Unlocking multilingual insights

Translating datasets with Python, LangChain, and Vector Database

elastic

Jessica Garson

@JessicaGarson

@JessicaGarson@macaw.social

Senior Developer Advocate

elastic

https://github.com/JessicaGarson/Translation-and-Vector-Search-with-LangChain-and-Elasticsearch

elastic


My band did a run of three shows in Shanghai at the beginning of the summer

I saw some local press about the shows but couldn’t understand much.

I wrote some code which translates the text of these clippings from Chinese to English using LangChain and Elastic's vector database to learn more about the translated documents.

elastic

Overview of the solution

Step 1: Load data in

Step 2: Use LangChain and DocTran to translate the data.

Step 3: Load the translated data into Elasticsearch

Step 4: Use vector search to learn more about more about the tour.

elastic

Let’s try this with a more general dataset

elastic

I decided to use a collection of news articles in Spanish, known as the DACSA corpus. This dataset is available from hugging-face.

Subset (2): spanish · 2.12M rows

Split (4): train · 1.8M rows

id string · *lengths* 64 64	summary string · *lengths* 40 2.99k	article string · *lengths* 512 733k
53f42b049dfe73a9d3f5c5684a97ed7330f4d0145ec5b94a6780ffb8b49323e4	El perfil se corresponde con el de un varón, con un nivel de formación de hasta secundaria,…	Más de 860.000 jóvenes menores de 25 años abandonaron el mercado laboral desde que…
c71ee83329b2480b0904e7433d489c22c1a2e90237df9277e627903502d4b0d2	La mujer, de 28 años, tiene quemaduras graves y está en la UCI. Sus dos hijas, de 5 y 3 años…	La madre de 28 años grave por quemaduras e inhalación de humo en el incendio de su casa e…
a7834fd4f309388af8be175800e89f982da0d8ff5c966d6925ee63271bc1c00e	El presidente del PP hace un llamamiento a los valencianos para que participen en las…	Barberá y Fabra se saludaban ayer antes de que el presidente impartiera una conferencia sobre…
3da811329c3cc0ae31b9c2111ee48d5497cc5e2f7381e3e52777f055a9a3e081	Diez de los doce meses del año batieron sus propios récord de temperatura.	El informe de la Agencia Nacional de Océanos y Atmósfera de Estados Unidos (NOAA) no deja…
7e43b72c8d38ac38ec14ea3cdfba6be5cf42a1490fe2bb9ce29969ce9e06e967	Hay que quedarse con la interpretación de Fanny Ardant, en un papel sorprendente de cabeza…	LOLA PATE. Dirección: Nadir Moknèche. Intérpretes: Fanny Ardant, Tewfik Jallab, Nadi…

elastic

Getting started

elastic

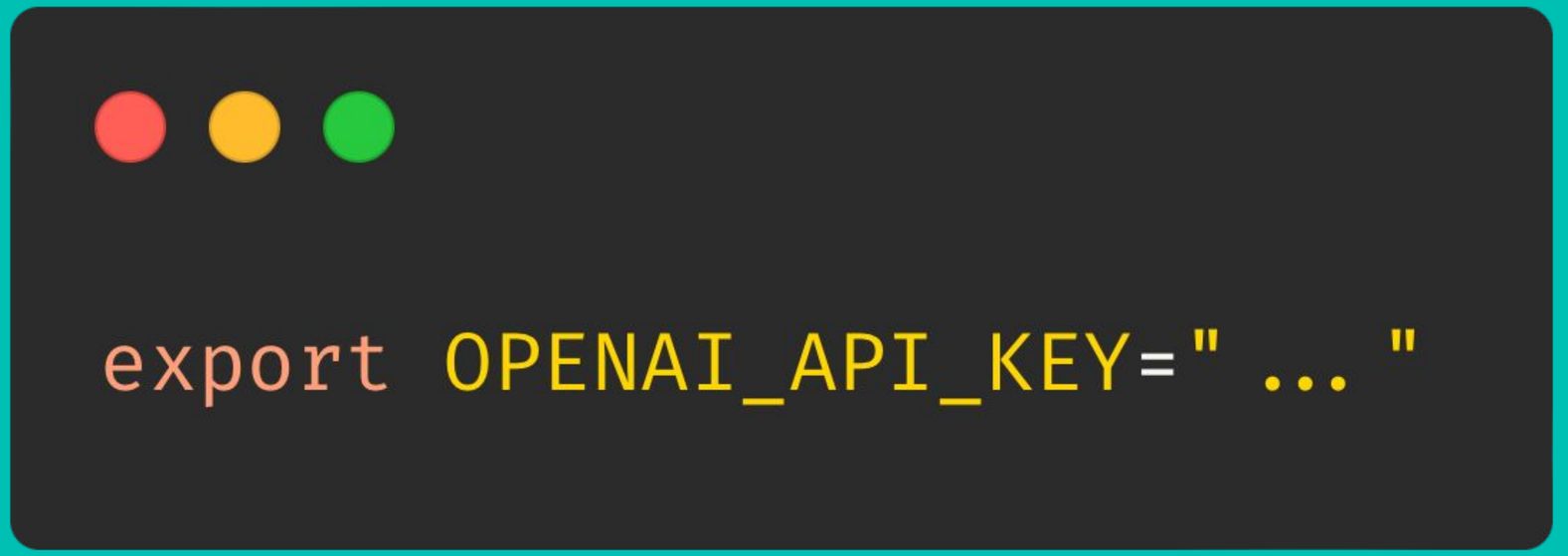
What version of Elasticsearch should I use?

