

Chennai Meetup



Someshwaran Mohankumar

𝕏 @som2396, someshwaran.me

Elastic Support Engineer II



“Why” I’m here?

Vector search using elasticsearch and azure openai model



Let's talk about, "What"?

Search



Lexical search



Semantic search



Vector search





AT&T 11:37 AM

For Christine

Elastic Where to? +

Add Home

Add Work

Saved Places >

1015 Folsom St
San Francisco, CA

113 S Mary Ave
Sunnyvale, CA

511 N San Mateo Dr
San Mateo, CA

643 Webster St

Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123 space Search



Pa CLEAR

Pakodas Dish

Pan-Asian Cuisine

Pancake Dish

Panini Dish

Papad Dish

Pasta Dish

Pastry

Search for Movies, Events, Plays, Sports and Activities

MOVIES CINEMAS Filter HINDI ENGLISH MARATHI GUJARATI TAMIL MALAYALAM + 2 MORE

HINDI

- 83 (U) 89% Doctor Strange: In The Multiverse Of Madness 82%
- Anek (UA) NEW Bhool Bhulaiyaa 2 (UA) 89% 3D | IMAX 3D | 4DX 3D | 2D | MX4D 3D
- Chitrakrat (A) NEW Chitrakrat (A) 79% Fantastic Beasts: The Secrets Of Dumbledore (UA) 76%
- Dhaakad (A) NEW Dilwale Dulhania Le Jayenge (U) 89% Jurassic World: Dominion (UA) 89%
- Doctor Strange: In The Multiverse Of Madness 82% Paw Patrol: The Movie (U) 89% 3D | IMAX 3D | 3D | 4DX 3D | MX4D 3D
- Heropanti 2 (UA) 81% Sonic: The Hedgehog 2 (U) 89% Jurassic World: Dominion (UA) 89%
- Jayeshbhai Jordaar (UA) 69% Jersey (UA) 81% The Bad Guys (UA) 89% 3D | 2D | MX4D 3D
- Jurassic World: Dominion (UA) 89% Top Gun: Maverick (UA) 89% 3D | 2D | 4DX
- K.G.F. Chapter 2 (Hindi) (UA) 93% Adrathy (UA) NEW Jersey (UA) 81% MARATHI
- Chandramukhi (Marathi) (UA) 87% Keedam (UA) NEW
- Kokka (UA) NEW Saunka Saunkne (UA) 81% PUNJABI

ENGLISH

GUJARATI

BENGALI

TAMIL

MALAYALAM

PUNJABI

Lexical Search?

What is Lexical search?



- 1: Winter is coming.
- 2: Ours is the fury.
- 3: The choice is yours.

1

fury



<u>term</u>	<u>freq</u>	<u>documents</u>
choice	1	3
coming	1	1
fury	1	2
is	3	1, 2, 3
ours	1	2
the	2	2, 3
winter	1	1
yours	1	3

Sample documents and resulting inverted index

How can we do Lexical search?



