Lecture 8

CSS 507 Data Collection, Wrangling, Analysis and Visualization

By: PhD, Andrey Bogdanchikov
Edited by: MSc, Elnura Nabigazinova

Content

Introduction

Unix shell have very powerful tools to manipulate text data. So it is easily accessible and no need for installation.

Unix shell pipelines are very useful to combine multiple commands into a series of transformations. We are going to see how to use this commands to extract data from archives, clean and transform to suitable formats.

Commands

Unix shell have lots of powerful commands that work with text data:

- tar
- gzip
- grep
- head
- tail
- cat
- cut
- sort
- sed
- awk
- wc
- uniq
- split
- join
- and many other ...

Compression

Sometimes it is very useful to compress data into archives for better mobility of the data. So most famous command to create and extract archives in Unix is tar:

- tar -cvf a.tar /etc
  Create an archive file without compression
- tar -cvfz a.tar.gz /etc
  Create an archive file with gzip compression
- tar -cvfj a.tar.bz /etc
  Create an archive file with bzip2 compression
- tar -xvfz a.tar.gz
  Uncompress and extract a gzip compressed archive

File inspection

Useful commands to inspect content of the file is:

- cat - concatenate files and print on the standard output
- head - output the first part of files
- tail - output the last part of files
- more - file perusal filter for crt viewing
- less - Less is a program similar to **more**, but it has many more features.
- etc.

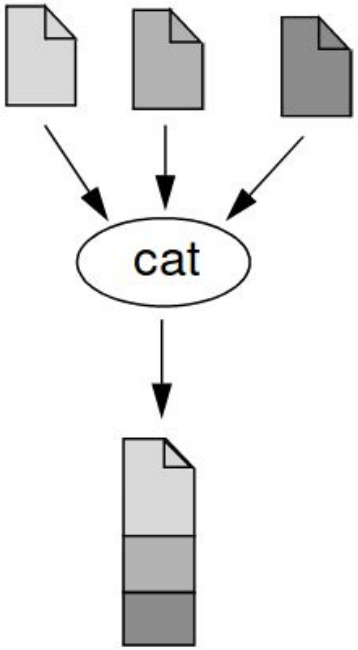
Search

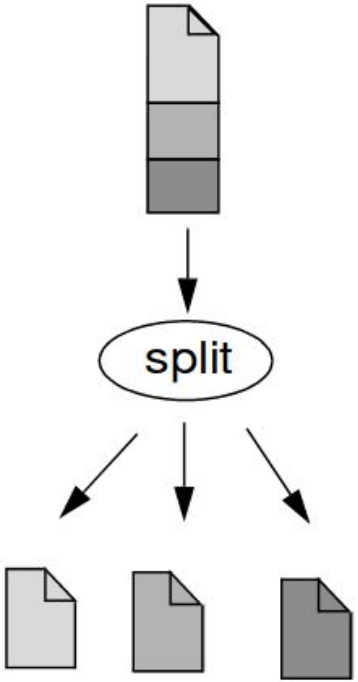
Commands to search for specific pattern in a text:

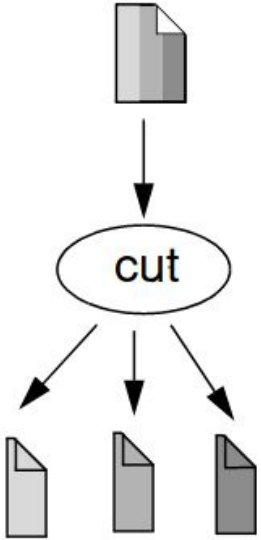
grep, egrep, fgrep, rgrep - print lines matching a pattern

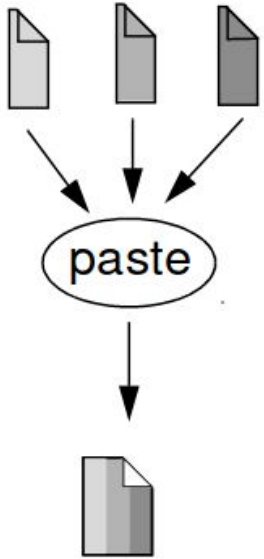
grep searches for PATTERN in each FILE. A FILE of "-" stands for standard input. If no FILE is given, recursive searches examine the working directory, and nonrecursive searches read standard input. By default, grep prints the matching lines.

Summary File operations









File transformation

To transform file there many commands:

- tr - translate or delete characters
- sort - sort lines of text files
- cut - remove sections from each line of files
- sed - stream editor for filtering and transforming text
- **awk**/mawk - pattern scanning and text processing language
- join - join lines of two files on a common field
- etc.

Practicing commands

Tutorial: Data Manipulation and Data Transformation using the Shell

To practice all these commands we will use tutorial provided by Andreas Schmidt, Steffen Scholz. Which were presented on The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA-2017)

- https://www.iaria.org/conferences2017/filesDBKDA17/tutorial_bash_data_handling.pdf
- http://www.smiffy.de/dbkda-2017/

Conclusion

Check following links for info:

- Info about Unix shell - http://hpc.ilri.cgiar.org/beca/training/ilri_addis/unix_linux_and_simple_tools_Ethiopia2017.pdf
- Tutorial: Data Manipulation and Data Transformation using the Shell, by Andreas Schmidt, Steffen Scholz - https://www.iaria.org/conferences2017/filesDBKDA17/tutorial_bash_data_handling.pdf

Assignment 6

For this lab you should perform exercises from the Tutorial and submit screenshot of the results:

- Exercise I (First contact) (50pt)
- Exercise II (Text processing) (50pt)
- Exercise III (sed & awk) (bonus 50pt)

Thank You