实验 04 sed 实验

班级：数据科学与大数据一班
学号：202026203005
姓名：张华
用户名：s13

一、实验目的

1．练习使用正则表达式匹配字符串
2．练习使用 sed 完成文本内容的转换

二、实验要求

1．填写实验报告，请将关键命令及其结果进行截图(请确保截图中的文字清晰可见)
2． 导出为 pdf 文件，文件名为用户名-姓名-lab04.pdf，在规定截止时间之前上传作业）
3．以下步骤中**所有 s01 请换成你自己的用户名。**

三、实验步骤

1. 复制学校首页的网页源代码并保存为 s01-jxnu.html 文件。
（1)提取网页中所有的绝对 URL(可能在一对单引号或双引号中)到文件 s01-jxnu.url 中。

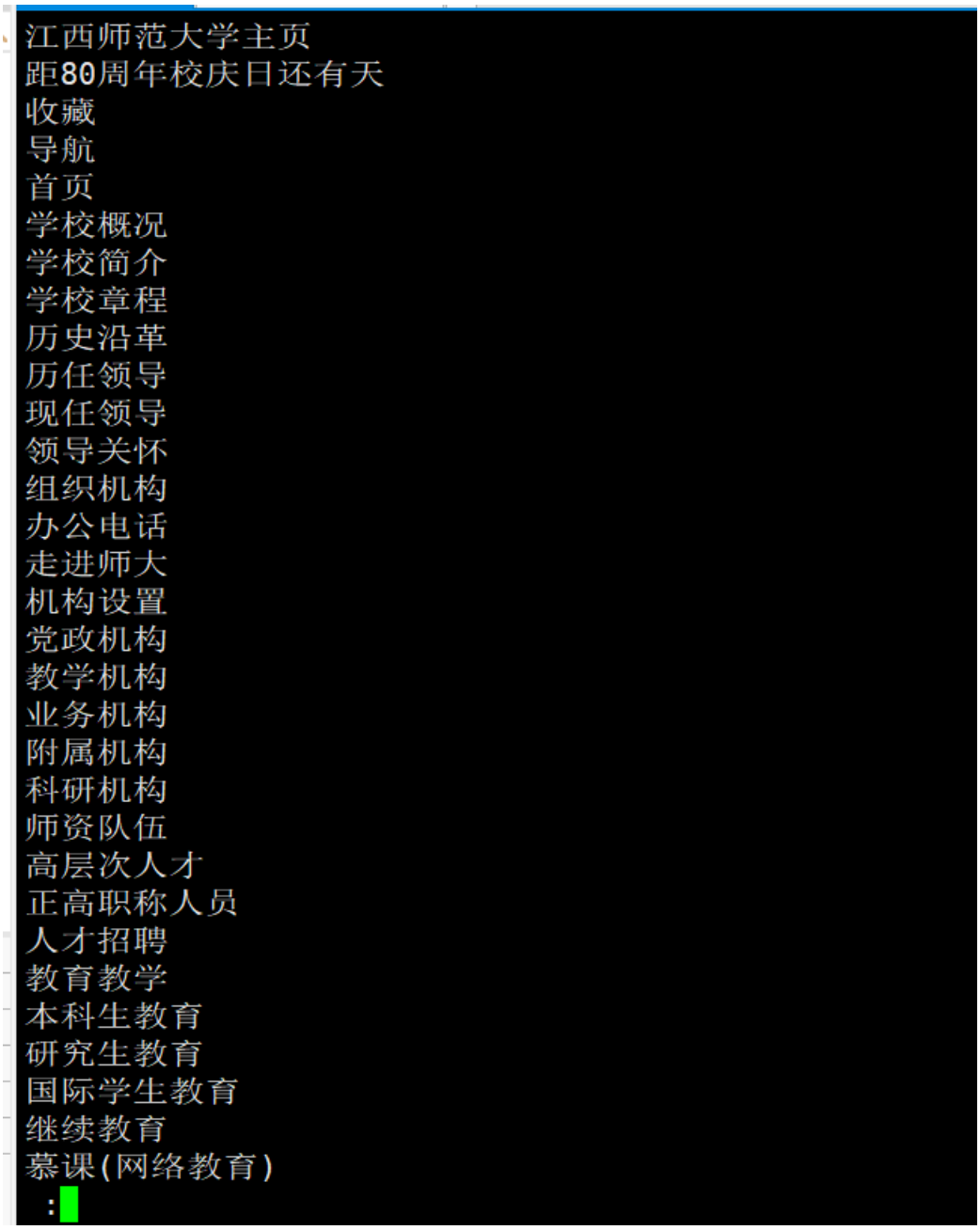
```
'http://xy.jxnu.edu.cn/'
'http://xy.jxnu.edu.cn/'
"http://bszs.conac.cn/sitename?method=show&id=40EA66A2295B384EE053012819AC9E2D"
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ grep -E -o "[\"|\'](http|https):\/\/[^\"\']+[\"|\']" s13-jxnu.html>s13-jxnu.url
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$
```