Dev Tools

Console

Console

Search Profiler

Grok Debugger

Painless Lab BETA

History Settings Variables Help

```
1 GET /_analyze
2 {
3   "char_filter": [ "html_strip" ],
4   "tokenizer": "standard",
5   "filter": [ "lowercase", "stop", "snowball" ],
6   "text": "These are <em>not</em> the droids you are looking for."
7 }
8
9 |
```



```
1 [
2   "tokens": [
3     {
4       "token": "droid",
5       "start_offset": 27,
6       "end_offset": 33,
7       "type": "<ALPHANUM>",
8       "position": 4
9     },
10    {
11      "token": "you",
12      "start_offset": 34,
13      "end_offset": 37,
14      "type": "<ALPHANUM>",
15      "position": 5
16    },
17    {
18      "token": "look",
19      "start_offset": 42,
20      "end_offset": 49,
21      "type": "<ALPHANUM>",
22      "position": 7
23    }
24  ]
25 }
```

```
{ "tokens": [{

    "token": "droid",
    "start_offset": 27, "end_offset": 33,
    "type": "<ALPHANUM>",
    "position": 4

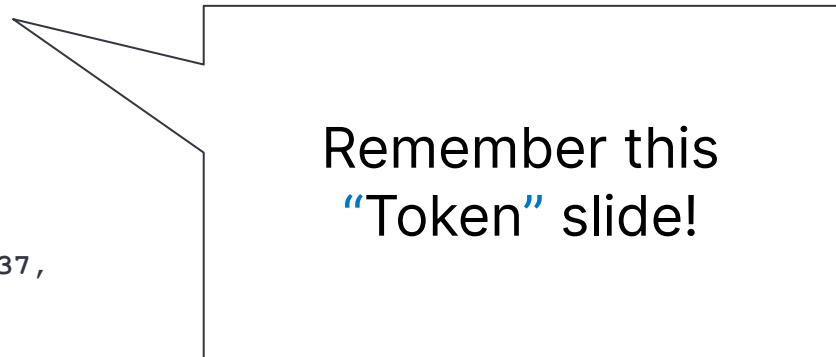
}, {

    "token": "you",
    "start_offset": 34, "end_offset": 37,
    "type": "<ALPHANUM>",
    "position": 5

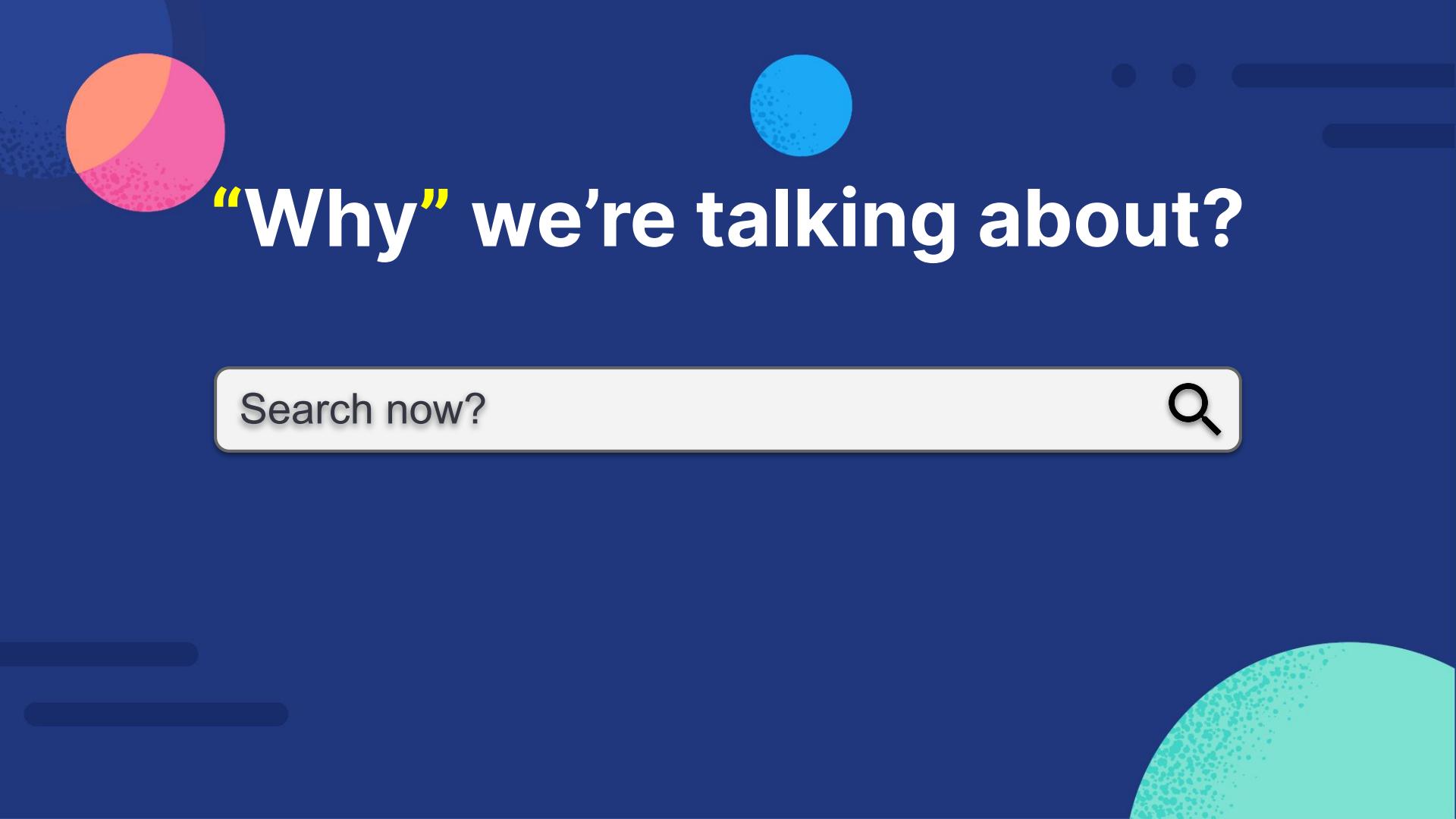
}, {

    "token": "look",
    "start_offset": 42, "end_offset": 49,
    "type": "<ALPHANUM>",
    "position": 7

}]}
```



