

Uso de web scraping en RRSS para su análisis

Por:

María Nanton (<u>mcnanton@gmail.com</u>) @bynans1



Adquisición de datos masivos

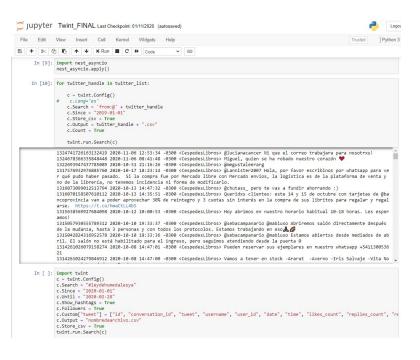
Herramientas y posibilidades

¿Qué es el Web Scraping?



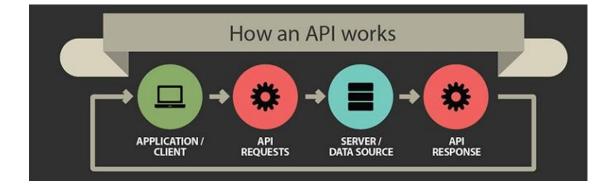
¿Cómo se scrapea un sitio?





¿Cuáles son las alternativas?





Instagram API

¿Cuáles son las alternativas?





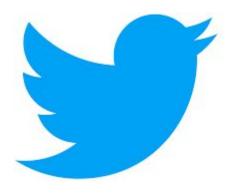




RRSS en contexto Covid-19

Análisis preliminar de estrategias de marketing y comunicación de librerías y editoriales

El corpus



25k de publicaciones de 26 usuarios de librerías y editoriales que comercializan libros en Argentina



41k de publicaciones de 40 usuarios de librerías y editoriales que comercializan libros en Argentina

