Esme Gonzalez
CIS 3120 Homework #1
Professor Jairam

Web Scraping

Web scraping is a useful tool when accessing or extracting the data. It gives the user the ability to access a lot of information that isn't available in the csv exports or accessible through an API. For instance imagine what you can extract from or what you can analyze from the websites with web scraping. For example, in this homework I will be scraping the data from the website of the Baruch college swimming and volleyball team for the corresponding genders: men's and women's.

To start I will give an example of an important tool from the python library named Beautifulsoup that I will import into my notebook. Beautifulsoup enables you to extract data from HTML, which is very useful for web scraping. When we scrape the website we also ensure that we have a way to import a request to the server. In my notebook, I then created a dictionary to collect the website's urls. A dictionary is used to store/collect data values in key:value pairs. In my code I put the keys as either mens or womens of volleyball or swim team, or for better terms the team_type. The values are named as the website's url. Next, I created a function called "*scraping_date(team_type, url):*" Once that's done you have to create two empty lists to be able to store the values of the heights. My first list is named "*raw_height*", I simply created a new list, an empty list. When creating my second list I named it "*height_1*" which is equal to zero because of the adding operator at the end of my notebook. I then write, "*page = requests.get(url)*" to begin to make my request to the server by using the requests function, where the four urls are being pulled from the dictionary. Then I write "*soup = BeautifulSoup(page.content, 'html.parser')*", to begin to import the raw html into beautifulsoup by retrieving the page.content through the BeautifulSoup library. Including the parser to collect the text from the tag. Then I write, "*all_relevant_td_tags = soup.find_all('td', class_ ='height')*" The reason I wrote this is because of the website to find all the tags/classes. From the website you can see that you will need the <td> tag for the height. Next I will be extracting the contents of the td tags using get_text(). I would then write *"for height in all_relevant_td_tags: raw_height.append(height.get_text())"* To create the loop to be able to retrieve the values of the heights. Next I begin to do the calculation and split of the data. The reason for the split is to separate the feet and inches to make them integers, if not it will look like this "5-7". Making it unable to calculate. Once I spilt the feet and inches, I will be able to convert the feet to inches Shown as "feet_to_inches=float(x.split('-')[0]*12". The float data in my code represents a real number in decimal form and the multiplication of 12 is to convert feet to inches. Then to get the average, I used "*height_1/len(raw_height)*". I also use the += to get the sum of the heights.

In the end, I put print "print('The average for the'" "+str(team_type)+" is "+str(average_height_team))" To get the output to say The average for the men's volleyball team is 73.27 inches. I will also put a loop to call the key value of my scraping_data. It would look like scraping_data(key, value)". To get the key and value from the dictionary.

The questions

Question 5 States, Compare the averages between the two men's teams.

Looking at the two men's teams you can see that the volleyball team is taller than the men's swimming team on average. Where the men's volleyball team is 73.27 inches and the men's swim team is 71.53 inches.

Question 6 states, Compare the averages between the two women's teams.

Comparing the two women's teams. The women's volleyball team is taller than the women's swimming team on average. Where the women's volleyball team is 66.33 inches and the women's swim team is 64.0 inches.

Question 7 states, Are you able to determine whether, in general, if the average swimmer is taller than the average volleyball player?

The two findings show that the female and male from the volleyball team were taller than the swim team on average.