In this Lab, we explore how “sed” is used for data preparation. We take the divide and conquer approach. The commands are simple, but when works together, it can solve a complex task.

What is “sed”? Let’s watch this video: “SED Tutorial Basic Substitution Linux Shell BASH”

(Source: <https://www.youtube.com/watch?v=32waL1Z9XK0> )

**Lab Introduction:**

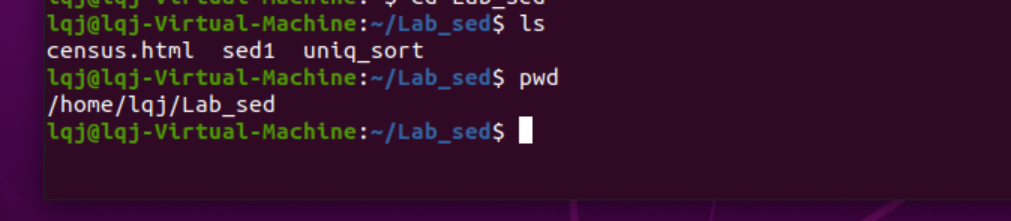
1. In this Lab folder, you will see the following files.

* “census.html”: This is the raw html file (US census population estimate website).
* “sed1”: An example of “sed” command file.

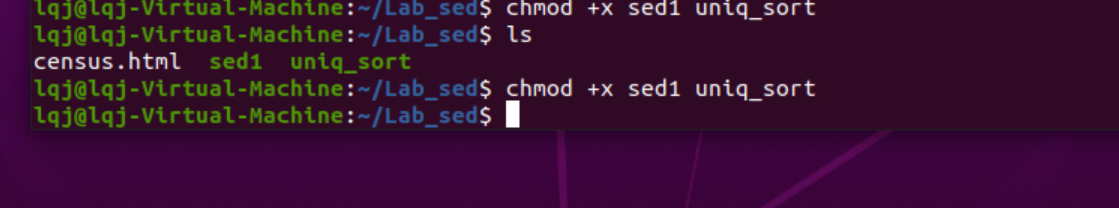
1. The objective of the scripts is to extract all the web-links found in “census.html”, remove redundancy, and presented the results in sorted order.
2. Please input your results to questions found in this document (Screen shot captures are helpful).

**Execution Steps:**

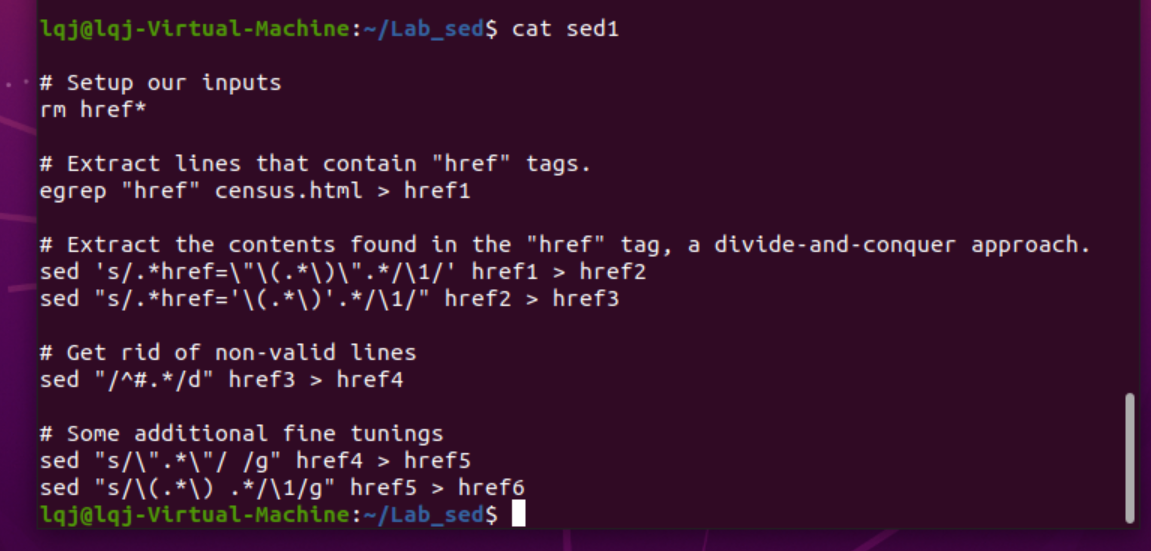
1. Download the Lab assignment and launch you Linux session. An example is show below.



