INF 141/CS 121
Information Retrieval

Assignment 4 Project 3

The HTML files

- A Zip file that contains crawl-able HTML files which you may parse/process for extracting tokens.

- The HTML files have been organized and stored in numbered directories. The file names are numbers as well.

- The bookkeeping.json and bookkeeping.tsv files represent the index of all the HTML files.

- The key value of the json file is essentially the relative file path of the HTML content. The value is the web URL of the HTML content.

- Do not confuse bookkeeping with the inverted index. It simply provides you a means to access the crawl-able HTMLs programmatically. The key values in bookkeeping can also be used to uniquely identify the files. This will be useful when you need to retrieve the web page and the content while displaying your search engine results.

Building the inverted index

- Now that you have been provided the HTML files to index. You may build your inverted index off of them.
- As most of you may already know, the inverted index is simply a map with the token as a key and a list of its corresponding postings.
- A posting is nothing but the representation of the token’s occurrence in a document.
- The posting would typically (not limited to) contain the following info (you are encouraged to think of other attributes that you could add to the index) :
    - The document name/id the token was found in.
    - The word frequency.
    - Indices of occurrence within the document
    - Tf-idf score etc

Inverted Index

- When designing your inverted index, you will think about the structure of your posting first.

- You would normally begin by implementing the code to calculate/fetch the elements which will constitute your posting.

- Modularize. For eg:- If you’re using python, use scripts that will perform a function or a set of closely related functions.This helps in keeping track of your progress, debugging, and also dividing work amongst teammates if you’re in a group.

Inverted Index

- You are free to choose any database system to store your inverted index.

- Some possible options - Redis, MongoDB, memcached, MySQL etc.

- Pro-tip : If you have a hard time choosing between the database systems. Read about their performance and learning curves of the libraries available with the language of your choice.

Search and Retrieve

- Once you have built the inverted index, you are ready to test document retrieval with queries.

- At the very least, the documents retrieved should be returned based on tf-idf scoring. This can be done using the cosine similarity method. Feel free to use a library to compute cosine similarity once you have the term frequencies and inverse document frequencies.

- You may add other weighting/scoring mechanisms to help refine the search results.

Tips

- Start as early as possible (esp. if you have slow computers/laptops).

- When asking questions on Piazza, please be as specific as possible. We cannot help you with very broad/open-ended questions as this assignment is pretty free-form.

- Read the other assignment documents carefully. If there are mistakes/ambiguities, please bring it to our notice.

- Google and Debugging are your best friends!