Nesse post, vou mostrar como foi feito um data scraping do site ‘https://www.classcentral.com/subject/data-science’ para obter os 50 cursos de data Science oferecidos na página.

Foi utilizado a biblioteca do BeautifulSoup para fazer o scraping do site.

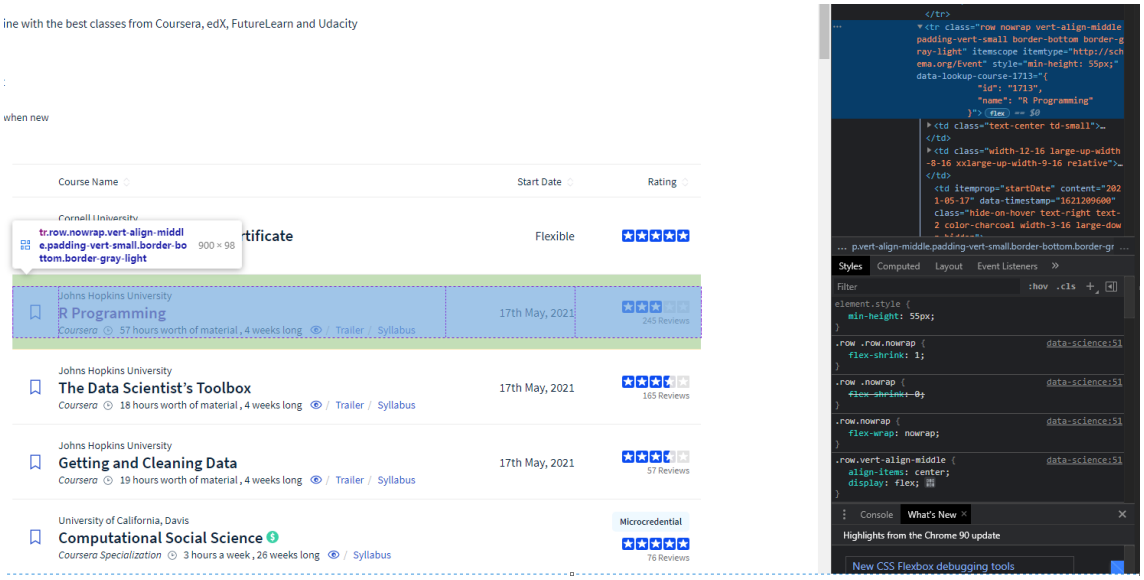

Figura 1 - Inspetor do chrome e as tags do HTML

Neste exemplo, queremos uma dataframe que tenha as informações sobre:

- Nome do curso
- Classificação
- Número de reviews
- Plataforma que oferece o curso
- Duração

O mais importante dessa técnica é se atentar ao HTML do site. Você consegue ver todo HTML utilizando a ferramenta de inspecionar do chrome. Neste exemplo, quando você localizar as informações do curso, basta ver no inspetor, qual é a tag e a class no HTML que contém o “pacote” todo. Na figura 1, vemos que a tag que queremos é a <tr> e a class “row nowrap vert-align-middle padding-vert-small border-bottom border-gray-light”. Com isso, atribuímos as informações de todos os cursos em uma variável.

Com essa variável, podemos iterar sobre ela até termos as informações específicas de cada curso, como qual é o nome do curso ou sua duração.

Tabela 1 – Dataframe dos cursos

	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

Depois de atribuir cada informação especifica para uma variável, podemos montar a dataframe da tabela 1.

Vale ressaltar que você vai obter as informações apenas disponíveis naquela pagina. Para obter informações sobre mais cursos, além dos 50 obtidos inicialmente, é preciso ir para a próxima pagina e refazer os passos até a parte da dataframe, na qual você apenas adiciona as novas informações obtidas.

Tabela 2 – Cursos com as melhores classificações

	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

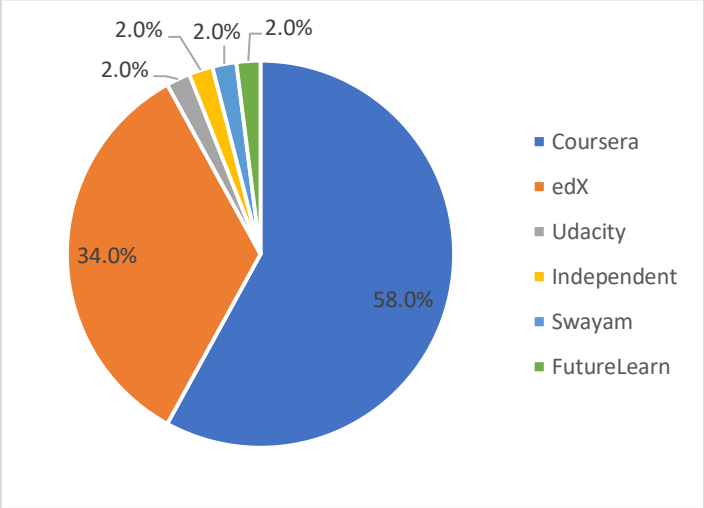

Figura 2 - Pie chart dos cursos

Com essas informações, você pode responder perguntas como: Qual é a plataforma com mais cursos ou quais cursos possuem as melhores classificações? Podemos responder essas perguntas utilizando python ou até mesmo excel. Como vemos na tabela 2, o curso com a melhor classificação é Genome Sequencing (Bioinformatics II), porém ele possui apenas 4 reviews, então temos que ter cuidado e saber interpretar as respostas obtidas. Já na figura 2,

temos que o Coursera detém mais da metade dos cursos oferecidos na pagina que fizemos o web scraping.

Por fim, essa técnica pode ser feita em qualquer site da internet, sempre se atentando ao HTML, e, assim, podemos gerar diversas dataframes para futuras análises.

Obrigado e até o próximo post!