1. Make the bash script files executable. An example is show below.



1. Learn the “sed” command by examining the script “sed1” and execute its codes 1-step at a time.



* rm href\*

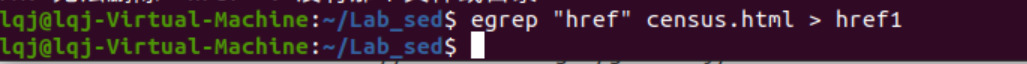
*What does “\*” stand for?*

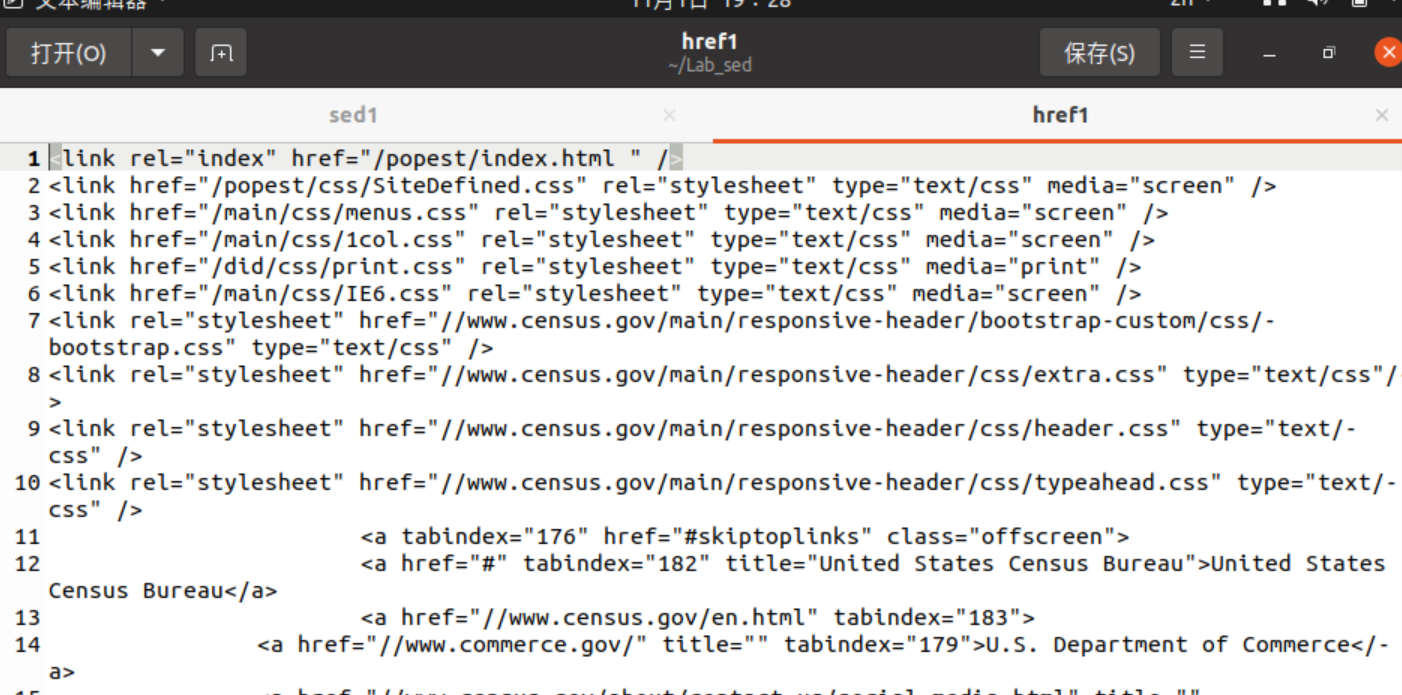
*All files in the directory.*

* Please run the sed command that creates “href1” file.

*What does the “egrep” command do?*

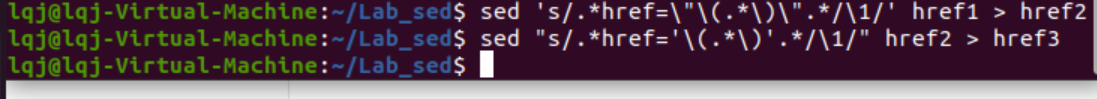
*Find out the specified string in a file*





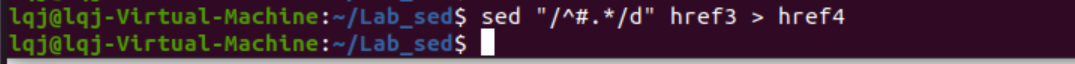
* Please run the sed command that creates “href2” and “href3” files.

