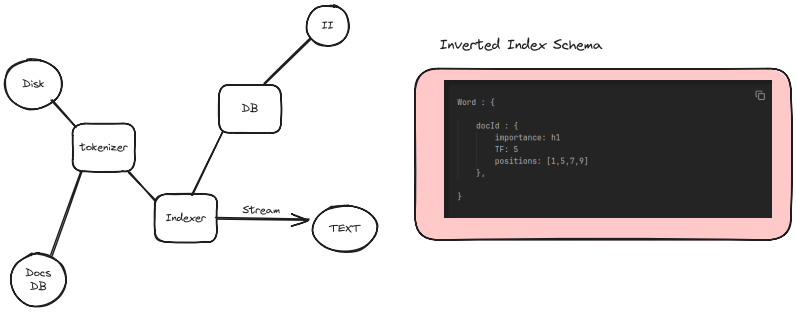
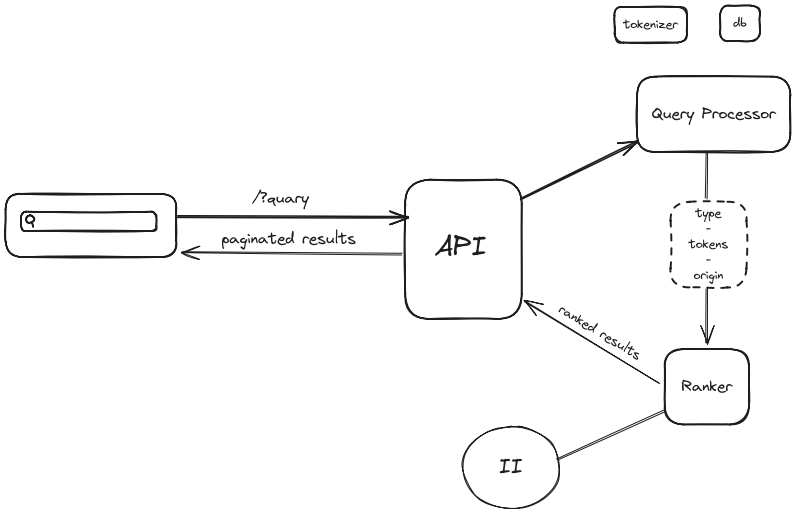
System Design

- Indexer



- Search Module



- Crawler

Components

- API
- Web Interface
- Ranker
- Indexer
- Crawler

Project Document

Search Engine Project.pdf

Crawler

Important Notes

> The crawler must maintain its state so that it can, if interrupted, be started again to crawl the documents on the list without revisiting documents that have been previously downloaded

Search Engine Project, page 1

> Provide a multithreaded crawler implementation where the user can control the number of threads before starting the crawle

Search Engine Project, page 1

> Use an appropriate data structure to determine the order of page visits.

Search Engine Project, page 1

Indexer

Notes

- Levels Of Importance: title - header - other

> words contained in each document and their importance (e.g., whether they are in the title, in a header or in plain text)

Search Engine Project, page 1

- Consideration

> consider how you will store your result by looking ahead on Ranker and Searching.

Search Engine Project, page 2

Query Processor

- It is just a module that the controller for the route will use

Notes:

- Two types of queries:
  - Normal
  - Phrases: with quotation marks

> Results obtained when searching for a sentence with quotation marks around them should be a subset of the results obtained when searching for the same sentence without the quotation marks.

Search Engine Project, page 2

- The horror

> the search query “travel” should match (with lower degree) the words “traveler”, “traveling” … etc

Search Engine Project, page 2

- It will use the tokenizer module

> performs necessary preprocessing

Search Engine Project, page 2

- It will use the DB-manager and Index Service

> searches the index for relevant documents.

Search Engine Project, page 2

Ranker

API

Web-interface

Unsolved Questions

- The Pagination Problem

If I searched for a word and query processor fetched 1 million results for pagination purposes we send 10 results. where should the rest go?