Remember this
“Token” slide!



“Why” we’re talking about?

Search now?



Elasticsearch: You Know, for Search

Elasticsearch: You Know, for **Vector** Search

What is Vector search?



Bag of Words

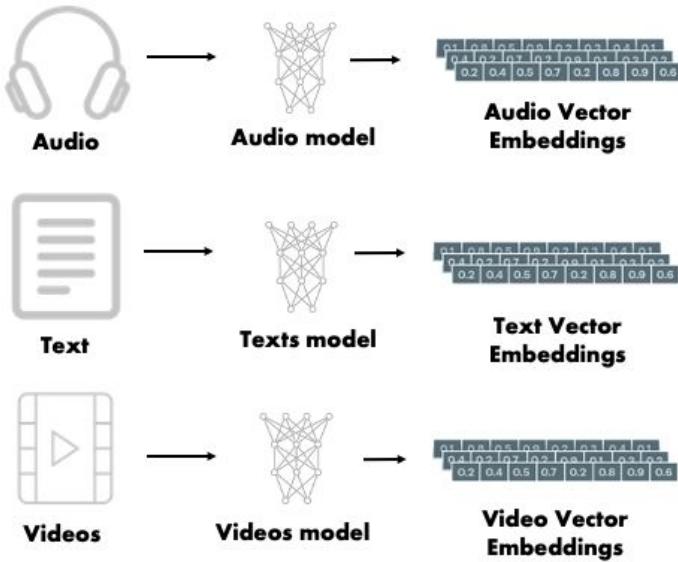
As a simplified illustration

These are not the droids you are looking for.

No. I am your father.

these: 1, are: 2, not: 1, the: 1, droid: 1, you: 1, look: 1, for: 1
no: 1, i: 1, am: 1, you: 1, father: 1

```
[these, are, not, the, droid, you, look, for, no, i, am, father]  
[1,      2,      1,      1,      1,      1,      1,      0,      0,      0]  
[0,      0,      0,      0,      1,      0,      0,      1,      1,      1]
```



- "Change":
 - Positive: "Change is the essence of progress."
 - Negative: "I'm not comfortable with this sudden change."
- "Challenge":
 - Positive: "Embrace the challenge and grow stronger."
 - Negative: "Dealing with this challenge feels overwhelming."