'http://www.baidu.com'
"http://www.jxnu.edu.cn/main.htm"
"http://www.jxnu.edu.cn/10/list.htm"
"http://www.jxnu.edu.cn/98/list.htm"
"http://www.jxnu.edu.cn/17/list.htm"
"http://www.jxnu.edu.cn/2018/1227/c56a1656/page.htm"
"https://rczp.jxnu.edu.cn/"
"http://jwc.jxnu.edu.cn/Portal/Index.aspx"
"http://graduate.jxnu.edu.cn/"
"http://laihua.jxnu.edu.cn/"
"http://jxjy.jxnu.edu.cn/"
"http://jxnu.fanya.chaoxing.com/portal"
"http://kjc.jxnu.edu.cn/"
"http://skc.jxnu.edu.cn/"
"http://www.jxnu.edu.cn/2018/1227/c54a1651/page.htm"
"http://zs.jxnu.edu.cn/"
"http://yz.jxnu.edu.cn/"
"http://jy.jxnu.edu.cn/"
"http://international.jxnu.edu.cn/"
"http://tsg.jxnu.edu.cn/"
"http://dag.jxnu.edu.cn/"
"https://xxgk.jxnu.edu.cn/"
"http://laihua.jxnu.edu.cn/"
'https://mp.weixin.qq.com/s/Hxznpr5xWmJynvDKICTW2Q'
'https://mp.weixin.qq.com/s/Q-FPrzR9-lvcMd-t_SugPQ'
'https://mp.weixin.qq.com/s/CqrFs-CiVC3MQpHcnXg_AA'
'https://mp.weixin.qq.com/s/7TJ1-S7WFc_CsZivFRf0fA'

（2）用一条管道命令去除该页面的所有 javascript 代码、网页标签以及多余的空白符和空行，提取出页面的文本内容并保存为 s01-jxnu.txt 文件。

```
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed "/<script.*>.*<\/script>/d" s13-jxnu.html |sed '/<script.*>/,/<\/script>/d'|sed "s/<[^<>]*>//g"|tr -s [
:space:] \\n >s13-jxnu.txt
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$
```

（3）用 less 查看 s01-jxnu.txt 文件的内容。



2. 现有文件 begin-end.txt 的内容如下：

$ cat begin-end.txt

good begin

Begin working on that

research you wanted

to, do not let it end

hi there

begin and strive to

reach the End

bye

（1）打印文件 begin-end.txt 中所有由行首为 begin 的行到行尾为 end 行所构成的行组（忽略大小写）之外的行。

```
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed -n '/^[B\|b]egin/,/[E\|e]nd$/!p' begin-end.txt
good begin
hi there
bye
```

（2）将文件 begin-end.txt 中包含长度为 5 的单词的行存入文件 five.txt，同时将包含长度为 6 的单词的行存入文件 six.txt。

```
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed -e '/[a-zA-Z]\{5\}/w five.txt' -e '/[a-zA-Z]\{6\}/w six.txt' begin-end.txt
good begin
Begin working on that
research you wanted
to, do not let it end
hi there
begin and strive to
reach the End
bye
```

```
good begin
Begin working on that
research you wanted
hi there
begin and strive to
reach the End
```

（3）将文件 hello.txt 的第 3-5 行替换为文件 begin-end.txt 的第 2-4 行。文件 hello.txt 的内容如下：

$ cat hello.txt
Hello World
How are you
This game is good
Today is sunny
12345
You are funny

```
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed -n '2,4p' begin-end.txt >1.txt
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed '5r 1.txt' hello.txt|sed '3,5d'
Hello World
How are you
Begin working on that
research you wanted
to, do not let it end
You are funny
```

（4）将文件 begin-end.txt 中所有包含 begin 或 end（不区分大小写）的行写入文件 flags.txt，其他行则写入文件 mid.txt。

```
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed -n '/^[B\|b]egin/,/[E\|e]nd$/w flags.txt' begin-end.txt
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$ sed -n '/^[B\|b]egin/,/[E\|e]nd$/!w mid.txt' begin-end.txt
[s13@iZuf6ixnt8107e8ns3k4tuZ ~]$
```

Begin working on that
research you wanted
to, do not let it end
begin and strive to
reach the End

good begin
hi there
bye

4. 转换文本文件内容，并利用 groff 从纯文本生成带图表和公式的 pdf 论文。

（1）修改虚拟机网络设置，使虚拟机能够正常访问外网。

（2）将实验文件 mypaper.txt 上传到虚拟机。

（3）执行以下命令安装 groff 软件包

```
sudo apt install groff
```

（4）请编写 sed 脚本 s01.sed，并执行命令

```
sed -f s01.sed mypaper.txt >s01-paper.ms
```

生成符合 groff 语法的 s01-paper.ms 文件

转换要求如下表（**左边单行内容一般转换成右边的多行，请注意换行**）

转换前内容	转换后变成
Wuji Zhang	你的姓名的汉语拼音
<title>…content…</title>	.TL …content…
<author>…content…</author>	.AU …content…
<institution>…content…</institution>	.AI …content…
<abstraction>…content…</abstraction>	.AB …content… .AE
…content……other content…	.I …content… …other content… （注意：.I 后面要留一个空格）
…content…	.NH …content…
 …content… 	.PP …content…
<foot> …content… </foot>	.FS …content… .FE
* …content…	.IP * …content…

