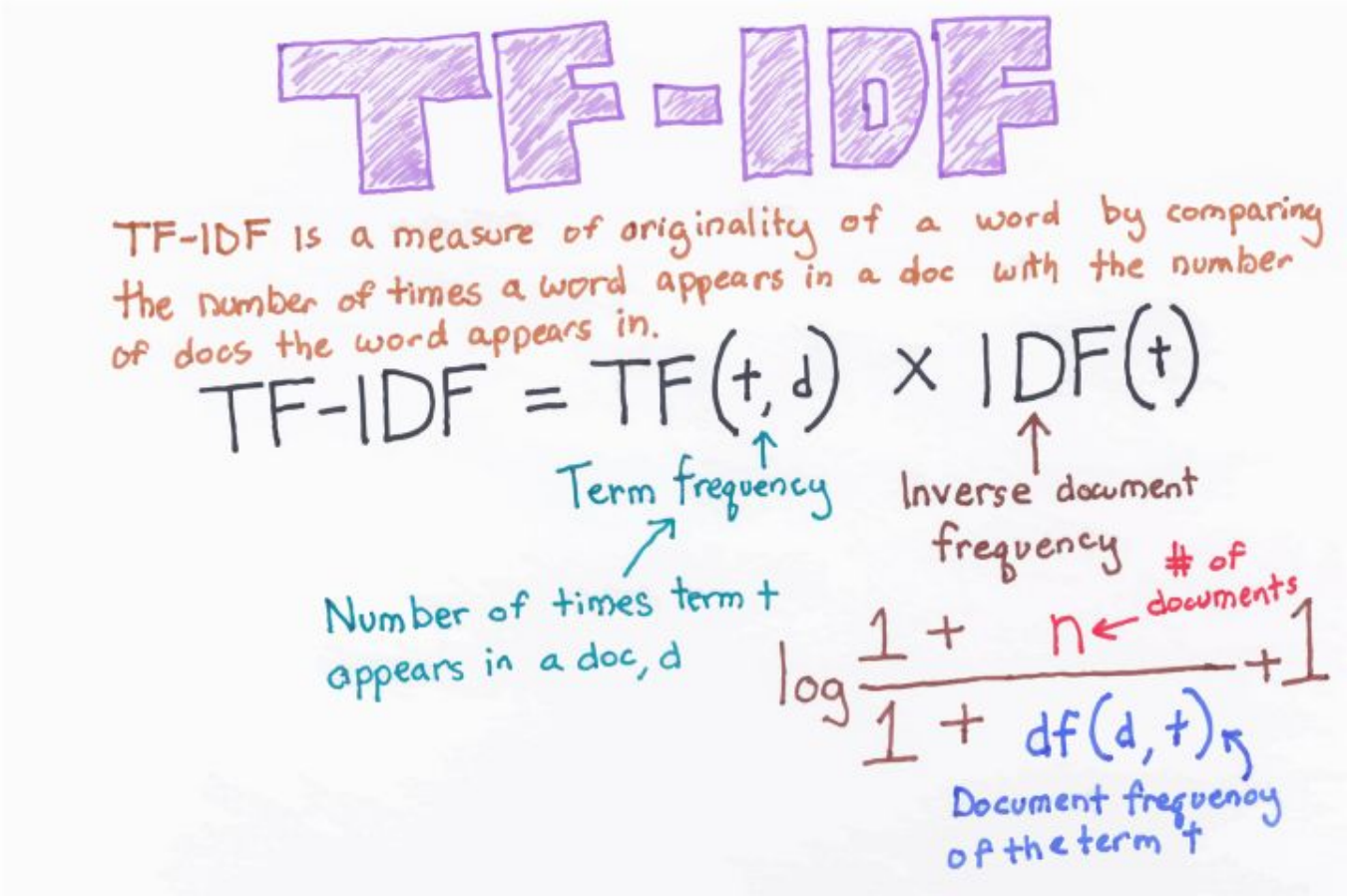
- limit the influence of token frequency on score
- consider document length

If a lot of documents contain the term, then the term doesn't contribute much ❌

Inverse Document frequency

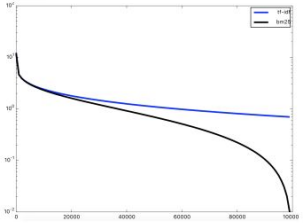
$$\frac{TF(t,d)}{TF(t,d) + k\left(1 - b + b * \frac{len(d)}{\text{average doc length}}\right)}$$

$$\log\left(1 + \frac{N - df(d,t) + 0.5}{df(d,t) + 0.5}\right)$$



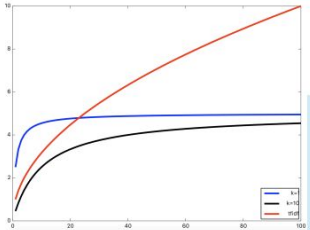
Credit : Chris Albon @ https://machinelearningflashcards.com/

BM25



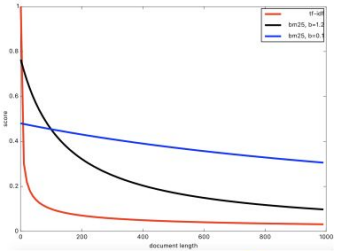
idf - how popular is the term in the corpus?

$$\text{bm25}(d) = \sum_{t \in q, f_{t,d} > 0} \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{f_{t,d}}{f_{t,d} + k \cdot (1 - b + b\frac{\text{l}(d)}{avgdl})}$$



saturation curve - limit influence of tf on the score

length weighing - tweak influence of document length



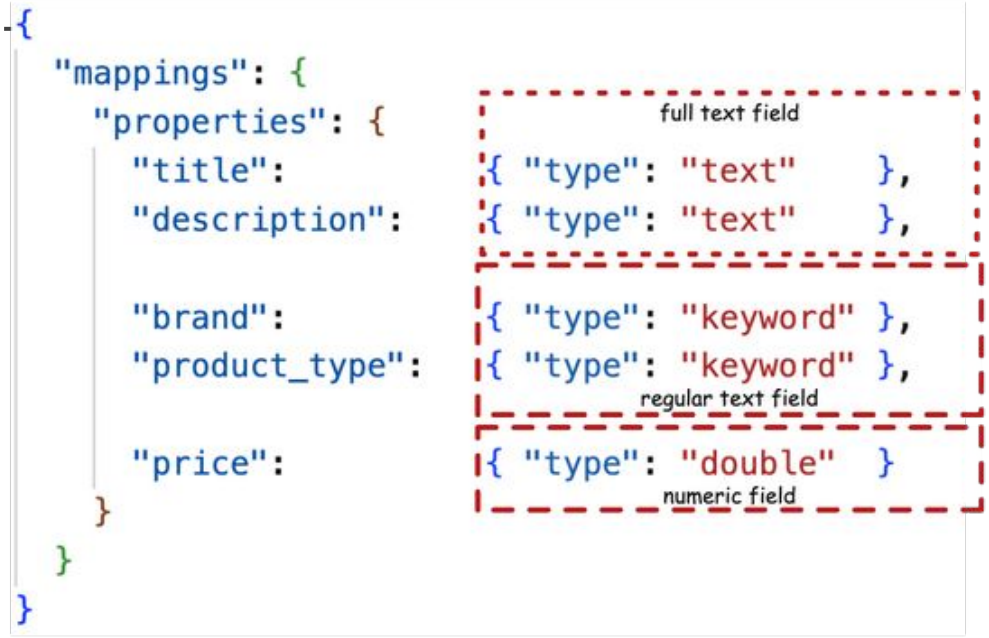
Weber, Britta. "Improved Text Scoring with BM25." Conference Presentation at ElasticCon 2016, https://speakerdeck.com/elastic/improved-text-scoring-with-bm25, February 11, 2016.

ElasticSearch

- Open source search engine based on Lucene library
- Supports BM25 and other similarities (link)
- Supports boosting , filtering , phrase match, autocomplete
- Distributed : index shards, replicas

elasticsearch

ElasticSearch Schema



Possible schema for an e-commerce item

ElasticSearch Query

Query: Nike shoes under 100$

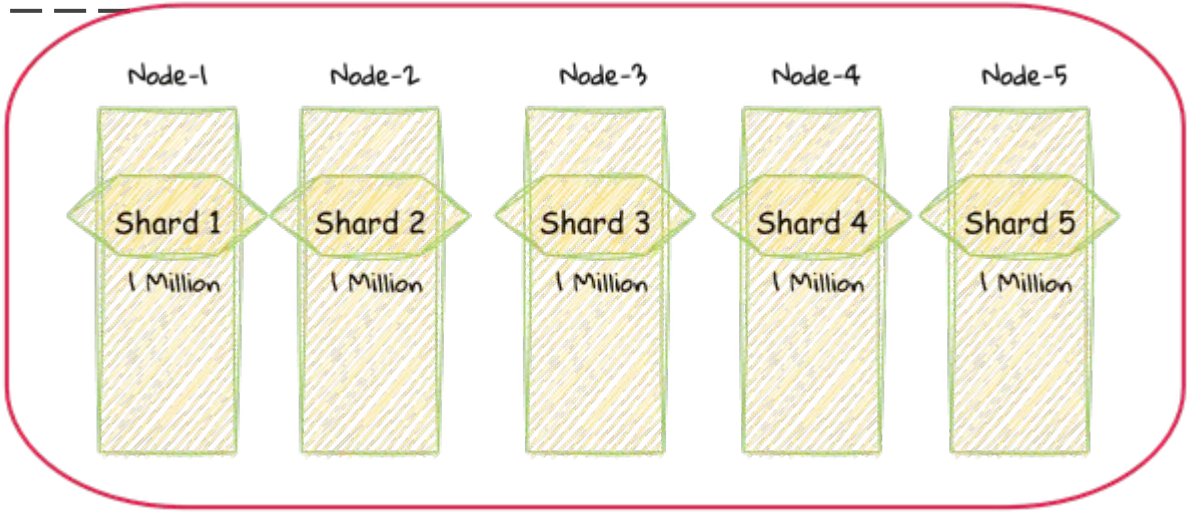
```
{
  "query": {

    "multi_match": {
        "query": "Nike shoe under 100$",
        "fields": ["title^2", "Description^1"]
    }

    ,"bool": {
      "filter": [
        { "term": { "brand": "nike" }}
      ]
    }
    ,"filtered": {
        "filter": {
            "range":  {
                "price" : {  "lte": 100 }
            }
        }
    }
}
```

Search for query tokens in title and description weight matches found in title double

filter items who has brand nike

filter items with price <= 100

Index with multiple shards



- Index can be composed of shards
- Reading / Writing can be faster

Lab

Jupyter Hub: https://hub.np.training
Repo:https://bit.ly/search-workshop-2022

Lab 1 Goals

- Understand different tokenization approaches
- Build a token based retrieval system using ElasticSearch and the BM25 algorithm