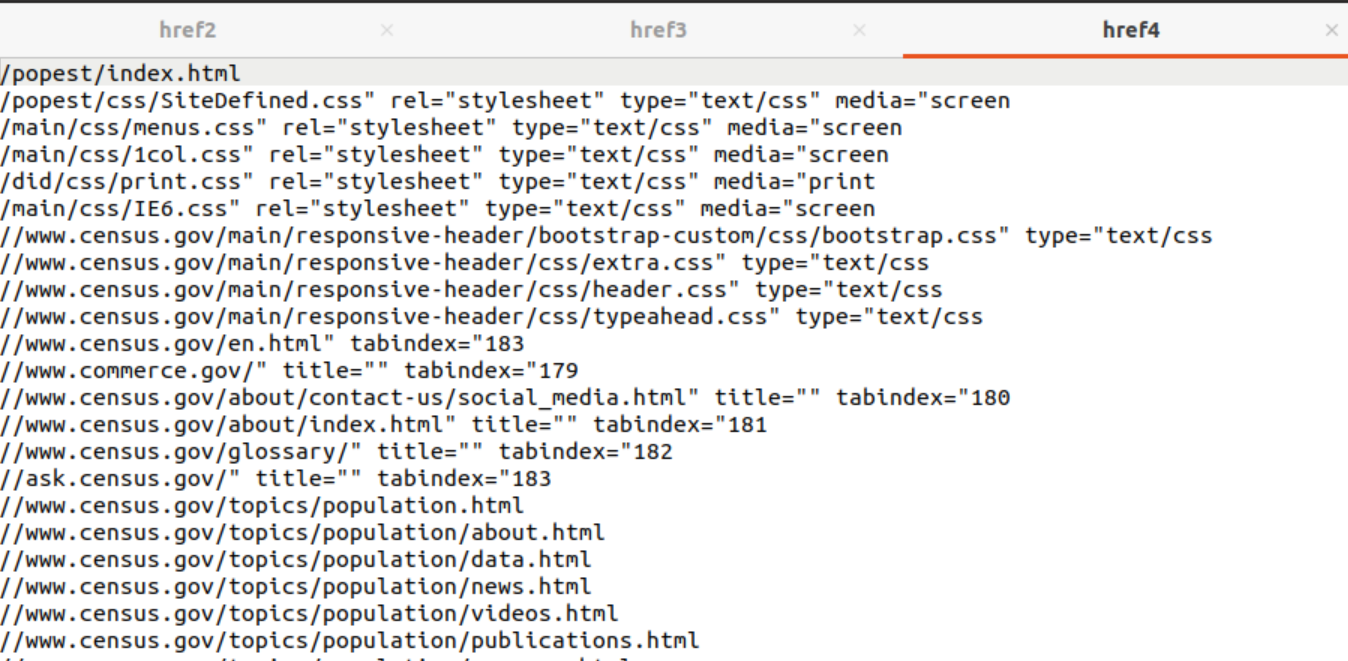
*What are the differences between the two files?*

