The Twitter Alliance



Descubriendo las relaciones entre los usuarios más seguidos de Twitter

David Arroyo Segovia Adrián Camacho Pérez Paula Muñoz Lago Carla Paola Peñarrieta Uribe

Facultad de Informática - Universidad Complutense de Madrid



Objetivos



- Encontrar insiders. Descubrir que cuentas son seguidas por personas clave en la red social Twitter. Estas cuentas suelen pertenecer a líderes de opinión que no tienen porqué contar con el mayor número de seguidores.
- **Desvelar las cuentas favoritas de un grupo determinado.** Qué cuentas son las más seguidas por una comunidad de usuarios, por ejemplo, el top 100 de España, políticos o periodistas.
- Identificar nodos influyentes. Determinar miembros clave de la comunidad que a menudo son desconocidos para los outsiders.

Primeras pruebas: T-Hoarder, Twecoll, Scrapy y Tweepy

T-Hoarder

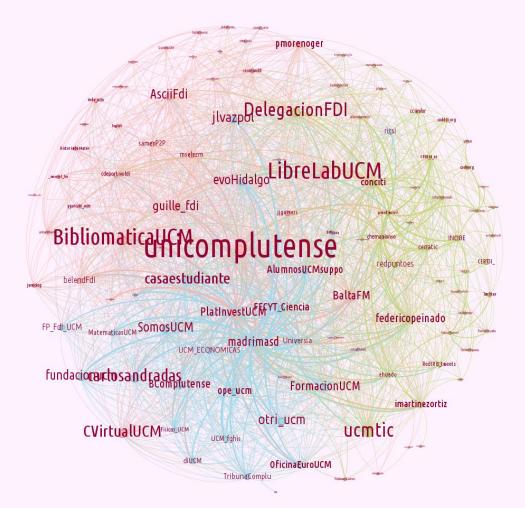


- T-Hoarder es una plataforma desarrollada por M. Luz Congosto (investigadora de la UC3M) para extraer información de Twitter a largo plazo. Esta información es procesada y visualizada automáticamente.
- Es una herramienta muy potente y con posibilidad de generar grafos.
- Sin embargo, no nos proporciona la suficiente flexibilidad para extraer la información que necesitamos.
- GitHub: https://github.com/congosto/t-hoarder_kit

Twecoll: Twitter Collection Tool



- Es una herramienta desarrollada en **Python** que se utiliza desde la línea de comandos para extraer datos de Twitter.
- También da la posibilidad de **generar grafos** a partir de la información extraída.
- Genera grafos de tus **relaciones de primer y segundo grado** en Twitter.
- Toma mucho tiempo debido a los API throttling. Una de las razones por las cuales no nos ha sido útil.
- Extrae todos los seguidores de una cuenta y los de todos los seguidores de los anteriores (ignorando los que no se encuentren relacionados con el primero).
- GitHub: https://github.com/jdevoo/twecoll



Followings: 114 # Followers: 1629

La solución

Scrapy

- Framework para scraping y web crawling desarrollado en Python.
- Librería Open Source que nos permite extraer información importante que se encuentra en internet. Es posible leer el contenido de una web y movernos en su estructura HTML para analizar sus datos y explotarlos según sean nuestras necesidades.
- Lo utilizamos para automatizar la extracción de las listas de cuentas más seguidas de cada país.
- https://scrapy.org/



