Building LLM Powered Solutions

Module 4: Let’s build a Search Engine

Hamza Farooq

Recap: What we have covered so far?

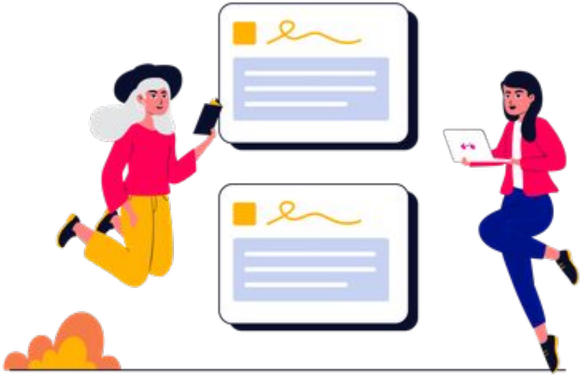
1. What is Self-Attention mechanism?
2. What is an Encoder?
3. What is Decoder?
4. What are Parameters?
5. Basic semantic search
6. Cosine similarity
7. Faiss
8. ANN

Learning Outcomes

We will be covering topics on:

- Why is search so important?
- Understand the building blocks of search, again
- Sentence Transformers
- BM25, Bi-encoder, Cross-Encoders
- Cosine Similarity, again!
- Fitting all of this into a ML System
- Coding, lots of it, with API
- Evaluation of Models
- Query intent model

A good search experience is key to a successful user journey



An estimated [50% of queries contain four or more words]() - search is no more just keyword based

62% of consumers will switch to a different brand or decide not to purchase from your brand at all after a bad customer experience — and poor site search is a bad customer experience.

Project Athena: Adding Semantic search to Hotel search

We want to build a hotel search using:

- Date check-in/check-out
- City
- Long text to get more granular choices such as:

```
Looking for a hotel in New York near Times Square with
free breakfast and cheaper than $100 for 2nd June which
is really kids friendly and has a swimming pool and I
want to stay there for 8 days...
```

Sentence transformers

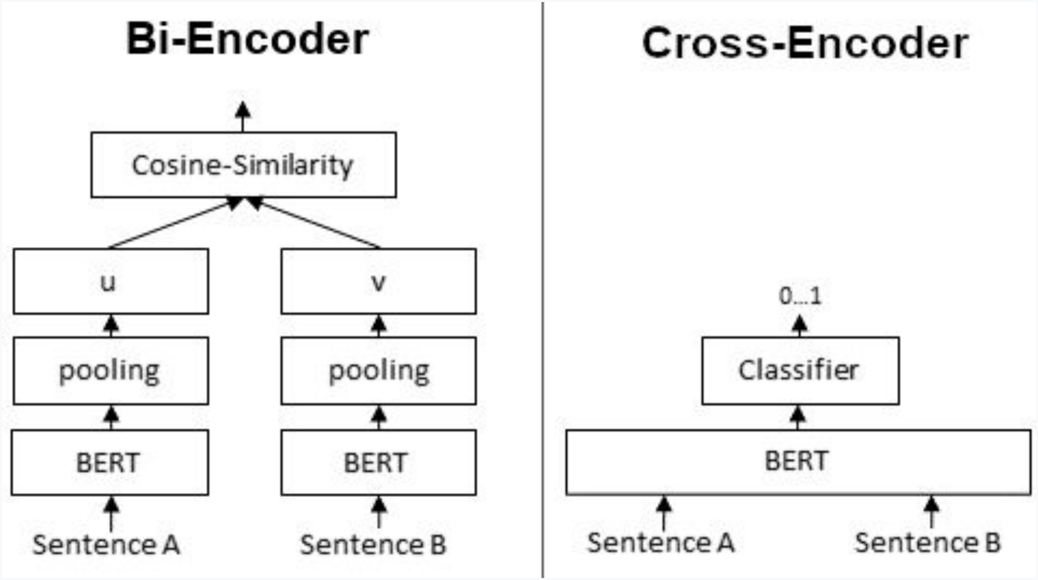
- Sentence transformers are models that encode the meaning of sentences using deep learning techniques
- They transform sentences into fixed-length numerical representations called embeddings
- Sentence embeddings capture semantic information and enable comparison and similarity calculations between sentences.
- A popular example is mpnet

Neural-IR Models

Numerous architectures are available for ranking: representation-focused, interaction-focused, all-to-all interaction(cross encoder), and late interaction.