El corpus

Users > PC > Documents > Untitled Folder 2 > base tw > III eternacadencia.csv id.conversation id.tweet.username.user id.date,time,likes_count,replies count,retweets count,hashtags,mentions,retweet,quote url,user rt id,source,retweet date 1322951033316921352,1322951033316921352,¿Y usted? https://t.co/BGSBExtHtk,eternacadencia,34765906,2020-11-01,14:18:00,30,4,5,[],[],False,,,, 1322924861090598913,1322924861090598913,"""Me han dicho que mis obras traducidas al español no han quedado muy bien, pero no sé, no hablo español"". https://t.co/QA90EQmDPI",eternacadencia,34765906,2020-11-01,12:34 322894721149186948,1322894721149186948,Leila Guerriero presenta su nuevo libro. La entrevista @LalaKarenina #agendalit,eternacadencia,34765986,2020-11-01,10:34:14,24,0,3,['agendalit'],"[{'screen name': 'lalaKarenina 285392384,1322874781285392384,"Las dependencias, un perfil de la escritora Silvina Ocampo realizado en 1999 https://t.co/MbYxQFCFP4",eternacadencia,34765906,2020-11-01,09:15:00,80,2,15,[],[],False,,,, 2444185600,1322697362444185600,"""La cantidad de cuentos que logró publicar parece interminable y por ellos desfilan, literalmente, miles de personajes"": tomado de Cuentos selectos (@EdhasaArgentina), aquí 7956569888,132266489795656988,Asesinada por su ex marido dejó una obra extraordinaria: fue admirada por Alfonsina Storni y Rubén Dario https://t.co/XtJKDS8vHH.eternacadencia,34765966,2020-10-31,19:21:00,88 322646778458509312,1322646778458509312,"""Yo soy una poeta que se mete, un poco intrusa, en otras áreas"" #MariaNegroni https://t.co/h2d0iDoHXN".eternacadencia,34765906,2020-10-31,18:09:00,53,0.9,['marianegroni'], 1322600725159976960.1322600725159976960."""Nadie debiera sentirse obligado a terminar un libro que no le gusta"" @FlaviaPitella https://t.co/4U1jnn5hKf".eternacadencia.34765906.2020-10-31.15:06:00.28.1.6.[]."[{'scre 1322571029462437890.1322571029462437890."""Poco a poco se iría transformando en una novela mía, y no sé si podría seguir tratándose de una traducción"" César Aira https://t.co/LKMmOrkWU1".eternacadencia.34765906.202 1322560711365570561.1322560711365570561."""Cuando vo era chico, una vez me regalaron una estación de bomberos de juguete, algo que no me interesaba particularmente, pero encontré que, si usaba mi imaginación, vo era 1322518432785616904,1322518432785616904,"Joyce Carol Oates: ""La soledad da origen a la literatura y otras veces puede ser un impedimento"" https://t.co/mGXE0FMBy3",eternacadencia,34765906,2020-10-31,09:39:00,22,0,3 1322379492766212096,1322379492766212096,"""El fin del concepto de publicación, tal como lo entendíamos hasta ahora"", por @GSchavelzon: https://t.co/6WhSwYsINV",eternacadencia,34765906,2020-10-31,00:26:54,13,0,6,[] 1322346298477932544,1322346298477932544,"""Mi mente es un suave espejo del mundo"". Dos poemas inéditos de Mircea #Cārtārescu en versiones de @LalaKarenina. https://t.co/CFWr1AlFXW",eternacadencia,34765906,2020-10 1322314337994887169,1322314337994887169,"""ya he visto cómo se pudre la carne se parece al color de algunas flores"" Tres #poemas de Lila Biscia (@Liluxb), tomados de La casa del tornado (@bajolaluna). https://t.c 1322299993462435842,1322299993462435842,"Virginia Woolf se negó a editar el Ulises: no quiso publicarla en el sello que mantenía con su marido, Hogarth Press. Las tiradas allí eran de 300 ejemplares. https://t.co/Ui 1322292460127739905,1322292460127739905,jSe viene la segunda! #leechild,eternacadencia,34765906,2020-10-30,18:41:04,2,0,0,['leechild'],[],False,https://twitter.com/blattyrios/status/1322197737937448961,,, 1322291880948846592,1322291880948846592,"La verdad increíble, de Tove Jansson, novedad de @CiaNavieraIltda recién llegada. #novedades https://t.co/rnsdmV0RcM",eternacadencia,34765906,2020-10-30,18:38:46,15,0,0,['nov 1322273317710000129,1322273317710000129,#LIJ y tabú: ¿Desde cuándo la muerte es tabú en los libros para chicos? Sandra Comino deja una guía generosa de lectura alrededor de este tema. https://t.co/ngUKxHuddk,eterna 1322240853830733825,1322240853830733825,"""La torta de naranja no tenía una receta legendaria, pero era rica y a los turistas les gustaba."" ""Torta de naranja"", un cuento de @taligoldman, tomado de su libro Larga 1322216658971332610,1322216658971332610, Qué novedad llegó de @las_afueras? #novedades https://t.co/ot0gEy3dhQ.eternacadencia,34765906,2020-10-30,13:39:51,16,3,1,['novedades'],"[('screen_name': 'las_afueras', 'name 1322214995418095617,1322214995418095617,"Cantinero @ChaiEditora, ¿me sirve Otro Manhattan? #novedades https://t.co/yc3ua7tuD4",eternacadencia,34765906,2020-10-30,13:33:15,32,2,2,['novedades'],"[{'screen_name': 'cha 1322210322011357185,1322210322011357185, "Esta semana, por suerte, salió El sol... #novedades https://t.co/artMjPs6vf",eternacadencia,34765906,2020-10-30,13:14:40,26,1,2,['novedades'],[],False,,,, 1322204111782191104.1322204111782191104."""Nunca hice carrera de nada, ni siguiera de la escritura, Arrangué con nada en el mundo pero una cierta pasión, un deseo impulsor. No sé de dónde vino y no sé por qué"". ht 1322191193107763202.1322185130555953152.