Sameer Reza

Dr. Leung

COMP 4321X

29/3/2023

Project Phase 1 Database Explanation

Schema:

* Page Table
  + Url
  + Title
  + lastModified
  + Size
  + childLinks (list of urls)
  + sortedKeywords (list)
  + sortedFrequencies (list)
  + parentUrl
  + Text

The page table is the table constructed on page reception by the spider. It contains the metadata and content needed for the page, on a per page basis. The pages are identified by their URL, as they should only be accessed by unique ones in future, when canonicalization is implemented fully (possibly up to comparing page contents to existing urls before usage, if the issue remains). The child and parent link information is to be used to connect document relations in future, and sorted keywords with their corresponding frequencies are in place to make it easier to display for end users.

The database also stores an inverted index, within such a structure:

* Per word:
  + Documents (a list of urls)
    - Positions (a list, as the value of the key:value pair with the url)

In this way, we can find, on a per word basis, the list of documents containing such, and the positions within such of words. The words themselves are used as the identifier of such, as they are normalized, stemmed, etc, before integration, and should thereby be unique, and I do not believe we will run into performance or storage issues opting to compare strings in such a small project. Future improvements include sorting the documents list in order of word frequency to aid searching within the list. I have also come to realize there may be issues with the usage of a dictionary as the overall storage structure of the inverted index, provided we index enough unique words and documents - to that end, this inverted index will likely be converted into a table within the database in future, even though this will not likely be an issue for the project, just for the sake of completeness.