CS242 Project: StackOverflow Search Engine

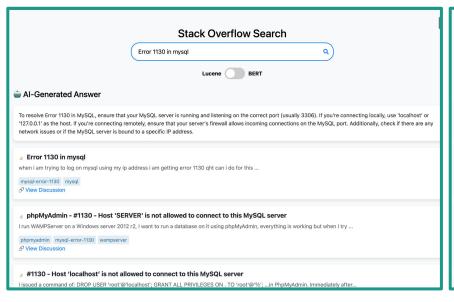
Winter 2025

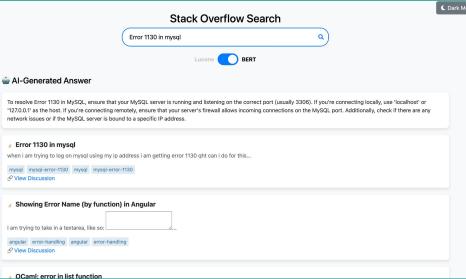
Team 06

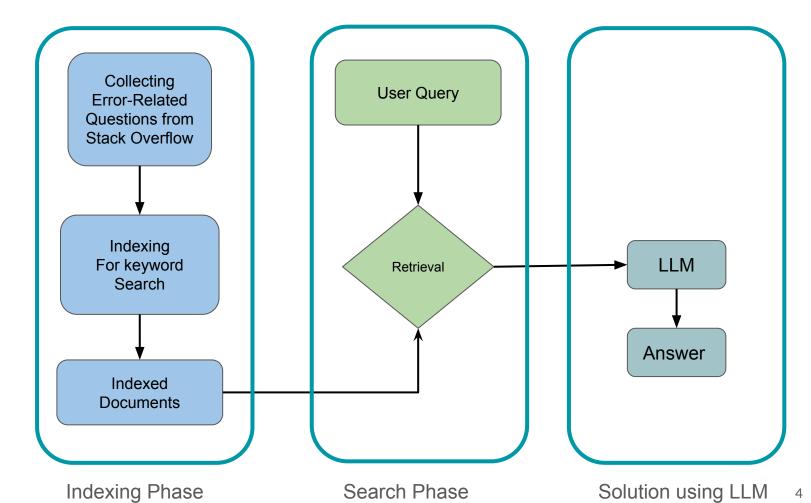
Team 06

- 1. G M Shahariar
- 2. Md Taukir Azam Chowdhury
- 3. Md. Olid Bhuiyan
- 4. Samiha Khan
- 5. Zabir Al Nazi

GUI Overview







Steps

Data Collection

Data Source:

- Scraped StackOverflow (224,046 error-related entries) using Scrapy
- Tags: error-handling", "compiler-errors", "runtime-error", "syntax-error",
 "linker-errors", "importerror"
- Collected questions, tags, answers, and comments, and saved those as json.

Challenges & Solutions:

- **Slow scraping** → Split workload across team
- Pagination Handling
- Applied "Download Delay" to avoid blocking from server
- Handling duplicates → Post Processing and filtering

Indexing Methods

- Pylucene (Sparse Retrieval):
 - Inverted index for fast keyword-based retrieval
 - Scoring
 - i. Vector Space Model (VSM)
 - ii. Probabilistic Models such as Okapi **BM25** and DFR
 - iii. Language models
 - **TextField**: title, body, tags, comments, answers.
 - StoredField: link, score, creation_date.
- - Transforms text into contextualized vector representations for efficient semantic matching.
 - Tokenized and truncated Stack Overflow question titles and bodies to first 512 tokens.
 - Used the [CLS] token from BERT's last hidden state for embedding extraction.
- FAISS (Efficient Similarity Search):
 - Stores embeddings for fast nearest neighbor retrieval
 - IndexFlatL2 exact nearest neighbor with euclidean distance.

LLM Integration

Large Language Model (Qwen 2.5 3B Instruct)

- Takes an error query and top-5 retrieved documents as inputs
- Generates **response** as a solution to the query

How It Works:

- 1. **User Query:** Enters programming error
- 2. **Retrieval:** Lucene or BERT fetches top-k relevant documents
- 3. **Prompting:**
 - Construct a closed book QA style prompt using query and documents
 - Prompt sent to model via Hugging Face API
 - o Model generates **an answer** to the query
- 4. **Result Displayed:** Shown in **GUI**

Challenges Solved:

- Improved query handling for ambiguous searches
- Better response generation

GUI Implementation

Features:

- User-friendly web interface
- Allows Lucene or BERT-based search
- Displays ranked solutions with Al-generated summaries
- Backend: Flask API connecting search & AI models

Hugging Face API Usage:

- Used pre-trained LLM via API call
- Requires API key authentication
- Sends query & retrieved answers for processing



http://169.235.31.51:8080/

Final Results & Achievements

- Fully operational Al-powered error resolution system
- Successfully implemented Lucene, BERT & FAISS
- Optimized search retrieval with Hugging Face API (Additional Contribution)
- Interactive GUI for real-time query processing (Additional Contribution)

Future Scope:

- Enhance UI for better UX
- Expand dataset for multi-language support
- Optimize Hugging Face API requests for speed & cost

Demo