COMP1011 Presentation Script

Group 8

May 2021

1 Introduction

Hello, we are group 8, and our project is implementing a searching engine for subjects offered by the department of computing. We, students, are eager to know the subjects that we can study, and a searching engine is very useful to us because we can find subjects that we are interested in and are eligible to study, which can save much time. To solve this problem and improve the studying experience of students, as well as practicing our programming skills, we develop this program not only to benefit university students but also to get self-improved.

2 Usage

Firstly let me introduce how to use this program. It is very easy to use, anyone who knows how to use google can use it. You just need to type keywords, and our program will automatically find related subjects to you. Then, you can choose to view the details and download the subject description form in PDF format. Meanwhile, you can know the relation index of subjects, which can help you understand how relative they are, and helping you make decisions.

3 Algorithm

Then let me introduce the algorithm of this project, in the other words, how can we calculate these relative indexes. We refer to a powerful and famous search engine in the word —— Elastic Search. We learned its algorithm and simplify it for small data amount, which can be understood easily. We understand that sometimes you know the subject code and you just want to find more information about the subject, so you can directly type the subject code for the exact match. If you do not know the subject code, you are free to type any words related to the subject.

Our program will first transform your input to a unique word ID by the red-black tree, which is very efficient and fast. Then it will search which subject contains this word ID. After that, it will find how many times the word has

appeared in each subject title and content separately, calculating relative index and display the subjects from the most related to the least related. Since most of the searching process uses a red-black tree, the searching algorithm is quite effective.

4 Data Structure

We have created two headers that can help us use object-oriented programming skills. the two headers contain two classes, the first class is a dictionary, its instance includes the word name, id, the document id which the word appeared in the title, and the document id which the word appeared in content. By using this dictionary, we can know the word's id and it can shorten the search time afterward. Meanwhile, by knowing where the document appeared, there is no need for us to search the document one by one, we only need to search the documents that the word appeared, which can save a lot of time.

The second class named course is used to store the information related to the subjects, it includes objects such as subject name, subject id, subject level, subject pre-requisite, subject title, and maps that store how many times the word appeared in the document. This can be useful to calculate the relative score of documents. Another thing we'd like to mention is that in course class, the data objects are set to private because it is beneficial for data security, and we can only access the data by calling public functions.

5 Implementation

We divided our work into some parts for distribution, which including data pre-process, data structure design, user input process, relative index calculation, and result display. Data download and data cleaning is done by Python because C++ does not have a web crawler function, and Python can finish this task conveniently. The data is in pdf format at the beginning, and we transform it to plain text, exact the important information, and saved all data in JSON files, which can immigrate to other platforms quickly, and it is also friendly to future development.

6 Conclusion

That's the end of our presentation, we use a lot of knowledge learned and we also learn more advanced skills and headers in C++. At the same time, we gain much learning experience and we are fully convinced that this will be helpful in our future development.