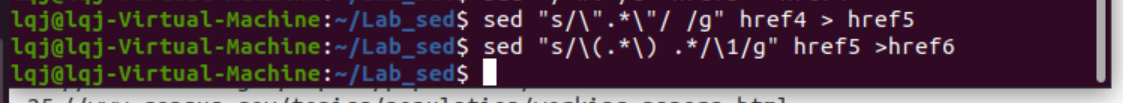


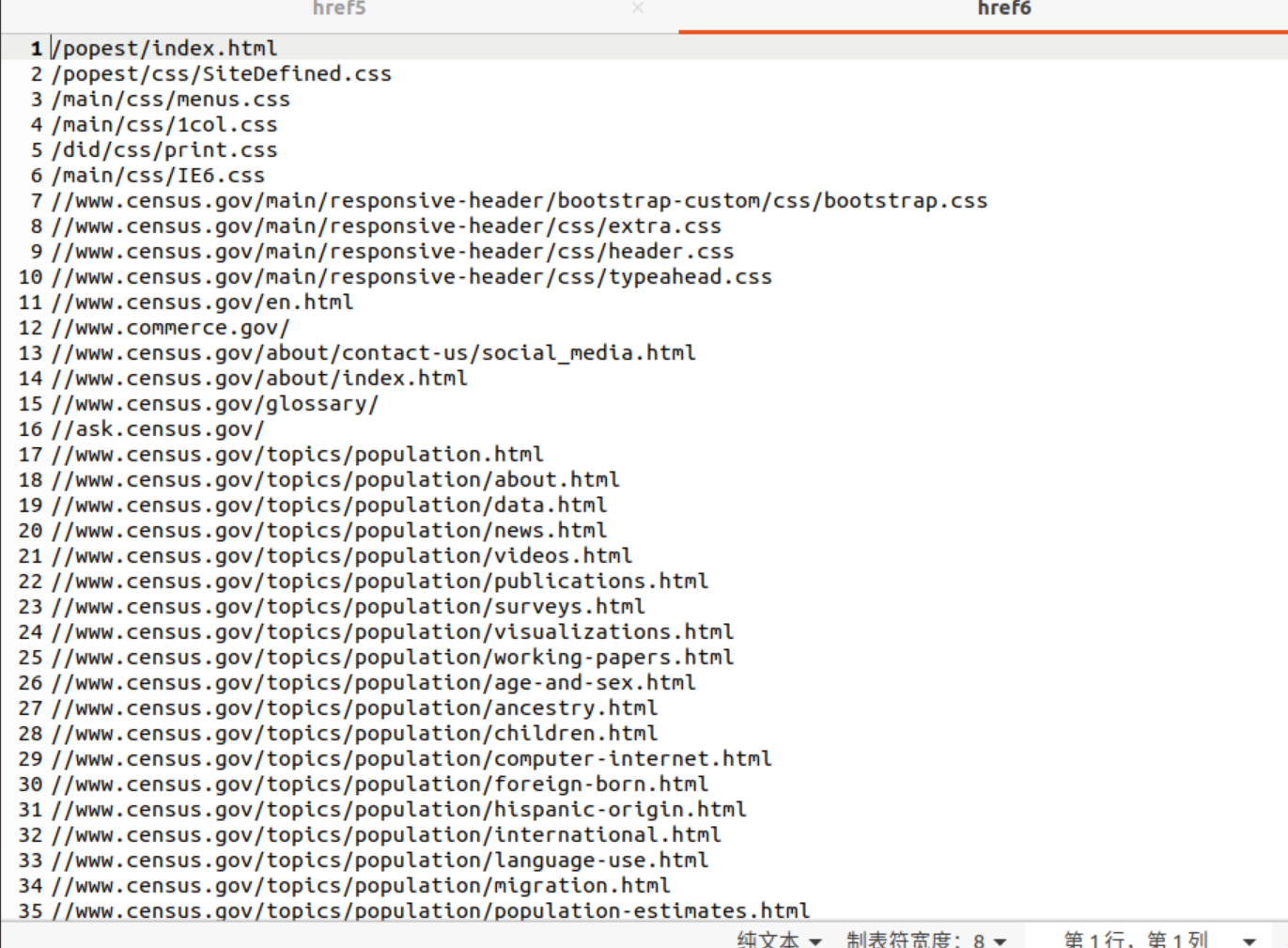
*Href3 gets the URL on the right of "/ /" based on Href2*

* Please run the sed command that creates “href4”, “href5”, “href6” files.

