資料工程 HW0

404410030 資工四 鄭光宇

系統需求

要執行這支程式，系統必須具備：

- 支援 ANSI C 的 gcc
- make 指令

如何編譯

在專案目錄下輸入指令：

```
make
```

如何執行

同作業要求

```
rsort 檔名 [參數]
```

```
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort
Usage:
rsort filename [-d delimeter | -k field | -n numeric comparison | -r reverse sort | -c case insensitive]
peter@peter-desktop:~/Data-Engineering/HW1$
```

`-d` 是分隔符號，`-k` 是要作為 key 的 pattern

而 `-c` 是不區分大小寫、`-n` 使用數值排序、`-r` 倒序（降序）排序

實作部分

前置處理

先合併檔

```
ta$ cat ettoday0.rec ettoday1.rec ettoday2.rec ettoday3.rec ettoday4.rec ettoday5.rec > fullfile.rec
ta$
```

主函數部份

讀取執行參數、讀檔、斷行、排序、輸出

```c
int main(const int argc, const char **argv) {
    int i=0;
    const char *filename=NULL;
    char **records = NULL;
    int records_cnt = 0;
    if (argc<2) {
        fprintf(stderr, "Usage:\nrsort filename [-d delimeter | -k field | -n numeric comparison | -r reverse sort | -c case insensitive]\n");
        exit(2);
    }
    get_args(argc, argv, parameters, set_parameters);
    records_cnt = reader(argv[1], &records); /* need to free! */

    /* now sort */
    qsort((void*)records, records_cnt, sizeof(char*), comp);

    for (i=0; i<records_cnt; ++i) {
        fputs(records[i], stdout);
    }
    fputs("\n",stdout);

    free(records); records=NULL;
    return 0;
}
```

讀檔部份跟前一次作業類似，就不放上來了

讀取參數部份

```c
rsort.c
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <ctype.h>

const char *parameter_patterns[5] = {"-d", "-k", "-c", "-r", "-n"};
const char *default_args[2] = {
    "\n", "" /* -d, -k */
};
const char *parameters[2]; /* -d, -k */
int set_parameters[3]={0}; /* -c, -r, -n */

int parse_parameter(const int argc, const char** args, const char* pattern) {
    int i=0;
    for (i=0; i<argc; ++i) {
        if(strstr(args[i], pattern)!=NULL) return i;
    }
    return -1;
}

void get_args(const int argc, const char** args, const char **parameters, int *set_parameters) {
    /** parse all parameters in following order
     * -d record_delimiter
     * -k key_pat
     * -c case_insensitive
     * -r reverse order
     * -n numerical comparison
     */
    int i = 0, argspos=-1;
    for (i=0; i<2; ++i) {
        argspos = parse_parameter(argc, args, parameter_patterns[i]);
        parameters[i] = argspos<0?default_args[i]:args[argspos+1];
    }
    for (i=0; i<3; ++i) {
        set_parameters[i] = parse_parameter(argc, args, parameter_patterns[i+2])<0?0:1;
    }
}
```

簡單解析從 main function 傳入的 argv 裏面的參數

`-d`, `-k` 後方需要接一個字串，其他3個參數則不需要輸入內容

排序用的比較函數

```
39 int comp(const void *a, const void *b) {
40     int e=0, f=0;
41     const char *c = *(const char**)a;
42     const char *d = *(const char**)b;
43     int val = 0;
44     if (parameters[1][0]!='\0') { /* has field */
45         /* not robust enough. need to handle more exceptions */
46         c = strstr(c, parameters[1]); /* jump to that field */
47         d = strstr(d, parameters[1]); /* ,, */
48     }
49     if (set_parameters[2]) { /* numerical comparison? */
50         while(c!=NULL&&*c!='\0'&&!isdigit(*c)) ++c;
51         while(d!=NULL&&*d!='\0'&&!isdigit(*d)) ++d;
52         e = atoi(c);
53         f = atoi(d);
54         val = e-f;
55     } else { /* lexical order */
56         val = set_parameters[0]?strcasecmp(c,d):strcmp(c,d); /* case insensitive? */
57     }
58     return set_parameters[1]?-val:val; /* reverse order? */
59 }
```

