

Manipulate Columns in Text Files

Saturday, 09 February 2008

Text files contain characters (ASCII characters) and some control characters. The control characters such as end-of-line, whitespace, and tab are used to organize data in the files. More info about text files can be found in Wikipedia page. This short article shows a quick and simple way to manipulate columns in a text file.

Suppose we have a file, called `equipment.lst`, with data organized in several columns. Each column is separated by a space. Assume also that the cardinality of the file is big.

```
1. Mouse 102 -29 45 -1
2. Harddisk 23 -34 21 0
3. LCD+Monitor 1234 -2 1 22 2
4. Laser+Printer 2222 -3 23 90 -2
5. Memori 23 -4 56 7
6. LCD+Projector 342 5 43 -1
:
:
900000. IBM+Thinkpad+Laptop 342 33 22 11
```

A quick and easy way to only take the name of the equipments, i.e. the second column, is to use `cut` command:

```
cut equipment.lst -f2 > equipment_name.lst
```

Another way is to use `awk` command:

```
cat equipment.lst | awk -F ' ' '{print $2}' > equipment_name.lst
```

Or, if we want to store only the unique name of the equipments, we can add `sort` command with pipelining:

```
cut equipment.lst -f2 | sort -u > equipment_unique_name.lst
```

But, how about if we want to store only the first 10000 or the last 5000 rows? Just use `head` or `tail` command:

```
head -10000 equipment.lst | awk -F ' ' '{print $2}' > equipment_name.lst
```

Or,

```
cut equipment.lst -f2 | head -10000 > equipment_unique_name.lst
```

Using `tail` command:

```
tail -5000 equipment.lst | awk -F ' ' '{print $1}' > equipment_name.lst
```

Or,

```
cut equipment.lst -f2 | tail -5000 > equipment_name.lst
```