Máster Data Science UOC



Práctica Web Scrapping

Autor: Curso 2020/2021

Kyrylo Morozov Fecha: 09/11/2020



1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Se trata de una red Social en continuo auge. El creciente uso de este tipo de webs, ha aumentado el tiempo en el que la sociedad hace uso de éstas. Esto conlleva que éstas dispongan de información útil que puede ser usada por empresas de marketing.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Instagram Dominant Color posts Torrevieja

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El Dataset extraido contiene información acerca de los posts de instagram en una región, en concreto Torrevieja. La finalidad de estos dataset es obtener los colores dominantes de las fotografías en esta zona.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente



	A	В	С	D	-	F	G	H		J
1	imgurl	usuario	nombre	tiempo	likes	seguidores	texto	hashtags	color1	color2
2	https://scontent	baby.montse	A Construyendo mi imperio 🛎		435	13516	Jamón □□	0	[251.9999999999994, 79.80392156862746, 14.563106796116	[211.57894736842104, 24.901960784313726, 44.881889763]
3	https://scontent	sarahbouzidi	MAMA	2020-10-22 21:37:50	200	816	for me and only for me	0	[30.0, 26.27450980392157, 11.940298507462686]	[22.22222222222207, 77.05882352941177, 69.23076923076
		photo_and_painting_art	Pablo Pérez	2020-10-25 16:16:31	303	4603		['mar', 'male', 'beach', 'boo	▶[206.12903225806454, 51.764705882352935, 25.20325203252	[194.1176470588235, 91.76470588235294, 80.95238095238
5	https://scontent	haglundnellie	Nellie Haglund	2020-10-26 17:53:29	524	3710	-0 €	П	[184.999999999997, 63.13725490196078, 6.3829787234042	[17.307692307692307, 26.27450980392157, 38.8059701492
							Night out			
6							20k so thankful for ye all! Like re		•	•
	https://sconten	oliviakrogh	OLIVIA KROGH	2020-10-21 07:09:43	3446	20194	#pool #summer #swimmingpoo	['poolside', 'billiards', 'bhfy	[36.206896551724135, 43.92156862745098, 25.892857142857	1 [210.0, 17.254901960784313, 20.4545454545454557]
7	https://scontent	mikademalten.	Mika De Maltén	2020-10-24 17:34:27	297	1424	0	D	[29.032258064516125, 27.254901960784313, 22.30215827338	[31.50000000000014, 73.72549019607844, 29.8507462686]
8	https://scontent	veeeronika		2020-10-29 16:12:48	327	4423	Happy birthday to me!□□	0	[292.00000000000006, 63.3333333333333, 8.0213903743315	[8] [233.68421052631575, 16.274509803921568, 22.8915662650
9	https://scontent	iammarieta	0000000000	2020-10-25 12:16:17	513	2231	Lo que digan de ella le importa	0	[38.75000000000001, 59.6078431372549, 23.30097087378640	ð[0.0, 11.96078431372549, 8.196721311475414]
	https://scontent		000000	2020-10-25 17:06:18	387	3619	Wearing all class _ lufestyl_	0	[15.00000000000053, 19.607843137254903, 4.0]	[14.736842105263165, 57.45098039215686, 26.26728110599
11	https://scontent	tammy.ga	Tammy ®	2020-11-08 22:45:16	22	2853	null	П	[23.999999999998, 86.6666666666667, 29.4117647058823	[19.310344827586203, 42.15686274509804, 40.4651162790
12	https://sconten	andreagarciaguerrera 199	300	2020-11-08 22:41:17	38	18008	#movies #toptags #theatre #	['movies', 'video', 'film', 'cir	[0.0, 74.50980392156863, 0.0]	[0.0, 9.803921568627452, 0.0]
13	https://sconten	vr. outlet torrevieja	V&R Outlet	2020-11-08 22:26:53	0	737	#ropa#torrevieja#outlet#ropper	[ropa#torrevieja#outlet#T	(30.00000000000014, 64.31372549019608, 34.065934065934	0 [240.0, 14.901960784313726, 21.052631578947386]
14	https://scontent	martine_rieth	00000000	2020-11-08 21:52:09	41	6006	My favorite place. ♥	0	[32.7272727272728, 64.50980392156862, 6.077348066298335	6 204.65753424657532, 36.47058823529412, 78.4946236559
15	https://scontent	fannyalvez	Eanny Alvez	2020-11-08 21:27:35	- 1	157	Quiero estar con vos. Quiero po		[45.40540540540541, 30.3921568627451, 23.87096774193547	
