Python pickle is used as database in this project, which is stored as \*.p file.

File name: url2pageID.p

Description:

It is an ordered dictionary that maps an url to an array. The array contain an unique page ID, its page title, last modified datetime, page size, word frequency list and its children links. Page tile, last modified datetime and page size are the meta data of the page. Word frequency list and children links are required in the “spider\_result.txt”. This data structure can facilitate getting the information by an url.

All other databases that are related to an url will be referenced by its page ID.

File name: forwardIndex.p

Description:

It is an ordered dictionary that maps a page ID to an array of word tokens of a page.

File name: word2wordID.p

Description:

It is an ordered dictionary that maps a word to a word ID. They are stored in key-value pair and both of them are unique.

All other databases that are related to a word will be referenced by its word ID.

File name: invertedIndex.p

Description:

It is an ordered dictionary that maps a word ID to an array. The array includes the page ID that contains the word and the word frequency.

It will be used later in the search engine.