Text Processing in Linux - the 'uniq' command - #2



Problem Statement

Introduction and References

In linux, the most vanilla version of 'uniq' eliminates consecutive repetitions of a line when a text file is piped through it.

Plain Uniq

For instance, if this is the file test.txt

00			
00			
01			
01			
00			
00			
02			
02			

This is the output on passing it through the 'uniq' command, either via pipes or as input via STDIN.

```
Command: uniq < test.txt

00
01
00
02
```

Since the first two lines of the original file are the same (00) and same for the next two (01) again followed by two repetitions of 00 and two repetitions of 02 - the 'uniq' command replaces consecutive repetitions by only one line in each case.

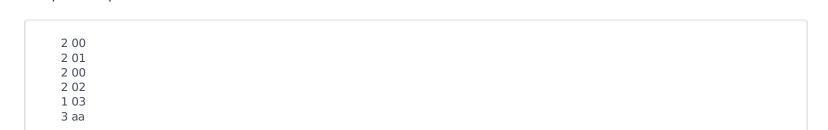
Uniq with counts

uniq -c < test.txt

This one also indicates the count of repetitions for each of the lines it collapses.

If this is the test file (say, testCounts.txt)

00			
00			
01			
01			
00			
00			
02			
02			
03			
aa			
aa			
aa			



The first number is the **count** of the number of repeated occurrences in the original file.

Printing only duplicate lines

uniq -c < input00.txt

The '-d' option prints only those lines which are followed by one or more repetitions immediately after them.

```
uniq -d < testCounts.txt
```

OR

cat testCounts.txt | uniq -d

OR

 $uniq - d \ testCounts.txt$

Printing only unique lines

The '-u' option printls only those lines which are succeeded and preceded by different lines.

```
uniq -u < testCounts.txt
```

OR

cat testCounts.txt | uniq -u

OR

uniq -u testCounts.txt

It is also possible to - limit comparison only to the first N characters (using he -w option)

- avoid comparing first N characters (using the -s option)
- ignore variations in case between lines (the -i option)
- avoid comparing the first N fields using the -f option.

(This may be useful while processing TSV files when you'd like to ignore the first column, if it has serial numbers.)

You might find these examples interesting and useful.

Current Task

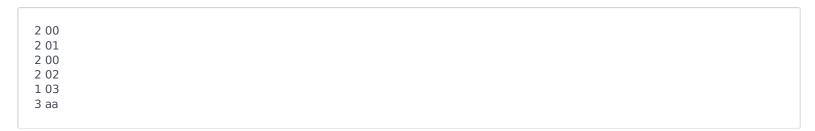
Given a text file, count the number of times each line repeats itself (only consider consecutive repetitions). Display the count and the line, separated by a space. There shouldn't be leading or trailing space. Please

note that the *uniq -c* command by itself will generate the output in a different format from the one expected here.

Sample Input

00			
00			
01			
01			
00			
00			
02			
02			
03			
aa			
aa			
aa			

Sample Output



Explanation

00 is repeated twice
01 is repeated twice
00 is repeated twice
02 is repeated twice
03 occurs once
aa is repeated thrice