Unix/Linux Tutorial for Beginners
Session IV – Text processing

Mandatory exercises

1. If we run sort on a file with this content:

```
10
2
19
22
6
```

   the output is:

```
10
19
2
22
6
```

   If we run sort -n on the same input, we get this instead:

```
2
6
10
19
22
```

   Explain why -n has this effect. Which statement is correct?

   (a) The option *-n* enables the sorting of real numbers, while the default options sorts only integers.

   (b) By default the command *sort* assumes the input is a string. While sorting it will look only at the first character to sort the file and if the strings have the same first character it will sort accordingly to the second one. The *-n* option tells the command *sort* that the input is numerical and it should be sorted accordingly.

   (c) The command *sort* can't be used without additional options. The option *-n* tells the command how to handle the input.

2. The file ~/data/others/employees.txt contain information about each employees in your company. Use the sort command and the corresponding options to sort the list first by employees name, then by their age. Which command is correct?

(a) sort employees.txt | sort -n employees.txt

(b) sort -k1,1 -k4,4 employees.txt

(c) sort -k1,4 employees.txt

(d) sort -k1 employees.txt | sort -k4

3. Cut out the 4th column of the tab-delimited file ~/data/others/employees.txt and sort it to only show unique lines in reverse order. Which command is correct? Tab characters are special characters and unprintable on the command line. To access their literally value the tab character needs to be escaped. There are several mechanisms to escape special characters and assign their values to variables, like \ or $'...'.

   (a) cut -d '\t' -f 4 ~/data/others/employees.txt | sort -ru

   (b) cut -d $'\t' -f 4 ~/data/others/employees.txt | sort -ru

   (c) cut -d ' ' -f 4 ~/data/others/employees.txt | sort -ru

4. FASTA files are plain text files used in bioinformatics to store nucleotide or peptide sequences. FASTA files contain several sequence records, whereas a sequence record consists of a single description line, followed by line(s) of sequence data. The first character of the description line is always a greater-than '>' symbol, followed (without any empty spaces) by the gene/protein name. How many sequences can you count in the file ~/data/fasta/wheat_PEP.fa?

   (a) 31029 sequences

   (b) 172095 sequences

   (c) 803 sequences

5. How many lines contain the word Luve (case insensitive) in the file ~/data/poetry/rosesRobertBurns.txt? Which commands will not return the correct answer?

   (a) 4 → grep 'Luve' rosesRobertBurns.txt | wc -l

   (b) 7 → grep -i 'Luve' rosesRobertBurns.txt | wc -l

   (c) 7 → grep -ic 'Luve' rosesRobertBurns.txt

   (d) 3 → grep -c 'luve' rosesRobertBurns.txt

6. How many sequences in the file ~/data/fasta/barley_CDS.fa begin with a start codon (ATG)?

   (a) 26159 sequences

   (b) 24892 sequences

   (c) 2423 sequences

7. Change all occurences of 'chr' to 'Chrom' in the file ~/data/fasta/brachy_CDS.fa and write the changed output to a new file. Use therefore the sed command. Which one is correct?

   (a) grep ">" brachy_CDS.fa | sed 's/chr/Chrom/' > brachy_CDS_processed.fa

   (b) cat brachy_CDS.fa | tr 'chr' 'Chrom' > brachy_CDS_processed.fa

   (c) cat brachy_CDS.fa | sed 's/chr/Chrom/' > brachy_CDS_processed.fa

8. Construct a command, which extracts from the file ~/data/fasta/myCDS.fa all fasta headers (lines starting with '>'), removes the '>' symbol and the word *contig* from the beginning of the gene names, and finally sorts the gene names in descending order. Redirect the output to a file called geneList.txt..

Optional exercises

1. What is the difference between:

```
$ wc -l < salmon.txt
```

   and

```
$ wc -l salmon.txt
```

2. What is the difference between:

```
$ echo hello > tetsfile01.txt
```

   and:

```
$ echo hello >> testfile02.txt
```

   Hint: Try executing each command twice in a row and then examining the output files.

3. We want to find the 3 files in the folder ~/data/others/ which have the least number of lines. Which command listed below would work?

   (a) wc -l *.txt > sort -n > head -3

   (b) wc -l *.txt | sort -n | head 1-3

   (c) wc -l *.txt | head -3 | sort -n

   (d) wc -l *.txt | sort -n | head -3

4. The command uniq removes adjacent duplicated lines from its input. For example, if the file ~/data/others/salmon.txt contains:

```
coho
coho
steelhead
coho
steelhead
steelhead
```

then uniq salmon.txt produces:

```
coho
steelhead
coho
steelhead
```

Why do you think uniq removes only adjacent duplicated lines? What other command could you combine with it in a pipe to remove all duplicated lines?

5. The file called ~/data/others/animals.txt contains the following data:

```
2012-11-05,deer
2012-11-05,rabbit
2012-11-05,raccoon
2012-11-06,rabbit
2012-11-06,deer
2012-11-06,fox
2012-11-07,rabbit
2012-11-07,bear
```

What text passes through each of the pipes and the final redirect in the pipeline below?

```
$ cat animals.txt | head -5 | tail -3 | sort -r > final.txt
```

   (a) 2012-11-05,deer; 2012-11-05,rabbit; 2012-11-05,raccoon; 2012-11-06,rabbit; 2012-11-06,deer

   (b) 2012-11-06,rabbit; 2012-11-06,deer; 2012-11-05,raccoon

   (c) 2012-11-05,raccoon; 2012-11-06,deer; 2012-11-06,rabbit

6. Extract the first and last 15 lines from the file ~/data/fasta/wheat_PEP.fa and redirect the output to two separated files ~/myLinuxProject/result/courseDay_01/head_selection.fa and ~/myLinuxProject/result/courseDay_01/tail_selection.fa. Which commands and redirects would you use to redirect the output of both commands to only one file (selection.fa)?

7. The command:

```
$ cut -d , -f 2 ~/data/others/animals.txt
```

produces the following output:

```
deer
rabbit
raccoon
rabbit
deer
fox
rabbit
bear
```

What other command(s) could be added to this in a pipeline to find out what animals the file contains (without any duplicates in their names)?

(a) sort | uniq

(b) uniq

(c) sort -u

(d) sort

8. Which command and options can you use to find all lines in the file ~/data/poetry/rosesRobertBurns.txt that contain the word 'the' and to number the lines that match? Is there any change in the number of matches if you make your search case insensitive?

   (a) command: less, searching for '/the' and '/The'. The command visualize all matches. No match for the search pattern 'The'.

   (b) *grep -n -w 'the' rosesRobertBurns.txt* → the command returns 5 lines: line 3, 9, 11, 12, and 14. For the case insensitive search the option -i can be added. This example does not show any changes in the number of matches while using the case sensitive and the insensitive search.

   (c) *grep -n 'the' rosesRobertBurns.txt* → The command returns 9 lines: line 3, 8, 9, 11, 12, 14, 16, and 17. For the case insensitive search the option -i can be added. This example does not show any changes in the number of matches while using the case sensitive and the insensitive search.

9. Which option can be used with grep to search for a pattern recursively in all directories?

   (a) grep –recursive

   (b) grep -r

   (c) grep all

Exercises are in part derived by material from ©Software Carpentry (`http://software-carpentry.org`, license: CC BY 4.0) that was adapted from me for this course. Another part is from a BILS course given by Martin Dahlö and used here by his kind agreement. Remaining exercises by M. Martis.