$$\sum_{i}^{n} IDF(q_i) \frac{f(q_i, D) * (k1 + 1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})}$$

BM25



Source: https://www.sbert.net/examples/applications/cross-encoder/README.html

BM25

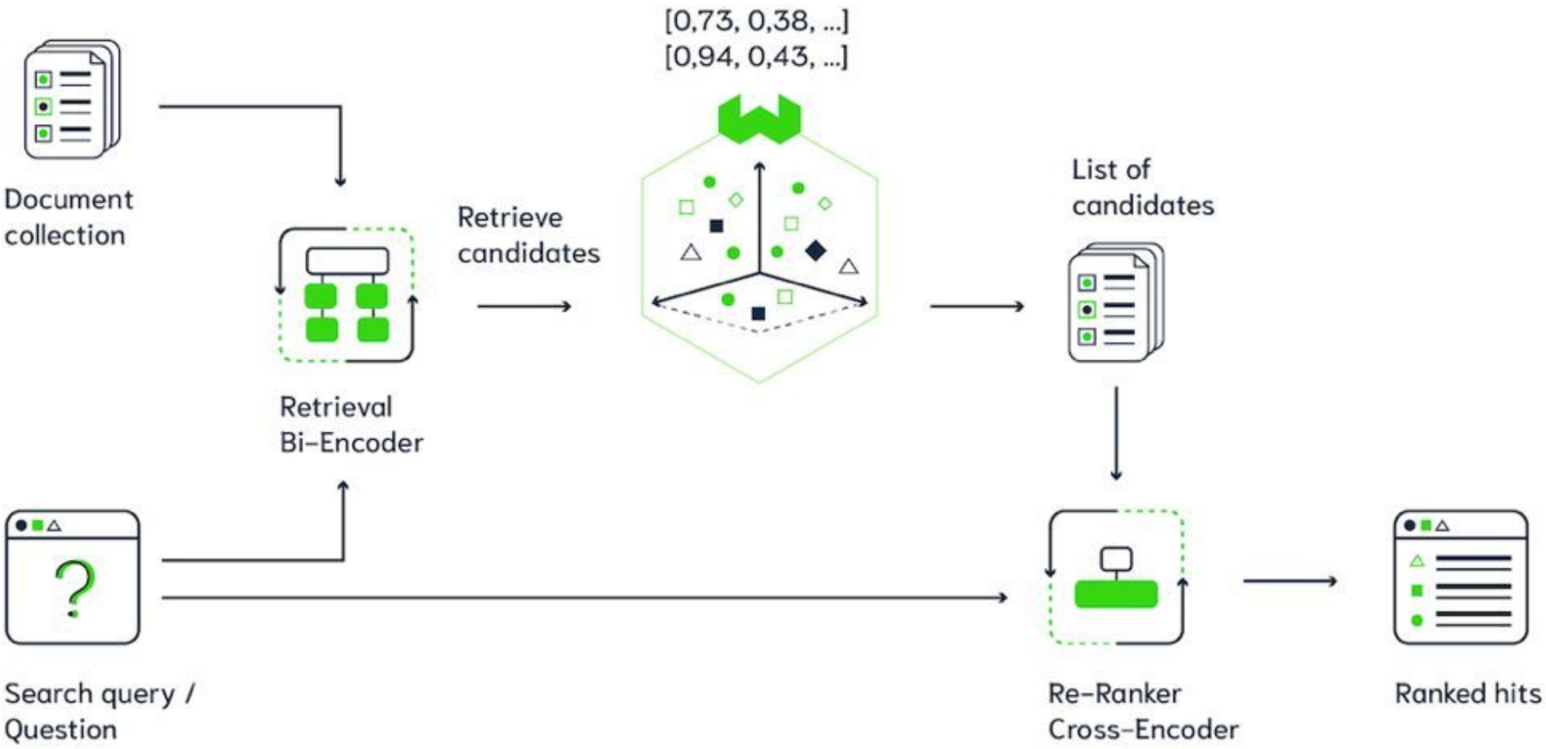
- BM25 stands for "Best Match 25" and is an information retrieval algorithm used to rank documents based on their relevance to a given query.
- It calculates a relevance score by considering the term frequency and document length in a collection of documents.

