# 文件搜索系統最終報告

## 系統特色

* 1. 做了比較友好的gui
  2. 打開文件夾后可以清楚的看到文件夾內所有文本的信息, 包括文件名稱, 文件路徑, 單詞總數, 索引大小, 建立索引的時間等信息
  3. 搜索關鍵詞系統會幫助您確定您的輸入是否符合搜索語法, 如不符合會在右邊的提示框中顯示語法錯誤
  4. 文件內容顯示框及索引顯示框會高亮顯示您輸入的關鍵詞

## 簡單用戶手冊

* 1. 首先雙擊FSE\_tset.exe文件執行文件搜索工具
  2. 點擊右上方”選擇文件夾”按鈕, 選定您所想要搜索的文件夾, 并按右下角的選擇文件夾按鈕進行確定(1)
  3. 在搜索關鍵詞中輸入搜索語句
     1. 所有的關鍵詞都必須用雙引號括起來, 如您要搜索it單詞的話, 請在搜索框中輸入 “it”
     2. 本系統接受多個關鍵詞搜索. 如果您想搜索同時存在it單詞及is單詞的文件, 請輸入 “it” AND “is” . 如果您想搜索出存在it的單詞的文件或者存在is單詞的文件, 請輸入 “it” OR “is” . 當然, 3個以上的關鍵詞也是可以進行搜索的, 但請注意, 在沒有添加括號的情況下, 我們默認邏輯運算是從左到右進行的.
     3. 若您想搜索所有不帶it單詞的文件, 請在搜索框中輸入 NOT “it”
     4. 若您想在搜索框中搜索it的個數為20個的文件, 請在搜索框中輸入 COUNT “it” = 20 . 同理, 您也可以以搜索出現次數小於20的文件, 您只需要把等於號改成小於號.
     5. NOT和COUNT關鍵詞都是可以與AND OR混合在一起使用的. 但我強烈建議您使用括號確定優先級確保搜索結果是您想要的
  4. 如果您的搜索語句不符合c中的語法規則, 那麼系統將會在左邊的提示框中顯示”語法錯誤”. 一旦系統檢測到輸入的搜索語句是有效的時候, 系統將會立刻進行檢索並將結果顯示在下面的顯示框中.
  5. 顯示框分為3部分. 最左邊的部分顯示的是本文件夾中有多少個文件時符合您的搜索語句的. 從列表中點擊您所想要查看的文件, 文件內容及索引內容將會在右邊的顯示框中顯示出來. 並且, 系統會將關鍵詞顯示為高亮.

1. 因為系統會掃描所有的文件夾內的文件, 而並不能分辨出其是不是有效文件, 所以請在進行搜索前對帶搜索的文件進行預處理(將他們拖到同一個文件夾中)

## 寫在後面

不是bug的系統缺陷

大概是今年做過最大最有趣的程序了吧. 雖然還有很多的地方自己都不是很滿意, 例如, 符號其實只能搜索到逗號, 句號, 單引號, 加號, 井號, 及減號. 前面的兩個符號我個人認為意義還是很大的, 可以用來判斷出這篇文章中到底有多少句話, 間接看出這篇文章的大小, 其他的符號實際上就意義不大了. 因此僅僅只添加的幾個測試文本中單詞包括了這些符號的符號(如c++, c#, k-12, students’)

關於符號的添加也是做的很差的一件事. 因為每次添加每個符號都需要去修改內部那個非常非常複雜的正則表達式, 感覺一點都不舒服. 所以原本想做成(如果用戶認為他們需要這個符號的搜索功能, 那麼便自己添加進去就好了呢)這樣的效果. 然而做成這樣其實還不如直接做成全符號搜索更好, 自己又還沒有那樣的具體想法去做, 因此一直沒能做出來

另一個比較失望的就是不能自動判斷文件是否是可搜索的文本文件. 原本是打算做成檢測txt文件作為搜索, 但後來自己用到了一些測試數據中有cpp文件, 有html文件, 有java文件, 而且測試提供的文件時單純的文件, 不是所謂的txt文件, 因此又把這個東西取消了. 感覺還是很糾結的, 因為不知道怎麼樣去做才比較好.

還有一個便是製表符的問題. 在QT內部顯示出製表符會亂碼, 這個實在是沒有辦法解決的事情. 目前想到的思路就是在測試文本讀入的時候把所有的製表符全部換成4個空格鍵, 不過因為最近身體太差, 加上提供的測試文本並沒有涉及到製表符, 因此一直沒有做出來.

這東西到底能做什麼

實際上一開始我就在想這個問題, 這個東西究竟能做什麼. 最終我把它定位為, 用於分析語言習慣和用詞概率. 還是有一定的實際意義的, 想想看把某個人所寫過的東西全部拖到文件夾中, 很快就能看出他洗一些語言習慣, 喜歡用什麼樣的詞之類的. 這樣的話這個東西要拓展也就簡單很多了, 如果要繼續做下去的話, 大概是將分析所得的書籍進行圖表化使他更加直觀吧.

然而我接到這個題目的時候, 一開會并不是這麼想的. 大家都用過chrome的f12抓包, 因為一些奇怪的個人興趣, 我非常喜歡分析抓到的網頁包, 有的時候為了弄清楚一個cookie的來龍去脈就需要非很大的功夫. Chrome抓下來的包如果是以文本形式保存下來, 那麼便是多個文本文件. 如果我能在這多個文本文件中搜索某個關鍵的詞(如cookie), 那麼便可以很清楚的知道它是什麼時候出現的. 這無疑是方便非常多的一個功能, 也是我一開始為什麼特別想時候py寫一個的原因(大概是因為py抓包很舒服).不過因為身體原因和其他的原因, 這個比較大的玩具看來只能拖慢一些進行了.

(PS.似乎自帶抓包和抓到的東西內的自動搜索是ie瀏覽器自帶的功能, 看來ie也算蠻良心的)