In this post, I'll show you how a 'https://www.classcentral.com/subject/data-science' website scraping data was done to get the 50 data science courses offered on the page.

The BeautifulSoup library was used to scraping the site.

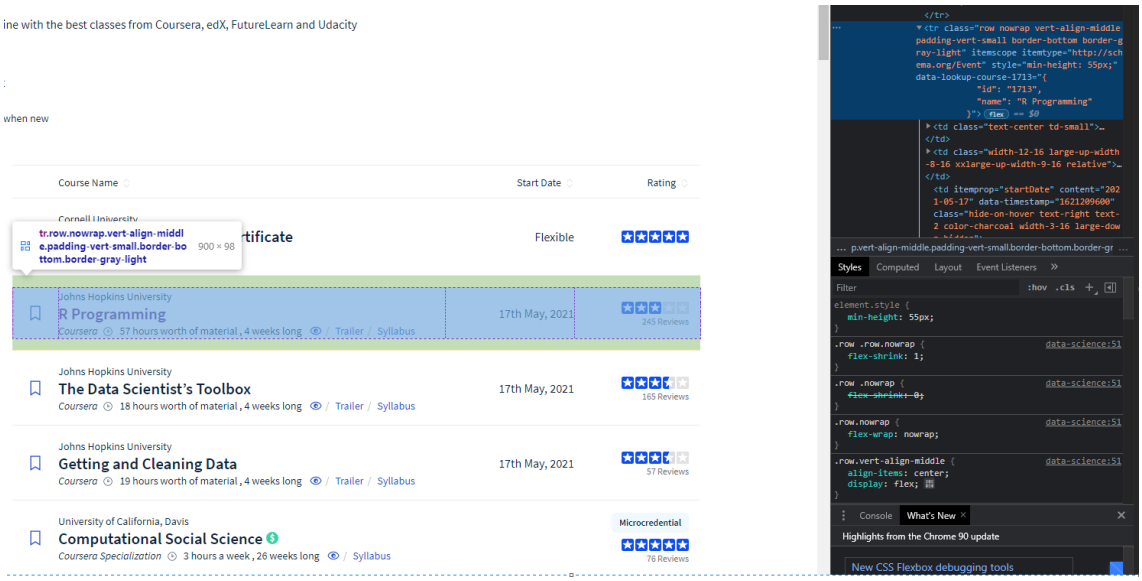

Figure 1 - Chrome Inspector and HTML Tags

In this example, we want a data frame that has information about:

1. Course name
2. Classification
3. Number of reviews
4. Platform offering travel
5. Duration

The most important of this technique is to take care of the HTML of the site. You can see all the HTML using chrome's inspect tool. In this example, when you find the course information, just see in the inspector, what is the tag and class in the HTML that contains the whole "package". In Figure 1, we see that the tag we want is the <tr>and the class "row nowrap vert-align-middle padding-vert-small border-bottom border-gray-light". With this, we assign the information of all courses in a variable.

With this variable, we can iterate on it until we have the specific information of each course, such as what is the name of the course or its duration.

Table 1 - Course data frame

	Curso	classificação	N de Reviews	Plataforma	Duração
0	R Programming	2.8	245 Reviews	Coursera	57 hours worth of material , 4 weeks long
1	The Data Scientist's Toolbox	3.3	165 Reviews	Coursera	18 hours worth of material , 4 weeks long
2	Getting and Cleaning Data	3.5	57 Reviews	Coursera	19 hours worth of material , 4 weeks long
3	Computational Social Science	4.8	76 Reviews	Coursera	3 hours a week , 26 weeks long
4	Introduction to Data Science in Python	2.4	46 Reviews	Coursera	29 hours worth of material , 4 weeks long
5	The Analytics Edge	4.7	80 Reviews	edX	10-15 hours a week , 13 weeks long
6	Exploratory Data Analysis	3.9	39 Reviews	Coursera	1 week long
7	Probability - The Science of Uncertainty and Data	4.9	32 Reviews	edX	10-14 hours a week , 16 weeks long
8	Become a Data Analyst	4.5	64 Reviews	Udacity	10 hours a week , 17 weeks long
9	Statistical Inference	2.8	34 Reviews	Coursera	54 hours worth of material , 4 weeks long
10	Introduction to Big Data	2.7	35 Reviews	Coursera	17 hours worth of material , 3 weeks long
11	Regression Models	2.5	33 Reviews	Coursera	53 hours worth of material , 4 weeks long
12	Python for Data Science	4.4	47 Reviews	edX	8-10 hours a week , 10 weeks long
13	Reproducible Research	3.9	27 Reviews	Coursera	7 hours worth of material , 4 weeks long
14	Mastering Data Analysis in Excel	1.8	26 Reviews	Coursera	21 hours worth of material , 6 weeks long
15	A Crash Course in Data Science	3.5	23 Reviews	Coursera	7 hours worth of material , 1 week long
16	Hadoop Platform and Application Framework	1.9	25 Reviews	Coursera	25 hours worth of material , 5 weeks long
17	Mining Massive Datasets	4.6	25 Reviews	edX	5-10 hours a week , 7 weeks long
18	Introduction to Computational Thinking and Dat...	4.5	31 Reviews	edX	14-16 hours a week , 9 weeks long
19	Digital Marketing Analytics in Practice	4.2	24 Reviews	Coursera	19 hours worth of material , 4 weeks long
20	Statistics and R	3.5	20 Reviews	edX	2-4 hours a week , 4 weeks long
21	Spatial Data Science: The New Frontier in Anal...	5.0	41 Reviews	Independent	2-3 hours a week , 6 weeks long
22	Developing Data Products	3.9	18 Reviews	Coursera	10 hours worth of material , 4 weeks long
23	Pattern Discovery in Data Mining	2.1	21 Reviews	Coursera	17 hours worth of material , 4 weeks long
24	Data Visualization	3.3	20 Reviews	Coursera	15 hours worth of material , 4 weeks long
25	Finding Hidden Messages in DNA (Bioinformatics I)	4.5	17 Reviews	Coursera	15 hours worth of material , 5 weeks long
26	Building a Data Science Team	3.6	14 Reviews	Coursera	5 hours worth of material , 1 week long
27	Whole genome sequencing of bacterial genomes -...	4.8	22 Reviews	Coursera	6 hours worth of material , 5 weeks long
28	Making Sense of Data in the Media	4.7	34 Reviews	FutureLearn	3 hours a week , 3 weeks long