同作業要求，實作了依照 `-k` 為key值排序、依照 `-n` 決定是否為數值排序、依照 `-c` 決定是否區分大小寫、依照 `-r` 決定是否降序排序。

實驗

使用上次作業的資料集的合併檔

照文章標題排序

```
peter@peter-desktop:~/Data-Engineering/HW1$ time ./rsort fullfile.rec -d "@GAISRec:" -k "@T:" > sort_by_title

real    0m19.200s
user    0m18.271s
sys     0m0.928s
peter@peter-desktop:~/Data-Engineering/HW1$
```

```
peter@peter-desktop:~/Data-Engineering/HW1$ grep "@T:" sort_by_title | head -n30
@TiTunes FESTIVAL   帶給你30個免費的音樂會夜晚!!!
@T曾銘宗：公公併不裁員，盼工會支持
@T:  3D解謎遊戲《Shadowmatic》趣味光影皮影戲
@T:  NBA／勞資雙方欲重啟談判  望合約和訴訟問題一併解決
@T:  《動新聞》口味？正妹童星洪岑成「AV優酪乳」代言人
@T:  「黑暗天空保護區」麥肯奇盆地  肉眼能看麥哲倫星雲
@T:  【寶靈老師】2012/2/02水瓶座運勢
@T:  【寶靈老師】2012/2/03雙子座運勢
@T:  【寶靈老師】2012/2/05金牛座運勢
@T:  【寶靈老師】2012/2/05雙子座運勢
@T:  台灣男最疼女友？跨國調查：戴套率全球第三
@T:  寶貝你最大！晶宴寶寶回娘家  愛心捐款Show活力
@T:  怪！沈玉琳喜搜「開運物」  林國基「芭比控」擁千隻
@T: 0425四大報頭版頭條  核四不玩了  馬政府讓步
@T: 0428四大報頭版頭條 馬宣布：核四全面停工
@T: 119專線您好！4歲男童：警察杯杯可以教我數學嗎？
@T: 12月31日新規定！「基改」食品強制標示  違者罰400萬
@T: 1417輛三菱汽車沉沒北海海域  5人死亡6人失蹤
@T: 18趴確定腰斬  「台銀年定存利率加7趴」
@T: 19歲女交換生赴日失蹤  家屬爆：學校要求勿聲張
@T: 200歲！  美釣獲最大尾「人瑞級」石斑
@T: 2012.03.12 NBA 賽前報導
@T: 2012/1/1射手座星座運勢
@T: 2012/1/1獅子座星座運勢
@T: 2012/1/2 雙子座星座運勢
@T: 2012/1/2 雙魚座星座運勢
@T: 2012/1/3 金牛座星座運勢
@T: 2012/1/3 雙子座星座運勢
@T: 2012/1/4 天蠍座星座運勢
@T: 2012/1/4 巨蟹座星座運勢
peter@peter-desktop:~/Data-Engineering/HW1$
```

