

作业二 Shell,Vim & data cleaning

课程: 系统开发工具基础

姓名: 吴啸天 专业: 计算机科学与技术 学号: 24020007137

Github: <https://github.com/Cheongfan/...>

日期: 2025 年 9 月 16 日

1 练习内容

1.1 内容简介

本次练习聚焦 Shell 命令行及脚本, Vim 编辑器方法, 以及一些数据整理方法, 包括 Shell 文件操作、函数编写与检索, Vim 基本编辑与配置, 以及使用 cut、awk、sed 等工具进行数据提取、过滤、清洗与统计的内容, 着重培养命令行工具的整合与实际数据处理能力。

1.2 实例清单

1. 【Shell】学习 ls 命令操作, 按需输出目录列表。
2. 【Shell】编写 marco 和 polo 函数对工作目录进行两种方式的保存。
3. 【Shell】学习 find 命令查找所需文件名。
4. 【Shell】示例 CSV 文件的创建与验证。
5. 【Shell】按不同规则查询文件信息。
6. 【Shell】使用 grep 搜索特定内容, 进行简单的搜索与过滤。
7. 【Shell】对文件进行复制与备份, 并验证备份成功。
8. 【VIM】完成 vimtutor 下载安装。
9. 【VIM】下载 vimrc。
10. 【VIM】使用 Vim 创建.txt 文件。
11. 【VIM】文件的搜索与替换 (单行)。
12. 【VIM】使用 Vim 进行全局搜索与替换。
13. 【Date_leaning】使用 cut 命令提取特定列。
14. 【Date_leaning】使用 awk 过滤特定行。
15. 【Date_leaning】使用 awk 处理缺失值, 将空年龄替换为 0。

16. 【Date_leaning】排序和去重操作。
17. 【Date_leaning】使用 awk 进行高级格式化。
18. 【Date_leaning】使用 sed 进行高级编辑，对标点符号进行操作，使用多个 sed 组合操作。
19. 【Date_leaning】结合上文的计算、统计、查找命令，进行系统的数据分析。
20. 【Date_leaning】创建和处理日志文件。

2 解题感悟

在完成本次作业的过程中，我系统性地实践了 Shell 脚本编写、Vim 文本编辑及数据清洗等相关操作，进一步掌握了命令行环境下的高效 workflow 构建方法。通过编写 marco/polo 函数、使用 find/grep 进行文件查找与过滤、结合 awk/sed 完成结构化数据处理，我体会到命令行工具在自动化任务与数据处理方面的强大表达能力。尤其是在处理 CSV 文件和日志数据时，灵活组合文本处理命令能快速实现提取、过滤、转换与统计等功能，显著提升数据预处理的效率。

此外，Vim 的多种编辑模式与快捷键操作也让我意识到文本编辑器不仅限于输入，更是一个可高度定制的高效编程环境。经过本次练习，我对 Linux 环境下常用开发工具的使用有了更深刻的理解，也增强了解决实际工程中文件操作与数据清洗问题的信心。

3 练习过程与成果

```
total 16K
drwx----- 2 ubuntu ubuntu 40 Sep 12 09:31 .gvfs
-rw-r--r-- 1 ubuntu ubuntu 807 Sep 12 09:31 .profile
-rw-r--r-- 1 ubuntu ubuntu 3.7K Sep 12 09:31 .bashrc
-rw-r--r-- 1 ubuntu ubuntu 220 Sep 12 09:31 .bash_logout
drwxr-x--- 17 ubuntu ubuntu 440 Sep 12 02:08 .
drwxrwxr-x 2 ubuntu ubuntu 40 Sep 12 02:08 .h
drwxrwxr-x 2 ubuntu ubuntu 40 Sep 12 02:08 .x
drwxr-xr-x 2 ubuntu ubuntu 60 Sep 12 01:57 Desktop
-rw----- 1 ubuntu ubuntu 117 Sep 12 01:56 .bash_history
drwxr-xr-x 13 ubuntu ubuntu 420 Sep 12 01:37 .config
drwx----- 9 ubuntu ubuntu 200 Sep 12 01:37 .cache
drwx----- 4 ubuntu ubuntu 80 Sep 12 01:34 snap
-rw-r--r-- 1 ubuntu ubuntu 0 Sep 12 01:33 .sudo_as_admin_successful
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Documents
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Music
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Pictures
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Public
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Templates
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Videos
drwxr-xr-x 2 ubuntu ubuntu 40 Sep 12 01:33 Downloads
drwx----- 4 ubuntu ubuntu 80 Sep 12 01:32 .local
drwxr-xr-x 1 root root 80 Sep 12 01:32 ..
```

