Text Processing in Linux - the 'uniq' command - #3

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
```

uniq -c < input00.txt

```
2 00
2 01
2 00
2 02
1 03
3 aa
```

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

Printing only duplicate lines

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 repetions). Display the count and the line, separated by a space. There shouldn't be leading or trailing spaces. Please

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

This time, compare consecutive lines in a **case insensitive** manner. So, if a line X is followed by case variants, the output should count all of them as the same (but display only the form **X** in the second column).

So, as you might observe in the case below: aa, AA and Aa are all counted as instances of 'aa'.

Sample Input

```
00
00
01
01
00
00
02
02
03
aa
AA
Aa
```

Sample Output

```
2 00
2 01
2 00
2 02
1 03
3 aa
```

Explanation

```
00 is repeated twice
01 is repeated twice
00 is repeated twice
02 is repeated twice
03 occurs once
aa is repeated thrice (if we ignore case - AA, Aa are the same as 'aa')
```