WHITE PAPER / SEARCH SYSTEMS

When the index speaks a different language

An engineer searching for “integer square root” may never type the identifier isqrt. We measured what happens when a small reference index can search the words used to explain an entry.

A catalog written for people who already know the name

Someone looking up a function by name can use an alphabetical index. Someone who remembers what the function does faces a different problem. An internal knowledge base has the same divide between the terminology in its titles and the language people bring to the search box.

An ICLR study of retrieval for code summarization prompted a smaller operational question. Before a team invests in a new ranker, how much can it gain by indexing material it already maintains? Our test begins at the point where search selects candidate entries for a reader.

Two indexes over the same 74 entries

We used public functions in the math and statistics modules shipped with Python 3.12. We wrote 16 short task descriptions and assigned each an intended function. For example, “greatest common divisor” points to math.gcd. The first index contained module and function names, with punctuation separated into words. The second also included each function’s built-in documentation.

Both used word-level TF–IDF with English stop words removed and cosine scoring. We counted only candidates with positive similarity. Ties followed alphabetical identifier order. The queries, candidate set and intended answers stayed fixed between runs.

More of the intended entries reached the reader

Retrieval results on the same 16 technical queries

Intended function	Names only	Names and documentation
Ranked first	5 of 16	12 of 16
In the first three	5 of 16	14 of 16
No positive match for the intended entry	11 of 16	1 of 16

The richer index put math.isqrt first for “integer square root.” Two cases illustrate the remaining gap. The intended entry for “most common value” appeared eighth. The query about a finite value gave math.isfinite no positive score. Even a correct reference description can phrase the concept differently from the reader.

Before changing the search model

Run separate evaluations for exact identifiers and descriptions of a task. Inspect the source fields on a missed entry. A glossary definition, unit of measure, abbreviation or editorial summary may supply the words the reader uses.

These 16 hand-written queries cover two small modules. For a knowledge portal, repeat the comparison with consented search logs, a broader catalog and multiple acceptable results per task. Keep a small set of queries held out while changing the index.

Research & data

- Retrieval-Augmented Generation for Code Summarization via Hybrid GNN · ICLR 2021
- Mathematical functions · Python 3.12 documentation
- Mathematical statistics functions · Python 3.12 documentation