图 1 实例 1

```
cf@ubuntu:~$ mkdir test
cf@ubuntu:~$ cd test
cf@ubuntu:~/test$ cat > marco.sh
cf@ubuntu:~/test$ source marco.sh
cf@ubuntu:~/test$ marco
save pwd /home/cf/test
cf@ubuntu:~/test$ cd ../
cf@ubuntu:~$ polo
cf@ubuntu:~/test$
```

图 2 实例 2

```
cf@ubuntu:~/test$ mkdir html_root
cf@ubuntu:~/test$ cd html_root/
cf@ubuntu:~/test/html_root$ touch {1..10}.html
bash: syntax error near unexpected token `}'
cf@ubuntu:~/test/html_root$ touch {1..10}.html
cf@ubuntu:~/test/html_root$ mkdir html
cf@ubuntu:~/test/html_root$ cd html/
cf@ubuntu:~/test/html_root/html$ touch index.html
cf@ubuntu:~/test/html_root/html$ find . -type f -name "*.html" | xargs -d '\n' tar -c
ar -cvzf html.zip
./index.html
cf@ubuntu:~/test/html_root/html$ cd ../
cf@ubuntu:~/test/html_root$ find . -type f -name "*.html" | xargs -d '\n' tar -c
vzf html.zip
./6.html
./html/index.html
./5.html
./10.html
./7.html
./3.html
./8.html
./2.html
./4.html
./1.html
./9.html
```

图 3 实例 3

```
cf@ubuntu:~/vim/pack/vendor/start$ mkdir -p ~/data_cleaning_practice/raw_data ~
/data_cleaning_practice/cleaned_data
cf@ubuntu:~/vim/pack/vendor/start$ cd ~/data_cleaning_practice
cf@ubuntu:~/data_cleaning_practice$ cd ~/data_cleaning_practice/raw_data
cf@ubuntu:~/data_cleaning_practice/raw_data$ cat > users.csv << EOF
> ID,Name,Email,Age,City
> 1,Alice,alice@example.com,28,"New York"
> 2,Bob,bob@example.com,32,"Los Angeles"
> 3,Carol,carol@example.com,, "Chicago"
> 4,Dave,dave@example.com,45,"Miami"
> 5,Eve,eve@example.com,29,"Seattle"
> EOF
cf@ubuntu:~/data_cleaning_practice/raw_data$ cat users.csv
ID,Name,Email,Age,City
1,Alice,alice@example.com,28,"New York"
2,Bob,bob@example.com,32,"Los Angeles"
3,Carol,carol@example.com,, "Chicago"
4,Dave,dave@example.com,45,"Miami"
5,Eve,eve@example.com,29,"Seattle"
```

图 4 实例 4

```
cf@ubuntu:~/data_cleaning_practice/raw_data$ head -n 3 users.csv
ID,Name,Email,Age,City
1,Alice,alice@example.com,28,"New York"
2,Bob,bob@example.com,32,"Los Angeles"
cf@ubuntu:~/data_cleaning_practice/raw_data$ tail -n 2 users.csv
4,Dave,dave@example.com,45,"Miami"
5,Eve,eve@example.com,29,"Seattle"
cf@ubuntu:~/data_cleaning_practice/raw_data$ wc users.csv
  6   8 209 users.csv
cf@ubuntu:~/data_cleaning_practice/raw_data$ cat users.csv
ID,Name,Email,Age,City
1,Alice,alice@example.com,28,"New York"
2,Bob,bob@example.com,32,"Los Angeles"
3,Carol,carol@example.com,, "Chicago"
4,Dave,dave@example.com,45,"Miami"
5,Eve,eve@example.com,29,"Seattle"
```

