Assignment #3: Additional Information

Understanding the data dump:

In Assignment-2, we crawled webpages belonging to the UCI ICS domain. We collected these webpages and their URLs and are providing them to you as ‘webpages.zip’ file. This zip file contains the following:

1. bookkeeping.json
2. bookkeeping.tsv
3. Folders 0 to 74

Folders:

The webpages are organized into 75 folders. Every file has the extracted HTML source code of a particular URL.

Bookeeping files:

bookkeeping.json and bookkeeping.tsv are two different formats of the same file. These files maintain a list of all the URLs that have been crawled. Every URL has an identifier associated with it. This identifier helps locate the HTML code of the URL. The identifier is of the format: “folder_number/file_number”

For example, consider the entry on line 13 of bookkeeping.json:

```
"0/108": "vision.ics.uci.edu/papers/RamananBK_ICCV_2007"
```

This means that the HTML code extracted for the link "vision.ics.uci.edu/papers/Ramanan BK_ICCV_2007" is located at folder 0, file number 108.

Understanding the content of the files:

To extract the content from the HTML tags, you will be using an HTML parser. There are many libraries available to achieve this task and we encourage you to compare the available options before selecting a library to perform HTML parsing for you (Suggestions: Beautifulsoup, HTMLParser)

Use of libraries:

It is strictly not allowed to use libraries that perform the entire task of index creation or ranking for you. Hence, libraries such as Lucene or Elastic Search are not allowed.

You may use libraries that help you achieve specific tasks. For example, you can use a tokenizer such as NLTK to tokenize your content.