Bi-Encoder

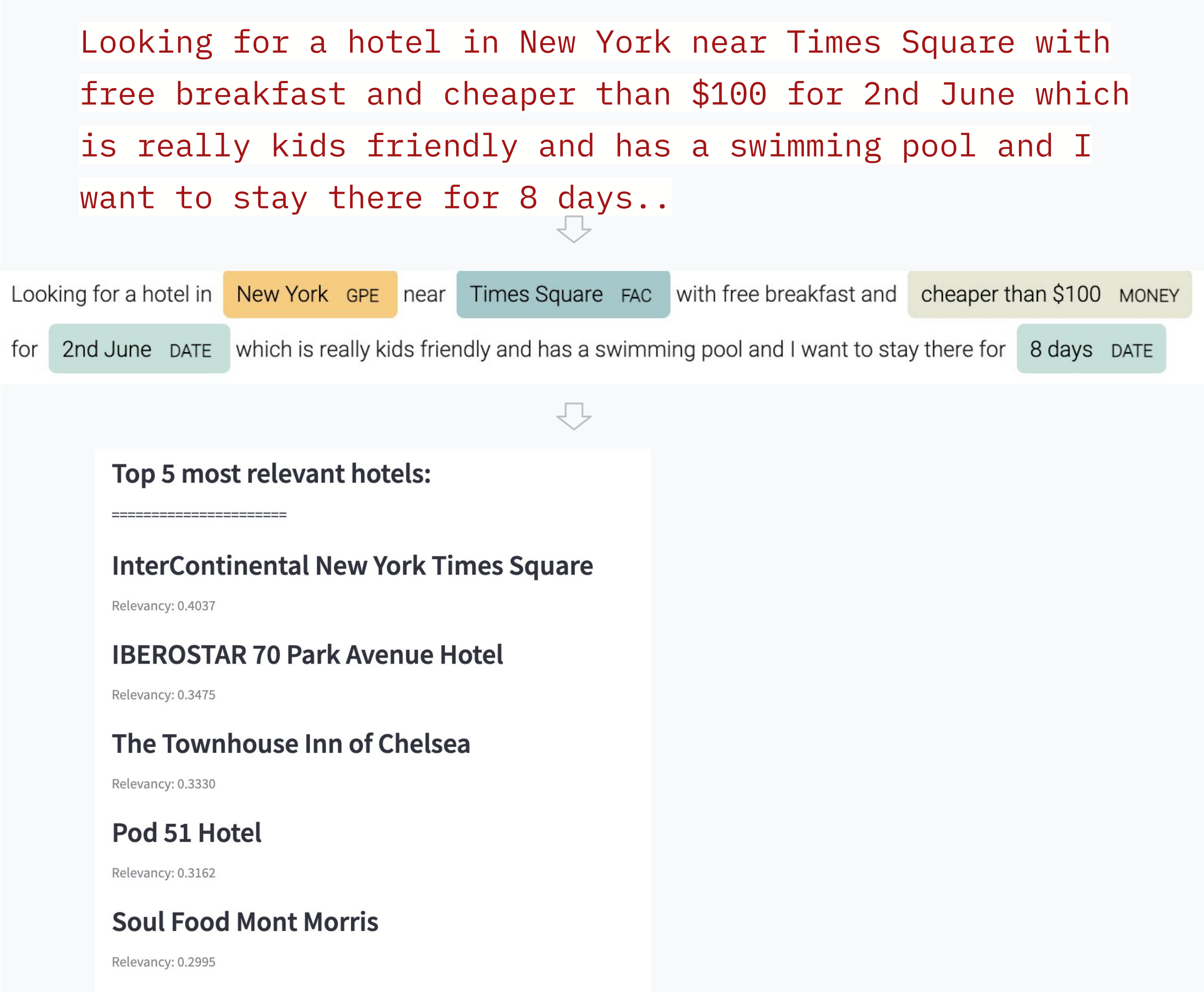
- A bi-encoder is a type of neural network architecture used in natural language processing (NLP) tasks.
- It consists of two encoders: one for encoding the input query and another for encoding the input document.
- Each encoder independently encodes the input into a fixed-length representation, often called an embedding.

Cross-Encoder

- A cross-encoder is another type of neural network architecture used in NLP tasks.
- It takes both the input query and document as a single input and encodes them into a fixed-length representation.
- Unlike the bi-encoder, the cross-encoder considers the interaction between the query and document when generating the representation.