Query: Two dogs playing in the snow

Photo title: #dog #dogs #snow
Distance: 9.53

../data/raw/images/5AyUHeSnxz8.jpg

Photo title: Dog in snow
Distance: 9.53

../data/raw/images/jtnH16_HfPE.jpg

Photo title: two cabins covered with snow
Distance: 9.06

../data/raw/images/T3WsIW4QwDU.jpg

Photo title: Two mountain peaks with snow
Distance: 9.06

../data/raw/images/oQgZrmpNwoY.jpg

Query: boy and girl on a beach

Photo title: Winter Girl
Distance: 8.75

../data/raw/images/rGL9bjHDqNM.jpg

Photo title: The girl
Distance: 8.75

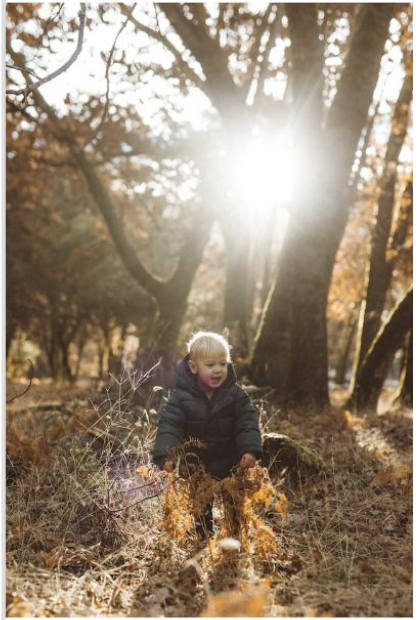
../data/raw/images/p5NcxQ6TE8I.jpg

Photo title: boy wearing goggles underwater
Distance: 8.55

../data/raw/images/0O0Cia89YLU.jpg

Photo title: boy stands under trees
Distance: 8.55

../data/raw/images/vJNe-sw36WU.jpg



Query: Seven Wonders of the world

Photo title: Cloud Seven | Instagram: @timmossholder
Distance: 10.35

../data/raw/images/O76IJx0xPb0.jpg

Photo title: Seven Seventeen: Thursday's Gone
Distance: 10.35

../data/raw/images/qKmkjx87OBg.jpg

Photo title: grayscale photography of seven moon illustration
Distance: 9.11

../data/raw/images/LD8FhvuLiEA.jpg

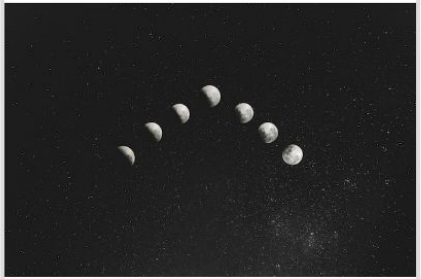


Photo title: World globe
Distance: 8.25

../data/raw/images/gEKMstKfZ6w.jpg

Takeaways

- How different tokenizers can affect the results
- BM25 retrieval is fast and is explainable
- ElasticSearch makes using BM25 accessible

Our Agenda

---

1. Token Based Retrieval
2. **Embedding Based Retrieval**
3. Approximate Nearest Neighbor (ANN)
4. Production Considerations

Embedding Based Retrieval

Issues with Token Based Retrieval

- Lexical GAP:
  - Covid vs Coronavirus vs Omicron variant
  - Car vs Automobile vs vehicle
- Ambiguity: bank (institution) vs bank (geography)

Issues with Token Based Retrieval

- Position matters: “river bank” vs “bank river”
- Lack of Contextualized embeddings

She will ***park*** the car so we can walk in the ***park***.

Dense Embeddings

Word Representation

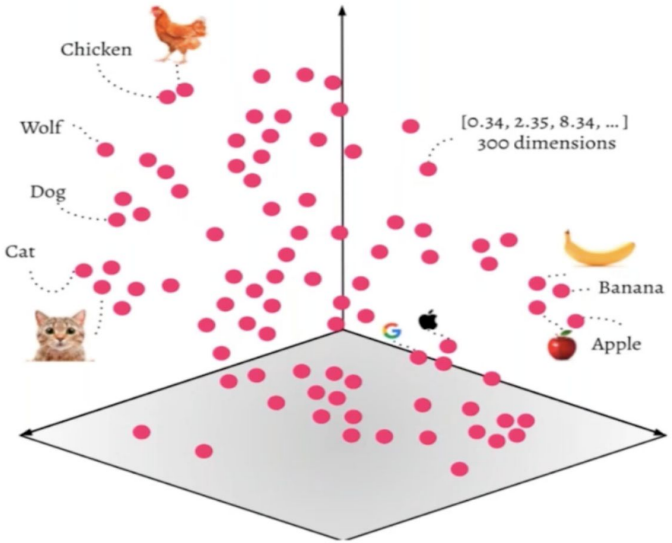
car: [0.2 , 0.3, 0.7]

automobile: [0.2 , 0.3, 0.7]

Similar concepts have similar embeddings

Regardless of content length, similar items should have similar embeddings

Size of embedding is independent of #tokens



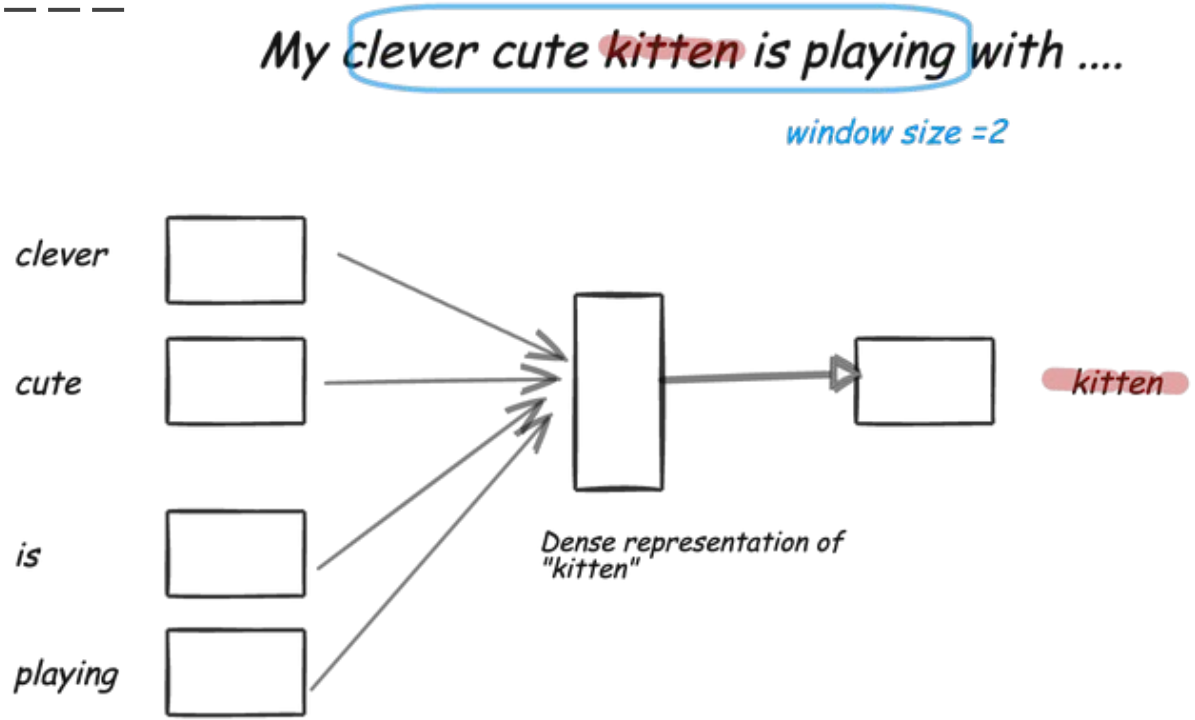
Review Representation

Review 1: 🍨 was great [0.5 , 0.1, … , 0.6]

Review 2: Chocolate ice cream was the best .. [0.5 , 0.1, … , 0.5]

Image from https://odsc.com/blog/getting-started-with-vector-based-search.

