

AWK Magic!

Tao Donglai

June 6, 2017

When we use something like `11`
or `ps`, what we get is usually
like:

When we use something like `ll` or `ps`, what we get is usually like:

```

l-ll
总用量 2.7G
drwxrwxr-x 7 thonastao thonastao 4.0K 3月 8 10:21 2016011322-陶东来
-rw-rw-r-- 1 thonastao thonastao 15K 2月 2 14:10 2016011322-陶东来.zip
drwxrwxr-x 6 thonastao thonastao 4.0K 4月 2 14:40 2016011322-陶东来
-rw-rw-r-- 1 thonastao thonastao 515 10月 11 2016 4.cpp
-rw-rw-r-- 1 thonastao thonastao 0 6月 28 2016 a.lisp
drwxrwxr-x 17 thonastao thonastao 4.0K 9月 24 2016 anaconda3
-rwxrwxr-x 1 thonastao thonastao 399M 3月 30 2016 Anaconda3-4.0.0-Linux-x86_64.sh
-rwxrwxr-x 1 thonastao thonastao 9.2K 10月 11 2016 a.out
drwxrwxr-x 7 thonastao thonastao 4.0K 1月 7 04:57 cdddd
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 3 04:14 class_talk
-rw-r--r-- 1 thonastao thonastao 445 9月 25 2016 dead.letter
drwxrwxrwx 8 thonastao thonastao 4.0K 12月 27 19:12 2016
-rw-r--r-- 1 thonastao thonastao 8.8K 6月 15 2016 examples.desktop
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 29 2016 fdu
-rwxrwxr-x 1 thonastao thonastao 7.3M 6月 29 2016 FTR_Launcher.jar
drwxrwxr-x 6 thonastao thonastao 4.0K 9月 25 2016 homework
-rw-rw-r-- 1 thonastao thonastao 284M 6月 29 2016 infTech2.rar
drwxr-xr-x 8 thonastao thonastao 4.0K 4月 1 2016 jdk1.8_0_22
-rw-r--r-- 1 thonastao thonastao 173M 6月 12 2016 jdk-Bu92-Linux-x64.tar.gz
drwxrwxr-x 2 thonastao thonastao 4.0K 11月 22 2016 mld_term
drwxr-xr-x 2 thonastao thonastao 4.0K 3月 21 08:45 pypy3-vs-7.0-linux64
-rw-rw-r-- 1 thonastao thonastao 23M 3月 30 22:54 pypy3-vs-7.0-linux64.tar.bz2
-rw-rw-r-- 1 thonastao thonastao 0 3月 16 10:09 setup.sh
-rw-r--r-- 1 thonastao thonastao 520K 6月 12 2016 shadowsocks-qt5-master.zip
drwxrwxr-x 3 thonastao thonastao 4.0K 6月 29 2016 Steam
drwxrwxr-x 29 thonastao thonastao 20K 4月 2 23:30 sup
drwxrwxr-x 2 thonastao thonastao 4.0K 10月 3 2016 Telegram
drwxrwxr-x 2 thonastao thonastao 4.0K 1月 15 22:29 test
-rw-rw-r-- 1 thonastao thonastao 421M 1月 25 01:50 tnp.bin
-rw-rw-r-- 1 thonastao thonastao 271 9月 16 2016 tnp.cpp
-rw-rw-r-- 1 thonastao thonastao 1.5G 2月 16 05:44 ubuntu-16.04.2-desktop-amd64.iso
drwxrwxr-x 3 thonastao thonastao 4.0K 6月 27 2016 VirtualBox Vms
drwxrwxr-x 2 thonastao thonastao 4.0K 9月 27 2016 vpn
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 16 2016 2016
drwxr-xr-x 16 thonastao thonastao 4.0K 1月 17 22:13 workspace
-rwxrwxr-x 1 thonastao thonastao 22K 12月 14 00:03 xxx
-rw-rw-r-- 1 thonastao thonastao 940 12月 14 00:04 xxx.cpp
drwxr-xr-x 2 thonastao thonastao 4.0K 6月 15 2016 公共
drwxr-xr-x 2 thonastao thonastao 4.0K 6月 15 2016 模板
drwxr-xr-x 2 thonastao thonastao 4.0K 6月 15 2016 视频
drwxr-xr-x 2 thonastao thonastao 4.0K 6月 15 2016 图片
drwxr-xr-x 2 thonastao thonastao 4.0K 6月 15 2016 文档
drwxr-xr-x 3 thonastao thonastao 12K 4月 30 03:37 下载
drwxr-xr-x 5 thonastao thonastao 4.0K 1月 7 04:19 音乐
drwxr-xr-x 2 thonastao thonastao 4.0K 10月 18 2016 桌面

```

Figure 1: After `ll`.