图 5 实例 5

```
cf@ubuntu:~/data_cleaning_practice/raw_data$ grep "New" users.csv
1,Alice,alice@example.com,28,"New York"
cf@ubuntu:~/data_cleaning_practice/raw_data$ grep ",," users.csv
3,Carol,carol@example.com,, "Chicago"
cf@ubuntu:~/data_cleaning_practice/raw_data$ grep "^[0-9]" users.csv
1,Alice,alice@example.com,28,"New York"
2,Bob,bob@example.com,32,"Los Angeles"
3,Carol,carol@example.com,, "Chicago"
4,Dave,dave@example.com,45,"Miami"
5,Eve,eve@example.com,29,"Seattle"
cf@ubuntu:~/data_cleaning_practice/raw_data$
```

图 6 实例 6

```
cf@ubuntu:~/data_cleaning_practice/raw_data$ cp ~/data_cleaning_practice/raw_data/users.csv ~/data_cleaning_practice/cleaned_data/
cf@ubuntu:~/data_cleaning_practice/raw_data$ mv ~/data_cleaning_practice/cleaned_data/users.csv ~/data_cleaning_practice/cleaned_data/users_backup.csv
cf@ubuntu:~/data_cleaning_practice/raw_data$ ls -l ~/data_cleaning_practice/cleaned_data/
total 4
-rw-rw-r-- 1 cf cf 209 Sep 16 08:00 users_backup.csv
```

图 7 实例 7

```
cf@ubuntu:~$ vimtutor
The program 'vimtutor' is currently not installed. You can install it by typing:
sudo apt install vim-runtime
cf@ubuntu:~$ sudo apt install vim-runtime
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  vim-common vim-tiny
Suggested packages:
  vim-gnome-py2 | vim-gtk-py2 | vim-gtk3-py2 | vim-athena-py2 | vim-nox-py2
  indent
The following NEW packages will be installed:
  vim-runtime
The following packages will be upgraded:
  vim-common vim-tiny
2 upgraded, 1 newly installed, 0 to remove and 192 not upgraded.
Need to get 5,718 kB of archives.
After this operation, 27.5 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
```

图 8 实例 8

```
cf@ubuntu:~$ wget https://missing-semester-cn.github.io/2020/files/vimrc
--2025-09-16 07:47:07-- https://missing-semester-cn.github.io/2020/files/vimrc
Resolving missing-semester-cn.github.io (missing-semester-cn.github.io)... 198.18.0.79
Connecting to missing-semester-cn.github.io (missing-semester-cn.github.io)|198.18.0.79|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3254 (3.2K) [application/octet-stream]
Saving to: 'vimrc'

vimrc                100%[=====>]   3.18K  --.-KB/s   in 0.02s
2025-09-16 07:47:08 (180 KB/s) - 'vimrc' saved [3254/3254]
```

图 9 实例 9

```
notes.txt + (~/.data_cleaning_practice) - VIM
Data Cleaning Practice Notes
- Created sample user data
- Practiced basic shell commands
- Next: Learn Vim editing
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
-- INSERT --                                4,26      All
```

图 10 实例 10


```
cf@ubuntu: ~/data_cleaning_practice/cleaned_data
cf@ubuntu:~/test$ cd ~/data_cleaning_practice/cleaned_data
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ cut -d',' -f2,5 users_backup.csv > name_city.txt
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ cat name_city.txt
Name,City
Alice,"New York"
Bob,"Los Angeles"
Carol,"Chicago"
Dave,"Miami"
Eve,"Seattle"
```

图 13 实例 13

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' '$4 > 30' users_backup.csv
ID,Name,Email,Age,City
2,Bob,bob@example.com,32,"Los Angeles"
4,Dave,dave@example.com,45,"Miami"
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' '$5 == "\"Miami\"" ' users_backup.csv
4,Dave,dave@example.com,45,"Miami"
```

图 14 实例 14

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' 'BEGIN {OFS=","} {if ($4 == "") $4 = "0"; print}' users_backup.csv
ID,Name,Email,Age,City
1,Alice,alice@example.com,28,"New York"
2,Bob,bob@example.com,32,"Los Angeles"
3,Carol,carol@example.com,0,"Chicago"
4,Dave,dave@example.com,45,"Miami"
5,Eve,eve@example.com,29,"Seattle"
```