[0,73, 0,38, ...]
[0,94, 0,43, ...]
Document collection
Retrieval Bi-Encoder
Retrieve candidates
List of candidates
Search query / Question
Re-Ranker Cross-Encoder
Ranked hits

The result…



```
Looking for a hotel in New York near Times Square with
free breakfast and cheaper than $100 for 2nd June which
is really kids friendly and has a swimming pool and I
want to stay there for 8 days..
```

Looking for a hotel in New York GPE near Times Square FAC with free breakfast and cheaper than $100 MONEY for 2nd June DATE which is really kids friendly and has a swimming pool and I want to stay there for 8 days DATE

Top 5 most relevant hotels:

=====================

InterContinental New York Times Square

Relevancy: 0.4037

IBEROSTAR 70 Park Avenue Hotel

Relevancy: 0.3475

The Townhouse Inn of Chelsea

Relevancy: 0.3330

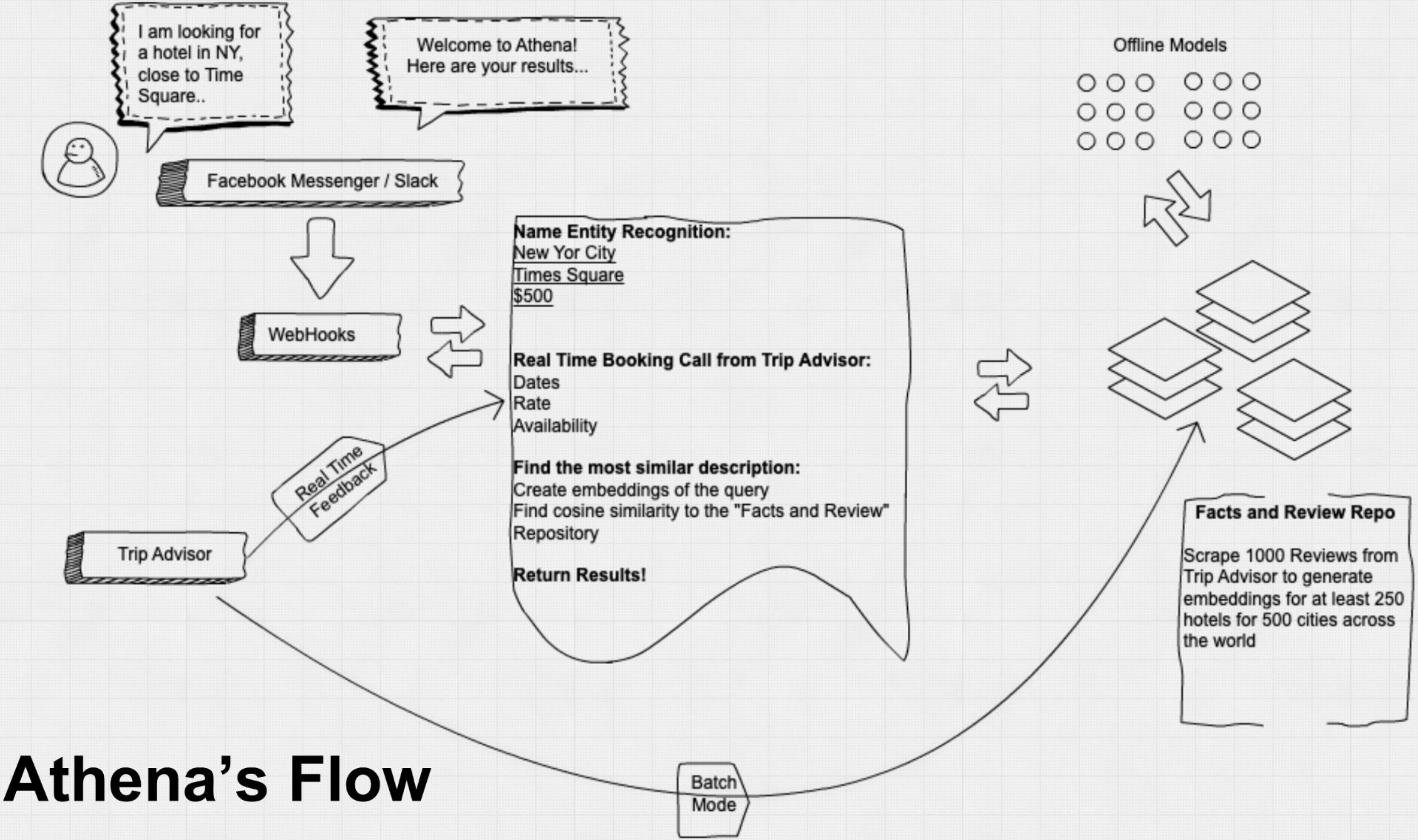
Pod 51 Hotel

Relevancy: 0.3162

Soul Food Mont Morris

Relevancy: 0.2995

Athena’s Flow



I am looking for a hotel in NY, close to Time Square..

