Assignment 3: Patterns

Task

Download your dataset and answer the questions about is contents. Write the report file in the specified format and submit it before the deadline as explained below.

Description

This assignment tests your ability to find, extract and process patterns in textual data.

Requirements

- Your submission file must be a valid gzipped tarball.
- The tarball must contain the path "report.txt" (no directories).
- 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.
- Part-a must be a grep command (Q1-10) using basic regular expressions.
- Part-a must be a sed command (Q11-16).
- The paths in your grep command must be as shown in the example to work in the testing environment
- When the grep command is run in the testing environment it must recreate the contents of the hits files for the question.

Percentage Correct	Grade
≥ 90%	A

≥ 80%	B
≥ 70%	C
≥ 60%	D
≥ 50%	E
< 50%	F

Examples

First I start by looking at q1data.txt to see how the source looks. Sometimes this will tell me something useful about the structure and sometimes it will not. Either way the struggle provides food for the soul.

```
<><><>??(null)<8628911>6787158<6183874>8372432??(skip)
??(void)<><><>??(void)??(void)??(null)<><><>
??(void)<><><>??(skip)<><><>
<9351931>9501104<6994546>4030208<><><><><><><><><><>
```

Next I look at q1hits.txt to see how the pattern looks. It is to be hoped that one of these steps, or some indescribable moment of insight causes me to split the air with a cry of pure inventive joy. Otherwise I would be trapped in a desperate cycle of searching, bouncing back and forward between the files until I understood the relation between them.

```
<8628911>6787158<6183874>8372432
<9351931>9501104<6994546>4030208
<8590952>4406544<0026381>1842688
<4157933>5933732<9717545>8385419
```

The important thing about a pattern is what changes and what stays the same. In this case I notice that all of the matches are the same length. The next thing to see is that there are four decimal numbers on each line, each with exactly seven digits. That seems to easy to extract so next I make a simple experiment:

```
cat datasets/q1data.txt | grep -o '[0-9]\{7\}'
```

The output starts with:

```
8628911
6787158
6183874
8372432
```

That looks like the numbers in the first match so now I need to join them together and add the angular brackets:

```
cat datasets/q1data.txt | grep -o '<[0-9]\{7\}>[0-9]\{7\}'
```

After checking the output I use that regex twice to get the whole match:

```
cat datasets/q1data.txt | grep -o '<[0-9]\{7\}>[0-9]\{7\}<[0-9]\{7\}>[0-9]\{7\}'
```

Obviously I work this way because I'm not afraid of cut and paste: typing in long regular expressions is cumbersome. Instead I could have jumped straight to something that is less precise, but also less work:

```
cat datasets/q1data.txt | grep -o '<[0-9]*>[0-9]*<[0-9]*>[0-9]*'
```

You will only be graded on whether your regex matches on this dataset, not on unseen data. So take shortcuts when you can... Final step is putting the answers in the report so that they match the requirements described above:

```
1a cat datasets/q1data.txt | grep -o '<[0-9]*>[0-9]*<[0-9]*>[0-9]*'
1b The structure that I found was four blocks of seven digits, separated
by angle brackets.
```

Questions / Tasks

Q	Basic regex handling in grep.	1.0 points
1a	Command to transform datasets/q1data.txt into an output matching datasets/q1hits.txt.	

1b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

2a Command to transform datasets/q2data.txt into an output matching datasets/q2hits.txt.

2b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

3a Command to transform datasets/q3data.txt into an output matching datasets/q3hits.txt.

3b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

4a Command to transform datasets/q4data.txt into an output matching datasets/q4hits.txt.

4b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

5a Command to transform datasets/q5data.txt into an output matching datasets/q5hits.txt.

5b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

6a Command to transform datasets/q6data.txt into an output matching datasets/q6hits.txt.

6b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

7a Command to transform datasets/q7data.txt into an output matching datasets/q7hits.txt.

7b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

8a Command to transform datasets/q8data.txt into an output matching datasets/q8hits.txt.

8b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

9a Command to transform datasets/q9data.txt into an output matching datasets/q9hits.txt.

9b Description of the pattern that you found.

Q Basic regex handling in grep. 1.0 points

10a Command to transform datasets/q10data.txt into an output matching datasets/q10hits.txt.

10b Description of the pattern that you found.

Q Text transformation in sed. 1.0 points

11a Command to transform datasets/q11data.txt into an output matching datasets/q11target.txt

11b Description of the pattern(s) that you found.

Q Text transformation in sed. 1.0 points

12a Command to transform datasets/q12data.txt into an output matching datasets/q12target.txt

12b Description of the pattern(s) that you found.

Q Text transformation in sed. 1.0 points

13a Command to transform datasets/q13data.txt into an output matching datasets/q13target.txt

13b Description of the pattern(s) that you found.

Q Text transformation in sed. 1.0 points

14a Command to transform datasets/q14data.txt into an output matching datasets/q14target.txt

14b Description of the pattern(s) that you found.

Q Text transformation in sed. 1.0 points

15a Command to transform datasets/q15data.txt into an output matching datasets/q15target.txt

15b Description of the pattern(s) that you found.

Q Text transformation in sed. 1.0 points

16a Command to transform datasets/q16data.txt into an output matching datasets/q16target.txt

16b Description of the pattern(s) that you found.