

中国人民大学数据挖掘中心出品

# Linux 系统及文本操作自学教程(1.0 版)

——Linux 入门资料整理推荐(含练习)



我们的网站:<http://rucdmc.net/>

2014.10

## 0. 写在前面

时隔半年，我们中国人民大学数据挖掘中心整理以往提供的 linux 教程以及学生们的实际使用 linux 中获得的经验与教训，发布了全新的基于 ubuntu 的 linux 自学教程。本教程旨在实现一下目的：

- 1) 提供 linux 完全零基础的程序初学者一个快速了解 linux 入手 linux 的通道；
- 2) 为想要学习使用 linux 进行数据清洗的数据分析人员一份实用简单的学习方案；
- 3) 整理并推荐优质的 linux 学习资料；
- 4) 给希望提高使用 shell 进行数据处理的学者一份 linux 练习资料。

为避免版权问题，本教程的内容多以链接的形式进行展现。需要注意的是，本教程主要供统计数据分析人员学习 linux 并使用 linux 进行数据分析，对于 linux 系统的网络配置等内容涉及较少，还有待扩充。希望通过我们和您的共同努力，能够实现您 linux 操作技术的提升。只要我们的努力能够令您得到哪怕一点点的收获，我们也会感到欣慰。最后，如果您对我们的教程有任何的意见和建议，都可以随时通过我们的网站（<http://rucdmc.net/>）联系我们，我们非常欢迎您的加入！

## 1. linux 简介与 linux 的安装 ubuntu

### 1.1 linux 简介：

- 1) 以下博客为比较精简的 linux 介绍：

<http://blog.csdn.net/furongkang/article/details/6974204?>

- 2) 稍微更加详细的可以参见：百度百科：linux

- 3) 作为初学者可以翻阅：

<http://wenku.baidu.com/view/c639e805cc175527072208c3.html?>

（十分入门的 linux 读物，比较详细也简单的讲述了什么是 linux 以及 linux 的好处，包含很多使用者的感悟与建议，适合快速阅读，了解 linux。）

## 1.2 Ubuntu 的安装

在用户已经安装 windows 系统的情况下，这里我们介绍以下几种非光盘的安装方法，也是广大学生比较常用的方法。

### 1.2.1 预备知识：Linux 文件格式简介及系统分区

在安装 Ubuntu 的过程中，我们会遇到建立系统分区的问题，在此摘选对 linux 文件存储格式的介绍（分区内容仅供参考，具体情况要按照安装方法以及是否是双系统来决定）：

<http://www.cnblogs.com/gylei/archive/2011/12/04/2275987.html>

注意：对于安装双系统用户，要事先准备好给 linux 进行安装的空间，在系统分区的一步中要十分的小心，否则会影响 windows 系统下的文件以及使用。

### 1.2.2 虚拟机安装方法

<http://wenku.baidu.com/view/1dad63dbce2f0066f53322b7.html?>

<http://wenku.baidu.com/view/d8888f33376baf1ffc4fad4a.html?>

### 1.2.2 U 盘安装方法

<http://wenku.baidu.com/view/6835607a01f69e314332943b.html?>

<http://blog.csdn.net/dreamthen/article/details/8765472>

### 1.2.3 Wubi 安装

<http://wenku.baidu.com/view/57fb1b6925c52cc58bd6be82.html?>

该文件为 linux 的一份基础教程，在教程的最开头详细介绍了如何使用 Wubi 进行 linux 安装的方法，以及 linux 文件格式和分区的内容。

## 1.3 linux 系统学习的经典教程:

鸟哥的 Linux 私房菜 基础学习篇(第三版)（这本书网上有很多电子版下载）

<http://zhidao.baidu.com/share/edc597271c68737eacfcfd737a454eb.html>

[http://www.7edown.com/soft/down/soft\\_24490.html](http://www.7edown.com/soft/down/soft_24490.html)