@orsenix En tal caso te dejo la sección de mi blog donde cargamos poesía cada miércoles. Oialá encuentres algo para vos: https://t.co/RijdknDi7R.eternacadencia.34765906.2020-1 1322185130555953152,1322185130555953152,"¿Quiénes son sus poetas favoritas vivas? ¿Y poetas favoritos? De cualquier parte del mundo, pregunto...", eternacadencia, 34765906, 2020-10-30, 11:34:34, 41, 43, 4, [], [], False,,,, 1322178946184679429,1322178946184679429,Novedades para despedir octubre 😚 Entramos en un nuevo mes con cinco libros geniales que te presentamos en #labandejadelosviernes. @LeteoEditor @altopogook @lapollera @Nordi 1322148746860371971,1322148746860371971, "Una conversacón con @paulavazquez , escritora, abogada y también librera, que por estos días inaugura junto a su socio la segunda sucursal de @LataPeinada, en Madrid. ""Sólo 1321983658962620416,1321983658962620416,"Lectura de martes: ""El principio de las cosas"", gran cuento gran de la autora inglesa Ali Smith, incluido en la antología La historia universal (@Nordica Libros). https:/ 1321951698592608263,1321951698592608263, "Un extracto de la novedad de @Ampersand ed: El libro expandido. Variaciones, materialidad y experimentos, de #AmaranthBorsuk. https://t.co/HFJ;Dblmkl",eternacadencia,347659 1321910678379257863,1321910678379257863,Los consejos para escribir que #Poe dejó como pistas en sus obras. Un taller literario imprescindible en 7000 caracteres. https://t.co/ijS3NDViSx.eternacadencia,34765906,2020 1321878214277537797,1321878214277537797,"""La lección mayor del infinito es dejar de ser a veces infinito."" Tres #poemas de Roberto #Juarroz, tomados de Poesía vertical (@VisorLibros). https://t.co/cefyf-3k2A",et 1321849021938331648,1321849021938331648,"Seguimos con fotos de nuestro paisaje favorito, las bibliotecas 🎏 Aquí la de Damián Huergo, quien dice acerca de su orden: ""El plan original fue por género. Luego el azar y 1321815048617037824,1321815048617037824,"""Ya no sabemos nada sobre nuestro futuro personal, mucho menos sobre el futuro global. Lo que me parece seguro es que ya nada volverá a ninguna normalidad"". Leé la introducc 1321785100934254592.1321785100934254592,"Leé el arranque de El gueto interior, de Santiago Amigorena (Amegustaleerarg) #novela finalista de los tres grandes premios de las letras francesas en 2019 y traducida por Amagorena 1321621019430301696.1321621019430301696."Nueve preguntas a Cristina García Morales, La escritora española, autora de Lectura fácil, responde nuestro cuestionario fijo. https://t.co/CIRZblg20y".eternacadencia.347659 1321591575500091392.1321591575500091392.Nueve preguntas a @francovaccarini. El autor de más de ochenta títulos para un público infantil y juvenil responde nuestras preguntas de siempre. https://t.co/FbFfZboull.eter 1321558356994215936.1321558356994215936."""El corazón de un hombre es algo vergonzoso"". #SeiShonagon #ElLibrodelaAlmohada. https://t.co/mwgC03DFKV".eternacadencia.34765906.2020-10-28.18:04:00.7.0.1."['seishonagon 1321532939184865286.1321532939184865286.""Tos niños al nacer saben leer el mundo, no los libros; nos hacemos lectores viendo leer""; Mercedes Calvo sobre mediación lectora, poesía e infancias #LIJ https://t.co/øf 1321526032084672513,1321526032084672513,";Felices 90, @Margo Glantz! 💠 https://t.co/MBevn05ZH1 https://t.co/DHbXehr7qW",eternacadencia,34765906,2020-10-28,15:55:33,56,0,5,[],"[{'screen name': 'margo glantz', 'name 1321500475296985088,1321500475296985088,"""Un cambio de orgánico"", un #cuento de #Fogwill. https://t.co/cyMr48kzoj",eternacadencia,34765906,2020-10-28,14:14:00,20,0,3,"['cuento', 'fogwill']",[],False,,,, 1321493680348368903,1321469521266442240,"@noralcira11 ;Nunca es tarde, Nora!",eternacadencia,34765906,2020-10-28,13:47:00,4,1,0,[],[],False,,,, 1321493409123762177,1321469521266442240, @GuadaMadrazo Graciela, enorme. https://t.co/xYeu9z0SyV",eternacadencia,34765906,2020-10-28,13:45:55,5,1,0,[],[],False,,,, 1321486634009858048,1321486634009858048,"El Observatorio de palabras de la @RAEinforma incluye el ""elle"": no determina la implementación de este término inclusivo, pero su incorporación abre el debate https://t.co 1321471781186785282,1321469521266442240,@misantapeluca 💠 https://t.co/YjIiaX0yE2,eternacadencia,34765906,2020-10-28,12:19:59,6,0,0,[],[],False,,,, 1321469521266442240,1321469521266442240,20al es el primer libro que recuerdan haber leido?,eternacadencia,34765906,2020-10-28,12:11:00.154,529,9,[],[],False,... 1321452660763947009,1321452660763947009,"""5i la no violencia parece una posición «débil», deberíamos preguntarnos: ¿qué se considera fuerza?"". Leé un fragmento de La fuerza de la no violencia, de #JudithButler (@Pl 1321421958295265280,1321421958295265280,"""[¿te acordás de aquella vez que paraste en la mitad de un poema para contarme de una ciudad sin sombras?]"". Tres #poemas de Marília Garcia, tomados de Cámara lenta (@Zind 221258380036485120.1321258380036485120."Compartimos uno de los textos incluidos en el Dossier descargable gratuito realizado en el marco de las Jornadas A Propósito de Juana Bignozzi (Buenos Aires. 1937-2015).