This demo uses Elasticsearch version 8.15, but you can use any version of Elasticsearch that is higher than 8.0.

elastic

What version of Python should I use?

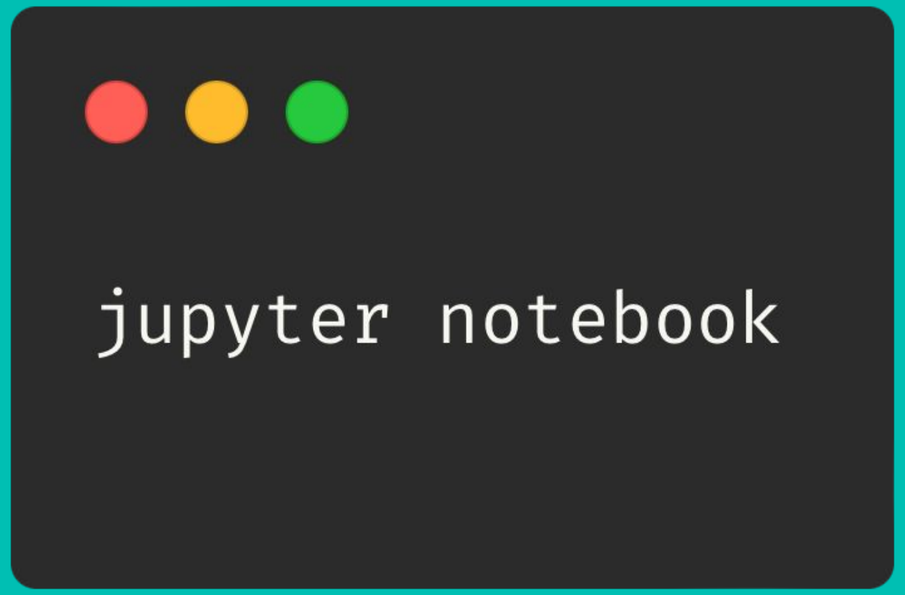
The version of Python that is used is Python 3.12.1 but you can use any version of Python higher than 3.9.



elastic

```
pip install jupyter pandas langchain nest_asyncio langchain-elasticsearch langchain-openai tiktoken
elasticsearch datasets
```

elastic



elastic

elastic


Using a vector
database
elastic

A vector database allows you to find similar data quickly. It stores vector embeddings, a type of vector data representation that converts words, sentences, and other data into numbers that capture their meaning and relationships.

elastic

What are embeddings?

Embeddings leverage a machine learning model to translate text into numbers, allowing you to perform vector searches.

What is RAG?

Retrieval-augmented generation (RAG) integrates external information retrieval into generating responses by Large Language Models (LLMs).

elastic

Vector Databases vs RAG

elastic

Demo

elastic

Can’t you do this in one step with a multilingual model?!?!?!

elastic

Next steps

elastic

This talk as a blog post

https://www.elastic.co/search-labs/blog/unlocking-multilingual-insights

elastic

Let me know if this talk inspires you to build anything. I'm @JessicaGarson on most platforms.

Let us know if you need if this demo inspires you to build anything or if you have any questions on our Discuss forums and the community Slack channel.

elastic

https://github.com/JessicaGarson/Translation-and-Vector-Search-with-LangChain-and-Elasticsearch

elastic


Thank you!
elastic