《Ubuntu 部落：初学者推荐用书》一本整理的不错的 ubuntu 教材，很适合初学者，网络上同样可以找到。

[http://wenku.baidu.com/link?url=36lursVTN0H-iQsB90ZunDXsHyQetpiGkdaKOGLI4bqptqeSW\\_0roGPDTUDifSsGRUpRCyYAHaNIw3LZzJBK8Twail61M0LwOzibSHWIFzq](http://wenku.baidu.com/link?url=36lursVTN0H-iQsB90ZunDXsHyQetpiGkdaKOGLI4bqptqeSW_0roGPDTUDifSsGRUpRCyYAHaNIw3LZzJBK8Twail61M0LwOzibSHWIFzq)

## 2. Linux 常用系统命令

### 2.1 系统的使用

#### 2.1.1 系统的登录和退出

由于 Linux 是一个多用户的操作系统，使用时需首先登录系统，使系统识别自己。Linux 允许用户随时修改自己的口令，修改口令是 `passwd`。退出系统口令是 `exit`

操作命令：

`-shutdown` 系统关机。`-r` 关机后重启,`-h` 关机后不重新启动, `-now` 立即关机

`--halt` 关机后关闭电源

`--reboot` 重新启动

#### 2.1.2 简单介绍图形界面 XWindow

Xwindow 有悠久的历史 and 传统。Xwindow 和 Xbox 中的“X”本意是不同的, X 只是 W 后的一个字母,差不多应该这样理解, Xwindow 是 Window 的接班人 (注意,Window 不是 Windows)。Xwindow 使用服务器-客户端架构。无论本地图形界面,还是远程图形界面,都以同样的流程工作。这样便不需要分别进行设计和维护,极大的提高了网络透明性。

本地 X 客户端

└ 键盘

远程 X 客户端 + X 协议 — X 服务器 — 硬件规范 + 鼠标

远程 X 客户端

└ 显示器

#### 2.1.3 linux 的路径

熟知 Linux 中的路径是进行 linux 命令操作的基础，一下文章摘自《Ubuntu 部落：

初学者推荐用书》:

路径分为绝对路径和相对路径。绝对路径的起始点为根目录 `/`，例如 `/usr/local/bin` 就是绝对路径，它指向系统中一个绝对的位置，不受其它因素影响。相对路径的起始点为当前目录，如果您现在位于 `/usr` 目录，那么相对路径 `local/bin` 所指示的位置为 `/usr/local/bin` 也就是说，相对路径所指示的位置，除了相对路径本身，还受到当前位置的影响。例如 Linux 系统中常见的目录 `/bin`、`/usr/bin`、`/usr/local/bin`，如果只有一个相对路径 `bin`，那么它指示的位置可能上面三个目录中的任意一个，也可能是其它目录。

如果我告诉您到 `bin` 目录寻找一个文件，您可能搞不清楚是哪一个 `bin` 目录。只有当前位置确定，相对路径指示的位置才能够确定。现在我说，`/usr/local` 目录下，它的相对路径 `bin` 中有某个文件，这样就比较明确了。在相对路径中 `.` 表示当前目录，`..` 表示当前目录的上一级目录。

假设您安装了一个程序，它的主程序没有被放置到上面三个 `bin` 目录中的任何一个，或者其它系统能够找到的地方，您就得告诉系统，它的可执行文件在哪里。可以使用绝对路径，例如：`/home/user/bin/可执行文件` 或者定位到 `/home/user/bin` 目录，使用相对目录来定位它 `./可执行文件`

如果您定位到了它的子目录，比如 `/home/user/bin/gui`，您可以使用 `..` 来表示它的上级目录 `../可执行文件`

有关路径的命令:

`cd (change directory)` 更改目录。

`pwd (print working directory)`显示当前路径。

`ls (list)` 显示当前目录中的文件列表。

请尝试以下操作:

`cd /etc` 进入 `“/etc”` 目录，这里使用的是绝对路径

`pwd` 显示当前路径，这个命令返回结果 `“/etc”`

`cd init.d` 进入 `“/etc”` 目录的子目录 `“init.d”`，这里使用的是相对路径

`cd ..` 进入上一级目录 `“/etc”`

`cd ../home` `“/etc”` 目录的上一级目录为 `“/”`，它的子目录 `“home”` 为 `“/home”`

`cd -` 回到上一次的目录，我们在 `“/etc”` 目录跳转到 `“/home”` 目录，所以这次是回到 `“/etc”` 目录

`cd ~` “~”代表当前用户的“\$HOME”目录，即“/home/{用户名}”目录。

`ls` 在任何时候您都可以使用“ls”命令，来了解当前目录下有哪些文件。

## 2.2 常用命令

Linux 的常用命令一般包括系统管理命令与文件目标操作命令两部分，这些命令往往都由单行代码实现，实现的语法规则一般为：

命令 -可选参数 命令对象

其中可选参数前面有一个“-”，如果没有或者不需要可选参数，那么语法规则就是：

命令 命令对象

在最开始学习这些命令的时候最好使用一本配图充分的书籍教材进行仿写，当熟悉了 linux 命令的语法之后看各种网络连接的时候就不会产生疑问了。

以下的三个网页链接给出了 linux 常用指令的列表，包含指令的名称、作用以及一些可选参数，大家可以进行逐条学习。由于每个命令都包含可能很多的可选参数，所以下面链接中的命令的可选参数可能不全，在大家实际使用某个函数的时候可以自行 baidu 搜索命令的可选参数，这里就不进行赘述了。

<http://blog.csdn.net/furongkang/article/details/6974468?>

<http://www.cnblogs.com/qg78292959/archive/2011/06/10/2077866.html>

[http://wenku.baidu.com/link?url=Jhi-kW5ro61mobGxn32mVc9S\\_pWmAk574gfQ5HBOXfK\\_EUEWt7OdJ6N15lgi9N6ly6xAyFv7eaHHZ3FtW\\_2NEmzU1XlrNbL-BbMGEa1crca](http://wenku.baidu.com/link?url=Jhi-kW5ro61mobGxn32mVc9S_pWmAk574gfQ5HBOXfK_EUEWt7OdJ6N15lgi9N6ly6xAyFv7eaHHZ3FtW_2NEmzU1XlrNbL-BbMGEa1crca)

## 2.3 linux 中使用 Python

一般的 Ubuntu 系统中都预装了 Python，其二进制可执行文件通常安装在 /usr/bin/python 目录中，如果运行它，就进入了 Python 的交互式解释器，可以往其中输入命令。和 Perl、PHP 等其他脚本语言一样，也可以添加 shebang 行（#!），指定用 /usr/bin/python 执行脚本文件。

特别地，在终端中直接执行 Python 命令后即进入 Python 的交互式解释器。（与 R 类似）

如果要编写 Python 脚本，可使用 vim 或 emacs 编辑器编辑 py 文件，并使用 Python 的脚本文件运行。编辑器的使用请参见后文。



## 2.4 常用快捷键

- 1、超级键（Win 键） - 打开 dash。
- 2、长按超级键 - 启动 Launcher。并快捷键列表。
- 3、按住超级键，再按 1,2,3 等数字键 - 从 Launcher 打开一个应用程序，当你按住超级键时，每个应用程序图标上都会显示一个数字，按下对应的数字就会打开应用程序。
- 4、超级键+A - 从 Launcher 打开应用程序窗口。
- 5、超级键+F - 从 launcher 打开文件和文件夹窗口。
- 6、超级键+M - 从 launcher 打开音乐窗口。
- 7、超级键+V - 从 Launcher 打开 Vedio 视频窗口。
- 8、超级键+W - 伸展模式，缩小所有工作空间中的窗口。
- 9、F10 - 打开顶部面板的第一个菜单，使用箭头键浏览菜单项。
- 10、超级键+T - 打开回收站。
- 11、超级键+S - Expo 模式，缩小所有工作空间，允许你管理窗口。
- 12、Ctrl+Alt+T - 启动 Terminal。
- 13、Ctrl+Alt+L - 锁住屏幕。
- 14、Ctrl+Alt+上/下/左/右键 - 移动到新的工作空间。
- 15、Ctrl+Alt+Shift+上/下/左/右键 - 将窗口放入新的工作空间。
- 16、Ctrl+Super 快捷键：
- 17、Ctrl+Super+Up 键：最大化当前窗口
- 18、Ctrl+Super+Down 键：最小化当前窗口
- 19、Ctrl+Super+D：最小化所有窗口
- 20、Ctrl+Super+Left：半最大化当前窗口（左边）
- 21、Ctrl+Super+Right：半最大化当前窗口（右边）
- 21、Alt+F1 - 将键盘焦点移到 Launcher 上，使用箭头键进行移动，按回车即可启动一个应用程序，按下右箭头键显示 Quicklist。
- 22、Alt+F2 - 以特殊模式打开 dash，以便运行任何命令。
- 23、Alt+F10 - 在最大化/非最大化之间切换当前窗口。
- 24、Alt+F9 - 最小化当前窗口。
- 25、Alt+Tab - 在当前打开的窗口之间切换。

