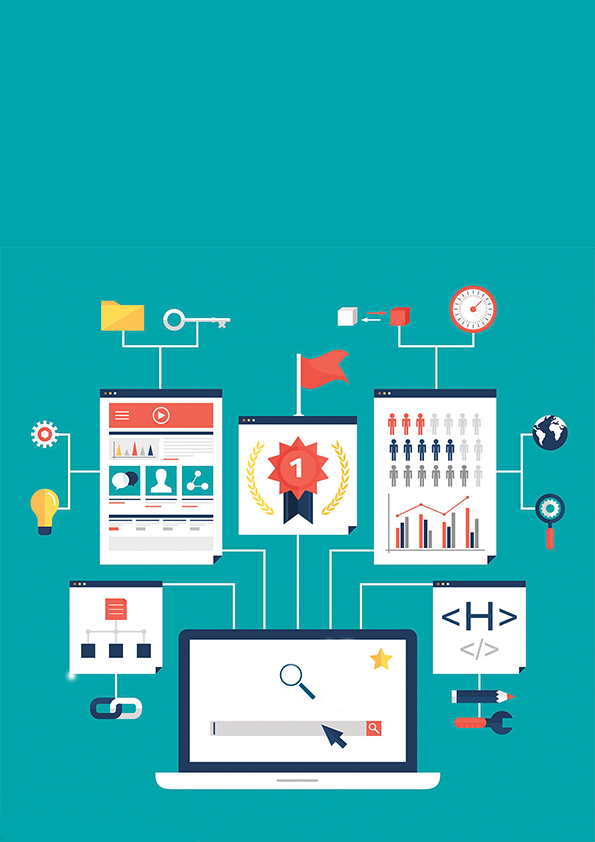
****

CS3343-

**Introduction**

Search Engine is a software system designed to search for information on the World Wide Web. In the century of information, people’s demand for information acquisition is getting higher and higher. Various search engines have emerged in recent years, based on the principle it seems extremely available to build a small size search engine. There is no need to connect the Internet all the time, and the dataset designed with requirements is more inclined to users’ own orientations.

The process of building search engine software can be roughly divided into two parts, preprocessing and querying. In the part of preprocessing, firstly obtaining an original webpages dataset in a certain formal saved locally. However, the pages information is not convenient to access randomly, so recording the index of webpage in the dataset by extracting keywords of the page. In order to inquire easier, reversing the index to obtain a hash map from keywords to URLs. Furthermore, the result of the obtained URLs will rank in a logical order, the webpage tends be more in line with users’ requirements is ranked in front based on TF-TDF.

Considering the input given by users freely, tokenizer is also included in the software.

Commonly saying, user inputs one or several words, a list of corresponding webpages’ URLs is returned. Then user can obtain the information from the obtained URLs.

**Design Constraints**

There are a few constraints we met while designing the program, a large part of the problems has been solved.

The first constraint is how to obtain the URLs of the download webpages. We cannot guarantee all html files of webpages containing their own URLs, because web front-end development engineers have their own styles. To solve this problem, we name the webpages files in the fixed format of URLs. And obtaining it by means of reading files in the designated folder.

The second constraint is the keywords extraction. Through extracting the fixed content part in html files and using regular expression, as well as some other processing, such as filtering the unnecessary modifier words.

Finally, the keywords can be roughly obtained to represent the webpage.

The third constraint is about the searched result. There seems blurred to obtain an unorder list of URLs.

Using TF-IDF to solve the problem and the results will be sorted from big to small in terms of relevance.

**Use Cases and Requirements Specifications**

This Search Engine Software tends to design for web engineer to implement site search and the users have demand for find information based on their own dataset.

Firstly, the program will ask users whether to add new webpages or not, if yes, input the webpage information of URL.

Then, users will be asked to input keywords, the number of keywords is unlimited, also there is no case sensitive.

Finally, the relevant URLs will be output in a correlation priority order. Users can look up the webpage from the given list to obtain the requested information.

A screenshot of a computer

Description automatically generated