sbsargentina -

Frecuencia de publicación

Comparación de publicaciones mensuales Año/Año



0	-0.52	-0.41	0.16	2.27	0.08	3.22	10.33	2	1.16	1
-	-0.43	-0.58	-0.5	-0.45	0.12	0	-0.47	-0.36	-0.5	-0.6
1	0.07	0.61	-0.46	0.1	0.13	0.81	0.52	0.94	0.28	0.16
10	1.33	-0.27	-0.44	-0.74	-0.84	-0.58	-0.73	-0.68	-0.58	-0.48
1		-0.29	-0.5	0	-0.9	0.43	-0.42	0		-0.5
ī	-0.11	-0.17	-0.57	-0.37	-0.04	-0.1/ -0.18	-0.37	-0.39	-0.47	-0.32
	-0.59 -0.75	-0.89 -0.5	0.17	0 77	0.29 0.83	W. 14	0.36 2.33	-0.38 2	0.28 0.6	0.11 1.17
	0.27	-0.52	-0.28	0.1	0.63	1.4	0.2	-0.22	-0.38	-0.17
j	-0.76		-0.26	-0.12	-0.81	-0.43	-0.72	-0.22	0.03	-0.17
81	V.10	0.33	-0.07	0.62	-0.15	-0.17	0.01	-0.24	-0.03	0
-	0.5	0	0.78	0	0.55	0.74	-0.11	-0.17	-0.34	-0.42
-	-0.39	0.46	0.1	0.03	-0.2	0.31	-0.02	-0.04	0.02	0.12
	Inf	11	12	2	3.33	Inf	1.33	9	11	Inf
		-0.82	-0.74	-0.89	-0.86	-0.8	-0.92	-0.93	-0.94	-0.85
2	-0.5	-0.56	-0.11	0.75	-0.57	-0.38	0.07	-0.33	0	-0.3
	0.33 1.5	-0.9 -0.6	-0.04	0.09	2.08 -0.86	0.89	0.33	0.05	0.65 2.33	0.09
	0.55	0.4	0.5	-0.6	-0.80 -0.78	0.39	0.65	0.33	_0.03	Inf 0.44
J	0.55	-0.5	-0.36	0	1	0.36	-0.6	0.55	0.67	-0.27
-	-1	0	-0.71	-0.54	1.67	2.33	1.67	Inf	0.07	1.75
-	1	4	8	Inf	1.5	4	Inf	-0.38	-0.11	-0.44
1	1.17	0.79	0.92	-0.23	-0.42	-0.22	-0.68	-0.67	-0.76	-0.76
1	Inf	-1	-0.33	3	Inf	-0.5	Inf	0.33	-0.75	0
7	0.04	0.08	0	0.28	0.65	0.95	1.13	-0.18	-0.91	-1
H	0.5	-0.38	5 -0.42	-1 -0.29	1.67	1.67	2	-0.33	0	8
	0.5 Inf	Inf	Inf	Inf		0.16 Inf	0.28	0.36	0.41	0.14
	-0.17	-0.27	0.19	0.06	Inf -0.21	_0.07	0.67	0.17	0	-0.25 -0.5
	-0.11	-0.17	0.19	-0.09	0	0	0.07	-0.03	0.07	0.29
	0.75	0.69	0.73	-0.25	0.02	0.16	-0.26	0.72	-0.34	-0.38
٠	2.75	1.12	2.5	0.5	0.86	0.69	0.83	0.35	0.81	0.27
- 1	-0.75	-1	-0.78	-1	-1	-1	7.5	6.6	13.5	5
1	-0.29	-0.07	0.42	-0.15	0.21	1.78	11.5	3.5	1	Inf
1	-0.43	-0.8	0.12	0.1	0.08	-0.22	0.55	0.25	0	1
ľ	0.08	-0.84	-0.21 -0.88	-0.87	-0.23 -0.88	0 71	-0.08	0.15	0.56	0
1	3.43	3.14	3.29	3.29	0.74	4.8	1.46	1.29	1.83	0.57 0.22
J	0.88	12	2	2.38	0.74	3.57	9	Inf	3.8	4.25
	1	17	T	- 1	1	- 1	1	- 1	1	
	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10