Word2Vec

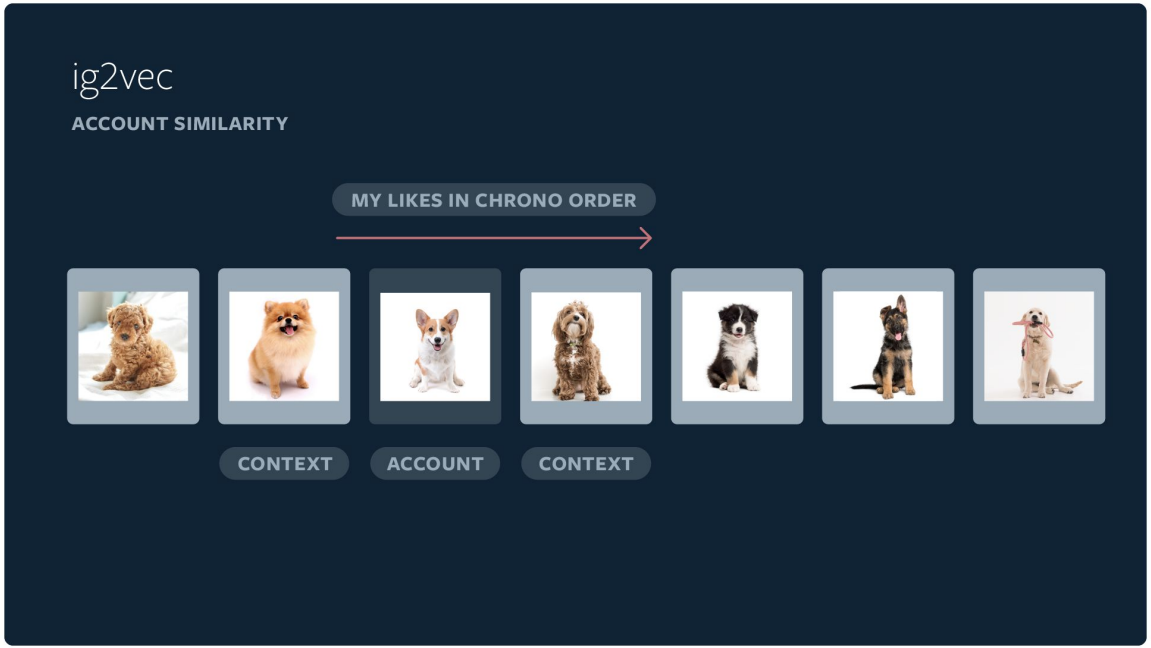


- Published in 2013
- Represent each word as a dense vector

- Uses a neural network model to capture linguistic concept of words

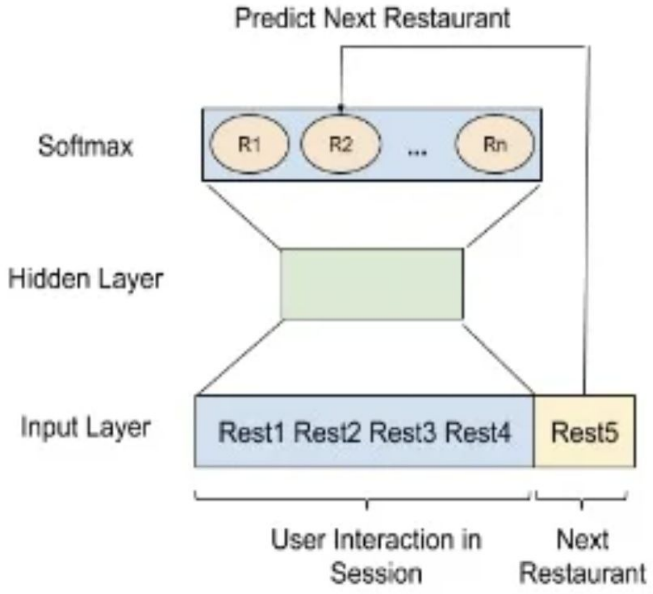
W2Vec Continuous Bag of Words

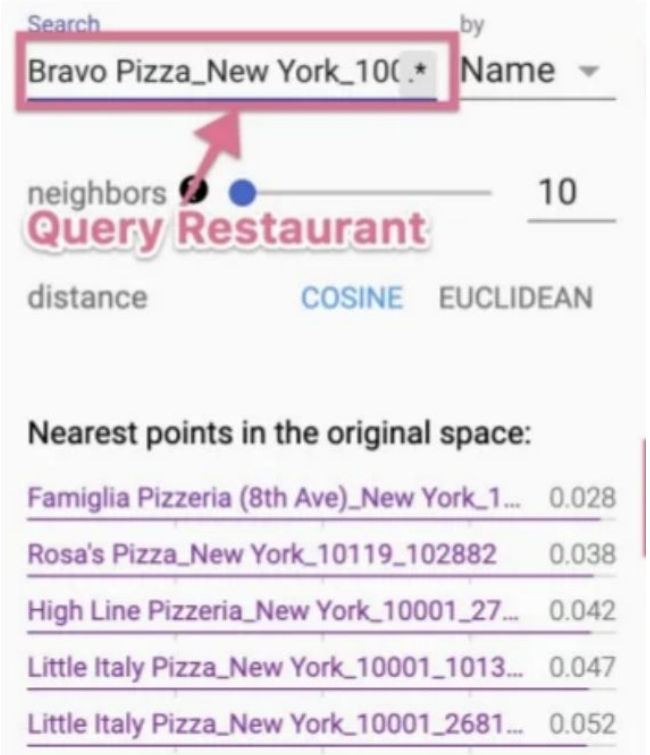
Instagram: ig2Vec



Medvedev, Ivan, Haotian Wu, and Taylor Gordon. “Powered by AI: Instagram’s Explore Recommender System.” Powered by AI: Instagram’s Explore recommender system, November 2019.
https://ai.facebook.com/blog/powered-by-ai-instagrams-explore-recommender-system/

Grubhub: Rest2Vec





Content from Pydata Talk “Alex Egg, Emily A Ray, Parin Choganwala: Discover your latent food graph with this 1 weird trick | PyData New York 2019”
https://www.youtube.com/watch?v=aRUaEt1q7BY&t=3s&ab_channel=PyData

Bidirectional Encoder Representations from Transformers (BERT)

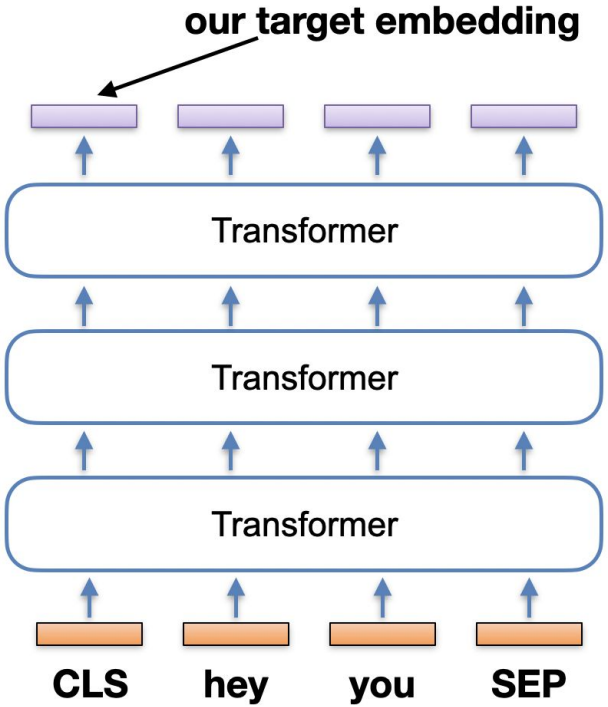
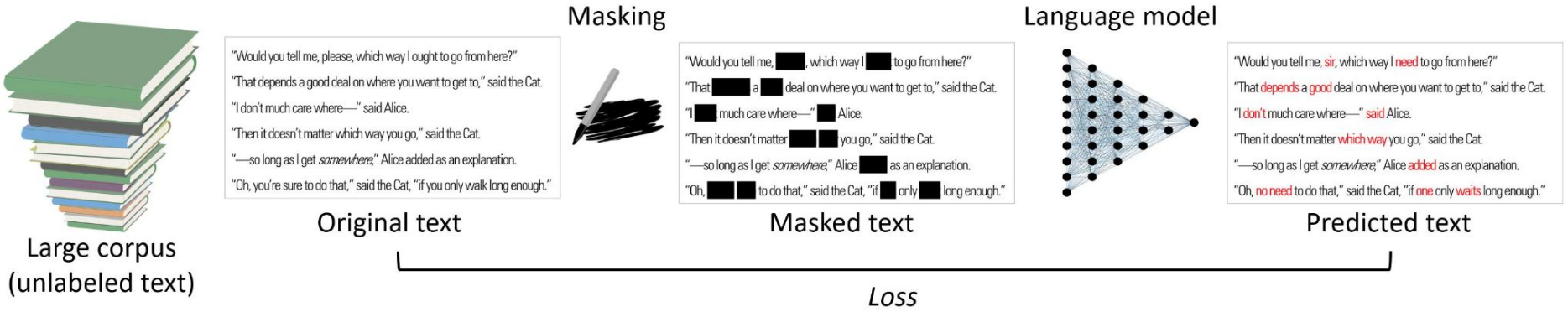


Image from https://www.usna.edu/Users/cs/nchamber/courses/nlp/f21/labs/lab6/index.html

BERT Highlights

- Train on large corpus in an unsupervised manner



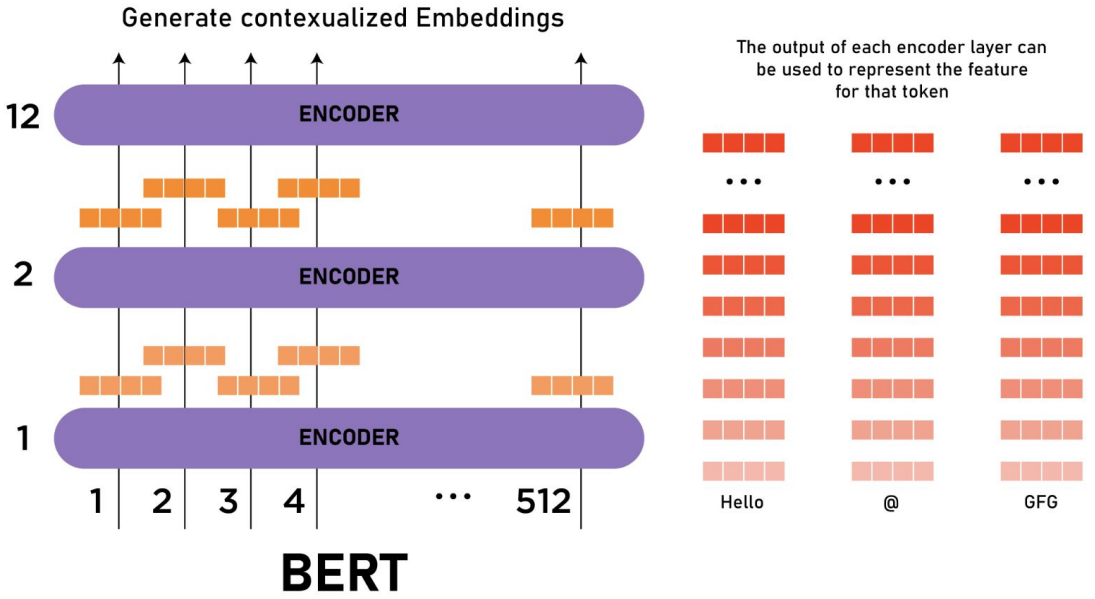
Ofer, Dan, et al. 'The Language of Proteins: NLP, Machine Learning & Protein Sequences'. *Computational and Structural Biotechnology Journal*, vol. 19, 03 2021, https://doi.org10.1016/j.csbj.2021.03.022.

BERT Highlights

- Sub Word Tokenization
  - running => run, ##ing
  - Vocab Size: Word2Vec (1M+) vs BERT (25K)

- Output contextualized embeddings for every word
  - Here is a list of classic english **novels**.
  - Steam Engine was a **novel** invention.