Tweepy



- Es una librería escrita en Python que te permite acceder a todos los métodos que proporciona la API REST de Twitter.
- Proporciona una interfaz muy sencilla encapsulando algunas partes que nos permiten agilizar el trabajo.
- Permite realizar esperas cuando se ha superado el rate limit que impone Twitter para no detener la ejecución del script.
- http://www.tweepy.org/

```
avid@parrot-server:~/Code/The-Twitter-Alliance/Code/TwCounterSpider$ python3 main.py --country spain
                                                                                 🔞 🖨 📵 david@parrot-server: ~/Code/The-Twitter-Alliance/Code/TwCounterSpider
                                                                                2018-01-14 21:58:53 [scrapy.core.scraper] DEBUG: Scraped from <200 https://twittercounter.com/pages/100/sp
     Map the relationships between the most followed accounts
                                                                                {'filename': 'Top100_spain.txt',
                                                                                  'user_accounts': ['@realmadrid',
918-01-14 21:58:51 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: TwCounterSpider)
018-01-14 21:58:51 [scrapy.utils.log] INFO: Versions: lxml 4.1.1.0, libxml2 2.9.7, cssse
                                                                                                           '@FCBarcelona',
1.3.1, w3lib 1.18.0, Twisted 17.9.0, Python 3.5.2 (default, Nov 23 2017, 16:37:01) - [G
                                                                                                           '@andresiniesta8'.
pyOpenSSL 17.5.0 (OpenSSL 1.1.0g 2 Nov 2017), cryptography 2.1.4, Platform Linux-4.4.
64-with-Ubuntu-16.04-xenial
                                                                                                           '@AlejandroSanz',
018-01-14 21:58:51 [scrapy.crawler] INFO: Overridden settings: {'SPIDER_MODULES': ['TwCo's'], 'BOT_NAME': 'TwCounterSpider', 'NEWSPIDER_MODULE': 'TwCounterSpider.spiders'}
                                                                                                           '@3gerardpique',
018-01-14 21:58:51 [scrapy.middleware] INFO: Enabled extensions:
                                                                                                           '@Pontifex es'.
'scrapy.extensions.memusage.MemoryUsage'
                                                                                                           '@RafaelNadal',
'scrapy.extensions.telnet.TelnetConsole'
scrapy.extensions.corestats.CoreStats
                                                                                                           '@enriqueiglesias',
scrapy.extensions.logstats.LogStats'
                                                                                                           '@SergioRamos',
018-01-14 21:58:51 [scrapy.middleware] INFO: Enabled downloader middlewares:
scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
                                                                                                           '@FCBarcelona_es',
'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware'.
                                                                                                           '@XabiAlonso',
scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware'
                                                                                                           '@Carles5puyol',
                                                                                                           '@Rubiu5'.
                                                                                                           '@D_DeGea',
                                                                                                           '@davidbisbal'.
                                                                                                           '@cesc4official',
                                                                                                           '@realmadriden',
                                                                                                           '@Guaje7Villa',
                                                                                                           '@muyinteresante',
                                                                                                           '@IkerCasillas',
                                                                                                           '@juanmata8',
                                                                                                           '@paugasol',
                                                                                                           '@el_pais',
                                                                                                           '@FCBarcelona_cat',
                                                                                                           '@aarbeloa17'.
                                                                                                           '@vegetta777',
                                                                                                           '@mangelrogel',
                                                                                                           '@WillyrexYT'
```

Extracción del top 100 de Twitter España de TwitterCounter.com

👹 🗇 🙂 david@parrot-server: ~/Code/The-Twitter-Alliance/Code/TwCounterSpider

Algoritmo: descubrir amistades

01100 10110 11110

- De cada una de las cuentas anteriormente extraídas se comprueba si existe un seguimiento mutuo con todas las demás.
- Si existe un seguimiento mutuo decimos que hay una **relación de amistad**.
- Se intentan hacer las menos peticiones a la API posibles: cacheo de peticiones pasadas, optimización del código, etc.
- El resultado se exporta a un fichero de texto plano que es posteriormente parseado con un script
 en C++ que genera la relación de nodos y aristas.
- El resultado final es un **grafo no dirigido**.

Algoritmo: all follows

01100 10110 11110

- A diferencia del procedimiento anterior, se **extraen los follows** que existen entre los usuarios de la lista aunque **no sean correspondidos** (sin ser mutuos).
- Se generan muchas más aristas que en el caso anterior.
- El resultado final es un grafo dirigido.