<quote> …content… </quote>	.QP …content…
<table> …content… </table>	.TS …content… .TE
<equation> (1.s01) …content… </equation>	.EQ (1.s01) …content… .EN
<picture> …content… </picture>	.PS …content… .PE
<pre> C …content… </pre>	.DS C …content… .DE
注意：存在一行转换成多行的情况！请严格按上表进行转换，包括空格和换行！	

（此处请贴上 s01.sed 的代码）

```
s#a01-Wuji Zhang#s13-Hua Zhang#
s#<title>.*</title>#.TL\n&#
s#<title>##
s#</title>##
s#<author>.*</author>#.AU\n&#
s#<author>##
s#</author>##
s#<institution>.*</institution>#.AI\n&#
s#<institution>##
s#</institution>##
s#<abstraction>#.AB#
s#</abstraction>#.AE#
s#<i>#.I #
s#</i>#\n#
s#<h1>#.NH\n#
s#</h1>##
s#<p>#.PP#
s#</p>##
s#<foot>#.FS#
s#</foot>#.FE#
s#<li> *#.IP *#
s#</li>##
s#<quote>#.QP#
s#</quote>##
s#<table>#.TS#
s#</table>#.TE#
s#<equation> (1.a01)#.EQ (1.a01)#
```

```
s#<li> *#.IP *#
s#</li>##
s#<quote>#.QP#
s#</quote>##
s#<table>#.TS#
s#</table>#.TE#
s#<equation> (1.a01)#.EQ (1.a01)#
s#</equation>#.EN#
s#<picture>#.PS#
s#</picture>#.PE#
s#<pre> C#.DS C#
s#</pre>#.DE#
/^$/d
```

（5）执行 groff 命令生成 pdf 文件

```
groff -t -e -p -ms s01-paper.ms | ps2pdf - s01-paper.pdf
```

（6）用 pdf 阅读器打开 s01-paper.pdf，并欣赏你生成的论文。

（此处请贴上 pdf 阅读器中显示的 s01-paper.pdf 的内容）

Using groff to prepare your paper

s13-Hua Zhang

College of Computer Information Engineering, Jiangxi Normal University

ABSTRACT

groff (GNU troff) is a typesetting system that reads plain text mixed with formatting commands and produces formatted output. Output may be PostScript or PDF, HTML, or text for display at terminal. Present on most Unix/Linux systems owing to its long association with Unix manuals, *groff* is capable of producing typographically sophisticated documents while consuming only minimal system resources.

1. History

The first version of UNIX was developed on a PDP-7 which was sitting around Bell Labs. In 1971 the developers wanted to get a PDP-11 for further work on the operating system. In order to justify the cost for this system, they proposed that they would implement a document formatting system for the AT&T patents division. This first formatting program was a reimplementation of McIllroy's 'roff', written by J. F. Ossanna.

James Clark began work on a GNU implementation of 'ditroff' in early 1989. It was declared a stable (i.e. non-beta) package with the release of version 1.04 around November 1991.

2. Why using groff?

Although WYSIWYG† systems may be easier to use, they have a number of disadvantages compared to 'troff':

** They must be used on a graphics display to work on a document.

** Most of the WYSIWYG systems are either non-free or are not very portable.

** 'troff' is firmly entrenched in all UNIX systems.

** It is difficult to have a wide range of capabilities available within the confines of a GUI/window system.

** It is more difficult to make global changes to a document.

"GUIs normally make it simple to accomplish simple actions and impossible to accomplish complex actions." -Doug Gwyn (22/Jun/91 in 'comp.unix.wizards')

3. Table

This is an example of how to create table using tbl.

Table a01: Some well-known TCP Ports

23	telnet
25	simple mail transfer protocal(SMTP)
80	hyper text transfer protocal(HTTP)
110	post office protocol(POP)

† WYSIWYG: What you see is what you get.

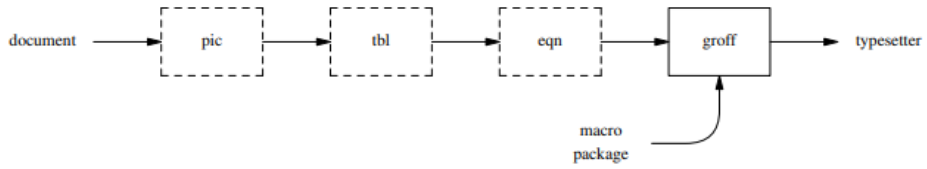
4. Equation

This is an example of how to write math equation using eqn.

$$f(\zeta) = \frac{1}{2\pi i}\int_{C}\frac{f(z)}{z-\zeta}\,dz \qquad (1.a01)$$

5. Picture

This is an example of how to draw picture using pic.



6. Convert it to pdf

Suppose your file has been saved as a01.ms, then you can convert it to a01.pdf using command as follows:

groff -t -e -p -ms a01.ms | ps2pdf - a01.pdf

Look, it is so simply, enjoy your work!