Team:

NetID: shchau2, Name: Chau Siu Hung (Captain)

Theme: Intelligent Browsing

Topic: Linked Page Search (Chrome Extension)

Description:

Index all pages that the current page links to and allow users to search over the collection of pages. For example, if we run the extension on `https://book.systemsapproach.org/` (table of contents of a textbook), then we can search over all chapter of the textbook, using a common retrieval function or other techniques (e.g. topic analysis).

Motivation:

Users would like to search over a online textbook or a blog website. Built-in tools are not always available, or are limited to exact keyword match. Google search with “site:domain.com” may be used. However, the linked pages may not be at the same domain (e.g. https://axisofordinary.substack.com/p/the-most-counterintuitive-facts-in has links to pages with different domain). Also, we might want to control the granularity of search (i.e. considering a sentence/paragraph as a document instead of the whole web page), or add some extra functions like topic analysis.

Data Sets:

Online textbook, example: Interactive SICP (xuanji.appspot.com), websites mentioned above.

Algorithms/techniques:

Web Scraping, Retrival function (BM25)

Evaluation:

Relevance judged by human, compare with Google search with “site:domain.com” option on some simple data sets (all linked pages within same domain).

Language:

JavaScript, Python

Workload:

Chrome Extension setup/UI (5-6h),

Web Scraping/Data Collection (3-4h)

Indexing (3-4h)

Retrievel Function (1-2h)

Extra function (e.g. spliting web pages into paragraphs, topic analysis) (> 6h)