Mes del 2020

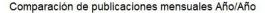
factor(valor)

negativo

nulo

positivo

Frecuencia de publicación





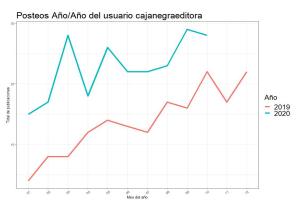
factor(valor)

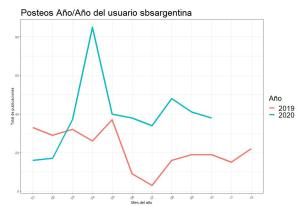
negativo nulo

positivo

Frecuencia de publicación





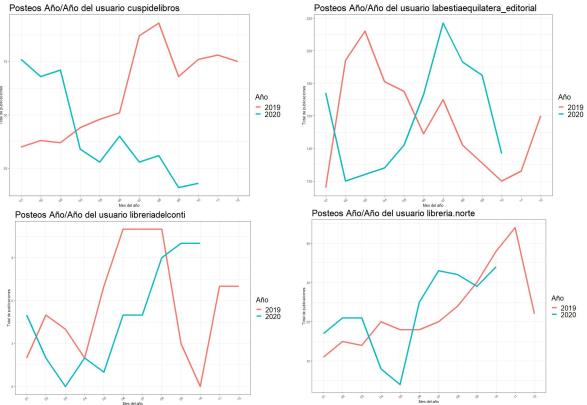






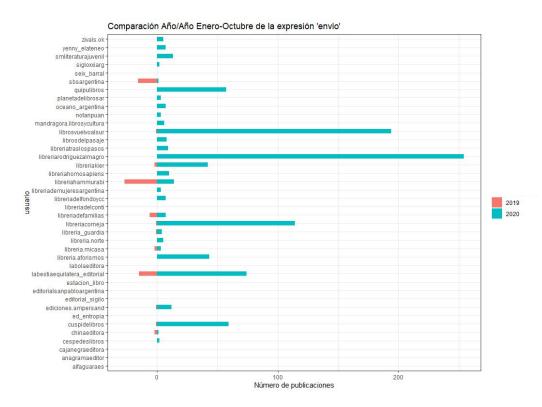
Frecuencia de publicación



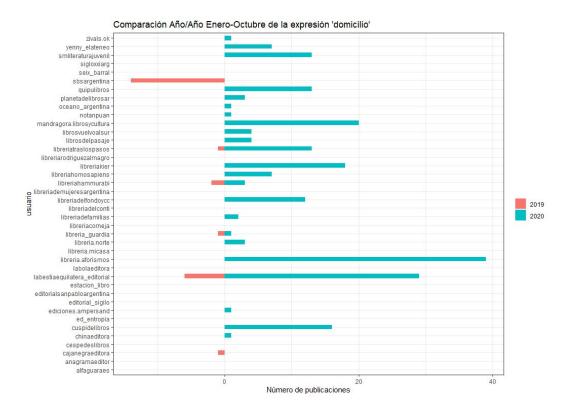


Tácticas de marketing - uso de RRSS como proxy

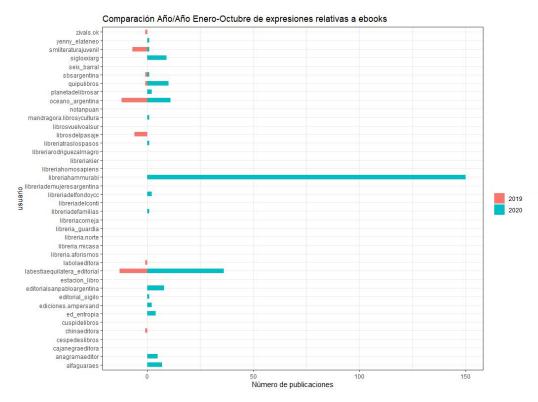




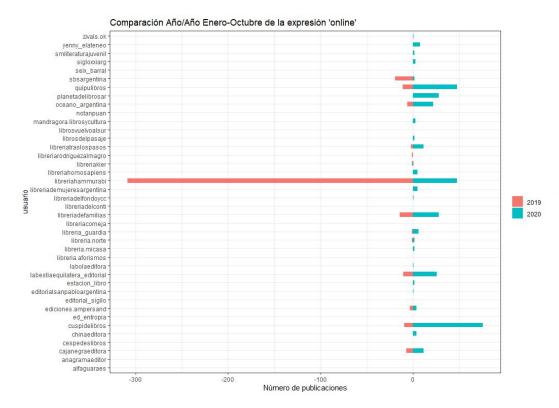