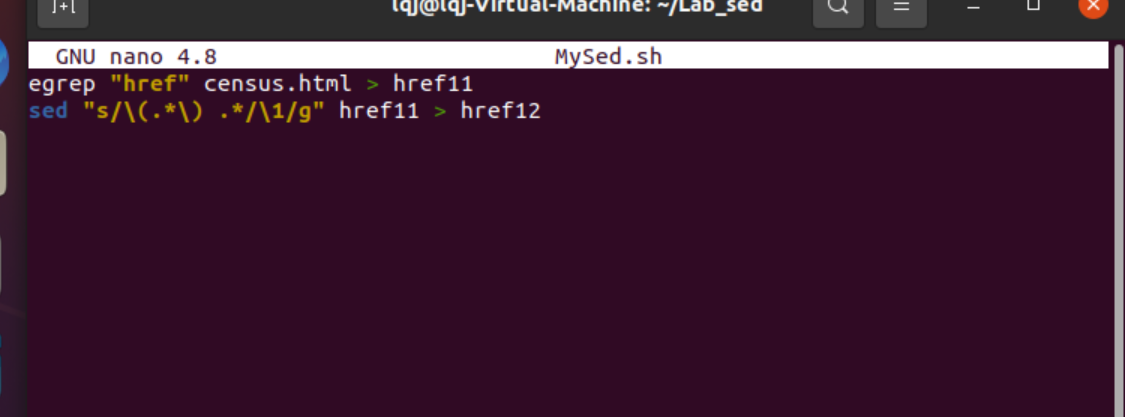


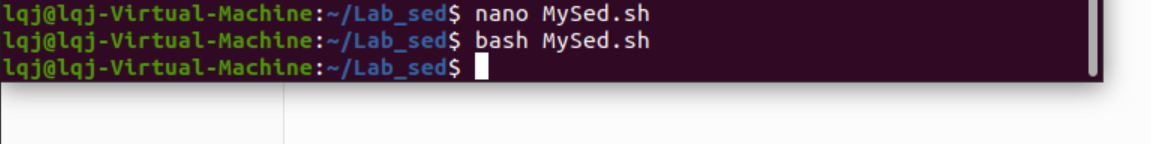


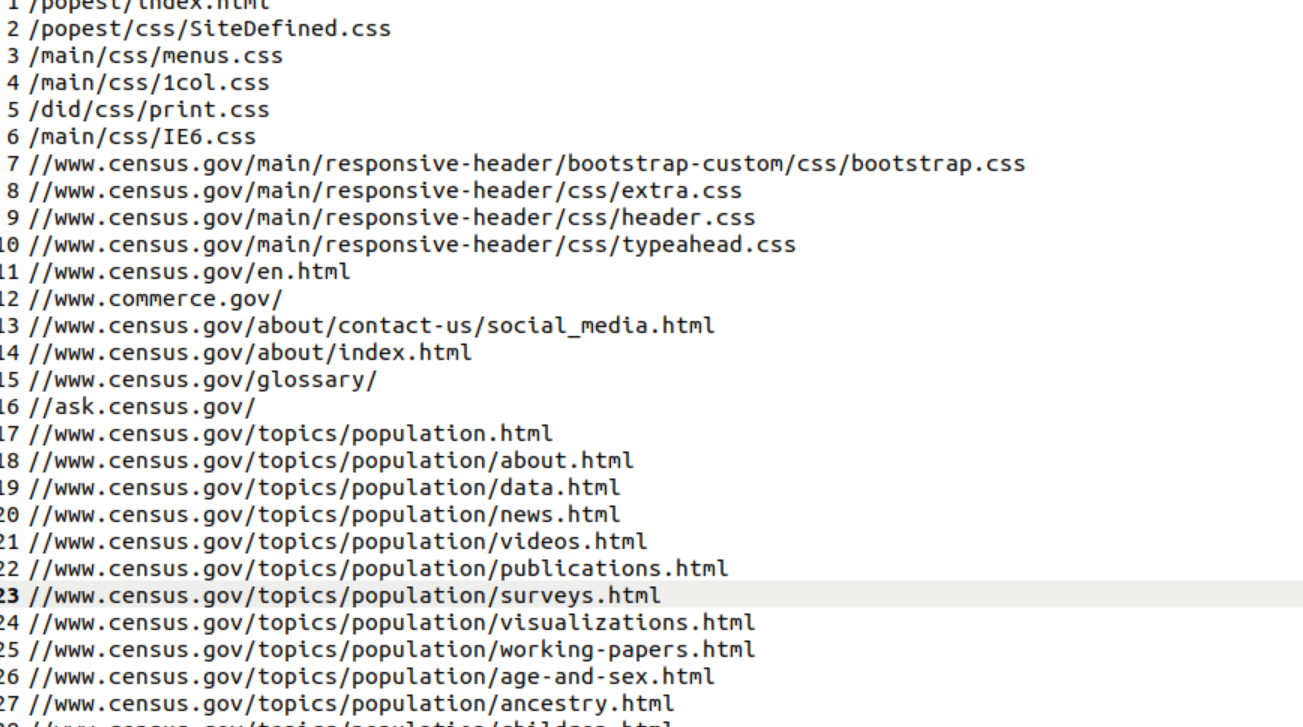




1. Make enhancements to the existing script “sed1”: Combine all the “sed” commands (seen in the existing script, there are 4+ of it), into 2 or less “sed” commands. Please name your enhanced script to “MySed.sh”. Show your results on the space below.