26、Alt+F4 - 关闭当前窗口。

27、Alt+F7 - 移动当前窗口（可以使用键盘或鼠标移动）。

## 3. Linux shell 文本处理命令

Linux shell 命令有着强大的文本处理功能，而且往往能够利用一行的代码进行很复杂的操作。该语句的入门学习并不难，只需要理解 shell 语句执行的一般过程，加上学会几个常用的强大的函数命令的使用方法就可以轻松解决一些在 window 系统下难以处理的问题。

对于 shell 语言的执行，可以直接在 terminal（快捷键 ctrl+alt+t）下进行输入，也可以先建立脚本再进行脚本的执行。在学习 shell 的过程中强烈建议从一本介绍全面的入门级别的基础用书开始，先学习 shell 的基本语法，在理解基础的过程下对于更加复杂的文本处理过程只需搜索各种资料（如 baidu 还有下面提供的资料）就可以得到答案。

推荐学习用书：

1、shell 从入门到精通 张春晓等 清华大学出版社。

（这是一本比较详细的 shell 语言学习教程，内容丰富配图多，同时配有教学视频，通过该书完全可以做到 shell 的入门与自学。也是本人实际自学过的一本书，非常抱歉没有找到电子版）

2、Shell 脚本学习指南

（非常经典的 shell 学习用书，具有第一本没有的一些知识，相对来讲配图稍少一些，但是讲解也很清楚。）

下载链接：链接：<http://pan.baidu.com/s/1gdDwFCJ> 密码：gy3a

### 3.1 shell 基本语法

[http://blog.sina.com.cn/s/blog\\_5b1acf750101g9gn.html](http://blog.sina.com.cn/s/blog_5b1acf750101g9gn.html)

（很好的一篇博客，首先列出了 linux 常用的 shell 语句，然后对于 shell 的基本语法，即变量，运算，结构都进行了详细的介绍与并提供举例。）



<http://www.doc88.com/p-5768780642620.html>

(对 linux 中的 shell 的结构语法讲解的很全面详细。)

<http://www.cnblogs.com/fhefh/archive/2011/04/13/2014967.html>

(对于 shell 语句中的结构（主要是 if 结构）进行了比较详细的解释与举例。)

## 3.2 常用文本处理命令列表

在下面链接中出现的命令中，我们将在下面一个模块中提供比较重要的命令的讲解文件，而对于其他命令，大家可以根据自己的需求直接进行 baidu 就能够很轻松的得到详细的使用方法。

<http://blog.csdn.net/forgotaboutgirl/article/details/6801525>

<http://os.51cto.com/art/201310/414325.htm>

## 3.3 重要命令详细解释

正则表达式 30 分钟入门教程: [http://www.oschina.net/question/12\\_9507](http://www.oschina.net/question/12_9507)

(仅仅作作为一个入门教程，从讲解来讲不够体系。)

正则表达式、Grep、awk、sed、sort、tr 的使用见附件:

链接: <http://pan.baidu.com/s/1i3mTYvj> 密码: wygg


# 4. VI 和 Emacs 编辑器简介:

## 4.1 VI 编辑器

vi 三种工作模式:

**【命令模式】:** 启动 vi 后进入的工作模式

**【文本输入模式】:** 命令模式下输入 i (插入命令), a (附加命令), o (打开命令) 或 s (替代命令) 等后进入文本输入模式; 按 ESC 键返回命令模式

**【最后行模式】:** 命令模式下按 “:” 键进入; 按 DEL 键或  键删除输入的命令回到命令模式; 多数文件管理命令是在最后行模式执行的

vi 的启动和退出:

【启动】 vi [目录] [文件]

【文本输入】 命令模式下输入 i (从当前光标位置开始输入), I (光标移动到当前行的行首), a (当前光标的下一个位置开始输入), A (光标移动到当前行行尾开始输入), o (光标所在行之下新增一行), O (光标所在行之上新增一行)

【打开文件】

:vi filename 打开或新建文件名为\*\*的文件到 vi 编辑器中,并将光标置于第一行首

:vi +n filename 打开文件, 并将光标置于第 n 行首

:vi + filename 打开文件, 并将光标置于最后一行首

:vi +/pattern filename 打开文件, 并将光标置于第一个与 pattern 匹配的串处

:vi filename...filename 打开多个文件, 依次进行编辑

:w 将工作缓冲区的变化写入默认文件中

:w filename 将工作缓冲区的变化写入名为\*\*的文件中

:e filename 打开文件 filename 进行编辑

:r filename 读取文件内容到当前 vi 编辑器

【保存文件】

:w 保存但不退出

:w [newfile] 保存为指定的文件

:ZZ 保存退出

:e 不保存当前修改, 回到初始版本文件

:q 退出不保存

:v 另存为文件 将 vi 编辑器中的内容另存为指定文件名

### 如果文件内容有改动需要使用命令:

:q! 不保存文件, 直接退出 vi

:wq (write quit) 存盘并退出 vi

其他可参考: <http://jingyan.baidu.com/article/9f63fb91c58387c8400f0eef.html>

## 4.2 EMACS 编辑器

参考: <http://www.cbi.pku.edu.cn/chinese/documents/csdoc/emacs/>

## 5. 实际练习

### 5.1 第一次任务：新浪 poi 数据的描述统计

POI 签到日志数据 <http://pan.baidu.com/s/1dD3pxNr>

字段：

```
["poiid","lat","lon","checkin_time","category_name","checkin_user_num","address","title","  
province","city"]
```

第一次任务主要使用 POI 数据，这个数据是一个超过千万行的微博签到数据。新浪微博用户每使用一次新浪微博的签到功能，就会返回一条数据记录，表现为我们数据中的一行。这次的任务主要是做描述统计，具体任务如下：

- 1、去掉没有记录的用户，也就是删掉像 “1906399327 NONE” 这样的行
- 2、查找所有城市在北京的行，并输出成文件。
- 3、生成一个无重复的用户 id 列表，这个列表里的用户在 poi 数据中有过记录，并统计有记录的用户个数
- 4、按以下顺序截取字段 “poiid” “city” “category\_name” ，字段之间用 “:” 分隔
- 5、保留 poiid 在一定范围内的字段（这个范围可同学们自己设定）

### 5.2 第二次任务：移动用户数据处理

此次数据是移动用户的资料、短信及通话清单的节选（数据量并不算太大）。下载地址为：

链接：<http://pan.baidu.com/s/1pJ2qZe7> 密码：xu3j

数据包中包含了对于字段的解释。下面进行以下几个任务

- 1、在短信信息数据中提取出接收号码为异网手机号码且接收号码运营商为移动的发短信数据。
- 2、汇总所有主叫号码的通话次数形成一个文件 call.txt。(两列：主叫号码；通话次数)
- 3、汇总所有发送号码的通话次数形成一个文件 message.txt。(两列：发送号码；发送次

数)

4、根据第 2 步和第 3 步生成的新文件中的主叫号码与发送号码为连接变量进行两个文件的连接。有两种不同的要求：

- 1) 对于无法连接的号码（即没有主叫或者发送的号码），直接删除该号码的数据
- 2) 对于无法连接的号码，依旧保存该号码，认为其通话次数或发送次数为 0。

