Homework 02 Instructions STAT/CS 287

Instructions

Please read these instructions completely before you begin.

To begin, rename the included `HW02_NETID.py` script and `HW02_writeup_NETID.docx` writeup by replacing `NETID` with your UVM NetID (ex: `HW02_jbagrow.py`, `HW02_writeup_jbagrow.docx`). Fill in the header information in the script file and then complete the assignment. As you complete the code portions of the problems below also complete your corresponding written answers.

- The `.py` file contains **headers** (as code comments) indicating where to place your functions and where to place the code you write to complete each problem. Please place your code in the relevant areas as indicated so that your code organization is easy to follow. Please do not delete the header comments.

Further, please note:

- Do not modify the block of imports near the top of your `HW02_NETID.py` script.
- Use no other import statements within your script, please.
- I will refer to the location on your computer containing the script file and the reports directory (see below) as the *working directory* for this assignment.
- You must **show your work** for all questions. This means answering questions using code, statistics, code *and* statistics, and it means describing and synthesizing your answers.

To submit:

1. Compress your HW02 working directory as per the "preparing and submitting your homework" slides. Make sure the zipped file includes your `.py` script file, your writeup, the reports directory and its contents, and any other files you may have generated while completing the assignment.
2. Upload your zipped file to Blackboard. No other files should be submitted.

Failures and risks at data centers

Before

After

You have been hired as a data scientist for one of the largest tech companies in the world. This company operates nearly 1500 separate data centers and server farms in the United States alone. Each data center is a building housing thousands of computers and the electrical and environmental equipment necessary to power and cool those computers. Every data center has a high-bandwidth, low-latency fiber optic connection to the internet. It costs your company approximately 13 billion USD per year to build and operate its data centers.

One of the challenges facing the owners of so much infrastructure is **maintenance**: computers fail all the time, electricity fluctuates or gets cut off entirely in a blackout, links to the internet get cut, and buildings may be damaged due to weather or other causes. Your company has developed a system of **error reporting** procedures to gather data on **error modes**, the types of errors and failures that occur within the company data centers.

Problem 1: Processing error reports

Provided with this assignment is an **error report dataset**. Each data center contains monitoring software that *automatically* logs errors and warnings within the running data center as they occur and sends them over the network to your data science group. Error reports are generated from these log files, one report per data center. Each error report is contained in a separate file.

Here is an example error report:

```
DATACENTER 000042
AUTOMATED REPORT PERIOD: FY 2015-2017

ERROR MODE                            COUNT
----------------------------------------------
A/C:                                 11
Fiber pipeline in:                   14
Fiber pipeline out:                   7
Misc. elec.:                         82
Operator (employee) error:            0
Operator (non-employee) error:        2
Physical intrusion (person):          0
Physical intrusion (water):          15
Power/generator loss:                 0
Power/generator reduction:            2
OPERATION NOTES:
e56770fe99159c4b7575409d0ed79acdd5a7(redacted)
45a632f05a81040440623fdb652bd9b109c8(redacted)
00db4fa51b2269b27f475197124dfddec263(redacted)
2ee07dfb1e2d03a950eee98bcadc74f81ce1(redacted)
550c3c2d859d678fb37e149f82c999176a03(redacted)
c4611464d15abeb6158b96572ed4a21d3ea0(redacted)
```

```
----------------------------------------------
END OF LINE
```

The primary data of interest are the **counts of errors of different types** (the error modes). In this example, we see that this data center has a lot of minor electrical problems.

- Unfortunately, not all reports match this exact format, due to different versions of the error reporting software and other problems. Such is life when managing billions of dollars of infrastructure. You need to handle this.

Prepare the files

1. Uncompress the included `reports.zip` file into a directory called `reports` and place this directory inside your HW02 working directory. In other words, alongside `HW02_NETID.py` should be a directory called `reports` and inside `reports` should be approximately 1400 `.dat` files, one for each data center. In the writeup, describe the organization of your working directory, and how you made sure that your Python code does not refer to the directory structure of your computer **outside** the working directory.

Process the files

2. Error reports have been gathered from a variety of different reporting systems, so not all reports follow the same format. Create a *function* called `load_report` where indicated that takes a report filename as input, reads the text of that file into Python, then extracts and returns the error mode counts as a Python dictionary. In the writeup, describe how `load_report` works and provide a brief narrative of how you wrote it.

3. Write a loop that calls your function on each report file and stores the error mode counts in an appropriate data structure.

    - Since different reports may look very different, you need to examine the report files both computationally (by writing, running, and rewriting Python code) and by eye (by reading the text files yourself) to determine how the reports are written. **Your function must automatically read all reports correctly**. In the writeup, describe how this loop works, what data structure you used to store your error mode counts and why, and how you determined that your `load_report` function works correctly on all report files.

Are the data consistent?

4. It appears that data centers have associated ID numbers. Are these IDs consistent? How should we define consistency? One way is to check if ID numbers are unique and if they are sequential or if there are gaps of missing ID numbers. How else should we define and measure consistency? Answer these questions with code in your `.py` file and written responses in your writeup.

5. Are the **error modes** consistent? Do you need to do any cleaning, such as merging different error mode names that represent the same thing or fixing typos in the names of error modes?

    - [**Bonus for ugrads/Required for grads**]: Make the error modes consistent or prove they are already consistent.

6. Are the error mode counts consistent? Define consistency and show if the counts are or are not consistent? If they are not consistent, make them consistent.

Remember these questions require both computational and written answers! And in your writeup, refer to the specific line numbers of your `.py` file for the code that supports your written answer(s)

Problem 2: Error rates and error statistics

We have now extracted the error mode counts for each data center, and we have these counts stored in one or more handy Python data structures.

Analyze these counts to report what are the different error modes and how often they occur. Specifically,

1. Do some data centers experience more errors than others?
2. What is the typical number of errors a data center should expect?
3. What are the most common types of errors experienced by data centers?
4. Being very careful as you interpret your analyses, what are the **riskiest** error modes in terms of shutting down data centers? Provide a written argument supporting your answer to this question.

For this problem, you should summarize the counts using basic statistics and visualizations, as needed to support your answers. Craft an appropriate argument to help your manager and the company's risk assessment office understand what is happening at the data centers. Provide both code and written responses to support each answer for this problem.

Problem 3: Flood risk

A particular risk factor the company faces is flooding. There is concern that some data centers are situated in areas that are especially at risk of floods. Computers and water do not mix well.

1. From the error reports, what data centers are most at risk due to flooding? Are some more at risk than others or is the risk of floods about the same across data centers. To answer these questions, look at errors related to water compared with overall error rates. What proportion of the errors at each data center is due to water?
2. Rank the N data centers according to flood risk, from a rank of $R = 1$ (most at risk) to $R = N$ (least at risk). Develop a visualization or other data presentation to display this information compactly.
    - [**Bonus for ugrads/Required for grads**]: When ranking the data centers according to flood risk, incorporate uncertainty in your risk assessments. Determine how best to do this, and **justify** your analysis choices.
3. [**Bonus for ugrads/Required for grads**]: Rank data centers by their flood risk both with and without uncertainty, then determine an appropriate statistics for comparing how the *ranks* changed when you incorporated uncertainty.

Notes

- Functions derived in class may be helpful for these questions. To use those functions, paste them in at the **top** of the `.py` file, in the section indicated for placing functions.
- Windows users: If you struggle to get Python to separate the different lines of text in the error report files, you may want to investigate "universal newlines" in Python 3.
- **Bonus (for all)**: The third image is taken from **what movie**?