Time Series: Predict the Web Traffic

HackerRank

Objective

In this challenge, we practice predicting time series. Check out the Resources tab for an overview of potential approaches to this problem.

Task

You are given the web traffic data for a particular website, which is measured in terms of user sessions. You are provided with the number of sessions for a time series of 1133 consecutive days starting from $\text{October } 1^{st}, 2012$. Your task is to predict the number of sessions for the next 30 days.

Input Format

The first row contains integer, N.
Each of the N subsequent lines contains an integer denoting the number of user sessions for day i (where $1 \leq i \leq N$).

Hidden Input File

The input file has 1134 rows ($N = 1133$), each containing an integer.
The first integer is the number of sessions on $\text{October } 1^{st}, 2012$.
The second integer is the number of sessions on $\text{October } 2^{nd}, 2012$. This pattern continues until we get to the last integer, which denotes the number of sessions on $\text{November } 11^{th}, 2015$.

Sample Input File

The Sample Input file has 500 rows ($N = 500$) from the same data for 500 consecutive days starting from $\text{October } 1^{st}, 2012$. Your output for this data will be calibrated against the next 30 days of data in the series, as predicted by you.

After inspecting the data (you may plot the first 500 rows from the Sample Input), you may observe large periods of somewhat periodic and stable trends, followed by certain abrupt dips and jumps. The abrupt dips and jumps typically occur when there is a major change or revamp to the site content. You may assume that no such drastic change was made to the website after $\text{November } 11^{th}, 2015$. So, any major trend observed for the first 1133 days of data will not be drastically disrupted in the next 30 days for which you must predict the values.

Constraints

Scoring

The final score obtained upon submitting your code will be based ONLY on the hidden test case.

We will compute the mean of the magnitude of the percentage difference by comparing your expected answers with the actual sessions for each of the missing records in all test cases, samples included:

$$d = \sum \frac{|\text{ expected sessions} - \text{computed sessions }|}{\text{expected sessions}} \times 100$$

Your final score on a scale of 100 will be: $5 \times max(20 - d, 0)$

If the mean value of d exceeds 20% (i.e. your predictions are off by 20% or more on average), you will score zero. If your predictions are right on target, you will score 100.

When you hit *Run Code* (instead of *Submit Code*), we will run your solution against the sample test only. At that time, the visible score will be normalized out of 1 rather than 100. In case your program throws an error (or has an incorrect output format) for a single test case, the overall score assigned will be zero.

Output Format

For each of the 30 days of predictions, print your number of predicted sessions for that day on a new line. For example, the first line should contain the predicted number of sessions on $\textbf{November } 12^{th}, 2015$, the second line should be the predicted number of sessions on $\textbf{November } 13^{th}, 2015$, and so on.

Sample Input

```
1339
1462
1702
1656
1439
1208
1613
1935
1964
2003
2023
1559
1274
1805
2051
2024
2049
...
...
...
```

Sample Output

```
1000
1500
....
....
....
(30 such integers, each on a new row)
```