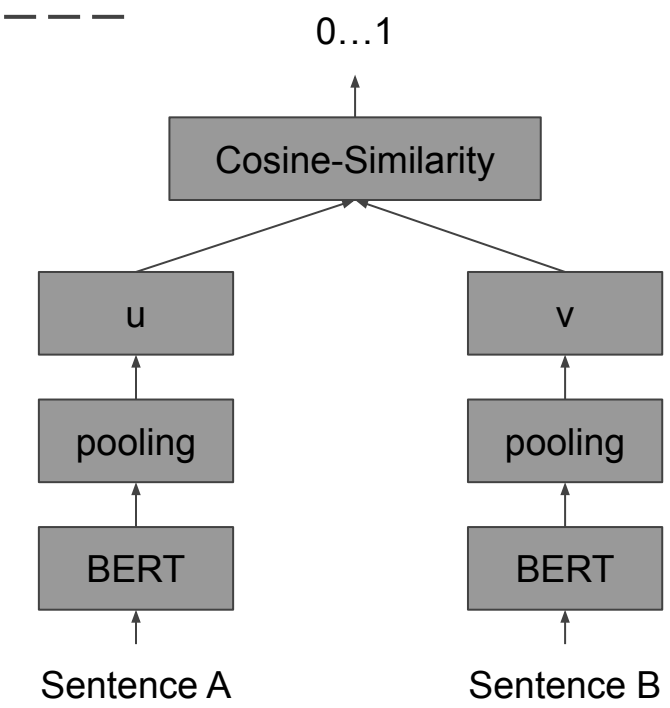
Bidirectional Encoder Representations from Transformers (BERT)



- BERT is composed of multiple encoders
- There is embedding for each token after each encoder
- Sentence embedding can be created from pooling
- Pooled embeddings are not best for similarity search

Image from https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/

Learning Similarity using Bi-Encoder



- Average Pooling Embeddings is not enough to capture similarity
- We can create a network that makes the network learn that two sentences are similar
- Each sentence is encoded separately
- Distance between two sentence embeddings can be measured

Lab

Jupyter Hub: https://hub.np.training
Repo:https://bit.ly/search-workshop-2022

Lab 2 Goals

- Understand subword tokenization
- Explore Text Embedding model

Query: Two dogs playing in the snow

Photo title: brown and black dogs running on snow
Distance: 0.73539

../data/raw/images/FAcSe7SjDUU.jpg

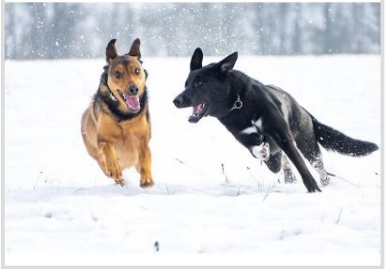


Photo title: white and black dog on snow
Distance: 0.68213

../data/raw/images/jzCmSH6en5c.jpg

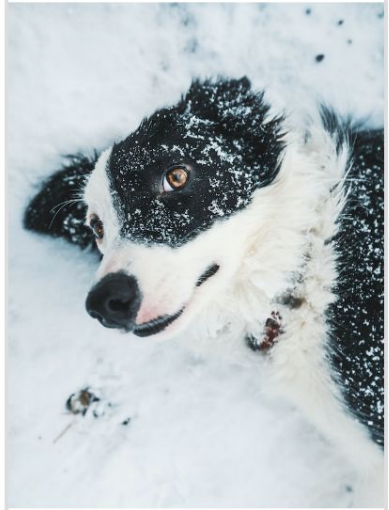


Photo title: Dog in snow
Distance: 0.66102

../data/raw/images/jtnH16_HfPE.jpg

Photo title: tan dog playing on snow
Distance: 0.64610

../data/raw/images/AVUX8QXnj4Y.jpg

Query: boy and girl on a beach

Photo title: children enjoying the beach
Distance: 0.70779

../data/raw/images/pV87YnElHow.jpg

Photo title: Couple on the beach
Distance: 0.69817

../data/raw/images/rl7mUDEUmVE.jpg

Photo title: people at beach
Distance: 0.68724

../data/raw/images/DtHchyQtyZ8.jpg

Photo title: two people on beach during daytime
Distance: 0.67832

../data/raw/images/ejhtQmbGZr8.jpg

Query: Seven Wonders of the world

Photo title: World globe
Distance: 0.49986

../data/raw/images/gEKMstKfZ6w.jpg

Photo title: Kingdom of Earth
Distance: 0.43012

../data/raw/images/o_-gToPk62c.jpg

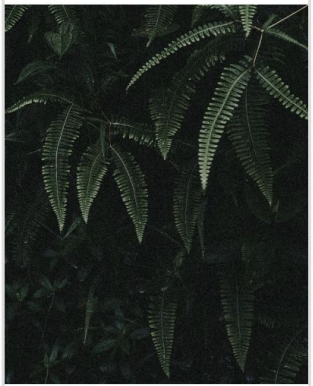


Photo title: Amazing ocean view #maldivesislands #maldives #maldivify #instago #maldives_ig #travelcaptures #nature #ig_global_life #oceanholic #moodhu #asiatravel #excusethehashtags #islandlife #beachesnresorts #bestplacestogo #travelinspiration #holidaypictures
Distance: 0.42829

../data/raw/images/BcqbKBVZ-Zk.jpg

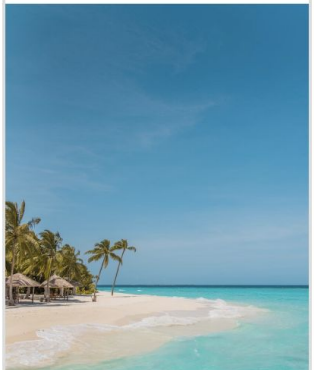
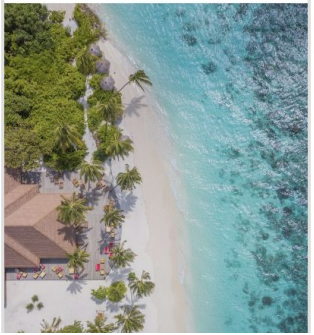


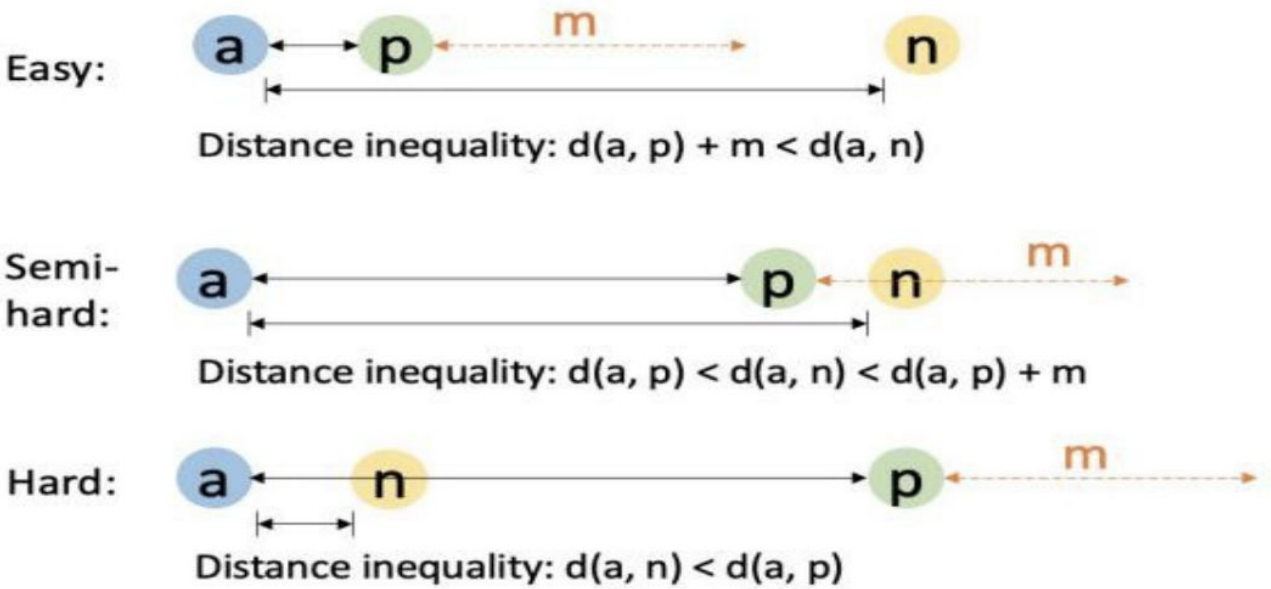
Photo title: Where the ocean meets the island. #maldives #maldivian #visitmaldives #beachlife #tropicalvibes #nature #dronestagram #discoverearth #earthoutdoors #earthfocus #seemaldives #ocean #visualsoflife #createcommune #travel #amazingplaces #beautifuldestinations #fromwhereidrone #bevisuallyinspired #agameoftones #theimaged
Distance: 0.40207

../data/raw/images/uw49OR2chh8.jpg



Different types of negatives

Anchor, Positive Ex, Negative Ex



Wang, Zeyu, Xiaoyu Liang, και Cheng Wang. 'Controllable Text-to-Image Generation with Enhanced Text Encoder and Edge-Preserving Embedding'. Journal of Physics: Conference Series 1856 (04 2021): 012003. https://doi.org/10.1088/1742-6596/1856/1/012003.

Multiple Modality: Vision Transformers (ViT)

- Vision Transformer models like CLIP model uses two encoders (text and image).
- These two models are trained in parallel and optimized via a contrastive loss function
- End result is where you can search by text or image

