Assignment 2: Text-processing

Task

Take copies of the data files as directed. Work through the questions deciding how to extract the relevant information. If it is not possible because the test-case asked for in the question does not occur in your data - work out how to create it and take new copies of the data files including the test-case. At the time of submission the data in the saved files must be in synchronisation with the answers in the report. You should check this using techniques described in the lectures.

Description

This assignment tests your knowlege of basic information sources and your ability to build pipelines that extract information.

Each question is testing your ability to perform the same task - identify which parts of the information in a file satisfies the question, work out how to extract only that information using the unix tools, and build a report containing answers that match the output of your commands. By doing this you are demonstrating knowledge of the course materials (i.e. the function of commands, and the content of the various files), and showing that you can use the tools to match an external constraint (that the answers in the report syntax match the output of your commands). The constraint is designed to be challenging to achieve manually, but easy to achieve if you use the same tools that you are being tested on and diff as a form of

verification. There are some hints in the lecture slides about how to automate putting the answers into the report. It is highly recommended that you use them.

Requirements

- Your submission file must be a valid gzipped tarball.
- The report may be either 7-bit clean ASCII or UTF-8 format
- If the report is UTF-8 then the machine-readable answers in part-a and part-b of each question must be 7-bit clean. Look these terms up if they are unfamiliar.
- The tarball must contain the path "report.txt" (no directories).
- The tarball must contain the path "pscopy.txt" (generated from "ps aux").
- The tarball must contain the path "netcopy.txt" (generated from "netstat -tWae").
- The tarball must contain the path "dircopy.txt" (generated from "ls -al").
- The tarball must contain the path "histcopy.txt" (generated from "history"). Please note the comments in lecture 3 about the minimum complexity of command-lines in this file.
- Part-c is a human-readable comment field, please use it if you want to write a message about the answer that is not meant to be fed into the shell in the testing machine - I will read these comments when I grade the work.
- If you used any sources of information external to the course then you must list them in the part-c comment, including URLs - and explain how the shell string you have used works.
- You may submit as many times as you want - only your most recent submission at the time of the deadline will be graded.
- After submission a validation page is available that will check that the machine-readable parts of your report make sense - a hex-dump is included in the validation page to help

you diagnose any issues with character sets and text editors
- When your shell command-lines are tested against the data files that you have included, they must regenerate the answers given in the report exactly. You should test that this happens before the submission deadline by extracting the answers, rerunning the shell commands and using diff to check they are the same.
- Each question asks you to identify particular records, or fields, in one of the data files - your submitted answer must match this data exactly to gain the point.

Percentage Correct	Grade
≥ 90%	A
≥ 80%	B
≥ 70%	C
≥ 60%	D
≥ 50%	E
< 50%	F

Examples

Q1 What is the set of owners for entries in your home directory?

1a Answer (multi/text)

1b Command-line use to find answer

1c Comments.

Contents of report.txt

```
1a amoss
1a root
1a backup
1b cat dircopy.txt | awk '{print $3}' | sort -u
1c The line format seems to use fixed positions for fields on my system
1c but the field extraction in awk seemed more robust.
```

Questions / Tasks

Q	Cropping a range from the history	1.0 points
1a	What are records 6-12 in the history? (count using the line-numbers and not the indices)	
1b	What shell command-line did you use?	
1c	Optional: any comments on your approach (including any external sources) ?	

Q	Cropping a suffix from the history	1.0 points
2a	What are the last 7 records in the history?	
2b	What shell command-line did you use?	
2c	Optional: any comments on your approach (including any external sources) ?	

Q	Filtering the history by keyword	1.0 points
3a	What are the records in the history that include ls? (include cases where the string is included in other words)	
3b	What shell command-line did you use?	
3c	Optional: any comments on your approach (including any external sources) ?	

Q	Cropping a prefix from the history	1.0 points
4a	What are the first 8 records in the history?	
4b	What shell command-line did you use?	
4c	Optional: any comments on your approach (including any external sources) ?	

Q	Extracting one field from network-connection records filtered by another.	1.0 points
5a	What is the set of inodes used by network-connections belonging to the root user?	
5b	What shell command-line did you use?	
5c	Optional: any comments on your approach (including any external sources) ?	
Q	Extracting hour-numbers from a field in process records.	1.0 points
6a	What is the set of hour-numbers in the Start Time field for processes that started today? (note that processes started on other days will have a different format and should be excluded from the result)	
6b	What shell command-line did you use?	
6c	Optional: any comments on your approach (including any external sources) ?	
Q	Extracting one field from the listing records filtered by another.	1.0 points
7a	What is the value in the name field for the listing record with the highest value in the link counter field?	
7b	What shell command-line did you use?	
7c	Optional: any comments on your approach (including any external sources) ?	
Q	Finding which network-connection record has the largest value in a field.	1.0 points
8a	Which network-connection record has the largest Send-Q?	
8b	What shell command-line did you use?	
8c	Optional: any comments on your approach (including any external sources) ?	

Q	Extracting the fields from process records filtered by another.	1.0 points
9a	What is the set of numbers in the PID fields of process records that have no controlling terminal? (note this implies that neither a tty or a pty)	
9b	What shell command-line did you use?	
9c	Optional: any comments on your approach (including any external sources) ?	
Q	Extracing a field value from the listing record that has the smallest value in another field.	1.0 points
10a	What is the name of the smallest symbolic link in your directory listing?	
10b	What shell command-line did you use?	
10c	Optional: any comments on your approach (including any external sources) ?	
Q	Extract a set of field values from multi-level records with both fixed-width and variable-width structures.	1.0 points
11a	What is the set of arguments for all commands in the history? (assume arguments are space-separated so ignore quotes and escaping, must be all arguments not just switches, must interpret semicolons and pipes correctly).	
11b	What shell command-line did you use?	
11c	Optional: any comments on your approach (including any external sources) ?	
Q	Extract a variable-length record from a field in a fixed-width record and extract part of it.	1.0 points
12a	From process-records that list absolute paths to the running binary, what is the set of directories that contain those binaries?	

12b What shell command-line did you use?

12c Optional: any comments on your approach (including any external sources) ?