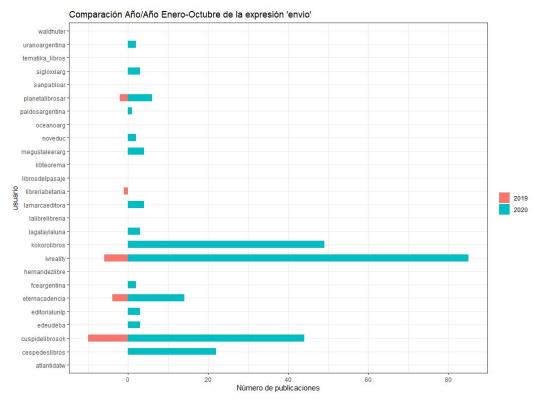




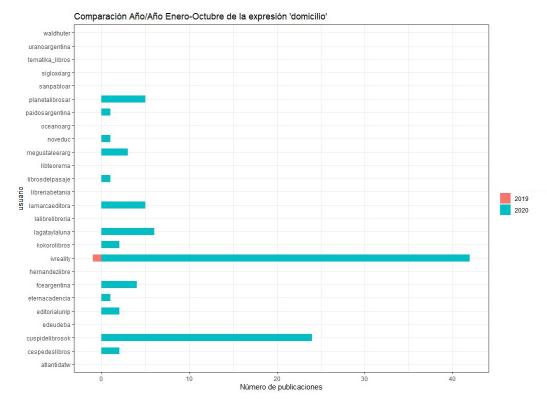




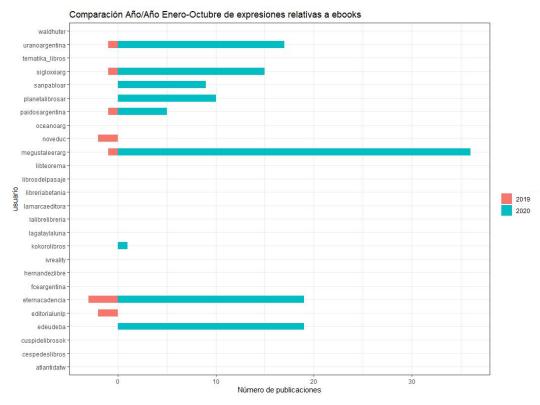




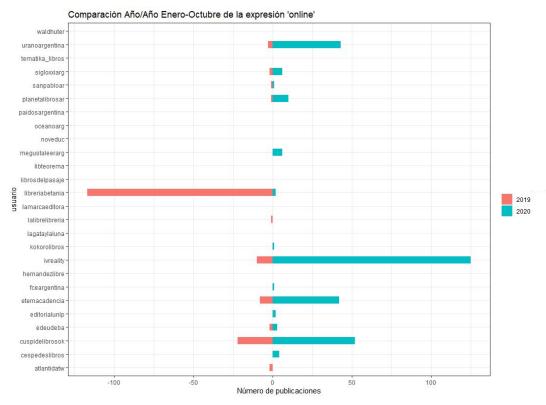












Próximos pasos

- Construcción de segmentos de usuarios
- Análisis de la efectividad de las comunicaciones
- Incorporación de usuarios de Facebook al corpus analizado
- Incorporación de algoritmos de detección de texto en imágenes



¡Muchas gracias!



Código y datasets <u>abiertos</u> disponibles en Github https://github.com/bynans/congreso-tendencias-2020/

María Nanton (<u>mcnanton@gmail.com</u>) @bynans1