MAMMAL



REALISTIC

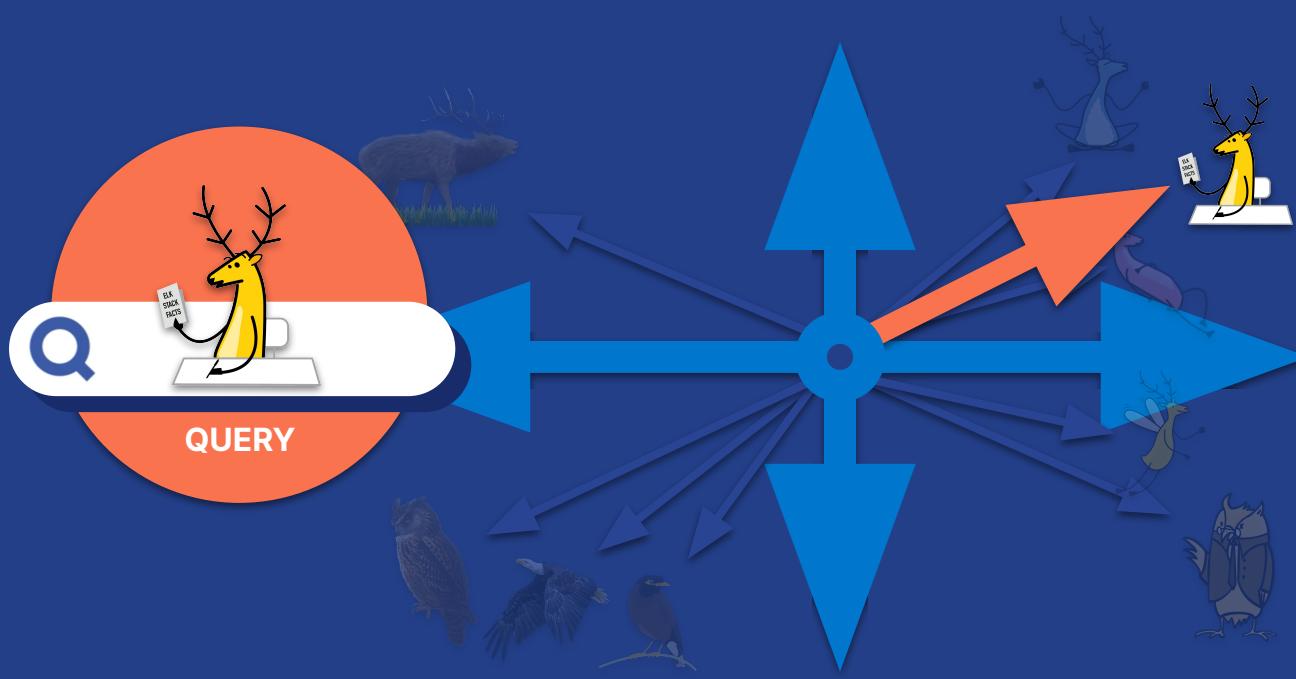
BIRD



CARTOON



Vector search finds similar objects as those close to the query



Relevance	Result
Query	
1	
2	
3	
4	
5	

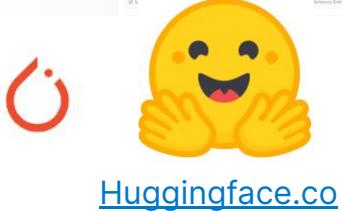
What is a Model?

Hugging Face NLP Libraries

Hugging Face Model	task-type
<u>Name-Entity recognition</u>	ner
<u>Text embedding</u>	text_embedding
<u>Text classification</u>	text_classification
<u>Zero shot classification</u>	zero_shot_classification
<u>Question & Answer</u>	question_answering

Eland Imports PyTorch Models

The screenshot shows the HuggingFace website's search interface. The search bar at the top contains the query "Models". Below the search bar, there are several categories: Tasks (e.g., Fill Mask, Question Answering, Summarization, Text Generation, TextClassification, TokenClassification, Translate, Zero Shot Classification), Libraries (e.g., PyTorch, TensorFlow, JAX), Datasets (e.g., common_gen, books100k, gutenberg, iit), Languages (e.g., en, es, fr, de, zh), Licenses (e.g., Apache-2.0, MIT, CC-BY-NC), and Other (e.g., AutoTokenizer, GPT2, GPT2CapGen, TextWithAnswer). The main content area displays a list of model cards, each with a thumbnail, name, description, and download count. Some visible models include: gpt2 (last updated May 21, 2023, 14.8M), bert-base-uncased (last updated May 21, 2023, 1.6M), distilbert-base-uncased (last updated May 21, 2023, 1.6M), roberta-base (last updated Jun 02, 2023, 1.6M), t5-base (last updated Jun 02, 2023, 1.6M), mmlu1-NLP/cross-ent-zen-1-en (last updated Jun 02, 2023, 1.6M), bert-base-chinese (last updated May 21, 2023, 1.6M), sentence-transformers/multiqa-MiSiLM-ls-cos-v2 (last updated May 21, 2023, 1.6M), bert-base-multilingual-cased (last updated Jun 02, 2023, 1.6M), distilbert-base-uncased-finetuned-ast-2-english (last updated Jun 02, 2023, 1.6M), albert-base-v2 (last updated May 21, 2023, 1.6M), roberta-large (last updated May 21, 2023, 1.6M), and sentence-transformers/all-MiSiLM-ls-v2 (last updated May 21, 2023, 1.6M).