图 15 实例 15

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ cut -d',' -f5 users_backup.csv | sort | uniq > cities.txt
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ cat cities.txt
"Chicago"
City
"Los Angeles"
"Miami"
"New York"
"Seattle"
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ sort -t',' -k4,4n users_backup.csv
3,Carol,carol@example.com,, "Chicago"
ID,Name,Email,Age,City
1,Alice,alice@example.com,28,"New York"
5,Eve,eve@example.com,29,"Seattle"
2,Bob,bob@example.com,32,"Los Angeles"
4,Dave,dave@example.com,45,"Miami"
```

图 16 实例 16

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' 'NR>1 {printf "%-10s (%d) is from %s.\n", $2, $4, $5}' users_backup.csv
Alice      (28) is from "New York".
Bob        (32) is from "Los Angeles".
Carol      (0) is from "Chicago".
Dave       (45) is from "Miami".
Eve        (29) is from "Seattle".
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' 'BEGIN {print "<table>\n<tr><th>Name</th><th>Age</th><th>City</th></tr>"}\nNR>1 {printf "<tr><td>%s</td><td>%d</td><td>%s</td></tr>\n", $2, $4, $5}\nEND {print "</table>"}' users_backup.csv > users.html
```

图 17 实例 17

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ sed 's/,/|/g' users_backup.csv
ID|Name|Email|Age|City
1|Alice|alice@example.com|28|"New York"
2|Bob|bob@example.com|32|"Los Angeles"
3|Carol|carol@example.com||"Chicago"
4|Dave|dave@example.com|45|"Miami"
5|Eve|eve@example.com|29|"Seattle"
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ sed 's/"//g' users_backup.csv
ID,Name,Email,Age,City
1,Alice,alice@example.com,28,New York
2,Bob,bob@example.com,32,Los Angeles
3,Carol,carol@example.com,,Chicago
4,Dave,dave@example.com,45,Miami
5,Eve,eve@example.com,29,Seattle
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ sed -e 's/,/|/g' -e 's/"//g' users_backup.csv > users_cleaned.csv
```

图 18 实例 18

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' 'NR>1 {print $5}' users_backup.csv | sort | uniq -c
      1 "Chicago"
      1 "Los Angeles"
      1 "Miami"
      1 "New York"
      1 "Seattle"
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' 'NR>1 {print $2 "," $4}' users_backup.csv | sort -t',' -k2,2n | head -1
Carol,
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk -F',' 'NR>1 {print $2 "," $4}' users_backup.csv | sort -t',' -k2,2n | tail -1
Dave,45
```

图 19 实例 19

```
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ cat > access.log << EOF
> 192.168.1.1 - - [01/Jan/2023:10:00:01 +0000] "GET /index.html HTTP/1.1" 200 1234
> 192.168.1.2 - - [01/Jan/2023:10:00:02 +0000] "POST /submit.php HTTP/1.1" 200 5678
> 192.168.1.3 - - [01/Jan/2023:10:00:03 +0000] "GET /about.html HTTP/1.1" 200 9012
> 192.168.1.4 - - [01/Jan/2023:10:00:04 +0000] "POST /login.php HTTP/1.1" 200 3456
> EOF
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ grep "POST" access.log
192.168.1.2 - - [01/Jan/2023:10:00:02 +0000] "POST /submit.php HTTP/1.1" 200 5678
192.168.1.4 - - [01/Jan/2023:10:00:04 +0000] "POST /login.php HTTP/1.1" 200 3456
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk '{print $1, $4}' access.log
| sed 's/[//]'
192.168.1.1 01/Jan/2023:10:00:01
192.168.1.2 01/Jan/2023:10:00:02
192.168.1.3 01/Jan/2023:10:00:03
192.168.1.4 01/Jan/2023:10:00:04
cf@ubuntu:~/data_cleaning_practice/cleaned_data$ awk '{print $1}' access.log | sort | uniq -c
      1 192.168.1.1
      1 192.168.1.2
      1 192.168.1.3
      1 192.168.1.4
```

图 20 实例 20

参考文献

- [1] https://blog.csdn.net/qq_51247028/article/details/121180700
- [2] <https://missing-semester-cn.github.io/missing-notes-and-solutions/2020/solutions/shell-tools-solution/>
- [3] <https://missing-semester-cn.github.io/missing-notes-and-solutions/2020/solutions/editors-solution/>
- [4] <https://missing-semester-cn.github.io/missing-notes-and-solutions/2020/solutions/shell-tools-solution/>