IR Project Proposal

Search engines for codes in c language

Arvind Deshraj S20160010007
Ajit Jadhav S20160010034
Junaid Nz S20160010036
Rohan S S20160010073

Abstract

In software development activities, source code examples are critical for understanding concepts, applying fixes, improving performance, and extending software functionalities. Previous studies have even revealed that more than 60% of developers search for source code every day. With the existence of super-repositories such as GitHub hosting millions of open source projects, there are opportunities to satisfy the search need of developers for resolving a large variety of programming issues. We propose to build a search engine for the same cause.

Features

Users may input the queries in natural language and relevant code snippets are returned. Currently planning to implement this for C language.

Datasets

1. GitHub repos
2. Stackoverflow

Flow

The user inputs a natural language query related to the code required. Based on the natural language query, we find a list of alternate queries which are similar to the input query.

Using the structured documents with ad-hoc XML tags from the Stackoverflow, we find out the code snippets related to the query and the list of alternate queries using the question which is in natural language and the code snippets related to the answer to that query.

Code samples will be indexed separately. We will implement k-gram index and normal positional inverted index and check their performances.

The IR concepts used:

1. Wildcard queries
2. K-gram index
3. Structured data retrieval