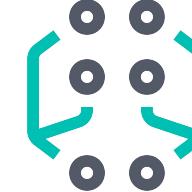
```
$ eland_import_hub_model  
--url https://Cluster_URL  
--hub-model-id bert_model  
--task-type text_embedding  
--start
```



Inference, not training

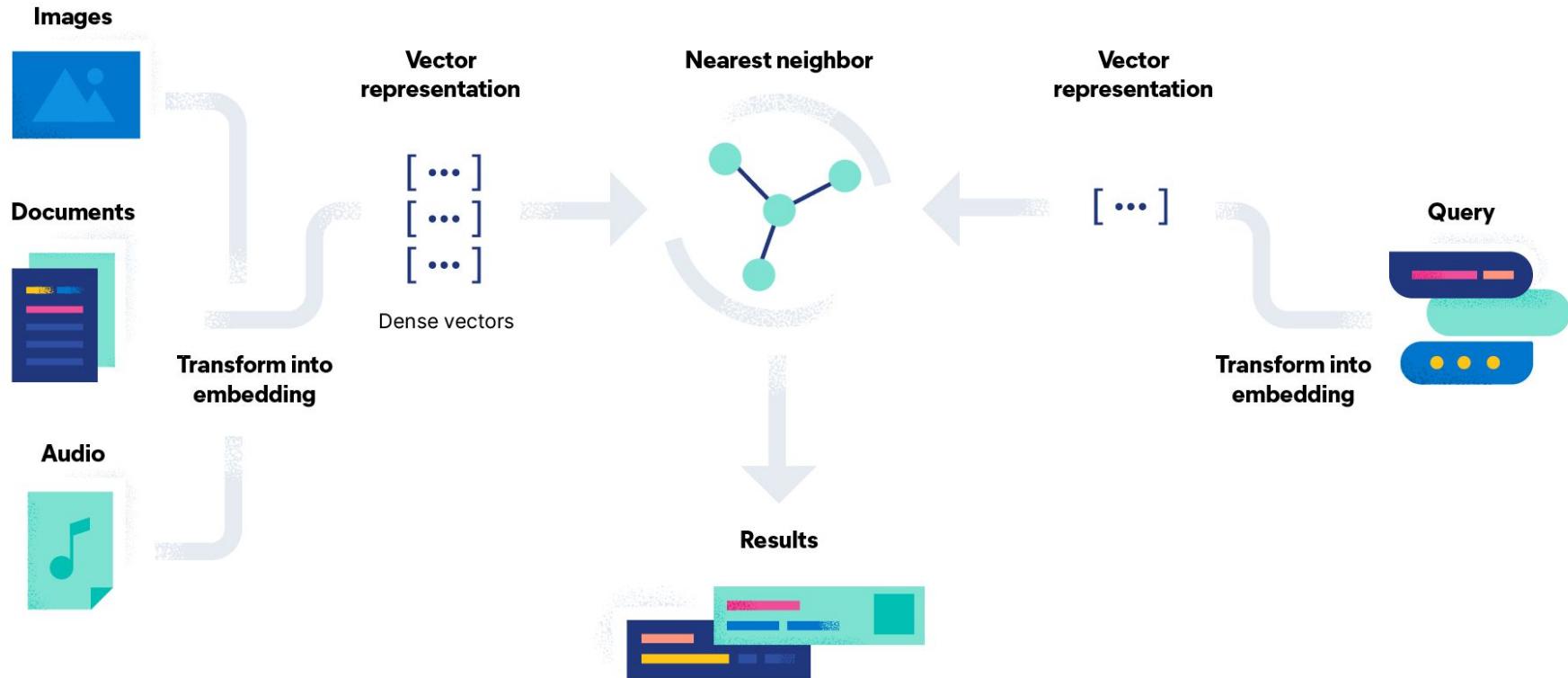
The screenshot shows the Elastic Model Management interface. The top navigation bar includes Code, File, Edit, Selection, View, Go, Debug, Terminal, Window, Help, and a timestamp (Mon 9:11 PM). The main menu has tabs for Overview, Anomaly Detection, Data Frame Analytics, Model Management (which is selected), Data Visualizer, and Settings. Below the tabs, a section titled "Trained Models" lists several entries:

Name	Description	Type	Status	Created at
bert-base-uncased	Model for identifying text uncased. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:10:31.366
distilbert-base-uncased	Model distilbert-base-uncased. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:26:18.460
roberta-base	Model roberta-base. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:11:03.166
t5-base	Model t5-base. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:09:17.040
mmlu1-NLP/cross-ent-zen-1-en	Model mmlu1-NLP/cross-ent-zen-1-en. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:11:03.166
bert-base-chinese	Model bert-base-chinese. Pre-trained on 2-gram Chinese for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:11:03.166
sentence-transformers/multiqa-MiSiLM-ls-cos-v2	Model sentence-transformers/multiqa-MiSiLM-ls-cos-v2. Pre-trained on 2-gram English for task type text_embedding.	pytorch - text_embedding	started	Jan 27, 2022 @ 13:10:31.366
bert-base-multilingual-cased	Model bert-base-multilingual-cased. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:11:03.166
distilbert-base-uncased-finetuned-ast-2-english	Model distilbert-base-uncased-finetuned-ast-2-english. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:26:18.460
albert-base-v2	Model albert-base-v2. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:11:03.166
roberta-large	Model roberta-large. Pre-trained on 2-gram English for task type text_classification.	pytorch - text_classification	started	Jan 27, 2022 @ 13:11:03.166
sentence-transformers/all-MiSiLM-ls-v2	Model sentence-transformers/all-MiSiLM-ls-v2. Pre-trained on 2-gram English for task type text_embedding.	pytorch - text_embedding	started	Jan 27, 2022 @ 13:10:31.366