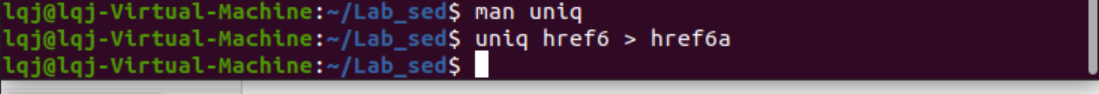


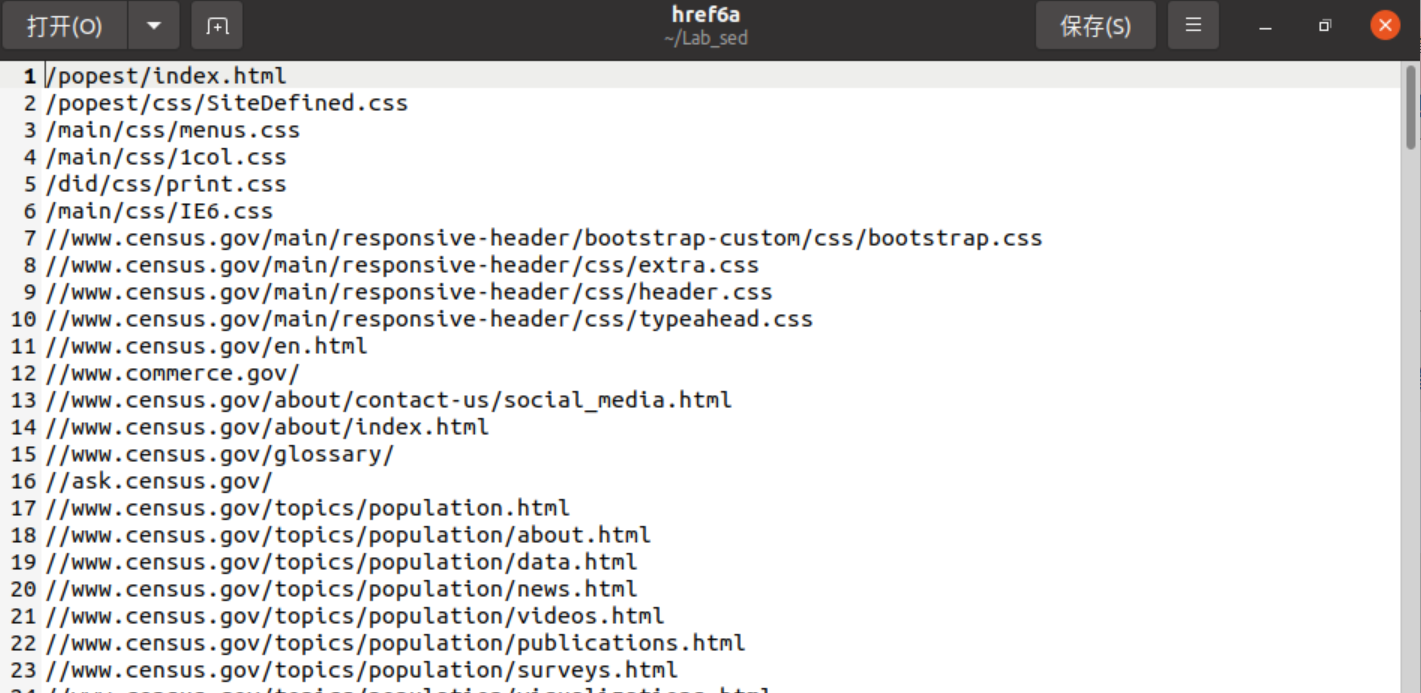
1. Learn the “uniq” command, and run it:

* man uniq
* uniq href6 > href6a

*Please explain the difference between “href6” and “href6a”?*

*Href6a remove duplicate lines from href6.(repeated lines must be continus)*





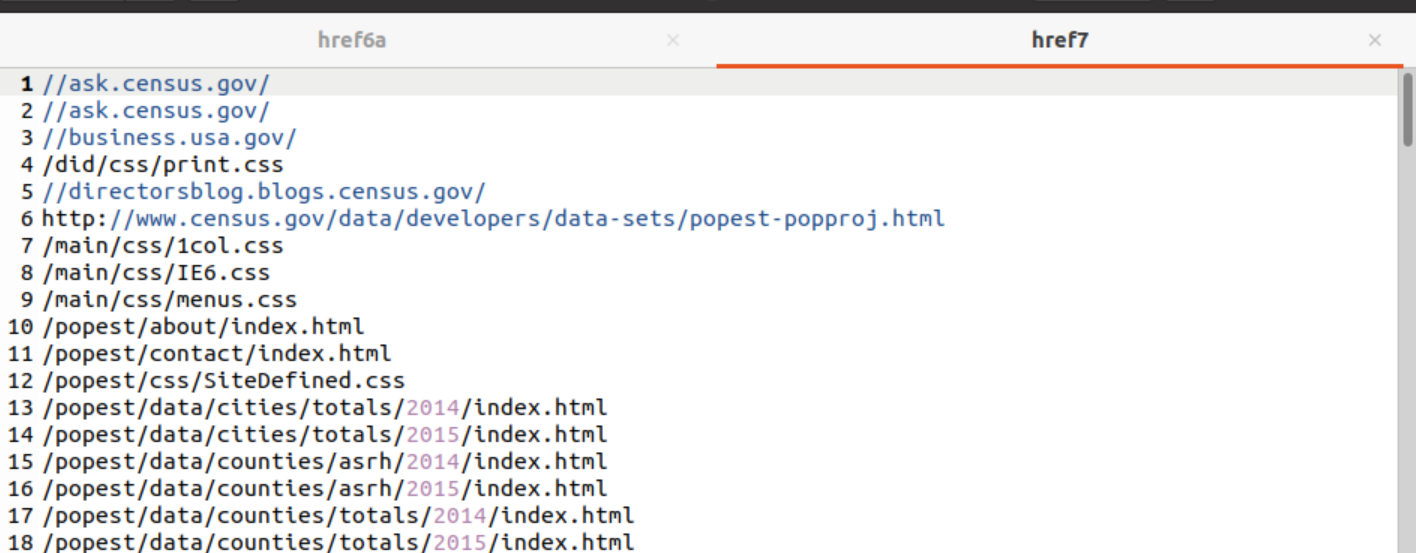
1. Learn the “sort/history” command, and sort it:

* man sort
* sort href6a > href7

*Please explain the difference between “href7” and “href6a”?*

*Href7 takes each line of href6a as a unit 、compares it and order the list alphabetically according to initial word.*

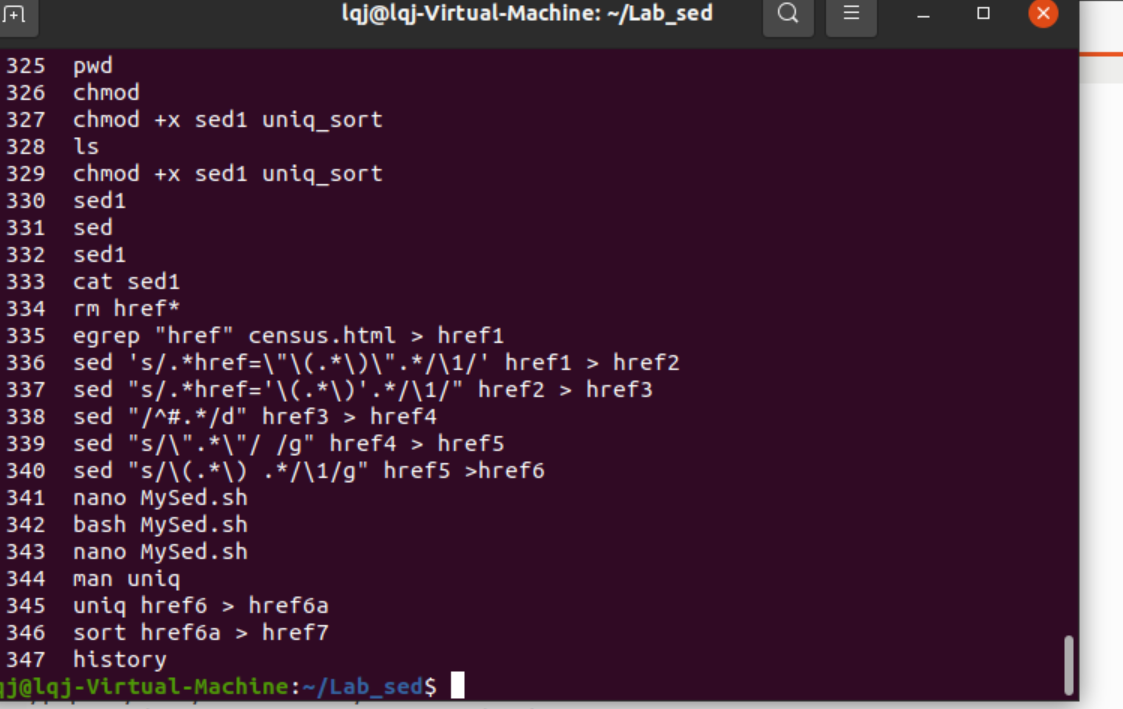




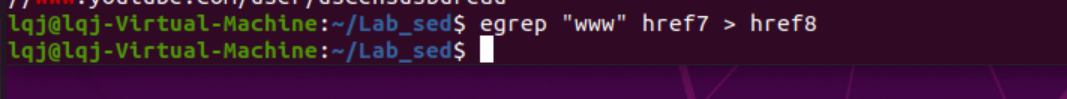
* history

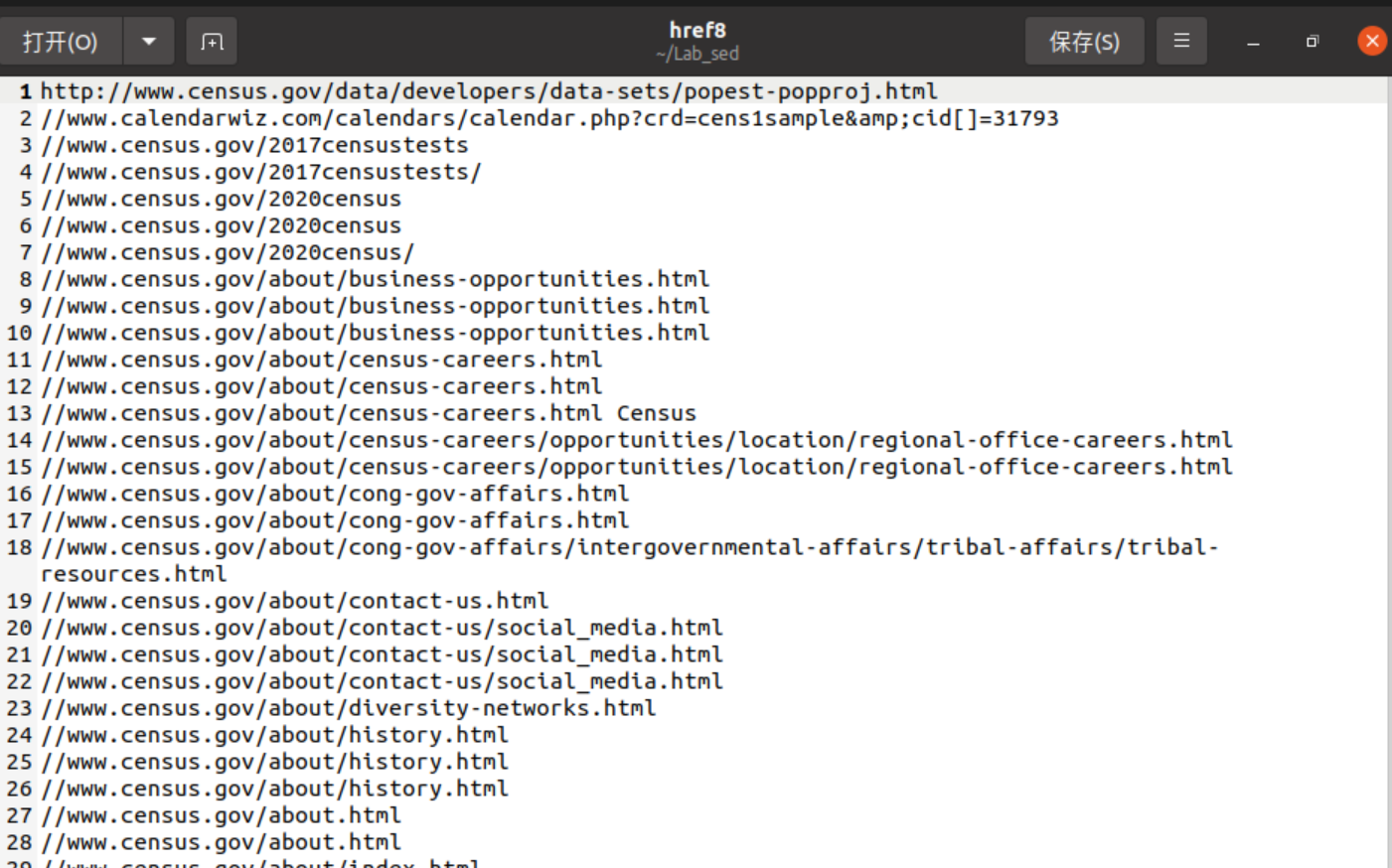
*What does the “history” command do?*

View your input command history



1. Refine the results: Please extract results of the last step above to contain entries with “www” prefix only (Hint: grep/egrep)?

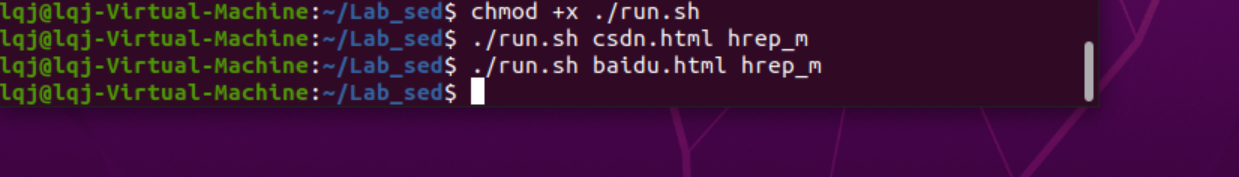




1. Create a “run” script that integrate all the tasks? Example on the script invocation:

./run my\_input.html my\_weblinks

* “my\_input.html” is the html file of the web site you visited most often (in other words, please select your favorite web portal, and save its html file as input to your script).
* “my\_weblinks” in the output/results of the script.



**Lab Result Submission:**

Please submit to online. If that fails, email your results to: [bangpanliang@gmail.com](mailto:jcc4018@qq.com).

The subject of the email should be:

[Your StudentID, Assignment Name]