After assigning each specific information to a variable, we can assemble the data frame from table 1.

It is worth mentioning that you will get the information only available on that page. For more courses information, in addition to the 50 initially obtained, you need to go to the next page and redo the steps to the data frame part, where you just add the new information obtained.

Table 2 - Courses with the best rankings

	Curso	classificação	N de Reviews	Plataforma	Duração
49	Genome Sequencing (Bioinformatics II)	5.0	4 Reviews	Coursera	17 hours worth of material , 5 weeks long
45	Causal Diagrams: Draw Your Assumptions Before ...	5.0	3 Reviews	edX	2-3 hours a week , 9 weeks long
20	Spatial Data Science: The New Frontier in Anal...	5.0	41 Reviews	Independent	2-3 hours a week , 6 weeks long
13	Probability - The Science of Uncertainty and Data	4.9	32 Reviews	edX	10-14 hours a week , 16 weeks long
23	Whole genome sequencing of bacterial genomes -...	4.8	22 Reviews	Coursera	6 hours worth of material , 5 weeks long
2	Computational Social Science	4.8	77 Reviews	Coursera	3 hours a week , 26 weeks long
24	Introducción a la Ciencia de Datos con Python	4.8	36 Reviews	edX	6-8 hours a week , 4 weeks long
46	Data Science: Visualization	4.7	3 Reviews	edX	1-2 hours a week , 8 weeks long
4	The Analytics Edge	4.7	80 Reviews	edX	10-15 hours a week , 13 weeks long
26	Making Sense of Data in the Media	4.7	34 Reviews	FutureLearn	3 hours a week , 3 weeks long
36	Data Science and Agile Systems for Product Man...	4.7	26 Reviews	edX	2-3 hours a week , 4 weeks long
14	Mining Massive Datasets	4.6	25 Reviews	edX	5-10 hours a week , 7 weeks long
29	Data Science: R Basics	4.6	11 Reviews	edX	1-2 hours a week , 8 weeks long
15	Introduction to Computational Thinking and Dat...	4.5	31 Reviews	edX	14-16 hours a week , 9 weeks long
5	Become a Data Analyst	4.5	64 Reviews	Udacity	10 hours a week , 17 weeks long
21	Python for Data Science	4.4	41 Reviews	Swayam	4 weeks long
8	Python for Data Science	4.4	47 Reviews	edX	8-10 hours a week , 10 weeks long
16	Digital Marketing Analytics in Practice	4.2	24 Reviews	Coursera	19 hours worth of material , 4 weeks long
33	Data Science Math Skills	4.1	10 Reviews	Coursera	13 hours worth of material , 4 weeks long
37	Mathematical Biostatistics Boot Camp 2	4.0	4 Reviews	Coursera	11 hours worth of material , 4 weeks long
39	Developing Data Products	3.9	18 Reviews	Coursera	10 hours worth of material , 4 weeks long
25	Reproducible Research	3.9	27 Reviews	Coursera	7 hours worth of material , 4 weeks long

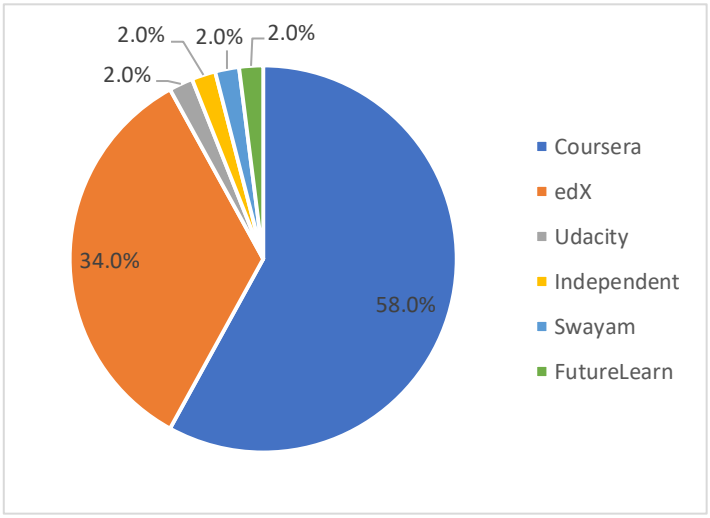

Figure 2 - Pie chart of all courses

With this information, you can answer questions such as: Which platform is the platform with the most courses or which courses have the best ratings? We can answer these questions using python or even excel. As we see in table 2, the course with the best classification is Genome Sequencing (Bioinformatics II), but it has only 4 reviews, so we have to be careful and know how to interpret the answers obtained. Already in Figure 2, we have that Coursera holds more than half of the courses offered on the page we did the web scraping.

Finally, this technique can be done on any website, always looking at HTML, and thus we can generate several data frames for future analysis.

Thanks, and until the next post!