2257 Analyzing Network Logs

As your company's network manager, you have become concerned about how slow your Internet connection seems to be in the afternoons. You have a fixed high-speed connection, so you suspect the problem is that too many employees are downloading music and video files, overloading your link. To test your theory, you have already hooked up a Linux box to your Internet connection, running the ¨tcpdump¨ program to monitor all network traffic. That program produced several huge text files, each with a detailed log of every network packet that passed between your company and the rest of the Internet during each tcpdump run.

Now your problem is to make sense of the logs, because looking at files with tens of thousands of records each makes your head hurt. Your task for this problem is to write a program that will read the information in the log files, and then print out a short, useful summary report. That report will be a list of the top ten Internet addresses, ranked by total network traffic, and a list of the top ten internal network users, again ranked by network traffic.

Input

The input file contains several tcpdump logs, each headed by a one-line description-line describing the tcpdump log, followed by several tcpdump-records. The file ends with a single line containing just the word `END`.

Every line begins in column 1.

Each description-line is in the format shown on the following line:

`TCP log from` $start-date-time$ `to` $end-date-time$.

The $start-date-time$ and $end-date-time$ are each in the format

'$YYYY$-MM-DD:HH:MM:SS'

where $YYYY$ is a four digit year, MM is a two digit month number, DD is a two digit day, HH is a two digit hour (in 24-hour clock format), MM is a two digit minute, and SS is a two digit second. For example, 10:30:15 AM on January 5, 2001 would be represented as "`2001-01-05:10:30:15`", and 10:30:15 PM on that day would be represented as "`2001-01-05:22:30:15`".

Each tcpdump-record is in a format shown in either of the following lines:

$timestamp$; $protocol$; $interface$; $size$; $between$
$timestamp$; $protocol$; $interface$; $size$; $between$; $optional$

Each tcpdump-record is a single line, no longer than 255 characters long, and terminated by an end-of-line marker. The fields $timestamp$, $protocol$, $interface$, $size$, and $between$ are all sequences of characters that do not contain end-of-line markers or semi-colons. If there is an $optional$ field, it may contain any character other than an end-of-line marker. The only fields you need to use are $size$ and $between$.

The $size$ field contains an integer in the range of 1 to 32767, followed by a single space, followed by the string '`bytes`' (not including the quotation marks.) For example, if the packet is 412 bytes long, the $size$ field will be "`412 bytes`". Even if the field is only 1 byte long, it will look like "`1 bytes`".

The $between$ field is in the form '`from` *source*:*port* `to` *dest*:*port*'. The $source$ and $dest$ fields are each distinct and valid IP addresses. An IP address consists of four integers in the range from 0 to 255 separated by periods (e.g., 127.12.255.0). The $port$ field is an integer in the range from 0 to 65535.

Processing:

Each time a new description-line is read, a new, independent log is started.

As your program reads tcp-dump records, it will keep track of the total number of bytes that passed to or from each IP address. It doesn't matter if an address is the *source* or *dest*; if it appears in a tcpdump-record, the value of the *size* field is added to the total being tracked for each of the two addresses.

After determining all the totals for all IP addresses, your program will determine the 10 external IP addresses that have the greatest number of bytes passed to or from them, and the 10 internal IP addresses that have the greatest number of bytes passed to or from them, and print those two "top 10" lists, suitably labeled and formatted. Each list should be printed in descending order by number of bytes.

An internal IP address is any address assigned to your company's own network. Those addresses are recognized by their prefixes. Any address in any of the following formats is internal: 192.168.x.x, 204.146.114.x, 206.199.79.x, where x is any integer in the range from 0 to 255. An external IP address is any IP address that is not an internal IP address.

You may assume that there are at least 10 internal and 10 external IP addresses represented in each section of the input data file. There will be at most 12000 distinct IP addresses in any tcp-dump.

You may assume that the total number of bytes sent to or from a single IP address in any single section of the input file is no more than 2147483647.

You may assume that no tcpdump-record begins with either 'END' or 'TCP'.

Output

The output file will contain a separate summary report for each tcpdump log in the input file. The first line of the summary report will be exactly the same as the tcpdump-description line read from the input file. That will be followed by a single blank line, then the following header line, beginning in column 1:

```
Top 10 External Sites Visited:
```

That header line will be followed by ten lines, each showing the total number of bytes seen passing to or from an external IP address, and the IP address. The total number of bytes should be printed in columns 1 through 10, right aligned in the field, followed by one space, followed by the IP address, left aligned. The list should be printed in descending order by number of bytes. If two sites have the same number of bytes, they should be ordered lexicographically by IP address.

There will then be a single blank line, followed by another header line, beginning in column 1:

```
Top 10 Internal Users:
```

That line will also be followed by ten lines in the same format as above, except using internal IP address. A blank line followed by a summary for the next tcpdump log will come next, until all logs have been reported on. Here is a short example of a properly formatted output file:

Sample Input

```
TCP log from 2000-03-16:15:06:33 to 2000-03-16:15:06:34.
Wed Mar 15 15:06:33 2000; TCP; eth0; 1296 bytes; from 204.146.114.10:1916 to 156.26.62.201:126
Wed Mar 15 15:06:33 2000; TCP; eth0; 625 bytes; from 204.146.114.30:289 to 188.226.173.122:13
Wed Mar 15 15:06:33 2000; TCP; eth0; 2401 bytes; from 192.168.5.41:529 to 188.226.173.122:31
Wed Mar 15 15:06:33 2000; TCP; eth0; 1296 bytes; from 206.199.79.28:1280 to 167.253.170.210:168;first packet
Wed Mar 15 15:06:33 2000; TCP; eth0; 625 bytes; from 192.168.5.72:1247 to 89.40.199.255:214
Wed Mar 15 15:06:33 2000; TCP; eth0; 2401 bytes; from 192.168.5.44:290 to 110.150.70.190:26
Wed Mar 15 15:06:33 2000; TCP; eth0; 2401 bytes; from 192.168.5.119:253 to 192.22.192.204:206;lost data
Wed Mar 15 15:06:33 2000; TCP; eth0; 1296 bytes; from 192.168.5.95:1646 to 156.26.62.201:12
Wed Mar 15 15:06:33 2000; TCP; eth0; 625 bytes; from 206.199.79.5:1566 to 6.234.186.83:145
Wed Mar 15 15:06:33 2000; TCP; eth0; 1296 bytes; from 204.146.114.14:2017 to 183.74.83.174:103
Wed Mar 15 15:06:33 2000; TCP; eth0; 2401 bytes; from 204.146.114.50:645 to 132.130.65.172:127
Wed Mar 15 15:06:33 2000; TCP; eth0; 2401 bytes; from 192.168.5.5:1184 to 83.141.167.38:64
Wed Mar 15 15:06:33 2000; TCP; eth0; 81 bytes; from 192.168.5.117:963 to 203.68.142.136:112
END
```

Sample Output

```
TCP log from 2000-03-16:15:06:33 to 2000-03-16:15:06:34.

Top 10 External Sites Visited:
      3026 188.226.173.122
      2592 156.26.62.201
      2401 110.150.70.190
      2401 132.130.65.172
      2401 192.22.192.204
      2401 83.141.167.38
      1296 167.253.170.210
      1296 183.74.83.174
       625 6.234.186.83
       625 89.40.199.255

Top 10 Internal Users:
      2401 192.168.5.119
      2401 192.168.5.41
      2401 192.168.5.44
      2401 192.168.5.5
      2401 204.146.114.50
      1296 192.168.5.95
      1296 204.146.114.10
      1296 204.146.114.14
      1296 206.199.79.28
       625 192.168.5.72
```