david@parrot-server: ~/Code/The-Twitter-Alliance/Code/TwitterAlliance anpaZurita', 'source_screen_name': b'@elmundoes') Tiempo ejecución script ≈ 13 horas 2018-01-15 11:44:36,939:INFO:Starting new HTTPS connection (1): api.twitter.com 2018-01-15 11:44:37.133:INFO:TOTAL COMPLETED: 29% 2018-01-15 11:44:37,133:INF0:PARAMS: {'target_screen_name': b'@instagram', 'source_screen_name': b'@elmundoes'} 2018-01-15 11:44:37,144:INFO:Starting new HTTPS connection (1): api.twitter.com 2018-01-15 11:44:37,408:INFO:TOTAL COMPLETED: 29% 2018-01-15 11:44:37,408:INF0:PARAMS: {'target_screen_name': b'@marcmarquez93', 'source_screen_name': b'@elmundoes'} 2018-01-15 11:44:37,419:INFO:Starting new HTTPS connection (1): api.twitter.com david@parrot-server: ~/Code/The-Twitter-Alliance/Code/TwitterAlliance 2018-01-15 11:44:37,627:INFO:TOTAL COMPLETED: 29% 2018-01-15 11:44:37,634:INF0:PARAMS: {'target_screen_name': b'@MesutOzil1088', 'source_sc2018-01-15 13:14:57.203:INF0:Fi @AnnaSimonMari - @danimartinezweb detected 2018-01-15 11:44:37,639:INFO:Starting new HTTPS connection (1): api.twitter.com @AnnaSimonMari - @3gerardpique detected! 2018-01-15 13:14:59.727:INFO: 2018-01-15 11:44:37,841:INFO:TOTAL COMPLETED: 29% 2018-01-15 13:15:08,224:INFO: @AnnaSimonMari - @flofdezz detected! 2018-01-15 11:44:37,841:INFO:PARAMS: {'target_screen_name': b'@TownGamePlay', 'source_scr'2018-01-15 13:15:14,729:INFO: @AnnaSimonMari - @El_Hormiguero detected! 2018-01-15 11:44:37,846:INFO:Starting new HTTPS connection (1): api.twitter.com 2018-01-15 13:30:00.276:INFO: @Reuters - @TheEconomist detected! 2018-01-15 11:44:38,086:INFO:TOTAL COMPLETED: 29% 2018-01-15 13:44:56,389:INFO: @andresiniesta8 - @lorenzo99 detected! 2018-01-15 11:44:38,086:INFO:PARAMS: {'target_screen_name': b'@rihanna', 'source_screen_ 2018-01-15 13:44:57,349:INFO: @andresiniesta8 - @Guaje7Villa detected! 2018-01-15 11:44:38.093:INFO:Starting new HTTPS connection (1): api.twitter.com 2018-01-15 13:44:58,813:INFO: @andresiniesta8 - @ristomejide detected! 2018-01-15 11:44:38.290:INFO:TOTAL COMPLETED: 29% @andresiniesta8 - @FCBarcelona_cat detected! 2018-01-15 11:44:38.290:INFO:CURRENT SOURCE USER: @MarcBartra 2018-01-15 13:44:58,813:INFO: @andresiniesta8 - @JordiAlba detected! 2018-01-15 11:44:38,290:INFO:PARAMS: {'target_screen_name': b'@MelendiOficial', 'source_s2018-01-15 13:45:01,610:INFO: 2018-01-15 13:45:01,852:INFO: @andresiniesta8 - @SSantiagosegura detected! 2018-01-15 11:44:38,297:INFO:Starting new HTTPS connection (1): api.twitter.com 2018-01-15 13:45:05,328:INFO: @andresiniesta8 - @llorentefer19 detected! 2018-01-15 11:44:38,545:INFO:TOTAL COMPLETED: 29% 2018-01-15 13:45:05,560:INFO: @andresiniesta8 - @FCBarcelona_es detected! 2018-01-15 11:44:38,545:INFO:PARAMS: { 'target_screen_name': b'@mario_casas_', 'source_scr₂₀₁₈₋₀₁₋15 13:45:05,858:INFO: @andresiniesta8 - @Carles5puyol detected! 2018-01-15 11:44:38,550:INFO:Starting new HTTPS connection (1): api.twitter.com 2018-01-15 13:45:07,237:INFO: @andresiniesta8 - @MarcBartra detected! 2018-01-15 11:44:38.770:INFO:TOTAL COMPLETED: 29% @andresiniesta8 - @aarbeloa17 detected! 2018-01-15 13:45:13.825:INFO: 2018-01-15 11:44:38.770:INFO:Friendship @MarcBartra - @lorenzo99 detected! @andresiniesta8 - @paugasol detected! 2018-01-15 13:45:14,451:INFO: 2018-01-15 11:44:38,770:INFO:TOTAL COMPLETED: 29% 2018-01-15 13:45:15,685:INFO: @andresiniesta8 - @3gerardpique detected! 2018-01-15 11:44:38,770:INFO:PARAMS: {'target_screen_name': b'@zaynmalik', 'source_scree 2018-01-15 13:45:18,112:INFO: @andresiniesta8 - @D DeGea detected! 2018-01-15 13:45:18,384:INFO: @andresiniesta8 - @SeFutbol detected! 2018-01-15 13:45:19,792:INFO: @andresiniesta8 - @ctello91 detected! 2018-01-15 13:45:21,543:INFO: @andresiniesta8 - @Thiago6 detected! 2018-01-15 13:45:21.759:INFO: @andresiniesta8 - @DANIROVIRA detected! 2018-01-15 13:45:22,670:INFO: @andresiniesta8 - @jordievole detected! 2018-01-15 13:45:26,896:INFO: @andresiniesta8 - @Buenafuente detected! 2018-01-15 13:45:27,122:INFO: @andresiniesta8 - @SergioRamos detected!