	https://scontent	ananae_danzaoriental	Ananae	2020-11-08 21:15:07	4	61	#danzaoriental #danzadelvien#	('clasessemanales', 'ejercio	• [99.9999999999999, 14.901960784313728, 7.8947368421052	[35.9999999999998, 71.17647058823529, 23.80952380952
17	https://scontent	rybackijacek	Rybacki Jacek	2020-11-08 21:14:22	72	1646	#salt #lake #pink #unesco #spal	"pink", 'unesco', 'salt', 'spai	F[17.647058823529356, 73.92156862745098, 12.781954887218	O [27.500000000000007, 25.098039215686274, 18.75]
18	https://scontent	carlatobar	c	2020-11-08 21:01:08	270	2321	you are my favorite song	0	[205.71428571428575, 72.94117647058823, 10.144927536231	[240.0, 13.333333333333338, 2.9411764705882346]
19	https://scontent	taichilibre	Taichi Libre	2020-11-08 20:46:22	4	42	Clases en Guardamar	0	[90.81081081081078, 73.92156862745098, 27.8195488721804	B[187.5, 20.0, 7.8431372549019805]
20	https://scontent	isabeleando	Rescue	2020-11-08 20:24:01	150	1237	Lo normal es ser diferente.	0	[5.454545454545443, 85.29411764705883, 14.66666666666666	5 32.50000000000001, 21.96078431372549, 21.42857142857
21	https://sconten	millie lilliellie	MILLIE	2020-11-08 20:20:49	13	588	#deaning #dean #deaningmo	('cleanbeauty', 'mamatips'	FI38.99999999999998. 65.49019607843137. 11.363636363636363	#147.999999999999986. 17.058823529411764. 17.2413793103
	https://scontent		Stine	2020-11-08 20:14:16	70	2239	Spent the morning in Torreview	"city#cityview#sundayfund	212.7272727272727, 31.176470588235293, 6.9182389937106	9 [29.589041095890405, 70.0, 47.7124183008538]
23	https://scontent	paulagm188	paula	2020-11-08 20:11:56	87	1146	#torreviela #nature #naturephot	"storm", 'naturephotograph		3 208.42105263157896, 80.58823529411765, 19.1919191919
	https://scontent		Esther Vega Coral	2020-11-08 20:11:03	61	1122	Hasta su lado oscuro está lleno!	n	[300.0, 41.764705882352935, 1.4084507042253538]	[189.67741935483875, 71.56862745098039, 21.3793103448]
	https://scontent		Aleksandr Karp	2020-11-08 20:08:09	55	498	От рассвета до заката	0	25.090909090909093, 61.76470588235294, 28.205128205128	2 309.2307692307692, 18.23529411764706, 13.97849462365
		aantoniomenaa.	A000000 M000	2020-11-08 20:07:27	76	822			[6.923076923076925, 7.8431372549019605, 65.0000000000000	
	https://scontent		Lars Joelsson	2020-11-08 20:02:45	10	50	Kvällsdrink på strandpromenads		[310.0000000000001, 68.23529411764706, 7.40740740740741	
		rocioversamador	Rocio Vera Amador	2020-11-08 20:02:38	11	167	Mira el mar y el te dará la respa		[17.9999999999995, 70.19607843137254, 6.57894736842104	
29	https://scontent	el_chriis	†_El Gufii.†	2020-11-08 20:00:10	4	680	La vida nos golpea y nos cierral	0	[27.096774193548388, 26.862745098039216, 22.62773722627	7 [28.965517241379317, 77.84313725490196, 25.66371681416
30	https://sconten	extraordinary.arch	ARCHITECTURE & DESIGN	2020-11-08 19:57:52	465	29969	#brutalism #architecture #bruta	['brutalistarchitecture', 'cor	18.46153846153845, 75.29411764705883, 20.6349206349206	5 [50.45454545454545, 23.52941176470588, 36.666666666666

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

- URL post
- Nick Usuario
- Nombre Completo
- Fecha publicación
- Likes
- Followers usuario
- Texto Post
- Hashtags
- Color Dominante 1
- Color Dominante 2



6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

No se ha hecho uso de ninguna fuente externa.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Como ya se ha mencionado al principio, las redes sociales pueden ser unas fuentes de información muy útiles pues éstas incluyen los gustos y las preferencias de la sociedad. Esto puede ser usado para un estudio de preferencias a la hora de vender productos y ofrecer servicios.

- 8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
- O Released Under CC0: Public Domain License
- 9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

https://github.com/kirilmn13/InstagramScraper/tree/main

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.



Todo el Trabaja ha sido Realizado por Kyrylo Morozov