Project 0.0

資工四 404410076 陳鐸元

Data Preprocessing

- Inputs(1GB): `ettoday0.rec`, `ettoday1.rec`, `ettoday2.rec`, `ettoday3.rec`, `ettoday4.rec`, `ettoday5.rec`
- Sorting: in-memory
- Output(890MB): `opt`
- Code: `hw0.c` (88 lines)
- Compilation: `gcc -Wall -Wextra -std=c11 -O2 -o hw0 hw0.c`

The input should be decoded with UTF-8. Thus, I use `setlocale` to set the encoding and `wchar_t` as the data type. I read one line at a time. If the line didn't start with `@`, I'll start parsing it. The delimiters are `，。！？\n`. Preceding characters which wasn't in between `0x4E00` and `0x9FBB` was discarded. This made each of our results don't start with an English character. And then, I saved the parsed lines into a dynamic string array(`wchar_t **`). At last, I sorted the array with `qsort` and write the result to `opt`.

Web Application

- URL: https://www2.cs.ccu.edu.tw/~cty104u/search.php
- Results per page: 100
- Code: `search.php` (103 lines), `yee.css` (42 lines), `out` (empty, everyone can write)

The main application was written inside `search.php`. All queries are sent with a `GET` request. If the query is empty, it will read and show all the data from `opt`; if not, it will `grep` the query from `opt` and dump the result to `out`. The tricky thing is that if `out` doesn't exist or the write permission isn't set correctly, we won't be able to use it as a temporary file.

Search

Search bar is located on the top of the page.

Paging

The range needed to be shown on the current page is calculated in advanced. And then, it will use `sed` to get the results from `opt` / `out` in the range.

At the bottom of the page, there are 10 buttons linked to the first page, nearby pages, and the last page. You can also navigate to a page by submitting the number in the text box below.

Live Demo

https://youtu.be/Y_qfqJ7pkYU

Code

https://github.com/Superdanby/Search-Engine/tree/master/Project%201