When we use something like `ll` or `ps`, what we get is usually like:

```

l- $ ll
总用量 2.7G
drwxrwxr-x 7 thonastao thonastao 4.0K 3月 8 10:21 2016011322-陶东来
-rw-rw-r-- 1 thonastao thonastao 15K 2月 2 14:10 2016011322-陶东来.zip
drwxrwxr-x 6 thonastao thonastao 4.0K 4月 2 14:40 2016011322-陶东来
-rw-rw-r-- 1 thonastao thonastao 515 10月 11 2016 4.cpp
-rw-rw-r-- 1 thonastao thonastao 0 6月 28 2016 a.lisp
drwxrwxr-x 17 thonastao thonastao 4.0K 9月 24 2016 anaconda3
-rwxrwxr-x 1 thonastao thonastao 399M 3月 30 2016 Anaconda3-4.0.0-Linux-x86_64.sh
-rwxrwxr-x 1 thonastao thonastao 9.2K 10月 11 2016 a.out
drwxrwxr-x 7 thonastao thonastao 4.0K 1月 7 04:57 cdddd
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 3 04:14 class_talk
-rw-rw-r-- 1 thonastao thonastao 445 9月 25 2016 dead.letter
drwxrwxrwx 8 thonastao thonastao 4.0K 12月 27 19:12 2016
-rw-rw-r-- 1 thonastao thonastao 8.8K 6月 15 2016 examples.desktop
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 29 2016 fdu
-rwxrwxr-x 1 thonastao thonastao 7.3M 6月 29 2016 FTR_Launcher.jar
drwxrwxr-x 6 thonastao thonastao 4.0K 9月 25 2016 homework
-rw-rw-r-- 1 thonastao thonastao 264M 6月 29 2016 infTech2.rar
drwxrwxr-x 8 thonastao thonastao 4.0K 4月 1 2016 jdk1.8_0_22
-rw-rw-r-- 1 thonastao thonastao 173M 6月 12 2016 jdk-Bu92-Linux-x64.tar.gz
drwxrwxr-x 2 thonastao thonastao 4.0K 11月 22 2016 mld_term
drwxrwxr-x 2 thonastao thonastao 4.0K 3月 21 08:45 pypy3-vs-7.0-linux64
-rw-rw-r-- 1 thonastao thonastao 23M 3月 30 22:54 pypy3-vs-7.0-linux64.tar.bz2
-rw-rw-r-- 1 thonastao thonastao 0 3月 16 10:09 setup.sh
-rw-rw-r-- 1 thonastao thonastao 520K 6月 12 2016 shadowsocks-qt5-master.zip
drwxrwxr-x 3 thonastao thonastao 4.0K 6月 29 2016 Steen
drwxrwxr-x 29 thonastao thonastao 20K 4月 2 23:30 sup
drwxrwxr-x 2 thonastao thonastao 4.0K 10月 3 2016 Telegram
drwxrwxr-x 2 thonastao thonastao 4.0K 1月 15 22:29 test
-rw-rw-r-- 1 thonastao thonastao 421M 3月 25 01:50 tnp.bin
-rw-rw-r-- 1 thonastao thonastao 271 9月 16 2016 tnp.cpp
-rw-rw-r-- 1 thonastao thonastao 1.5G 2月 16 05:44 ubuntu-16.04.2-desktop-amd64.iso
drwxrwxr-x 3 thonastao thonastao 4.0K 6月 27 2016 VirtualBox Vms
drwxrwxr-x 2 thonastao thonastao 4.0K 9月 27 2016 vpn
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 16 2016 2016
drwxrwxr-x 16 thonastao thonastao 4.0K 1月 17 22:13 workspace
-rwxrwxr-x 1 thonastao thonastao 22K 12月 14 00:03 xxx
-rw-rw-r-- 1 thonastao thonastao 940 12月 14 00:04 xxx.cpp
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 15 2016 公共的
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 15 2016 模板
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 15 2016 视频
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 15 2016 图片
drwxrwxr-x 2 thonastao thonastao 4.0K 6月 15 2016 文档
drwxrwxr-x 3 thonastao thonastao 12K 4月 30 03:37 下载
drwxrwxr-x 5 thonastao thonastao 4.0K 1月 7 04:19 音乐
drwxrwxr-x 2 thonastao thonastao 4.0K 10月 18 2016 桌面

```

Figure 1: After `ll`.

or something even longer.