花費時間約為18秒

照URL中的編號排序

升序

```
peter@peter-desktop:~/Data-Engineering/HW1$ time ./rsort fullfile.rec -d "@GAISRec:" -k "@U:" -n > sort_by_url_number

real    0m18.488s
user    0m17.800s
sys     0m0.688s
peter@peter-desktop:~/Data-Engineering/HW1$ grep "@U:" sort_by_url_number  | head -n 20
@U:http://travel.ettoday.net/article/1.htm
@U:http://travel.ettoday.net/article/3.htm
@U:http://travel.ettoday.net/article/7.htm
@U:http://travel.ettoday.net/article/8.htm
@U:http://travel.ettoday.net/article/9.htm
@U:http://travel.ettoday.net/article/10.htm
@U:http://travel.ettoday.net/article/11.htm
@U:http://travel.ettoday.net/article/12.htm
@U:http://travel.ettoday.net/article/13.htm
@U:http://travel.ettoday.net/article/14.htm
@U:http://travel.ettoday.net/article/15.htm
@U:http://travel.ettoday.net/article/16.htm
@U:http://travel.ettoday.net/article/17.htm
@U:http://travel.ettoday.net/article/18.htm
@U:http://travel.ettoday.net/article/19.htm
@U:http://travel.ettoday.net/article/20.htm
@U:http://travel.ettoday.net/article/21.htm
@U:http://travel.ettoday.net/article/22.htm
@U:http://travel.ettoday.net/article/23.htm
@U:http://travel.ettoday.net/article/24.htm
peter@peter-desktop:~/Data-Engineering/HW1$
```

降序

```
peter@peter-desktop:~/Data-Engineering/HW1$ time ./rsort fullfile.rec -d "@GAISRec:" -k "@U:" -n -r > sort_by_url_number_reversed

real    0m18.596s
user    0m17.823s
sys     0m0.772s
peter@peter-desktop:~/Data-Engineering/HW1$ grep "@U:" sort_by_url_number_reversed  | head -n 20
@U:http://travel.ettoday.net/article/584389.htm
@U:http://travel.ettoday.net/article/584388.htm
@U:http://travel.ettoday.net/article/584387.htm
@U:http://travel.ettoday.net/article/584381.htm
@U:http://travel.ettoday.net/article/584379.htm
@U:http://travel.ettoday.net/article/584378.htm
@U:http://travel.ettoday.net/article/584377.htm
@U:http://travel.ettoday.net/article/584375.htm
@U:http://travel.ettoday.net/article/584374.htm
@U:http://travel.ettoday.net/article/584372.htm
@U:http://travel.ettoday.net/article/584371.htm
@U:http://travel.ettoday.net/article/584369.htm
@U:http://travel.ettoday.net/article/584367.htm
@U:http://travel.ettoday.net/article/584365.htm
@U:http://travel.ettoday.net/article/584364.htm
@U:http://travel.ettoday.net/article/584363.htm
@U:http://travel.ettoday.net/article/584362.htm
@U:http://travel.ettoday.net/article/584360.htm
@U:http://travel.ettoday.net/article/584359.htm
@U:http://travel.ettoday.net/article/584357.htm
peter@peter-desktop:~/Data-Engineering/HW1$
```

其他實驗

數字排序

```
peter@peter-desktop:~/Data-Engineering/HW1$ cat 123
 222 111 333 6 7 3333 44 444 55
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort 123 -d " "
 111 222 333 3333 44 444 55 6 7
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort 123 -d " " -n
 6 7 44 55 111 222 333 444 3333
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort 123 -d " " -n -r
 3333 444 333 222 111 55 44 7 6
peter@peter-desktop:~/Data-Engineering/HW1$
```

大小寫字母排序

```
peter@peter-desktop:~/Data-Engineering/HW1$ cat abc
 a b c d e f g A B C D E F G
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort abc -d " "
 A B C D E F G a b c d e f g
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort abc -d " " -r
 g f e d c b a G F E D C B A
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort abc -d " " -c
 a A b B c C d D e E f F g G
peter@peter-desktop:~/Data-Engineering/HW1$ ./rsort abc -d " " -c -r
 g G f F e E d D c C b B a A
peter@peter-desktop:~/Data-Engineering/HW1$
```

總結

程式行為符合預期，如果是針對數字就必須使用數值排序。

不然預設就是字典序，這點可以從 `3333 44 444 55` 和 `7 44 55 111 ... 3333` 兩種排序看出來。 `-n` 這個選項是方便的功能。

GitHub

https://github.com/peter0749/Data-Engineering/tree/master/HW1