# Project 1

資工四 404410076 陳鐸元

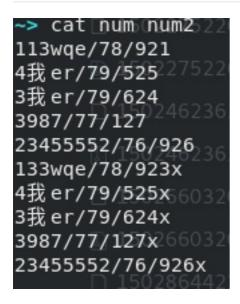
#### **Features**

```
-d: delimiter(defaults to '\n')
-sd: in-record delimiter(defaults to delimiter)
-k: sub-record selection key(regex)
-p: ignore preceding(regex)
-n: numerical sort
-i: ignore case
-s: sort by size
-m: memory usage(defaults to 4Gb)
-j: number of cpus(defaults to all cpus - 1)
-J: external sort cpu reduction mode(default: use 1 cpu only)
```

#### Structure

- Input: read at most 10Mb of data at a time (single thread)
  - After hitting memory limit, create a LWP to parse, sort and output (multi-thread)
- Internal sort: qsort
- External sort: merge sort with winner tree, O(log(k) \* cmp \* N)
  - -J = 3 : use 1 cpu
  - -J = 0 : reduce half of cpus each time
  - -J = 1: recude cpus by square root
  - -J = 2: reduce cpus to 1 after the first merge
- Compare function
  - locate key -> ignore preceding -> sort by size -> numericalize -> ignore case
  - Perl Compatible Regular Expression: key, preceding, numerical

### Example



```
0[danny@localhost HW1]$ 2018/10/23 08:31:45
  ./recordesort -sd / -k ^.2 -p ^. -n num num2
0
file: num
thread token: 0
Finished reading file: num
      necord_sort
file: num2
thread token:d-port_in
Finished reading file: num2
Finished internal sort
start: 0, reccnt: 2, target: 2, final: result
Bye~
      C record_sort_v1.c
#0[danny@localhost HW1]$ 2018/10/23 08:32:45
cat Cresultd_sort.c
113wqe/78/921
133wqe/78/923x md
3我er/79/624
3我 er/79/624x
4我 er/79/525+
4我 er/79/525x
23455552/76/926
23455552/76/926x
3987/77/127 C
3987/77/127x
```

## Compile

```
gcc -Wall -Wextra -std=c11 -pthread -o record_sort record_sort.c -lpcre2-8 -lm
```

#### Benchmark

### Computer Hardware

- CPU: i7 6700HQ (4C8T)
- Memory: 16GB (2x8GB)
- Storage: 512Gb M.2 PCIE Gen 3 x4 SSD (Plextor M8pegn)

#### **Test Data**

18 Gb Youtube data

```
-d $'\n'@$'\n' -s
```

./record\_sort 1501651417.rec -d \$'\n'@\$'\n' -s 1501937141.rec 1502275226.rec 1502462361.rec 1502660320.rec 1502864422.rec 1503161947.rec 1503740416.rec 1503863552.rec -J = 0

: 13m6.023s

```
-d $'\n'@$'\n' -sd $'\n' -k published -p @ -n
```

./record\_sort 1501651417.rec -d \$'\n'@\$'\n' -sd \$'\n' -k published -p @ -n 1501937141.rec 1502275226.rec 1502462361.rec 1502660320.rec 1502864422.rec 1503161947.rec 1503740416.rec 1503863552.rec -J = 0

: 21m21.992s

# Code(907 lines)

https://github.com/Superdanby/Search-Engine/blob/master/Project%202/record\_sort.c