Welcome to Athena! Here are your results...

Facebook Messenger / Slack

WebHooks

Name Entity Recognition:
New Yor City
Times Square
$500

Real Time Booking Call from Trip Advisor:
Dates
Rate
Availability

Find the most similar description:
Create embeddings of the query
Find cosine similarity to the "Facts and Review" Repository

Return Results!

Trip Advisor

Real Time Feedback

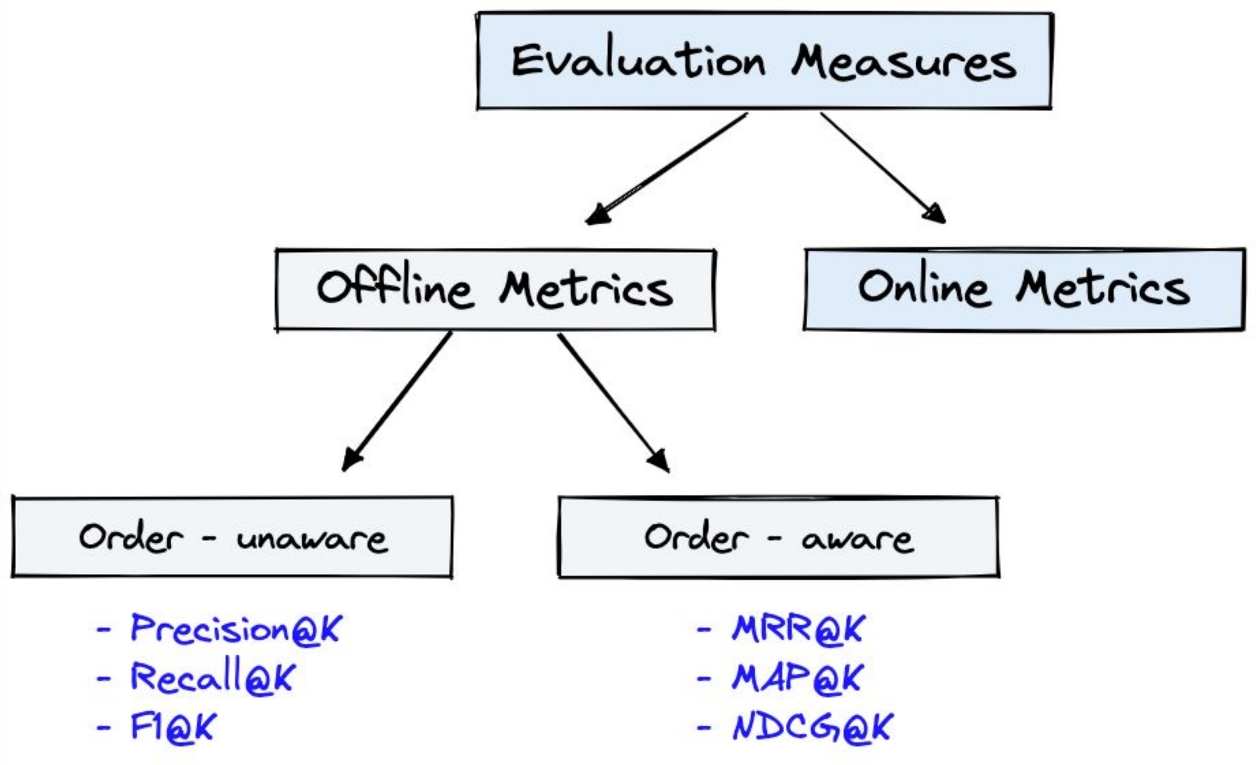
Batch Mode

Offline Models

Facts and Review Repo

Scrape 1000 Reviews from Trip Advisor to generate embeddings for at least 250 hotels for 500 cities across the world

