Programming Assignment #1 - Page Rank

Methods Description Report B02901035 電機五 高紹芳

Data Structures:

We used the class “hash” in Ruby to read in all the data in the “web-search-file” folder, with the “key” being the filename of the page (format: page###, ### are numbers), and “values” being the content inside the files. Content in the files is split into two parts using Regular Expression, links and words, with links being the pagenames that it links to, and words being the ones after the hyphens. Hyphens are eliminated through RegExp.

https://docs.ruby-lang.org/en/2.0.0/Hash.html

Algorithms:

The algorithm follows the psuedo code given.

We define a function called PageRank. A hash named “pr” is initiated and returned to save page rank values with different stopping differences and d-values.

A hash named “word_hash” is initiated with a pair of keys: word and values: an array of pagenames. We sort both hashes by turning them into arrays, and write them into the designated format of file.

The search engine with a prompt and response function is also built according to instructions on prog1.pdf.

Performance Analysis:

The time complexity of randomly accessing a Ruby hash can be assumed as O(1) with the hash map implementation.

When building “pr”, we run over several stages of calculation to get an output that is close enough to be the final result. For each stage, we need to calculate the pagerank value for each page. To calculate the pagerank for a single page, we need to iterate over all the pages to get a subset of pages that links to the specific page (this implementation does not use the trick of building a reverse index for looking up all the pages that links to a specific page), and sums up their pagerank values, which has a time complexity of O(n). So the time complexity of doing a stage to produce a pr will be O(n^2).

On the other hand, building the reverse map will need to loop over each page, to build a map of words as the key, and arrays of page names as the value. The time complexity will be O(n).