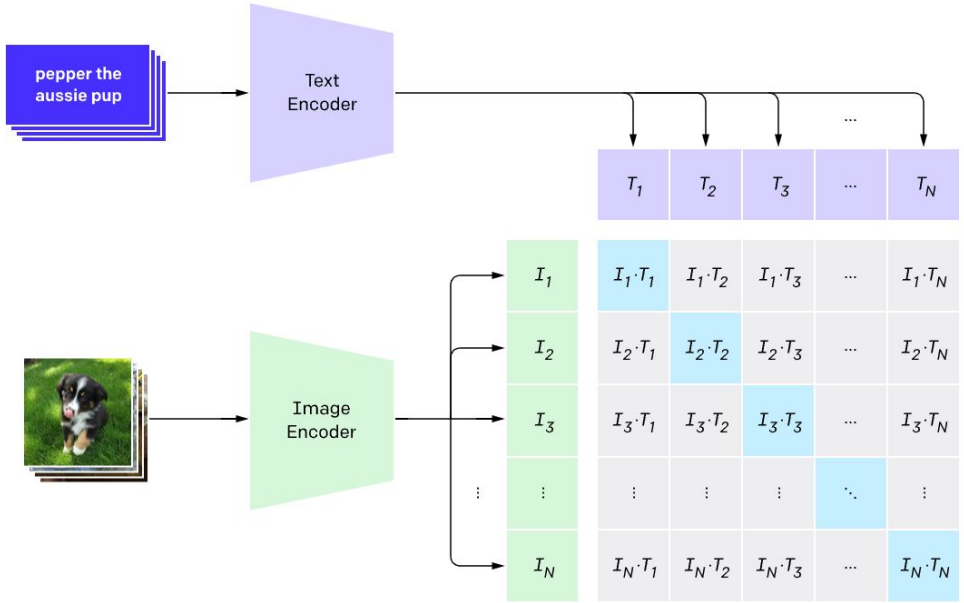


Image from OpenAi CLip blog
https://openai.com/blog/clip/

Encoder can include multiple features

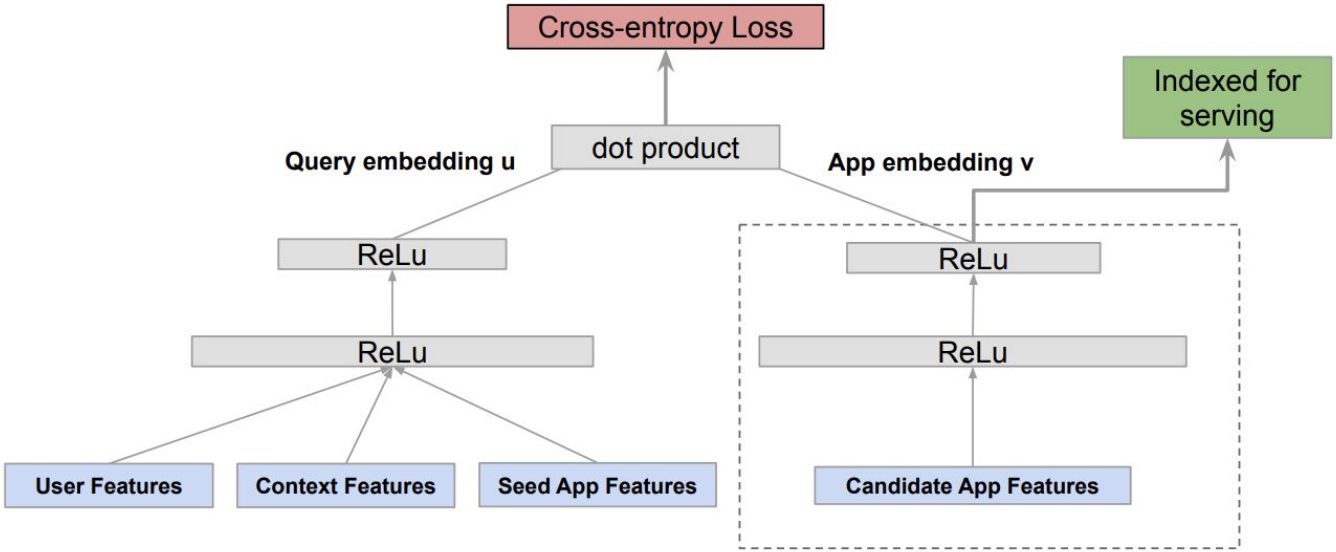


Figure 5: Two-tower model architecture for Google Play app recommendation.

Yang, Ji, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, και Ed H. Chi. 'Mixed Negative Sampling for Learning Two-Tower Neural Networks in Recommendations'. Στο Companion Proceedings of the Web Conference 2020, 441–47. WWW '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3366424.3386195.

Lab

Jupyter Hub: https://hub.np.training
Repo:https://bit.ly/search-workshop-2022

Lab 3 Goals

- Building a simple in-memory retrieval system using a multi-modal model like CLIP

Query: Two dogs playing in the snow

Photo title: brown and black dogs running on snow
Distance: 0.31937

../data/raw/images/FAcSe7SjDUU.jpg

Photo title: Friends will be friends
Distance: 0.30999

../data/raw/images/lyStEjlKNSw.jpg

Photo title: Caramel-brown-white-coloured dog in the snow.
Distance: 0.30256

../data/raw/images/JAjrbewMbFQ.jpg

Photo title: arctic fox fight...
Distance: 0.30226

../data/raw/images/Hb6nGDgWztE.jpg

Query: boy and girl on a beach

Photo title: White sands
Distance: 0.30794

../data/raw/images/3xUnaShh5SQ.jpg

Photo title: aerial view of seashore
Distance: 0.30317

../data/raw/images/dNyPi0PKWo0.jpg

Photo title: children enjoying the beach
Distance: 0.30200

../data/raw/images/pV87YnElHow.jpg

Photo title: Dusk on beach shore
Distance: 0.29762

../data/raw/images/vCloc-Wha1Q.jpg

Query: Seven Wonders of the world

Photo title: Machu Picchu - Peru
Distance: 0.28726

../data/raw/images/SVE2I-_vPcI.jpg

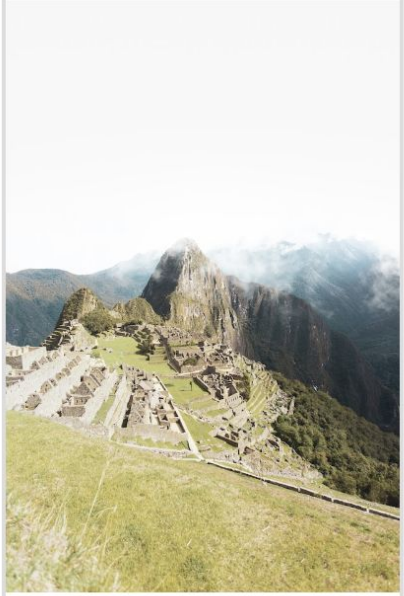


Photo title: Blue sky over Machu Picchu
Distance: 0.28449

../data/raw/images/8xpkIKrfsG4.jpg

Photo title: aerial photo of Machu Picchu, Peru
Distance: 0.28406

../data/raw/images/PO7CGnoDFUI.jpg

Photo title: Machu Picchu - Perú
Distance: 0.27667

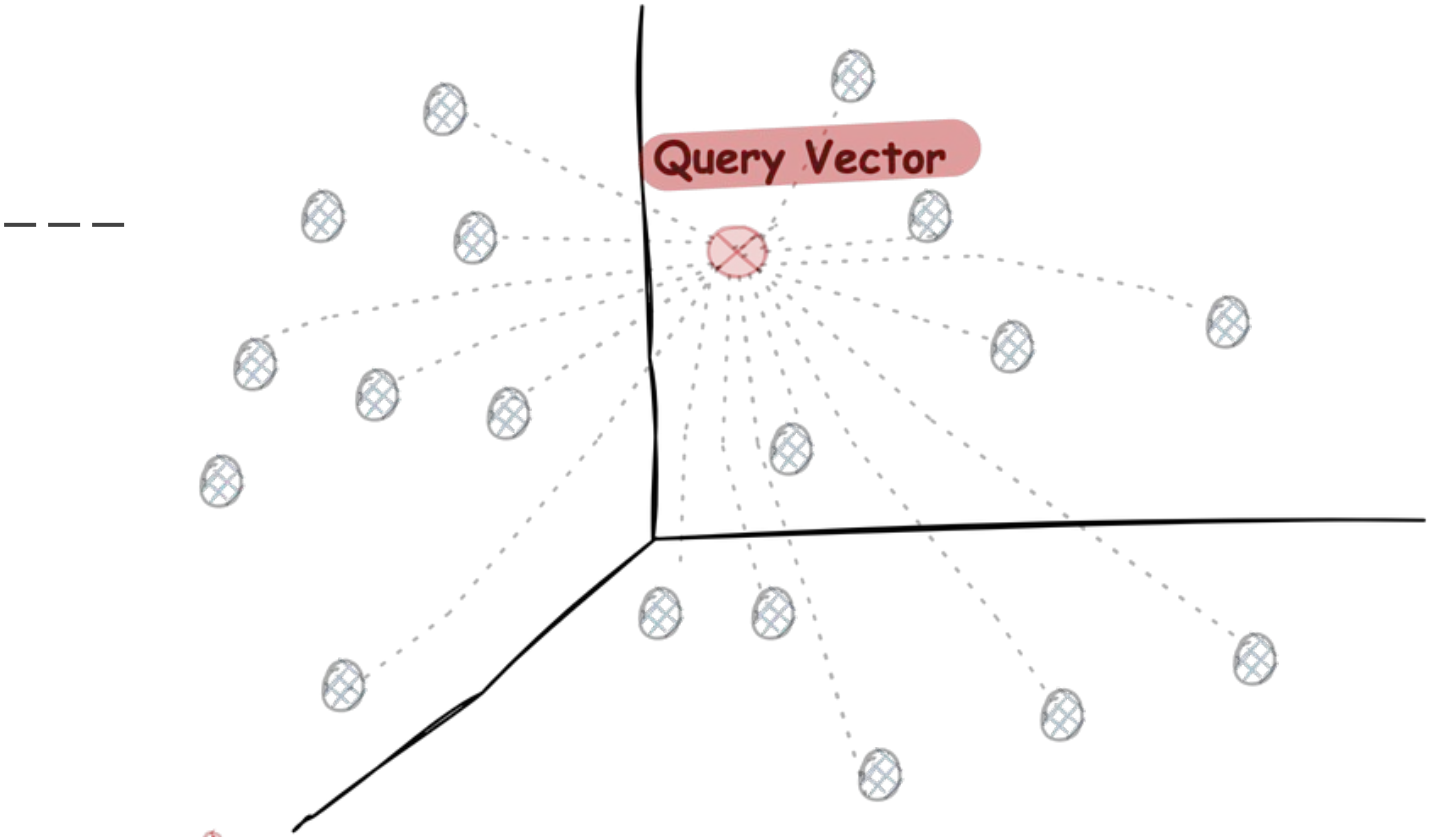
../data/raw/images/h3IxnaFjr0M.jpg

Our Agenda

---

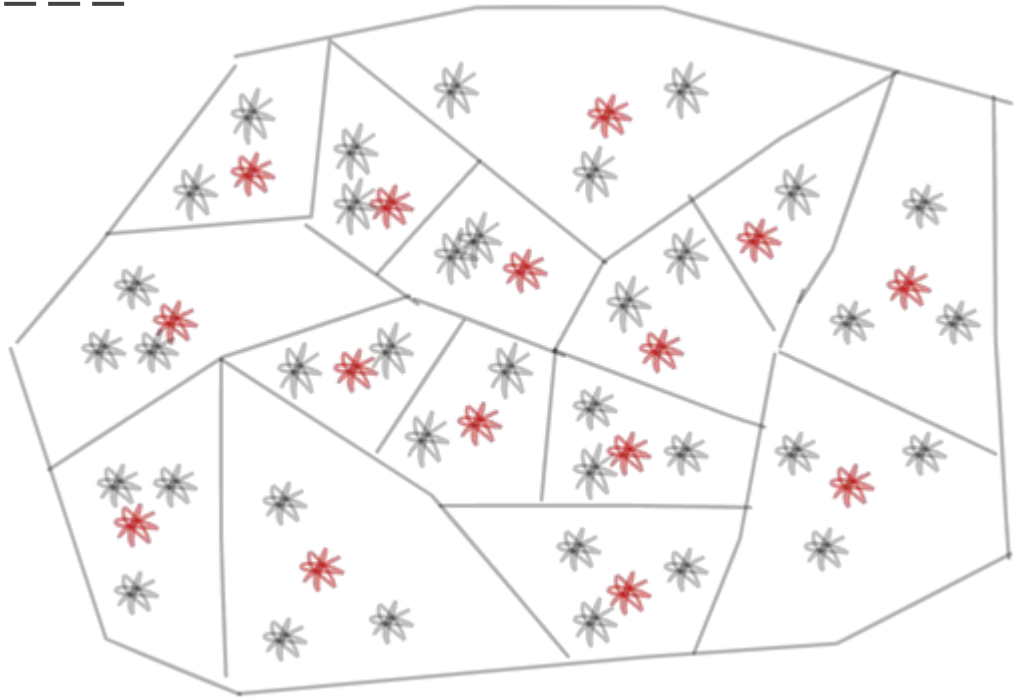
1. Token Based Retrieval
2. Embedding Based Retrieval
3. **Approximate Nearest Neighbor (ANN)**
4. Production Considerations