Evaluation Criteria

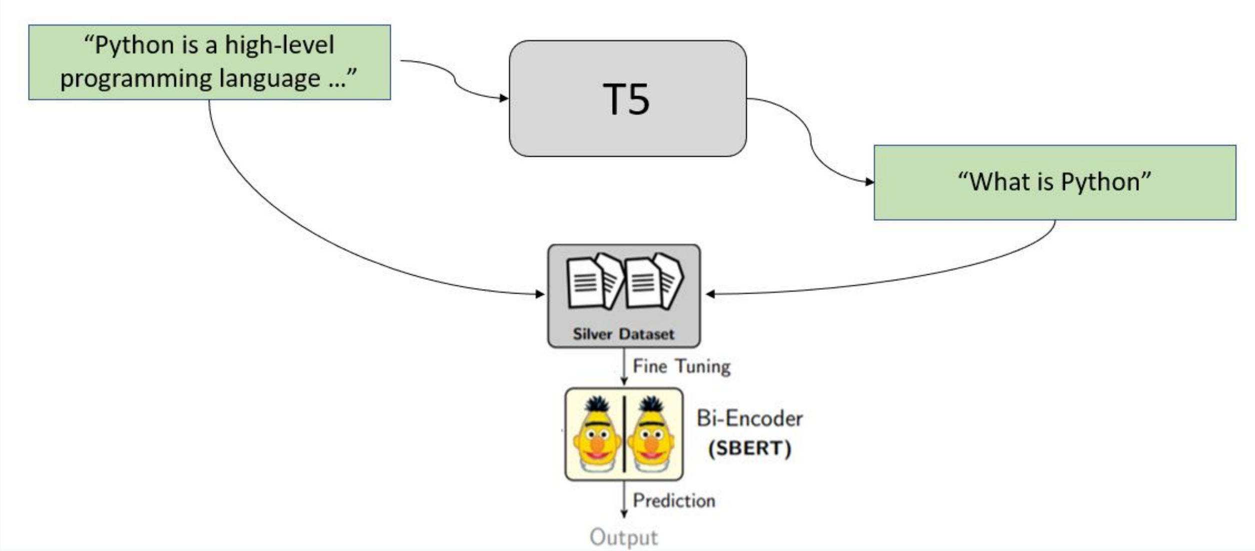


Source:

https://www.pinecone.io/learn/offline-evaluation/

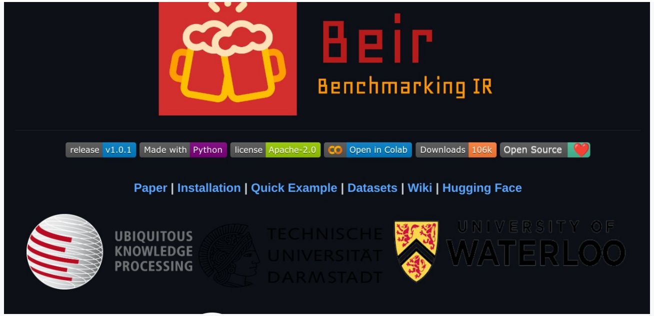
What happens when we don't have ground truth?

Creating ground-truth from scratch



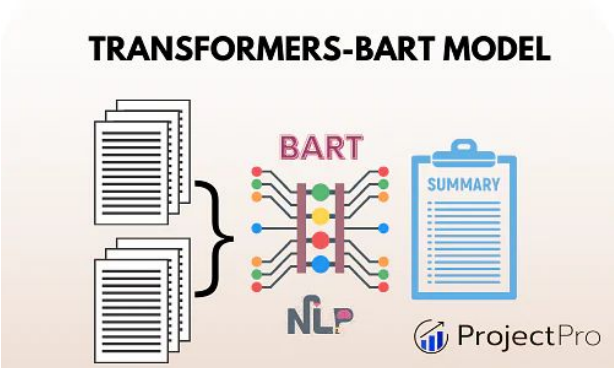
BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych



Paper | Installation | Quick Example | Datasets | Wiki | Hugging Face

Summarization



Demo

- BART is a sequence-to-sequence model trained as a denoising autoencoder.
- A fine-tuned BART model can take a text sequence (for example, English) as input and produce a different text sequence at the output (for example, French).
- This type of model is relevant for machine translation question-answering, text summarization or sequence classification
- Also, given two or more sentences, evaluates whether the sentences are logical extensions or are logically related to a given statement.

