Spring 2024

BEM 106: Homework #5 Description

Goal: The goal of this homework is twofold: (1) to get you started with web scraping using Python's Selenium drivers and (2) to get you using Great Expectations to validate and clean messy datasets.

Part 1 - Selenium

From the Selenium website, download the .exe file for the Chrome driver. Refer to class slides for instructions of how to launch Chrome from Python using Selenium.

Your tasks:

1. Using Selenium, navigate to https://en.wikipedia.org/wiki/Main_Page.
2. In the search box, type in "Caltech", and click "search"
3. Find the "Sporting affiliations" link (SCIAC) in the right gray panel and click on it.
4. Scrape the "chronological timeline" and assemble a dataframe with "years" as one column (e.g. 1915, 1920, 1926, etc) and the "description" of the event in another column. For example, the first row of this dataframe is "1915" (in column "year") and "The Southern California Intercollegiate Athletic Conference (SCIAC) was founded. Charter members included Occidental College, Pomona College, the University of Redlands, Throop College of Technology (now California Institute of Technology) and Whittier College, effective beginning the 1915-16 academic year" (in column "description")
5. Go back to the previous Wikipedia entry of Caltech from step 2.
6. Find the image titled "Richard C. Tolman and Albert Einstein at Caltech, 1932" and download it as a .jpg file.
7. Navigate to the website https://tineye.com/
8. Upload the image of Tolman and Einstein that you saved before.
9. Scrape and organize into a list the 88 results (109 urls) that use this image.

Part 2 – Validating and Cleaning data

The data file “hw5_dataset.xlsx” contains 673 data scientist job postings scraped from Glassdoor.

According to your industry knowledge, the following expectations should hold:

1) Salary Estimate are between 50k and 400k
2) Salary Estimate are listed as {A } – {B } where A is strictly smaller than B.
3) Rating is between 0 and 5
4) Location of the company is either remote or in the United States
5) Size variable contains the pattern {integer} – to – {integer} employees
6) Founded to be later than the year 1800
7) All firms founded in 2019 should not have Revenue (i.e. Unknown / Non-Applicable)

Due to bad data entry errors at Glassdoor, some of these do not hold. Feel free to create additional variables as you see fit. Using the Great Expectations package in Python, validates these 10 expectations. Please change all “-1” to missing values in all columns before running the validation procedure.

You task is to output the **index** of any rows that contain unvalidated data.