Introducing txtai, an AI-powered search engine built on Transformers

Add Natural Language Understanding to any application

Search is the base of many applications. Once data starts to pile up, users want to be able to find it. It's the foundation of the internet and an ever-growing challenge that is never solved or done.

The field of Natural Language Processing (NLP) is rapidly evolving with a number of new developments. Large-scale general language models are an exciting new capability allowing us to add amazing functionality quickly with limited compute and people. Innovation continues with new models and advancements coming in at what seems a weekly basis.

This article introduces txtai, an AI-powered search engine that enables Natural Language Understanding (NLU) based search in any application.

Introducing txtai

txtai builds an AI-powered index over sections of text. txtai supports building text indices to perform similarity searches and create extractive question-answering based systems. txtai also has functionality for zero-shot classification. txtai is open source and available on GitHub.

txtai and/or the concepts behind it has already been used to power the Natural Language Processing (NLP) applications listed below:

- paperai — AI-powered literature discovery and review engine for medical/scientific papers
- tldrstory — AI-powered understanding of headlines and story text
- neuspo — Fact-driven, real-time sports event and news site
- codequestion — Ask coding questions directly from the terminal

Build an Embeddings index

For small lists of texts, the method above works. But for larger repositories of documents, it doesn't make sense to tokenize and convert all embeddings for each query. txtai supports building pre-computed indices which significantly improves performance.

Building on the previous example, the following example runs an index method to build and store the text embeddings. In this case, only the query is converted to an embeddings vector each search.