Aspect-based Sentiment Analysis

Utilize transformers and other ML models to generate sentiment for various aspects of an entity

```
@misc{YangL2022,
  title = {PyABSA: Open Framework for Aspect-based
Sentiment Analysis},
  author = {Yang, Heng and Li, Ke},
  doi = {10.48550/ARXIV.2208.01368},
  url = {https://arxiv.org/abs/2208.01368},
  keywords = {Computation and Language (cs.CL), FOS:
Computer and information sciences, FOS: Computer and
information sciences},
  publisher = {arXiv},
  year = {2022},
  copyright = {arXiv.org perpetual, non-exclusive license}
}
```

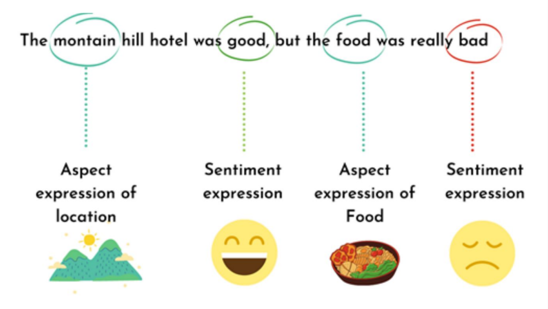
Consider these reviews:

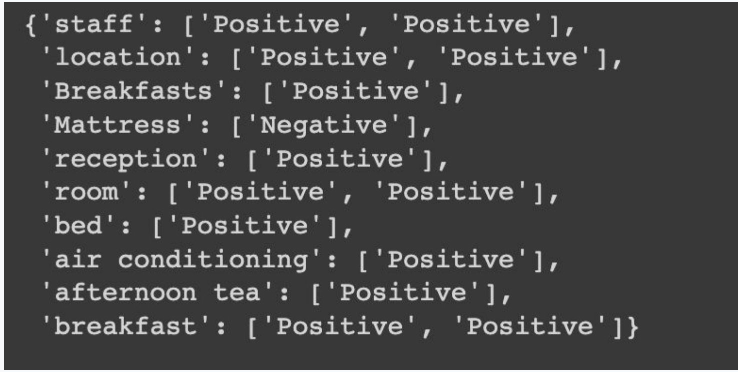
Friendly and accommodating staff helpful with transportation, restaurants and directions. Great location for all activities. Easy walk to Louvre. Breakfasts exceeded expectations. Mattress was too soft to my liking.

The reception was friendly and professional and speedy. The room was ready and perfect. The bed was very comfortable and the air conditioning was silent and potent. The free afternoon tea was amazing and open until 2am. The breakfast was one of the very best you could find in Paris

The room was awesome

Stayed here for two nights after a work trip in the city. I made an error in my booking and the hotel were very gracious and sorted it out for me. Kindly offered breakfast on the morning of my arrival. Very good selection for breakfast. Excellent location and fab staff would recommend


The montain hill hotel was good, but the food was really bad
Aspect expression of location
Sentiment expression
Aspect expression of Food
Sentiment expression



```
{'staff': ['Positive', 'Positive'],
 'location': ['Positive', 'Positive'],
 'Breakfasts': ['Positive'],
 'Mattress': ['Negative'],
 'reception': ['Positive'],
 'room': ['Positive', 'Positive'],
 'bed': ['Positive'],
 'air conditioning': ['Positive'],
 'afternoon tea': ['Positive'],
 'breakfast': ['Positive', 'Positive']}
```