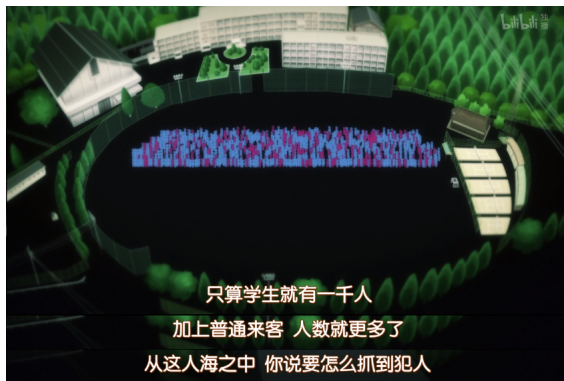


Figure 2: From a bangumi *Hyouka*.

Two examples.

Two examples.

- 1 Count the lines of a file (or several files).

Two examples.

- 1 Count the lines of a file (or several files).
- 2 Find out all the records can be accepted by a certain regexp(e.g. 3rd field - /l..e/, 5th field - /..u/).

Two examples.

- 1 Count the lines of a file (or several files).
- 2 Find out all the records can be accepted by a certain regexp(e.g. 3rd field - /l..e/, 5th field - /..u/).

Solutions will be given out later. (in awk!)

Key concepts.

Record & field.

Key concepts.

Record & field.

- A file is a sequence of records, separated by `RS(\n` defaultly).

Key concepts.

Record & field.

- A file is a sequence of records, seperated by `RS(\n` defaultly).
- A record is a sequence of fields, seperated by `FS(' '` defaultly).

Key concepts.

Record & field.

- A file is a sequence of records, seperated by `RS(\n` defaultly).
- A record is a sequence of fields, seperated by `FS(' '` defaultly).

AWK is a “record-oriented” programming language!

Language stucture

Pattern-action pairs.

PROGRAM ::= ϵ | PA-PAIR PROGRAM

PA-PAIR ::= CONDITION ACTION

CONDITION ::= BEGIN | END | ϵ

 | range-expression

 | condition-expression-set

ACTION ::= { commands }

condition-expression ::= VARIABLE TILDE-OP

/regexp/

 | simple-compare-expression

TIDLE-OP : ~ ! !~

TIDLE-OP : ~ ! !~

‘Hello world’ in awk:

```
awk 'END {print "hello , world !"}'
```

Variables and syntax

VARIABLE: [A-Za-z_][A-Za-z0-9_]*

e.g.

```
{  
  a=3; # a is a variable  
  print $a; # output the a-th field of the current record  
           # $0 == the whole record  
           # a > NF, then $a is an empty string  
  b="13:123:1"; split(b,array,/:/); # b is an array  
  c["wtf"]="yes_i_can!" # actually c is more a DICT than an ARRAY  
}
```

Built-in variables

NF: Number of fields.

NR: Number of records.

FS: Field separator.

RS: Record separator.

OFS: FS for output.

ORS: RS for output.

FILENAME: Filename of current input file.

see <https://www.tutorialspoint.com/awk/> for more
about built-in functions & variables.

User defined functions

This is a bit away from our everything-in-a-line principle.
so see <https://www.tutorialspoint.com/awk/>.

Solution

```
1 cat *.txt | awk 'END {print NR}'  
2 awk '$3~/1..e/ && $5~/..u/ {print $0}'
```

Conclusion

Conclusion

- Tiny but powerful.

Conclusion

- Tiny but powerful.
- Suitable for well-structured forms of data.

Conclusion

- Tiny but powerful.
- Suitable for well-structured forms of data.
- When it comes to more complicated data, e.g. html, json, awk is not that powerful.

Conclusion

- Tiny but powerful.
- Suitable for well-structured forms of data.
- When it comes to more complicated data, e.g. html, json, awk is not that powerful.

Well...

Conclusion

- Tiny but powerful.
- Suitable for well-structured forms of data.
- When it comes to more complicated data, e.g. html, json, awk is not that powerful.

Well...

```
$ cowthink "Life is short. I use Python."
( Life is short. I use Python. )
-----
      ^ ^
      (oo)\
      (  )\
      | |  ---w |
      | |      |
```

Reference & quotation

氷菓 第十五话 十文字事件. see

<http://bangumi.bilibili.com/anime/3398/play#84791>.

<https://en.wikipedia.org/wiki/AWK>.

<https://www.tutorialspoint.com/awk/>.