Final Project

資工四 404410076 陳鐸元

Web Application

- Search rules:
    - - MUST_CONTAIN
    - - MUST_NOT_CONTAIN
    - REGULAR (error = 0 for length < 5, error = 1 for 5 <= length < 9, else error = 2)
    - e.g. + 雞肉飯 + 台灣
- Keyword highlight
- Page navigation
- URL: https://www2.cs.ccu.edu.tw/~cty104u/final/search.php

Structrue

1. Record split(multi-thread)
2. Calculate scores(multi-thread)
    1. If error = 0, use Boyer Moore(Bad Character + Good Suffix)
    2. Else, use Levenshtein Distance
3. Sort
4. Write to file
5. Get contents for specified page
6. Apply highlights

Code

Record Split

```
int uidx = unparsed.find(args.url, idx) + args.url.size();
int tidx = unparsed.find(args.title, uidx);
int cidx = unparsed.find(args.content, tidx + args.title.size());
inp.url = unparsed.substr(uidx, tidx - uidx);
tidx += args.title.size();
inp.title = unparsed.substr(tidx, cidx - tidx);
cidx += args.content.size() + 1;
int endidx = unparsed.find(args.delimiter, cidx + args.content.size());
inp.content = unparsed.substr(cidx, endidx - cidx);
idx = endidx + 1;
```

Boyer Moore

```
unordered_map<char, int> occurred;
for (unsigned int i = 0; i < A.size(); i++)
{
        occurred[A[i]] = i;
}

// bmPreprocess1();
vector<int> border_front(A.size() + 1);
vector<int> shift(A.size() + 1);
size_t i = A.size(), j = A.size() + 1;
border_front[i] = j;
while (i > 0)
{
        while (j <= A.size() && A[i-1] != A[j-1])
        {
                if (shift[j] == 0)
                        shift[j] = j - i;
                j = border_front[j];
        }
        border_front[--i] = --j;
}

// bmPreprocess2();
j = border_front[0];
for(i = 0; i <= A.size(); i++)
{
        if(shift[i] == 0)
                shift[i] = j;
        if(i == j)
                j = border_front[j];
}

// Boyer Moore Search
ll cnt = 0;
int Aidx = 0, Bidx;
while (Aidx <= int(B.size()) - int(A.size()))
{
        Bidx = A.size() - 1;
        while(Bidx >= 0 && A[Bidx] == B[Aidx + Bidx])
                --Bidx;
        if (Bidx < 0)
        {
                ++cnt;
                Aidx += shift[0];
        }
        else
                Aidx += max((shift[Bidx + 1]), Bidx - (occurred.find(B[Aidx + Bidx]) != occurred.end() ? occurred[B[Aidx + Bidx]] : -1));
}
return cnt;
```

Levenshtein Distance

```
vector<int> table(A.size() + 1);

for(size_t i = 0; i <= A.size(); i++)
        table[i] = i;

double score = 0;

int ubound = min((int)table.size(), tolerance + 2);
for(size_t j = 0; j < B.size(); j++)
{
        for(int i = 1; i < ubound; i++)
        {
                int prev = table[i];
                table[i] = min(table[0] + (A[i - 1] != B[j]), min(table[i] + 1,
table[i - 1] + 1));
                table[0] = prev;
        }
        table[0] = 0;
        while(table[--ubound] > tolerance);
        ubound = min((int)table.size(), ubound + 2);
        score += table.back() <= tolerance ? double(tolerance - table.back() + 1) /
double(table.back() + 1) : 0;
}
return score;
```

Sort

```
sort(all.begin(), all.end(), [](const Record &A, const Record &B){return A.score >
B.score;});
```

Keyword highlighting

Regex Preparation

```
$plus_stripped = preg_replace('/[+-][\s]+/', '', $trimmed);
$query_pattern = str_replace(' ', '|', $plus_stripped);
```

Substitution

```
echo preg_replace('/(' . rtrim($query_pattern) . ')/ui', '<span
class="highlight">$1</span>', $entry);
```