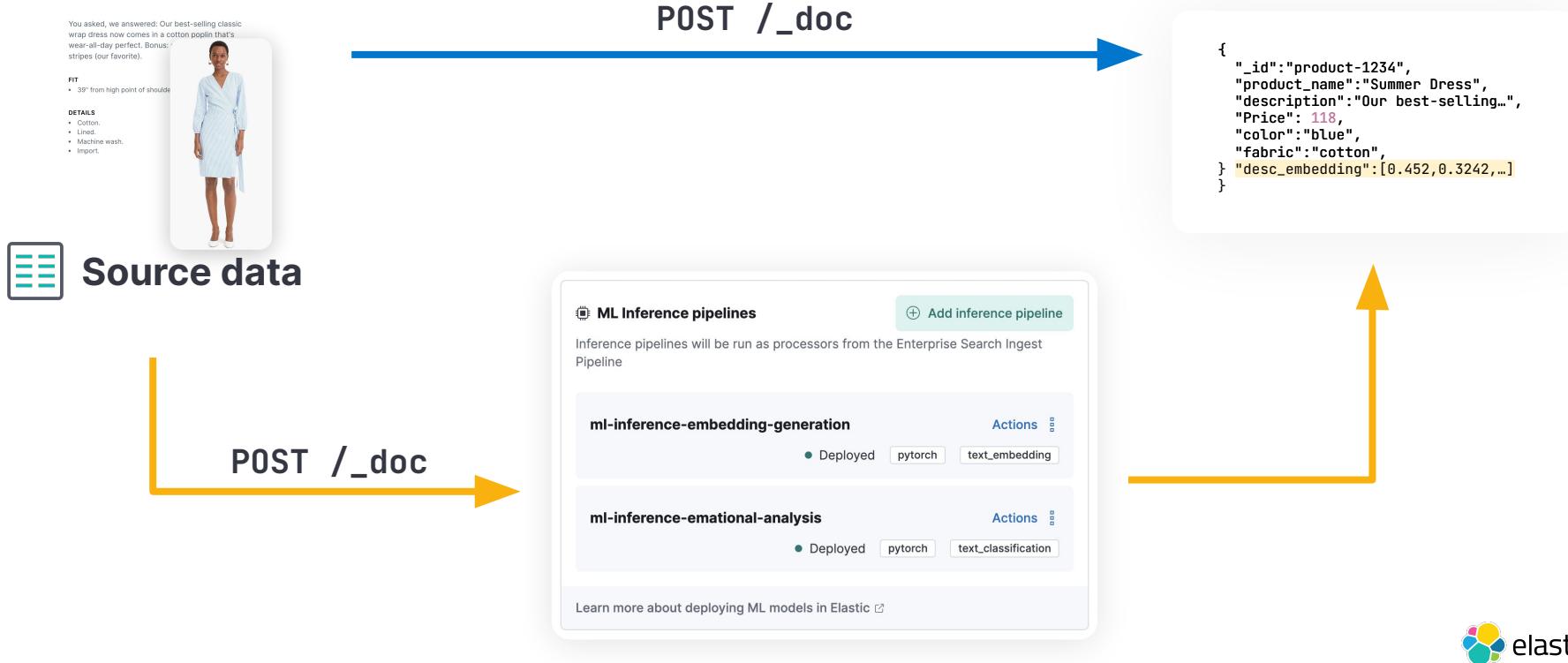


How Do You Search Vectors?

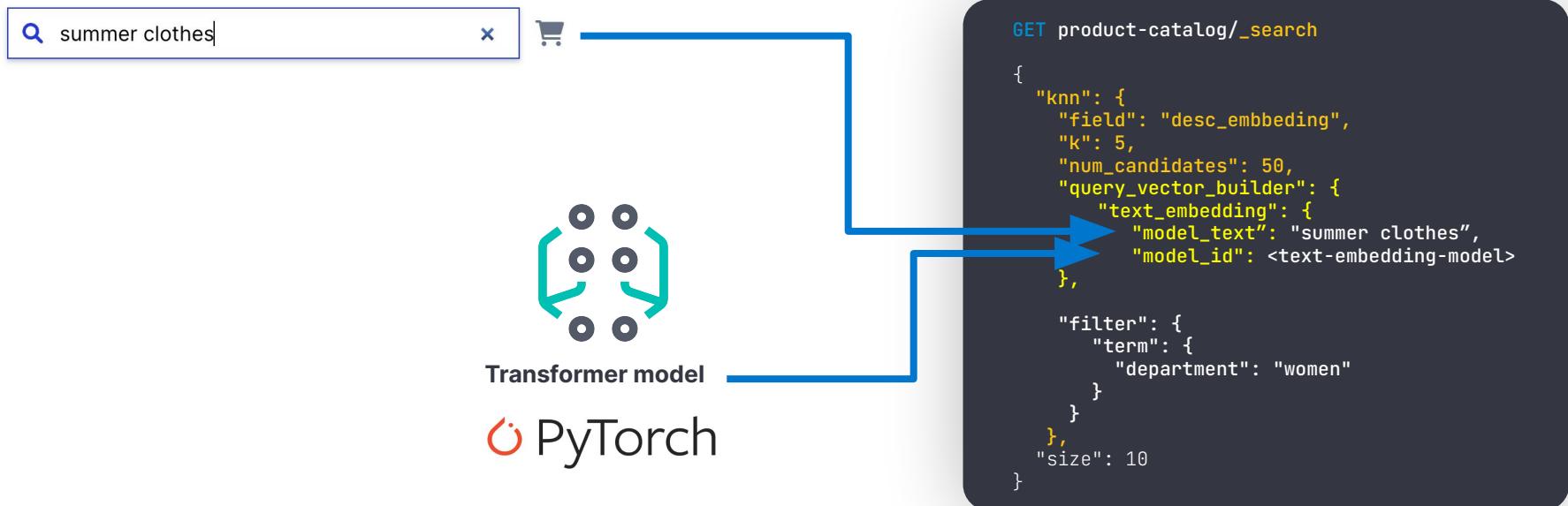
Vector Search



Data Ingestion and Embedding Generation



Vector Query



But How Does It Really Work?

