Problem 5: Fake Data – A CLASS-BASED PROBLEM

Overview

The term “fake news” is often associated with Donald Trump, who seems to use (or tweet) the phrase every day. Indeed, he has even given out the “Fake News Awards.” Some people think that he invented the term – but that would be “fake news”. President-elect Trump’s first tweet with the phrase was sent out on January 11, 2017, but his opponent in the presidential election had used the term (at least) twice on December 8, 2016, one time referring to “the epidemic of malicious fake news and false propaganda that flooded social media over the past year.” Lesson: even learning about fake news can be difficult due to fake news. Fortunately, things are a bit easier when it comes to fake data.

Problem

Statisticians can use many tests to root out fake data sets. One interesting fact about “random” data is that it isn’t always as “random” as we think. If a data set has certain properties, then it will obey Benford’s Law. Benford’s law states that the probabilities of a digit, , arising as the first digit of a number in the data set is NOT 1/9, but rather depends upon the itself, and can be computed as \( \log_{10}(d+1)-\log_{10}(d) \). Data sets that conform to Benford’s law include powers of two, county populations, amounts of reimbursements for business expenses, and surface areas of rivers in a geographic region, among many others. In this problem, you will be given a data set and you will need to determine whether or not it conforms to Benford’s Law. For purpose of this problem, we will say that a distribution is plausibly valid if each digit (1-9) occurs with a frequency that is within 20% of what Benford’s Law would predict. So, if we have 1000 data points, and the Benford probability of a 5 is 0.0792, then we would expect 79.2 5’s in our distribution. A twenty percent bracket around this is 63 and 96 items inclusive. [We round generously, down at the bottom of the range, up at the top of the range.] If all of the digits fall within their expected ranges, then the data is plausibly valid; if any digit falls outside of its intended range, then we will deem this data set to be “fake data”.


AS FREQUENTLY HAPPENS IN SCIENCE, BENFORD WAS NOT THE FIRST TO DISCOVER HIS LAW; IT WAS FIRST PUBLISHED BY SIMON NEWCOMBE 57 YEARS BEFORE BENFORD RE-DISCOVERED IT AND PUBLISHED IT AGAIN.

ARTIFICIALLY CONSTRAINED, FIXED LENGTH DATA SETS – SUCH AS PHONE NUMBERS OR SOCIAL SECURITY NUMBERS – SHOULD NOT BE EXPECTED TO OBEY BENFORD’S LAW.
**Input**

The input consists of a four-letter, upper case only, name of a data source. [For the record, the data source TRUE produces plausibly valid data and the data set FALS produces fake data.]

**Output**

The output is either “FAKE DATA” or “PLAUSIBLY VALID DATA” according to whether or not the data passes the test described above.

**About the class**

You must use the `DataSource` class in your solution to this problem. Otherwise, you will not be able to access the data. This class also includes a method to help you with the probability computation. All of the necessary methods are provided for you (although you may not need to use everything that is provided). Be sure to pay close attention to documentation of the class, including the pre- and post-conditions for all of the methods and constructors.

**Example 1**

Input

    BONA

Output

    PLAUSIBLY VALID DATA

**Example 2**

Input

    CONW

Output

    FAKE DATA