Paging

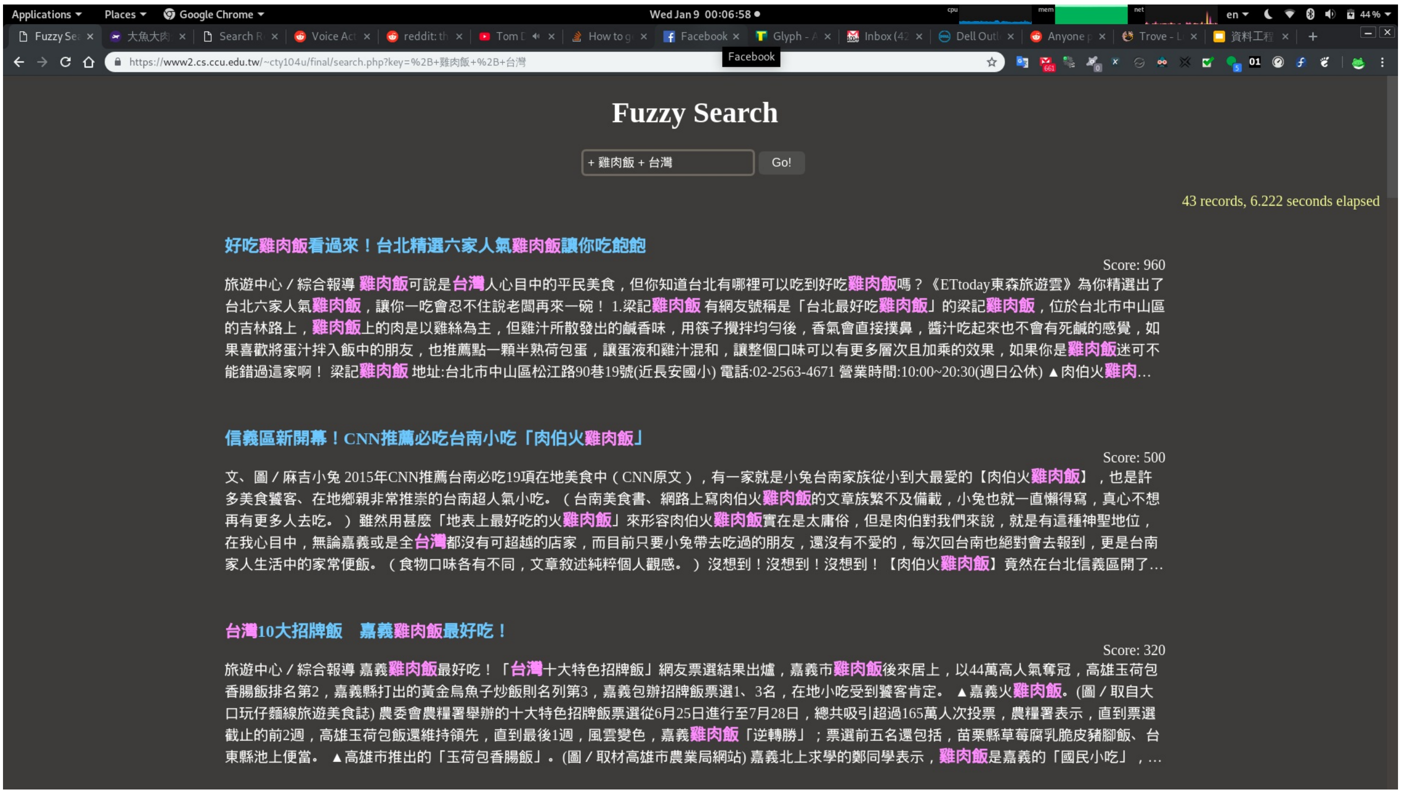
The range needed to be shown on the current page is calculated in advanced. And then, it will use `sed` to get the results from `opt` / `out` in the range.

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.