Hierarchical Navigable Small Worlds (HNSW)

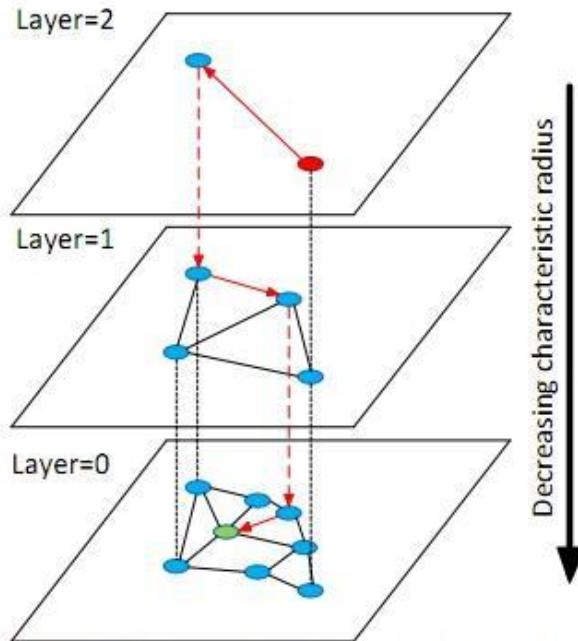
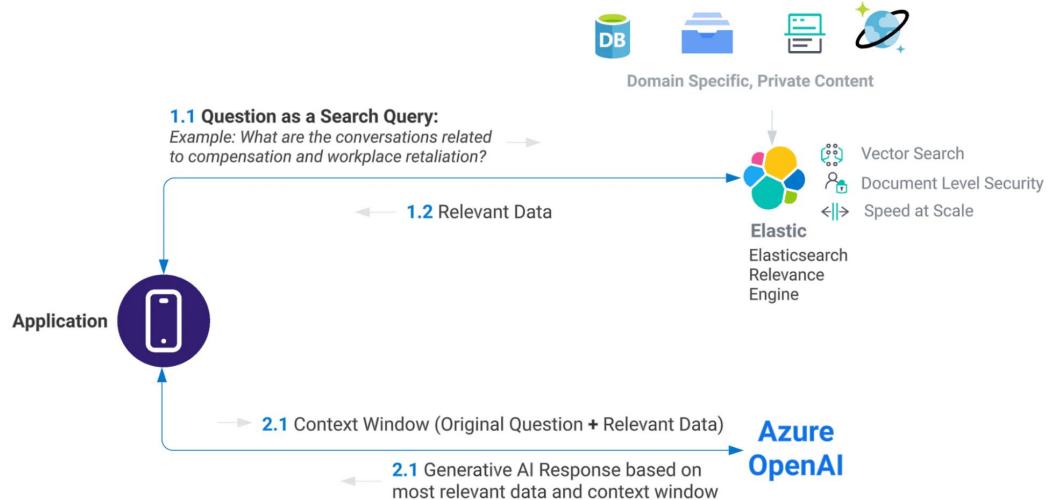


Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

<https://zhuanlan.zhihu.com/p/98028479>

Demo Time!



The Elastic search platform is for everyone



YOUR PERFECT BANKING PARTNER



TECHNOLOGY

FINANCE

TELCO

CONSUMER

HEALTHCARE

PUBLIC SECTOR

AUTOMOTIVE /
TRANSPORTATION

RETAIL

THANK YOU

*Please Feel Free To Connect
Twitter @som2396,
LinkedIn somdevsupport
someshwaran.me*