Jordan Machita
Jm8ux
CS 5010

Homework 3: Python and Web Scraper

Overview

The web scraper portion of the program uses the URL of a NASA website to read a table that lists all of the solar eclipses that will occur throughout this decade (2020-2030) as well as some of their characteristics, including duration and eclipse type. The program stores the content of the table in a csv file called "Eclipse.csv."

The analysis portion of the program first focuses on cleaning the data. It stores the csv file in a pandas data frame and removes unnecessary headings and rows. The program also manipulates some of the columns to extract more usable data such as pulling the number of minutes out of a string that stores the length of time in minutes and seconds with characters in between. This allows us to append columns with new values for further data manipulation and analysis. The result is that we have quick and useful ways to visualize and analyze the data about the upcoming eclipses.

Approach

My approach to the assignment began with determining what type of data to scrape from the web and how to go about doing so. As I am currently re-watching Avatar: The Last Airbender (now on Netflix), in which information about a solar eclipse holds significant importance, I thought it would be interesting to learn about the upcoming solar eclipses. One of the hardest parts of the program was building the web scraper and analyzing the html from the NASA website to figure out how to correctly pull the contents of the table I wanted.

Once the web scraper successfully stored the data in a csv, it took me longer than expected to determine how to clean and manipulate the data as a pandas data frame. Admittedly, I still find the indexing logic of the data frame confusing, but this assignment helped me work through it and forced to confront my confusion. I then focused on how to visualize the data in ways that provided meaningful information. I did this by using the context of the data to determine what would be useful then figured out the best way to graph it. In addition to plotting, I also filtered through the data looking for specific requirements.

Data

While all of the data was stored as string types in the csv, not all of the items were strictly categorical or qualitative. The date and duration were stored as strings but by manipulating those strings, I was able to pull relevant discrete data, such as years and minutes, and continuous data, such as duration (minutes + seconds/60). After some creativity and experimentation, I was able to transform data that could not be compared into useful inputs for graphing and calculations.

Algorithms

I used the Beautiful Soup library and the tutorial from Module 4.8 in my algorithm to develop the web scraper. This also required importing requests and csv for holistic functionality. Once the web scraper was complete and the data was stored in a csv, I used the pandas and matplotlib libraries through Jupyter Notebook to clean, manipulate, and visualize the data. I developed my own algorithm for cleaning and storing the data in a data frame because the tasks were specific to the quality of my dataset. However, for calculating the average duration for each month, I watched several tutorials and scanned through the postings on the pandas and numpy discussion board to determine how to iterate through data frames and successfully call specific indices as well as how to filter based on specific requirements. While this taught me the general structure, I still had to develop the algorithms for what I specifically wanted to accomplish.

I also ran into challenges when attempting graph data from the data frame using matplotlib so I found some useful guides, especially this matplotlib article. I learned how to plot different types of bar graphs and histograms as well as how to improve the visualization by adding labels, adjusting tick marks, and more.

Use of Code

While this program may not seem useful to the average person, it would hold significant importance in the universe of Avatar: The Last Airbender. In short, firebenders draw their power from the sun, so during a total solar eclipse, they are unable to firebend because the moon blocks the sun. During eras of peace, this may not be any cause for alarm as they are only left powerless for a matter of minutes. However, in the setting of Avatar, the fire nation has waged a war on the rest of the world and their leader must be defeated by the Avatar and his friends. Thus, the solar eclipse presents an opportunity for the Avatar to strike while the fire lord is unable to firebend. This program could be used for good by alerting the Avatar of future solar eclipses or for evil by alerting the fire nation of when they will be vulnerable.

Let us suppose that the Avatar has hired a data scientist to build this program and provide meaningful information from the data that becomes knowledge as the Avatar leverages it to restore peace and balance to the world. In the Jupyter Notebook file of the program, the data scientist can run the program and several significant visualizations will already be produced. The graph of "Average Duration of Eclipse by Month" can help him determine what months or seasons tend to have the longest eclipses. The "Frequency of the Type of Eclipse" plot can tell him how often the firebenders will be left totally defenseless versus how often their firebending will just be weaker during a partial or annular eclipse. The "Frequency of Duration of Eclipses" chart provides useful information on the typical duration of the eclipse and thus how long they have to execute their mission (in the show, the eclipse lasted 8 minutes, which according to the chart is actually longer than most solar eclipses). The "Frequency of Eclipses by Month" plot shows which months tend to have the most eclipses (rather than the longest eclipses shown earlier). The last plot, "Frequency of Eclipses by Time of Day", illustrates what time the eclipse tends to occur, which trends toward the late afternoon.

Expansion

The data obtained by the web scraper only includes solar eclipses for this decade (2020-2030). A fire nation historian may be interested in analyzing historical eclipses over the past 200 years. This may especially be true if data on future eclipses is not available (because NASA does not exist in the Avatar universe). In this case, the historian may need to use past data to predict when future eclipses are likely to occur and how long they will last. This would have been a likely investment by the fire nation after "The Darkest Day in Fire Nation History" when they suffered a major defeat due to a solar eclipse. To avoid the repetition of history, the fire nation can use their analysis on historical eclipses to prepare accordingly and take precautions for an unexpected solar eclipse.

Additionally, the Avatar and his team may need data on eclipses that go further into the future. Afterall, the war with fire nation has already lasted 100 years so ending it in the next 10 years may not be doable. There are also other elements that could be incorporated into the program, such as data on comets. While a solar eclipse leaves firebenders powerless, a comet gives them immense strength. In the show, the team learns about an upcoming comet and must defeat the fire nation before its arrival because otherwise they will be unstoppable. Thus, the program could be more robust by accounting for other meteorological factors that influence bending.

Extra Functionality

While it is useful to visualize the dataset by plotting specific elements, it may also be useful to obtain specific information of interest. Toward the end of the Jupyter Notebook file, the program gives several useful outputs. The first filters the dataset for only those solar eclipses that have an eclipse type of total. This subsection of the data only includes six eclipses rather than the entire 22. As firebenders are only weakened during partial, annular, and hybrid eclipses, this significantly reduces the number of opportunities for the Avatar to strike when they will not be able to firebend at all. We can also filter for the eclipse with the longest duration, which is 10.45 minutes (or 10 minutes 27 seconds). However, this is an annular eclipse so it still may not be as useful as a shorter total eclipse. Thus, we can combine these filters as is done in the final lines of the program to output the data of the longest total eclipse, which will occur on August 2, 2027 for 6 minutes and 3 seconds. The Avatar could similarly filter the data in various ways to limit the year, geographic location, time of day, etc. in order to determine which eclipse best meets his specific criteria.