

文本处理命令（一）



Linux命令之CT机

`grep`

Linux命令之切片机

`cut`

Linux命令之粘合剂

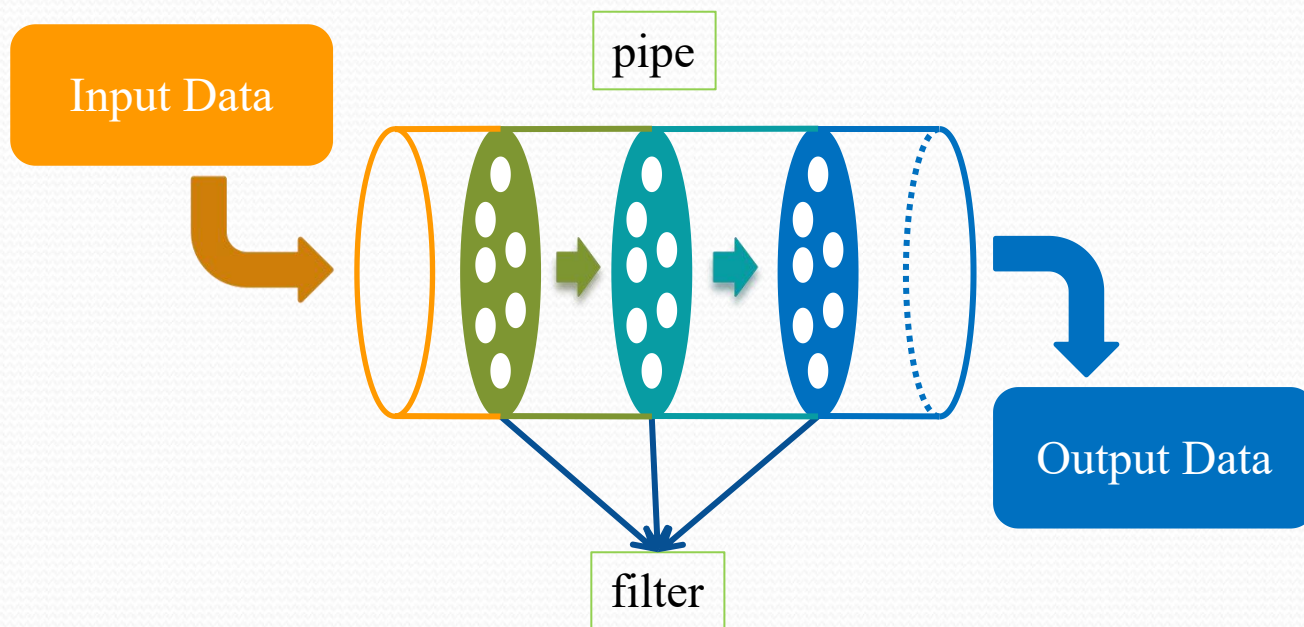
`paste`

井井有条的`sort`命令

独一无二的`uniq`命令

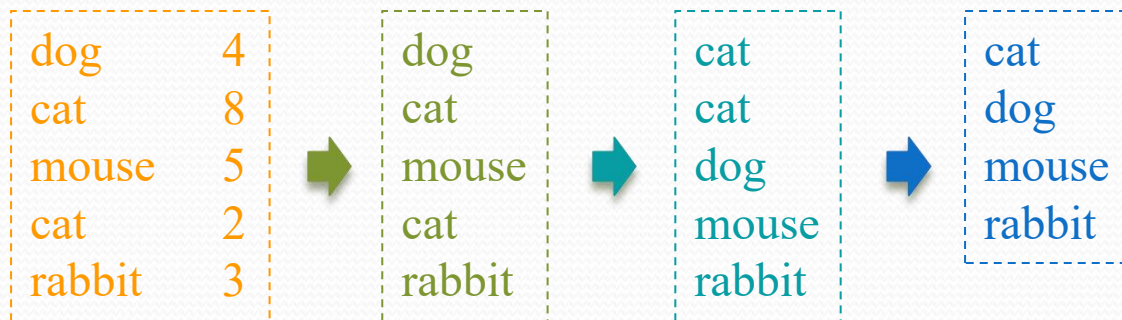
擅长计数的`wc`命令

Linux下的文本处理原理



命令管道: `zcat input.gz | cut -f1 | sort | uniq >output.txt`

数据变化:



Linux命令之CT机--grep命令



grep命令档案



命令全名： globally search a regular expression and print

命令用途： 查找并显示包含指定模式的行

命令格式： grep [option] pattern file

常用选项：

- c 统计结果行数
- E 使用扩展的正则表达式
- i 忽略大小写
- v 反向查找
- A n 显示找到的行及其后面n行内容

grep命令用法

∞ 查找文件中包含一个模式的行

```
$grep 'NP_173616.2' at_pep.fa
```

∞ 查找文件中包含一个模式的行及其后面几行内容

```
$grep -A 3 'NP_173616.2' at_pep.fa
```

∞ 查找当前目录中名字包含特定字符串的文件/目录

```
$ls | grep at
```

∞ 去掉文件中包含某个字符的行

```
$grep -v '#' ~/bin/wgetr
```


Linux命令之切片机--cut命令



cut命令档案



命令全名： cut

命令用途： 取出文本中指定的列

命令格式： cut [option] file

常用选项：

- d 指定列分隔符，默认是制表符
- f 指定取出的字段
- b 指定取出的字节
- c 指定取出的字符

cut命令的用法

取出一个文件中的一列，分隔符是制表符

```
$cut -f1 at_gene_partial.gff
```

取出一个文件中的中的两列，分隔符不是制表符

```
$cut -d: -f1,6 /etc/passwd
```

取出压缩文件的几列

```
$zcat at_gene_partial.gff.gz | cut -f1,3-5
```


cut命令选项示例

```
$ cat char.txt
```

China

中国

```
$ cat char.txt | cut -b1
```

C

中

```
$ cat char.txt | cut -b1-2
```

Ch

中

```
$ cat char.txt | cut -b1-3
```

Chi

中

```
$ cat char.txt | cut -c1
```

C

中

file命令查看一个文件的字符编码方式

```
$ file char_chs.txt
```

char.txt: UTF-8 Unicode text

字符编码

“a”的UTF-8编码:

0	1	1	0	0	0	0	1
---	---	---	---	---	---	---	---

“中”的UTF-8编码:

1	1	1	0	0	1	0	0
1	0	1	1	1	0	0	0
1	0	1	0	1	1	0	1

Linux命令之粘合剂--paste命令



paste命令档案



命令全名: paste

命令用途: 按列合并文件

命令格式: paste [option] file1 file2 ...

常用选项:

- d 指定合并后的列分隔符，默认的列分隔符是制表符

paste命令的用法

∞ 按列合并两个文件

```
$paste animal number
```

∞ 按列合并两个文件，分隔符是冒号

```
$paste -d: animal number
```

井井有条的sort命令



sort命令档案



命令全名: sort

命令用途: 按行排序文本

命令格式: sort [option] file

常用选项:

- r 反向排序
- n 按数字大小排序
- g 按科学计数法数字大小排序
- k 指定参与排序的字段

sort命令用法 (1)

∞ 按ASCII码值升序排序

```
$sort num
```

∞ 按数字大小升序排序

```
$sort -n num
```

∞ 按数字大小降序排序

```
$sort -nr num
```

∞ 按数字大小排序，包括科学计数法的值

```
$sort -g num
```


sort命令用法 (2)

指定字段排序

`$sort -k3 price`

`-kn1[.m1][,n2[.m2]]`

开始字段 开始字符 结束字段 结束字符

按指定字段的指定字符排序

`$sort -k1.6,1n price`

按多个字段排序

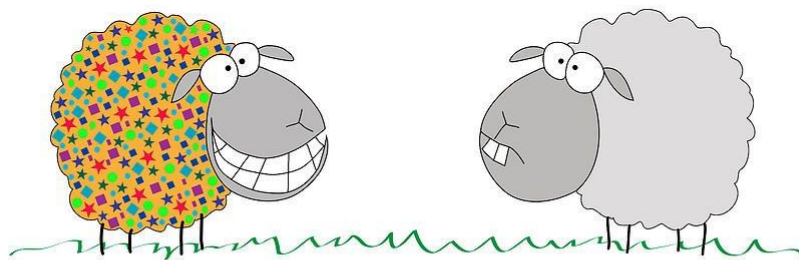
`$sort -k1.6,1n -k3 price`

sort的-k选项用法:

-k1	:	<u>fruit2</u>	<u>6.5</u>	<u>China</u>
-k2	:	<u>fruit1</u>	<u>5.4</u>	<u>US</u>
-k2,2	:	<u>fruit5</u>	<u>7.7</u>	US
-k2.1,2.2	:	<u>fruit6</u>	<u>3.2</u>	UK
-k2.2,2.4	:	<u>fruit1</u>	<u>4.7</u>	China
-k1.6,2	:	<u>fruit10</u>	<u>6.2</u>	China

独一无二的uniq命令

BE UNIQUE



uniq命令档案



命令全名: **unique**

命令用途: 检查和删除重复的行（经常用在sort后面）

命令格式: **uniq** [option] file

常用选项:

- c 统计行重复的次数

uniq命令用法

∞ 去掉文件中相邻的重复的行

```
$uniq animal
```

∞ 去掉文件中所有的重复的行

```
$sort animal | uniq
```

∞ 统计文件中行的重复次数

```
$sort animal | uniq -c
```


擅长计数的wc命令



WC命令档案



命令全名: word count

命令用途: 统计文本的行数、单词数、字符数或字节数

命令格式: wc [option] file

常用选项:

- l 只输出行数
- w 只输出单词数（空格或制表符分隔的字符串）
- m 只输出字符数
- c 只输出字节数

WC命令用法

∞ 统计文件的行数、单词数及字节数

```
$wc animal
```

∞ 统计文件的行数

```
$wc -l animal
```

∞ 用在管道中，统计其它命令输出的文本的行数

```
$sort animal | uniq | wc -l
```

课后作业3 (HW3)

(1) 统计文件`at_gene_partial.gff.gz`中类型是mRNA的序列的个数，并将结果写入到本目录下的文件`mRNA_num`中；

(2) 取出文件`at_gene_partial.gff.gz`中类型是mRNA的行，按序列起始位置（第4列）从大到小排序（反序），并将结果写入到本目录下的文件`mRNA_sorted`中。



The End