2018-01-15 13:45:31,179:INFO:

2018-01-15 14:00:00,209:INFO:F

avid@parrot-server:~/Code/The-Twitter-Alliance/Code/TwitterAlliance\$

Logs generados al ejecutar el algoritmo sobre el top 100 de Twitter España



@andresiniesta8 - @_Pedro17_ detected!

@andresiniesta8 - @marcmarquez93 detected!



Resultados preliminares: primeras redes

@fifacom_es @CesarAzpi @LaLiga @realmadrid @SeFutbol @fifaworldcup_es @19SCazorla @AnderHerrera @nachofi1990 @juanmata8 @DaniCarvajal92 @ctello91 @realmadriden @SergiRoberto10 @FCBarcelona @AlvaroMorata @ Pedro17 @JeseRodriguez10 @Torres @aarbeloa17 @Thiago6 @XabiAlonso @FCBarcelona es @llorentefer19 @Rafinha @D DeGea @JordiAlba @isco alarcon @officialpepe @FCBarcelona cat @SergioRamos @Guaje7Villa @taniarin @21LVA @3gerardpique @GUTY14HAZ @MarcBartra @VogueSpain @2010MisterChip @Carles5puvo @enriqueiglesias @cesc4official @lkerCasillas @Pontifex_es @andresiniesta8 @mundodeportivo @BBV#world @diarioas @danimartinezweb @RefactNadat @rickynubio9 @policia @ danielmartin @paugasol @ifilosofia @AbrahamMateo @ActualidadRT @marcmarquez93 @A3Noticias @MelendiOficial @AlejandroSanz @ristomeiide @EUuanpaZurita @davidbisbal @ MaluOficial @aLexBY11 @lorenzo99 @alo oficial @BoseOfficial @LuzuVlogs @LuzuGames @AnnaSimonMari @_anapastor_ @pabloalboran @WillyrexYT @mangelrogel @SSantiagosegura @Buenafuente @elmundoes @vegetta777 @jordievole @auronplay @El Hormiguero @Rubiu5 @TownGamePlay @CristiPedroche @flofdezz @Cosmopolitan es @DANIROVIRA @Wismichu

