Building a natural language code search tool with vector embeddings and LLM

Dima Yanovsky

Context

Initial Frustrations

❌ Can’t follow all the functions in a chain

❌ Can’t read codebase like a book

❌ Limited access to the code authors

How to learn fast?

⁉️ Ask senior SWE a lot of questions

❌ Seniors have limited time

📉 Learning velocity decreases

You need to understand

✅ How processes are implemented

✅ Workflows and business logic

✅ Codebase on a conceptual level

Solving the problem

🤖 Build code knowledge tool

❓ Receives a question in natural language

👍 Answers the question and provides relevant code from the codebase

Background

⚙️ **Code elements**	Functions, interfaces, structs and other parts of the code that serve a purpose (aka chunks)
✂️ **Code parsing and splitting**	A technique to split it codebase into code elements.
↗️ **Vector Embeddings**	Representation of some object in a multi-dimensional space. Neural nets capture the semantic meaning of the object and express it as a vector. Objects with similar meanings are close to each other
🫙 **Vector Storage Database**	Database that stores vector embeddings and is used to find closest vectors to an input vector.
🤖 **Large Language Model**	Model that can understand and generate human-like text based on the vast amount of data it's been trained on

Parsing and Splitting

Code Parser Program

def foo(): int x =

def bar(): for i

def foobar(): foo()

“filename” : “important_file.py”,
“file_description”: “This file contains…”,
“chunk_type” : “function”,
“code”: “def foobar(): int f = foo() int b = bar()”,
“functions_called”: [foo, bar]

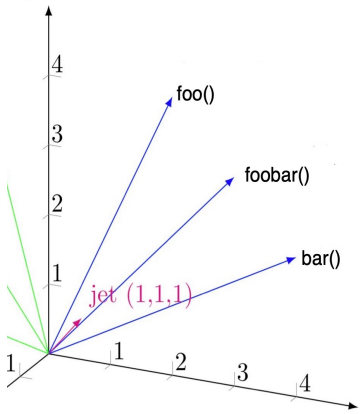
Chunk (code element)

Standard Approach

def foo(): int x =

def foobar(): for i

Embedding Model



Modified Approach

def foo(): int x =

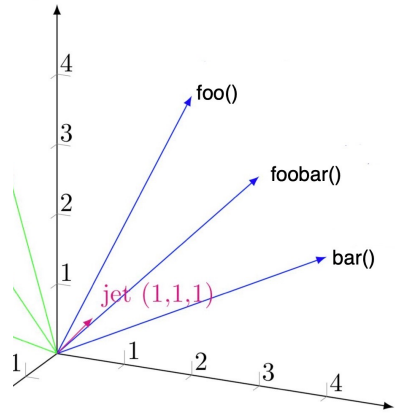
def foobar(): for i

LLM

Q: What does function foo() ..?

Q: How foobar() calculates .. ?

Embedding Model



User: “Explain foo() function”

Embedding Model

[123.553, 54.22, -34.5, …, 76.21]

find(v)

def foo(): int x =

def foobar(): for i

…

Prompt

User asked this {query}
Answer using context:
{chunk_1}
{chunk_2}
{chunk_3}
{chunk_4}
{chunk_5}

LLM

Answer: Function foo() in your code is used for calculating…