Approximate Nearest Neighbors



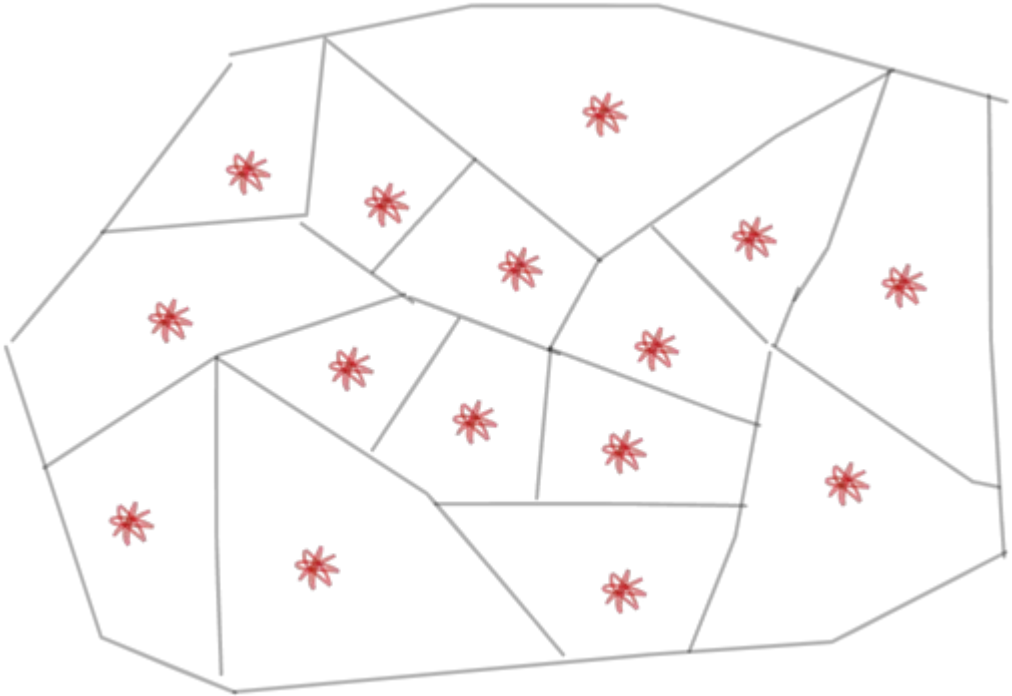
In a flat index / no index, our query vector is compared against all the vectors in the database.

Inverted File Index: Building



- Find centroid and create Voronoi Cells
- Number of centroids is determined by **nlist** parameter

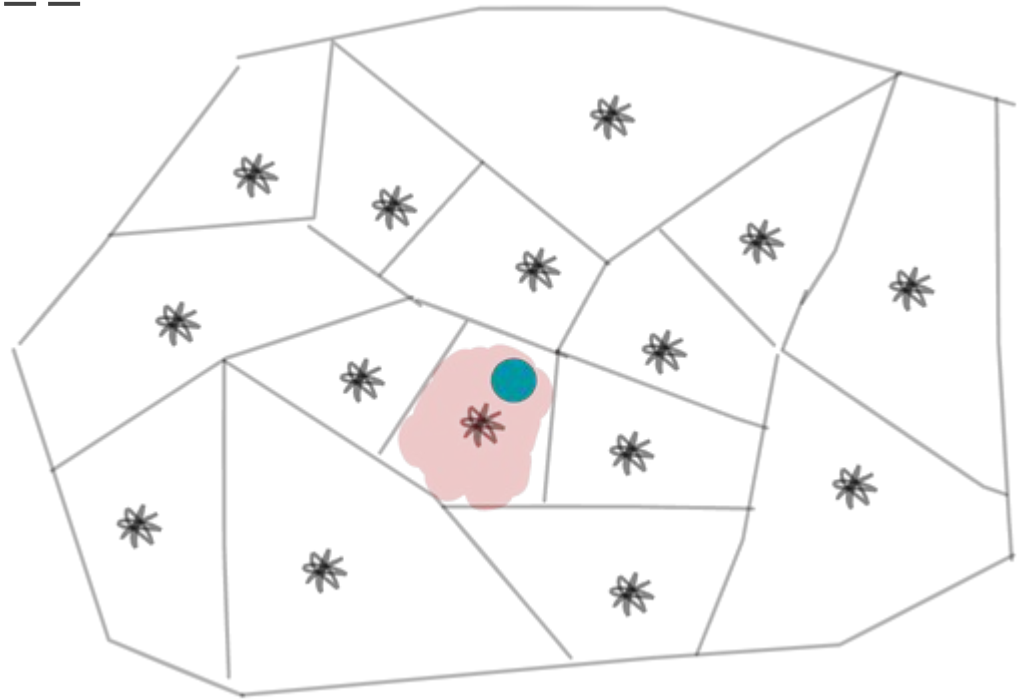
Inverted File Index: Building



- Built Voronoi Index with 15 centroids

- Memory usage is not reduced
- But retrieval is faster

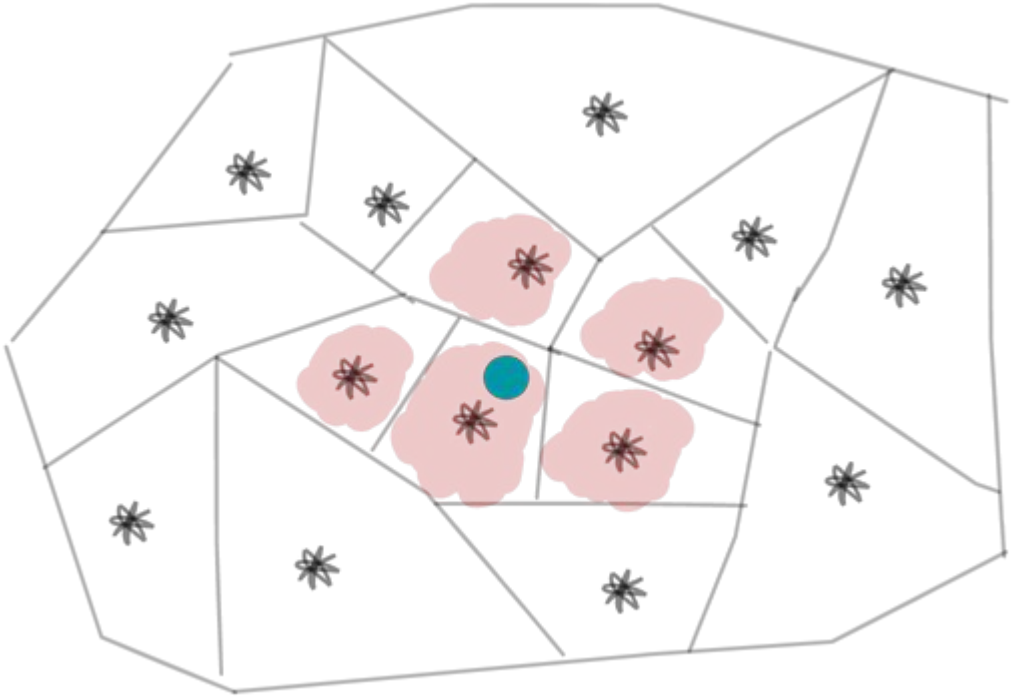
Inverted File Index: Searching



nprobe = 1

- Find the distance between query vector and all the centroids
- Query all the elements inside the n closest cluster determined by **nprobe** parameter

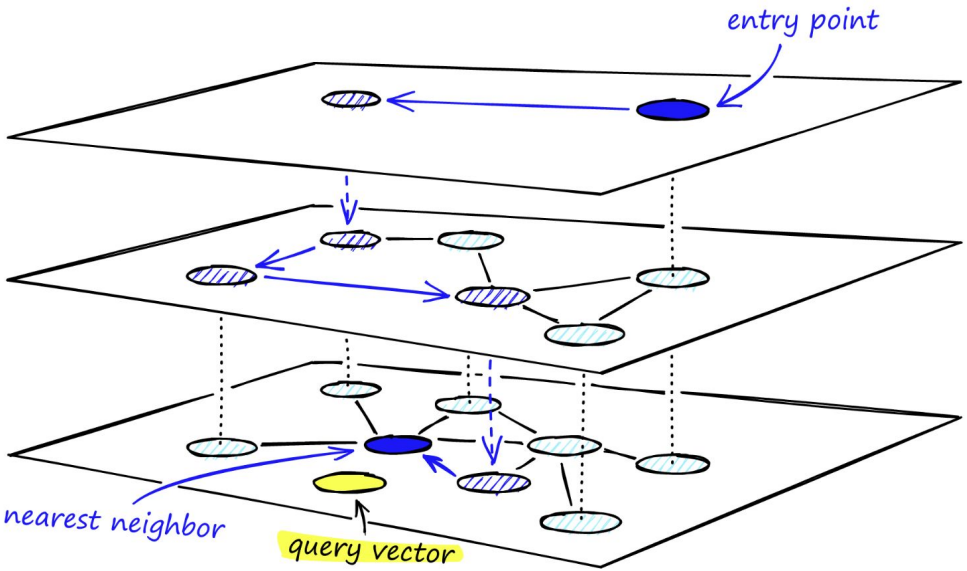
Inverted File Index: Searching



nprobe = 5

- Increasing the value of **nprobe**, improves recall but increases latency
- If nprobe=nlist, similar to flat index

Hierarchical Navigable Small Worlds (HNSW)