@Pablo_Iglesias_

@perezreverte

@muyinteresante

@Berto_Romero

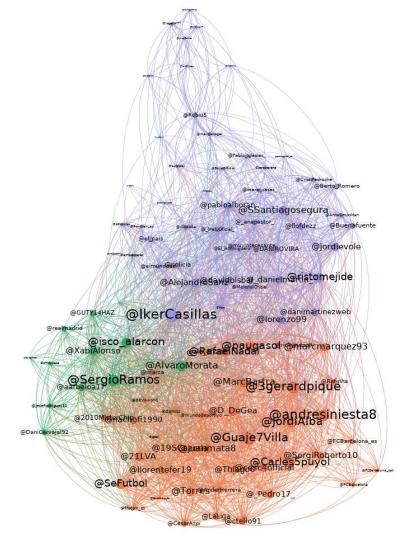
@mario casas

Red del top 100 de Twitter España



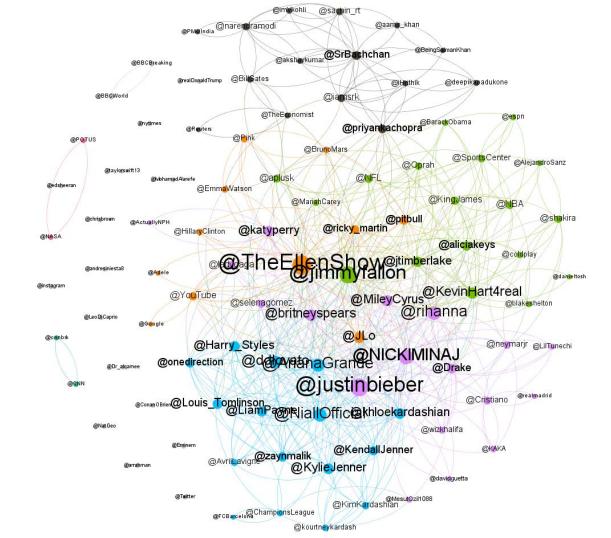
Red dirigida del top 100 de Twitter España



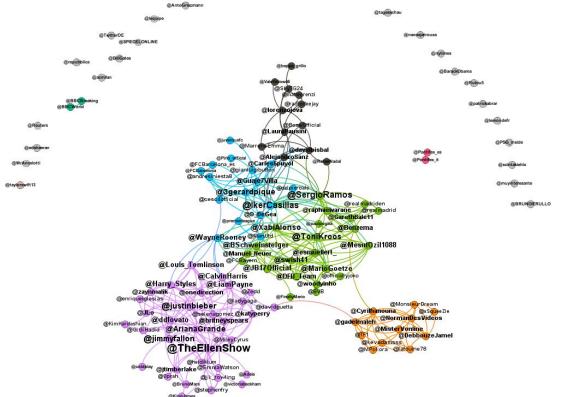


Red del top 100 de Twitter





Red de relaciones entre el top 20 de Twitter ESP, UK, USA, FR, IT, GER





@bepge grillo @lorenzojova @Laura Pausini **Odavidb**isbal @AlejandroSanz @Pirlo official @Carles puvol @FCBarcelona es @giantuigibuffon @andresiniesta8 @Guaje7Villa @3gerardpique @delperoale @cesc4official @lkerCasillas @raphaelvarane @realmadrid @D DeGea @GarethBale11 ®paulpogea @Benzema @XabiAlonso @ToniKroos @WayneRooney @Manufo @MesutOzil1088 @BSchweinsteiger @esmuellert @Manuel Neuer @swish41 @PCBayern @Louis Tomlinson @JB17Official @MarioGoetze @DFB Team @officiallyjoko @CalvinHarris @Harry Styles @LiamPayne @woodvinho @zaynmalik / @onedirection @FinallyMario @enriqueiglesias @ladygaga @MonsieurDream @justinbieber @davidguetta @Cyrilhanouna @xSquee Zie @selenagomez @katyperry @gadelmaleh @NormanDesVideos @ddlovato @britneyspears @GiGiHadid @ArianaGrande @Mister Vonline @DebbouzeJamel @immyfallon @MileyCyrus @kevadamsss @TheEllenShow @MPokora @lafouine78 @heidiklum @jtimberlake@EmmaWatson @Optah @ik rowling @BrunoMars @victoriabeckham

Red de relaciones entre el top 20 de Twitter ESP, UK, USA, FR, IT, GER



Trabajo por hacer

- Determinar sobre qué más grupos o comunidades realizar el análisis.
- Analizar los datos y comprender las relaciones.
- Interpretar y comparar las redes obtenidas.