Code

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

Screenshots


Applications
Places
Google Chrome
Wed Jan 9 00:06:58
https://www2.cs.ccu.edu.tw/~cty104u/final/search.php?key=%2B+雞肉飯+%2B+台灣
Facebook
+ 雞肉飯 + 台灣
Go!

Fuzzy Search

43 records, 6.222 seconds elapsed

好吃雞肉飯看過來！台北精選六家人氣雞肉飯讓你吃飽飽

Score: 960

旅遊中心／綜合報導 **雞肉飯**可說是**台灣**人心目中的平民美食，但你知道台北有哪裡可以吃到好吃**雞肉飯**嗎？《ETtoday東森旅遊雲》為你精選出了台北六家人氣**雞肉飯**，讓你一吃會忍不住說老闆再來一碗！ 1.梁記**雞肉飯** 有網友號稱是「台北最好吃**雞肉飯**」的梁記**雞肉飯**，位於台北市中山區的吉林路上，**雞肉飯**上的肉是以雞絲為主，但雞汁所散發出的鹹香味，用筷子攪拌均勻後，香氣會直接撲鼻，醬汁吃起來也不會有死鹹的感覺，如果喜歡將蛋汁拌入飯中的朋友，也推薦點一顆半熟荷包蛋，讓蛋液和雞汁混和，讓整個口味可以有更多層次且加乘的效果，如果你是**雞肉飯**迷可不能錯過這家啊！ 梁記**雞肉飯** 地址:台北市中山區松江路90巷19號(近長安國小) 電話:02-2563-4671 營業時間:10:00~20:30(週日公休) ▲肉伯火**雞肉**…

信義區新開幕！CNN推薦必吃台南小吃「肉伯火雞肉飯」

Score: 500

文、圖／麻吉小兔 2015年CNN推薦台南必吃19項在地美食中（CNN原文），有一家就是小兔台南家族從小到大最愛的【肉伯火**雞肉飯**】，也是許多美食饕客、在地鄉親非常推崇的台南超人氣小吃。（台南美食書、網路上寫肉伯火**雞肉飯**的文章族繁不及備載，小兔也就一直懶得寫，真心不想再有更多人去吃。）雖然用甚麼「地表上最好吃的火**雞肉飯**」來形容肉伯火**雞肉飯**實在是太庸俗，但是肉伯對我們來說，就是有這種神聖地位，在我心目中，無論嘉義或是全**台灣**都沒有可超越的店家，而目前只要小兔帶去吃過的朋友，還沒有不愛的，每次回台南也絕對會去報到，更是台南家人生活中的家常便飯。（食物口味各有不同，文章敘述純粹個人觀感。）沒想到！沒想到！沒想到！【肉伯火**雞肉飯**】竟然在台北信義區開了…

台灣10大招牌飯　嘉義雞肉飯最好吃！

Score: 320

旅遊中心／綜合報導 嘉義**雞肉飯**最好吃！「**台灣**十大特色招牌飯」網友票選結果出爐，嘉義市**雞肉飯**後來居上，以44萬高人氣奪冠，高雄玉荷包香腸飯排名第2，嘉義縣打出的黃金烏魚子炒飯則名列第3，嘉義包辦招牌飯票選1、3名，在地小吃受到饕客肯定。 ▲嘉義火**雞肉飯**。(圖／取自大口玩仔麵線旅遊美食誌) 農委會農糧署舉辦的十大特色招牌飯票選從6月25日進行至7月28日，總共吸引超過165萬人次投票，農糧署表示，直到票選截止的前2週，高雄玉荷包飯還維持領先，直到最後1週，風雲變色，嘉義**雞肉飯**「逆轉勝」；票選前五名還包括，苗栗縣草莓腐乳脆皮豬腳飯、台東縣池上便當。 ▲高雄市推出的「玉荷包香腸飯」。(圖／取材高雄市農業局網站) 嘉義北上求學的鄭同學表示，**雞肉飯**是嘉義的「國民小吃」，…