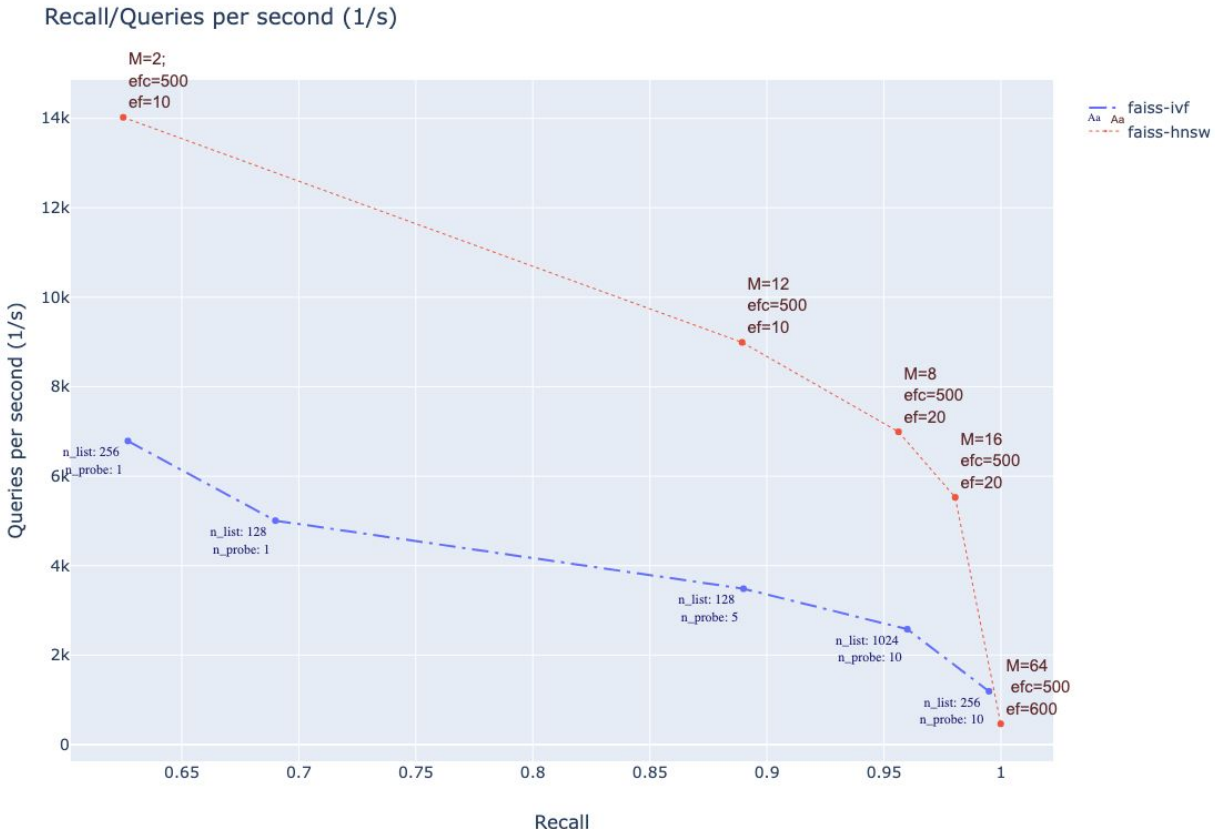
The search process through the multi-layer structure of an HNSW graph.

- Hierarchical Graph
- Every layer stores neighbors
- Earlier layers are sparse

- Common implementation in most production ANN databases

Image from https://www.pinecone.io/learn/hnsw/

ANN Benchmarks



Source:
[ANN Benchmarks]

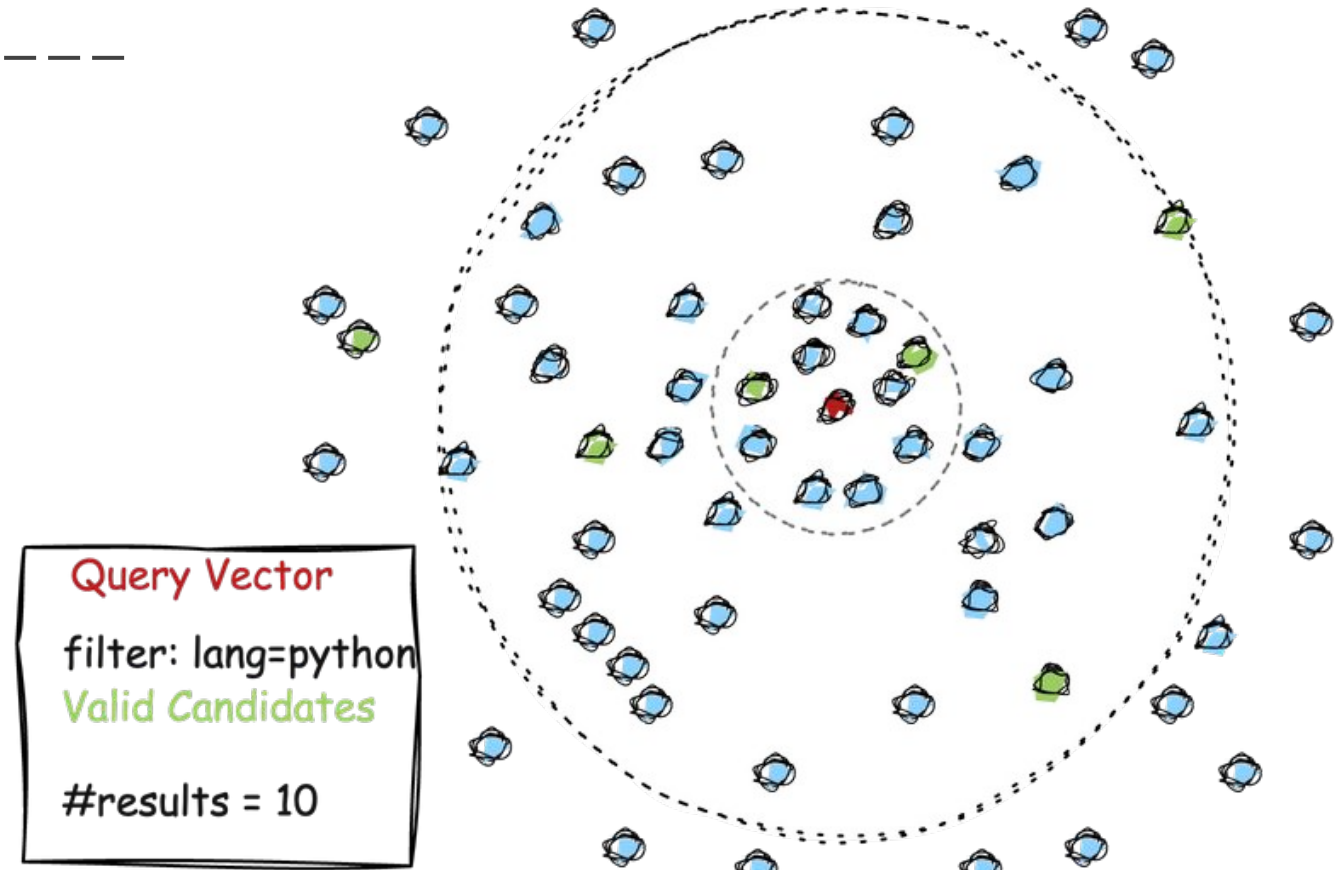
[Subset]

Lot of ANN Option

Evaluation Considerations

- Managed vs Self-hosted
- Performance
- Update Embeddings / Partial Updates
- Metadata Filtering
- Filtering / Hybrid Retrieval
- Plugins

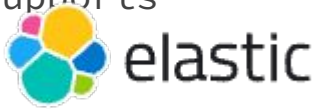
Filtering



Closest 10 candidates don’t meet our filter

Hybrid / Full Retrieval

Supports



- In ES, **disjunction** of knn and bm25 match.
- The score of each hit is the sum of the knn and query scores. A boost can be specified

```
POST image-index/_search
{
  "query": {
    "match": {
      "title": {
        "query": "mountain lake",
        "boost": 0.9
      }
    }
  },
  "knn": {
    "field": "image-vector",
    "query_vector": [54, 10, -2],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1
  },
  "size": 10
}
```

ElasticSearch example of Hybrid Retrieval

Ex from Elastic Doc link

Lab

Jupyter Hub: https://hub.np.training
Repo:https://bit.ly/search-workshop-2022

Lab 4 Goals

- Build a flat ANN index
- Optimize retrieval by building an IVF Index

Our Agenda

---

1. Token Based Retrieval
2. Embedding Based Retrieval
3. Approximate Nearest Neighbor (ANN)
4. **Production Considerations**

Production Examples

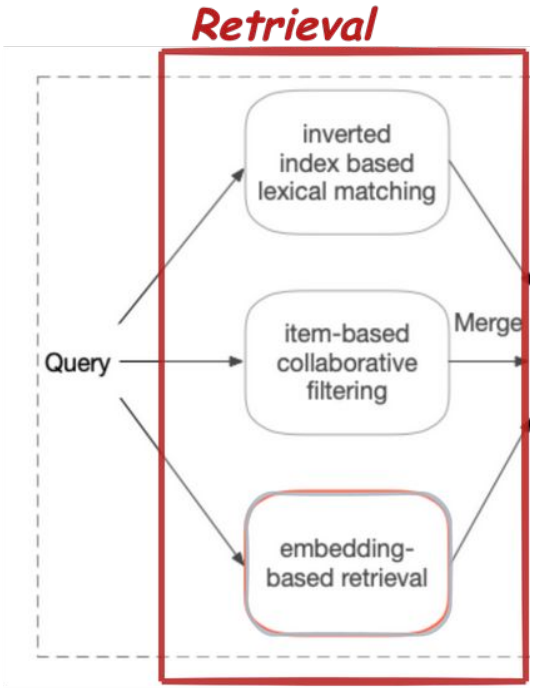
Embeddings for out of domain corpus

Dataset	BM25	Dense
Baseline (Microsoft Search)	0.228	0.408
Quora	0.789	0.835
TREC-COVID (Medical)	0.656	0.481
Signal-1M (Tweets)	0.330	0.289

- Do models trained on MS MARCO work for different datasets ?
- Dense Retrieval performs well on similar domain
- COVID/Tweets are out of domain

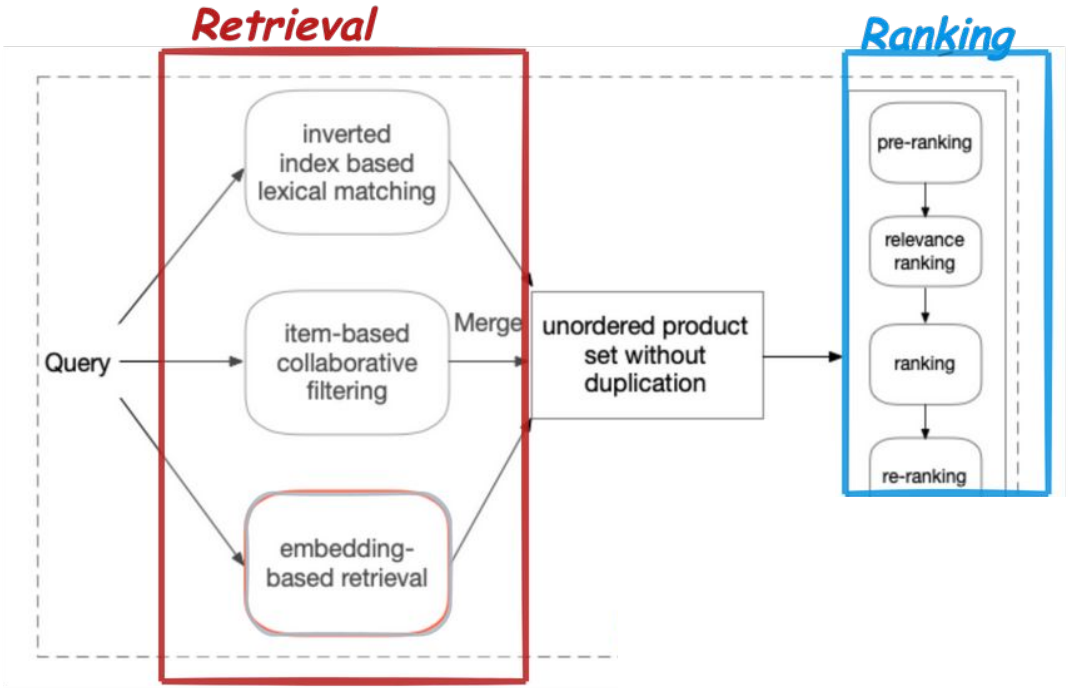
Thakur, Nandan, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, και Iryna Gurevych. 'BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models'. arXiv, 2021. https://doi.org/10.48550/ARXIV.2104.08663.

