Web Search Engine – Project Proposal

Search Overflow

Narasimman Sairam (ns3184), Manasa Kunaparaju (mk5376)

Objective of the project:
A Query Engine to leverage the body of knowledge created by the socio-professional media, to recommend high quality, embeddable code.

Category: Specialized search engine: Implement a search engine that gives high quality results for some particular class of queries.

Q&A services like stack overflow are filling archives with millions of entries that contribute to the body of knowledge in software development" and they often become the substitute of the official product documentation.

When a developer is writing code and has to check online for a syntax or api related information, typically the developer would:

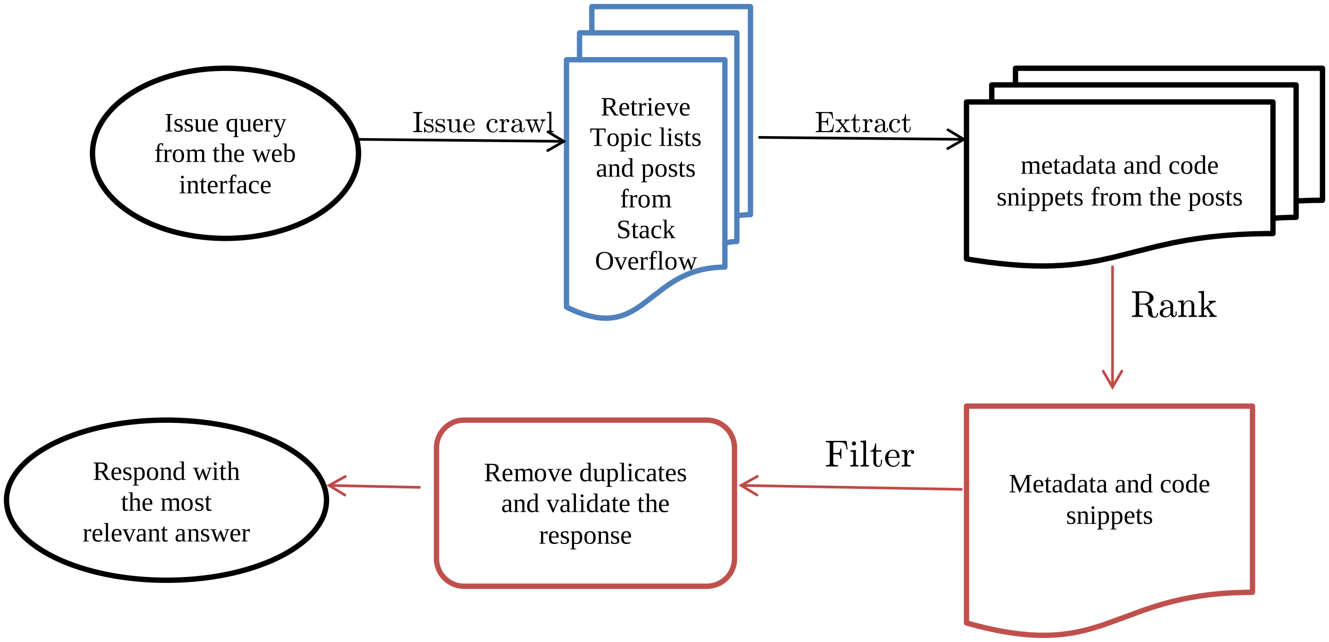
- Open the browser
- Google for the particular question (Ex: how to iterate over a map in Java)
- From the list of responses, (usually stack overflow) open couple of tabs of stack overflow pages that answers the query.
- Look for the best answer (Top voted in stack overflow website)
- Close the other tabs
- Copy paste the top rated answer to the development environment and continue working

Our search engine will do all the hard work of querying for the best answer and suggest the developer with the top rated response from the stack Overflow website. The answer to the query will be exactly one response (top rated).

Searching for code examples is possible using Stack Overflow directly. However, using designated code search tools on top of Stack Overflow may provide better results in terms of streamlining the various activities involved in example centric development (search, evaluation, and embedding).

Sketch of the architecture:

- A web interface that accepts query and returns the result
- A Crawler that crawls the web pages relevant to the query (On demand)
- A Page rank algorithm that ranks the stack overflow pages based on the relevancy and maximum score of the responses
- Additionally, an Indexer that indexes the official documentation of the programming language. (This can be used as a fail back response if there are no acceptable answers on stack overflow)


Issue query from the web interface
Issue crawl
Retrieve Topic lists and posts from Stack Overflow
Extract
metadata and code snippets from the posts
Rank
Metadata and code snippets
Filter
Remove duplicates and validate the response
Respond with the most relevant answer

This is a very broad overview of architecture of the system. We plan to maintain a repository of high quality content which will be ranked and synced with the stack overflow web content.

List of external software to be used:

Apache Lucene - for indexing the pages
Other tools and software to use - Yet to be decided

List of web resources to be used:

Stack Overflow API
Crawling the stack overflow website on demand
Upon issuing a query, crawl a programming language official documentation.