Keyword creation using Transformers

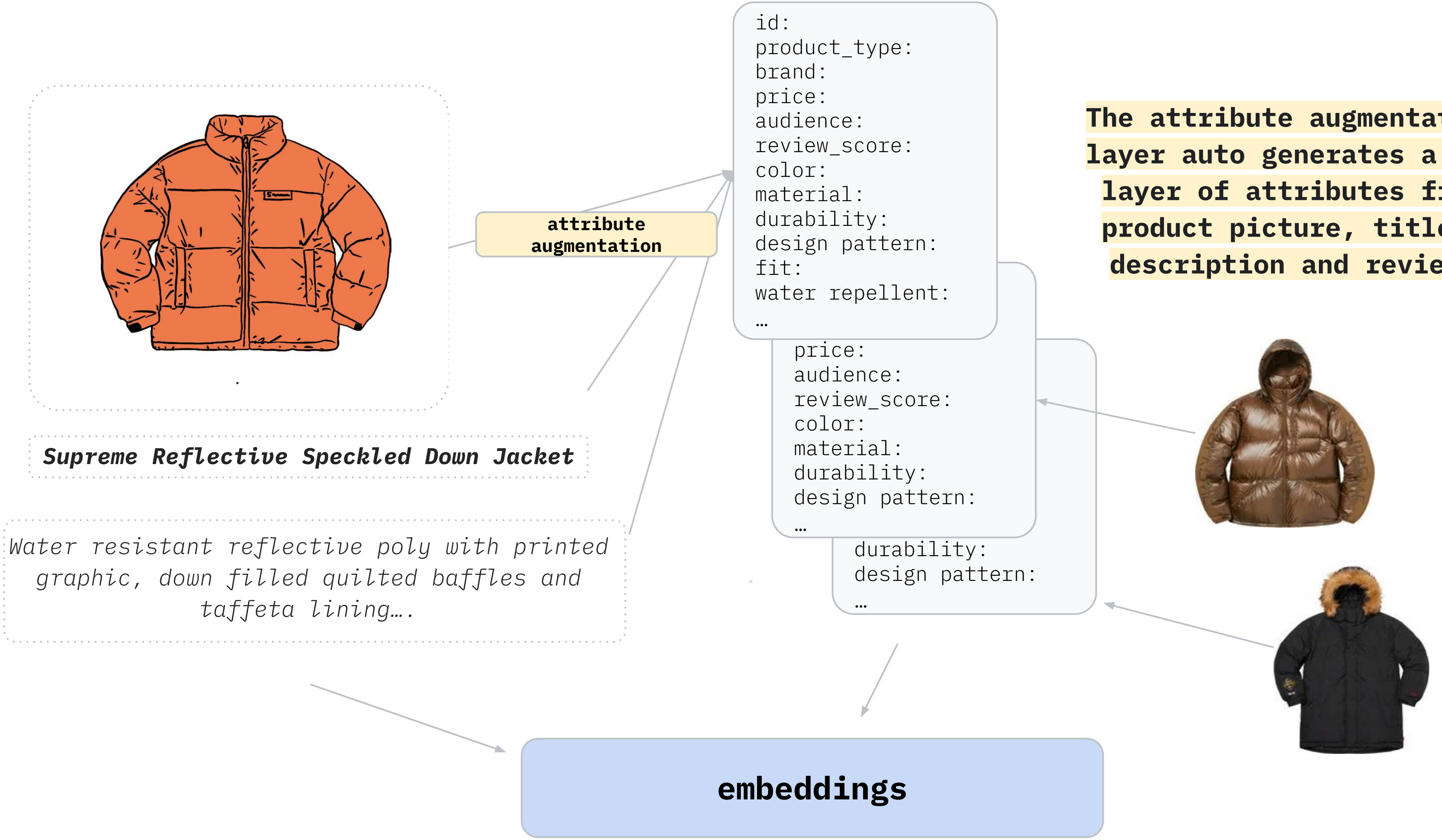
KeyPhraseTransformer is built on T5 Transformer architecture, trained on 500,000 training samples to extract important phrases/topics/themes from text of any length.

Source: https://github.com/Shivanandroy/KeyPhraseTransformer

- *Hotel staff were very helpful and friendly.*
- *I was very happy with the room and bathroom.*
- *I was very happy with my stay at the hotel.*
- *I would highly recommend this hotel to anyone who is looking for a place to stay.*
- *Hotel staff is very friendly and helpful.*
- *I was so happy to stay at this hotel... it was amazing!*
- *Louvre and many other locations*
- *Hotel staff were very friendly and helpful.*
- *Breakfast and afternoon snacks*
- *I know where i will be staying on our next trip to paris*

Query Intent Models

Queries need special handling and interpretation due to their tendency to be short, and too often imply more than they state explicitly



Supreme Reflective Speckled Down Jacket

Water resistant reflective poly with printed graphic, down filled quilted baffles and taffeta lining….

attribute augmentation

id:
product_type:
brand:
price:
audience:
review_score:
color:
material:
durability:
design pattern:
fit:
water repellent:
…

price:
audience:
review_score:
color:
material:
durability:
design pattern:
…

durability:
design pattern:
…

embeddings

The attribute augmentation layer auto generates a new layer of attributes from product picture, title & description and reviews

water resistant orange down jacket supreme..|

query intent

Product Type: Jacket
Color: Orange
Brand: Supreme
Water Repellent: Yes

Use knn model to reduce the sample space

similarity measure

embeddings

Use a combination of BM25 with Bi-Encoder and Cross-Encoders

Homework

- Implement similar hotel search engine for Miami hotels – feel free to apply any of the methods mentioned for retrieval and additional methods to improve your modeling
- Step 1:
  - Create a hotel summary/ encompasses a large amount of hotel info
- Step 2:
  - Create your search
  - Are the results similar to what you're searching for?
- Submission is a notion doc with the colab notebook and a writeup:
  - What you did?
  - How did you collate the data on the hotel
  - Simple feedback on your search

Thank you.

Appendix