Overview of Taobao Search (Chinese Online Shopping)



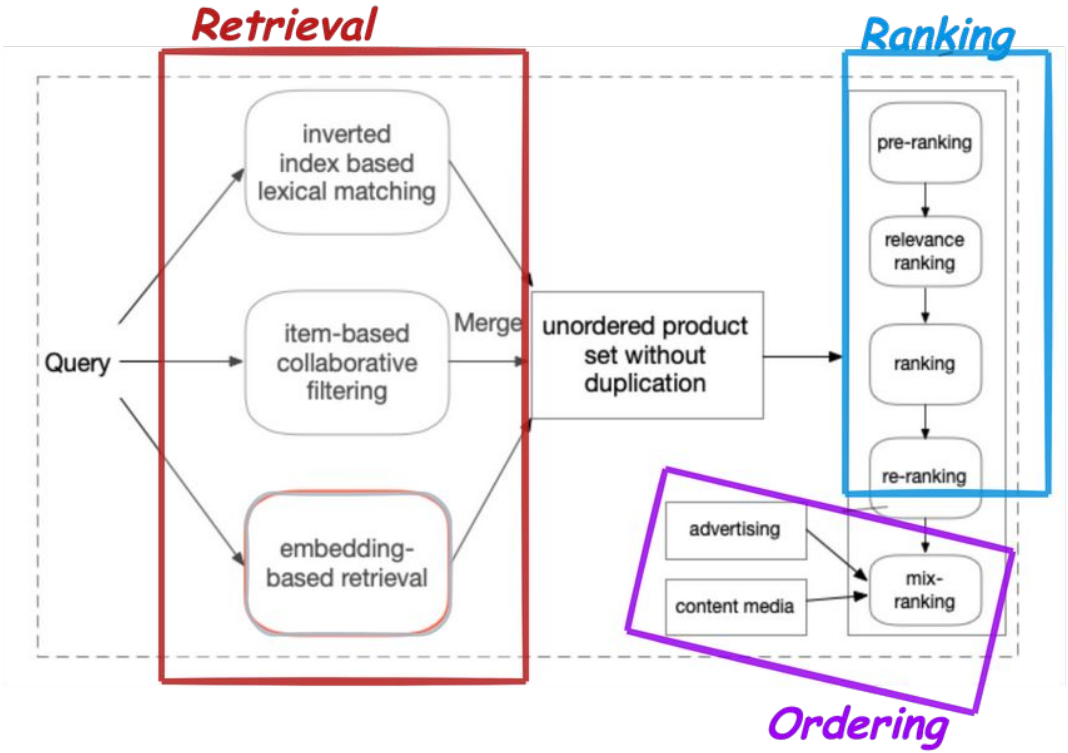
Li, Sen, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, και Qianli Ma. 'Embedding-based Product Retrieval in Taobao Search'. arXiv, 2021. https://doi.org/10.48550/ARXIV.2106.09297.

Overview of Taobao Search (Chinese Online Shopping)



Li, Sen, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, και Qianli Ma. 'Embedding-based Product Retrieval in Taobao Search'. arXiv, 2021. https://doi.org/10.48550/ARXIV.2106.09297.

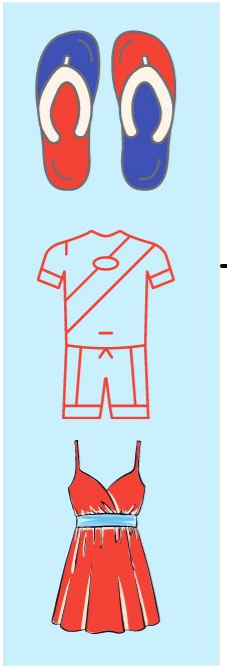
Overview of Taobao Search (Chinese Online Shopping)



Li, Sen, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, και Qianli Ma. 'Embedding-based Product Retrieval in Taobao Search'. arXiv, 2021. https://doi.org/10.48550/ARXIV.2106.09297.

Learning to Rank

Original Order



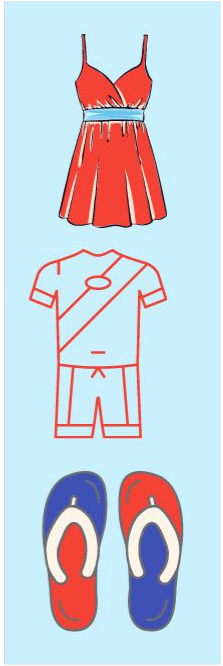
Learning to Rank
dmlc
XGBoost

Query Intent Features
- product type
- color
- brand

Item Features
- customer score
- price
product type
- color
- brand

Context Features
-previous queries
- season

New Order



Conclusion

Takeaways

- Use ElasticSearch to implement a Token Based Retrieval
- Use SentenceTransformer Library to use pre-trained models for embedding retrieval
- Understand how to scale Embedding Retrieval using FAISS

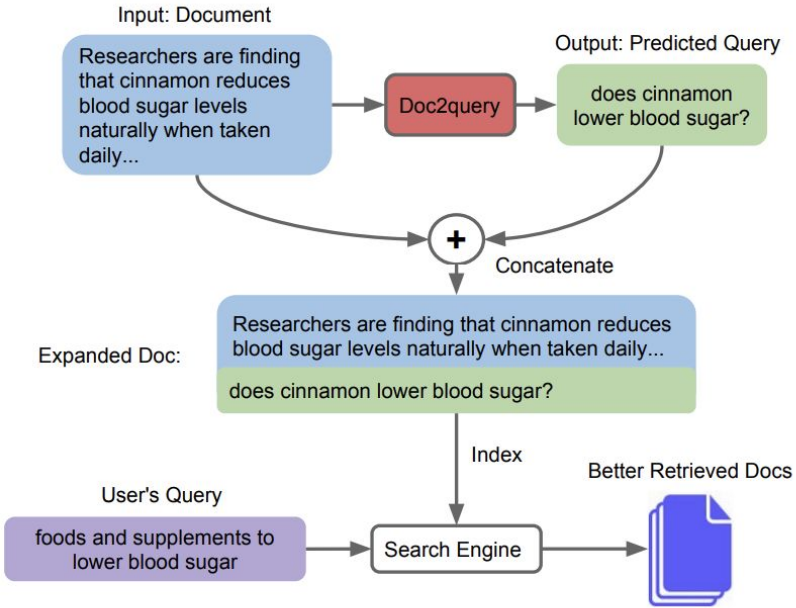
Resources

- Natural Language Processing (NLP) for Semantic Search (Pinecone)
- CIKM 2021 Tutorial: IR From Bag-of-words to BERT

- Haystack US 2021 - Semantic Product Search - Vector Search for E-Commerce - Simon Hughes
-
- Faiss Missing Manual (Pinecone)

Other

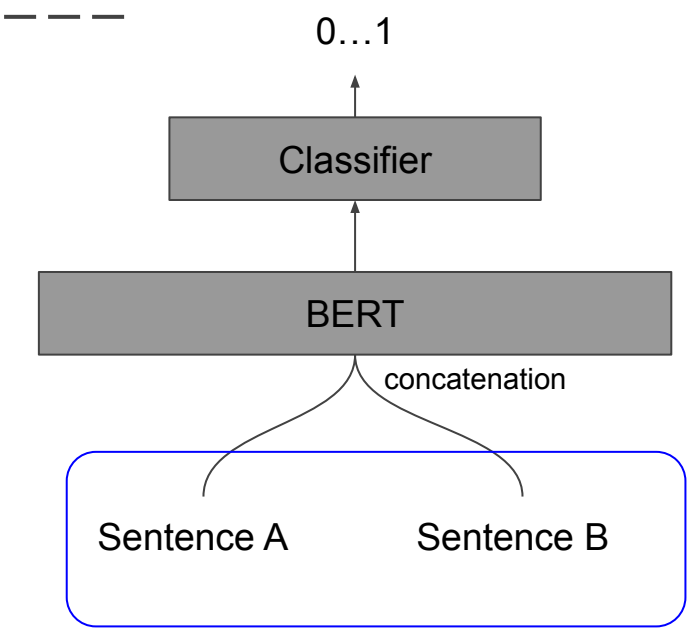
Enhancing Sparse Index: Doc2Query / Doc2T5Query



- Use a causal language model to generate additional text to add to documents when indexing.
- At retrieval time, use BM25

Nogueira, Rodrigo, Wei Yang, Jimmy Lin, και Kyunghyun Cho. 'Document Expansion by Query Prediction'. arXiv, 2019. https://doi.org/10.48550/ARXIV.1904.08375.

Sentence Pair Encoder - Cross-encoder



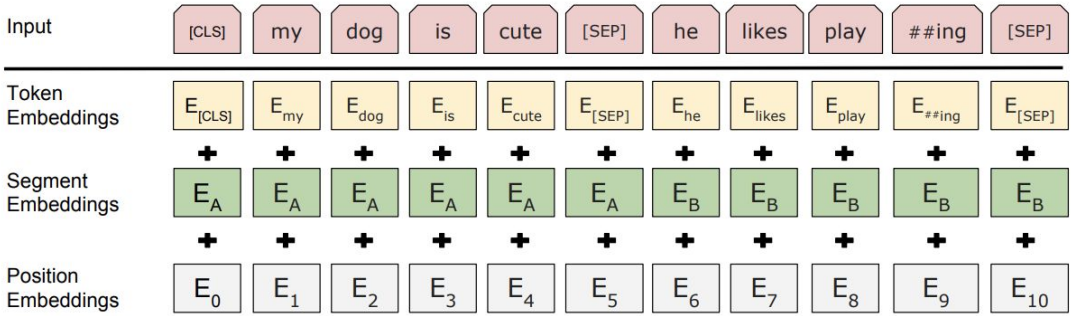
Pros

- Accurate classification

Cons

- Need new encoding for each pair
- Computationally inefficient for information retrieval.

Bidirectional Encoder Representations from Transformers (BERT)



Input to the model, contains token , segment and position embedding

Image from BERT paper https://arxiv.org/pdf/1810.04805.pdf