Author Profiling for Identifying Robots in Twitter

francisco.rangel@autoritas.es

Descargar el dataset de:

http://ow.ly/vWRiT

Al descomprimir se obtienen dos elementos:

- index/lukeall-1.0.1.jar ---> Visor de índices Lucene
- index/robots/* ---> Índice Lucene (+ contenido de los campos) creado con las librerías en su versión 3.0.1 y/o 3.0.2

El índice contiene los campos:

- id: Identificador único del tuit proporcionado por Twitter
- user: Username de Twitter
- date: Fecha del tuit en formato YYYYMMDDhhmmssl
- location: Localización del perfil del usuario
- content: Contenido del tuit
- followers: Número de followers en el momento de generación del tuit
- friends: Número de followees en el momento de generación del tuit
- favourites: Número de favoritos a tuits del usuario
- listed: Número de listas en las que aparece el usuario
- statuses: Número de tuits desde la creación de la cuenta de usuario
- username: Nombre completo del tuitero
- bio: Biografía del tuitero
- userurl: Url personal del tuitero, si la informó
- gender: Género del tuitero obtenido a partir de su nombre. Si no está infromado es porque el nombre no se encuentra disponible en el gazeeter, posiblemente, perfil institucional
- created: Fecha de creación del perfil, en formato YYYYMMDDhhmmssl
- dayssincecreation: Días desde la creación de la cuenta
- project: Proyecto del que se extraen los tuits y que se corresponde con las listas de robots del excel adjunto

El campo proyect puede tener cinco valores y se debe filtrar por él para obtener los datos concretos del proyecto sobre el que se desee trabajar:

- gasofa: 339.734 tuits en el rango 01/10/2013 – 31/12/2013
- castellon: 1.824 tuits en el rango 05/01/2013 – 28/02/2013
- senado: 64.014 tuits en el rango 10/03/2014 – 13/03/2014

- comite: 323.692 tuits en el rango 01/03/2014 – 03/04/2014
- rato: 323.692 tuits en el rango 01/03/2014 – 03/04/2014

Los listados de robots se encuentran en la hoja de cálculo de la carpeta “Datos Robots”.

La detección de estos robots se hizo manualmente por un consultor experto de Autoritas. Las características principales de estos usuarios serían las siguientes:

- Volumen muy elevado de publicaciones por día. Por encima de 30, incluso algo menos. El 40%, aprox, de los usuarios publican entre 1 y 10 tweets al día. Sólo e. 4,5% está por encima de 20. http://www.slideshare.net/fullscreen/adigitalorg/adigital-estudio-usotwitterenespaa2012def/10
- Suelen utilizar medios de publicación directa. Es decir, sindican varias fuentes de noticias. De ahí el alto volumen de publicación y que sean perfiles automatizados. Suelen utilizar TwitterFeed http://twitterfeed.com/
- Se pueden identificar redes específicas por seguir una determinada estructura. Por ejemplo los identificados en “gasofa” tienen en común que todos sus nombres de usuario empiezan por Gasofa. Otra característica es que utilicen similar diseño en sus perfiles como avatares o imagen de fondo.
- Muchos ni se molestan en “personalizar” el perfil, por lo que algunos cuentan con el avatar de serie. Son huevos.
- Los detectados recientemente suelen ser todos con avatares de chicos sudamericanos y con bios bastante inteligibles y con tipografía extraña: ڮ♈Jєα ה'Cαяℓч_Traductor - @Olle_Carly ->‹--- Λτι tє Fαℓτα Mucнσ Ƥαяα SєCσмσ γσσ ;:) ;; Sαиgяє @FlakiitoSexy_x3 . PiiN 22784E46 . . 28/3/13
- Las redes suelen publicar todos lo mismo de manera sincronizada.
- Hay bastantes usuarios de este tipo que suelen incorporar siempre un hashtag a modo de identificación. Muchos utilizan sus iniciales o nombre. Les servirá para llevar la monitorización o la atribución de publicación más fácilmente.

Gasofa

En el proyecto Gasofa es una constante la aparición de un grupo de bots que publican sobre en qué gasolinera de la zona se encuentra el carburante más barato. El listado se encuentra en la pestaña “gasofa” de la hoja de cálculo.

Castellón

En el proyecto Castellón, en el informe realizado con los datos entre el 5 de enero de 2013 hasta el 28 de febrero de 2013 se detectaron una red de bots que publicaba noticias de El País. La lista se encuentra en la pestaña Castellón.

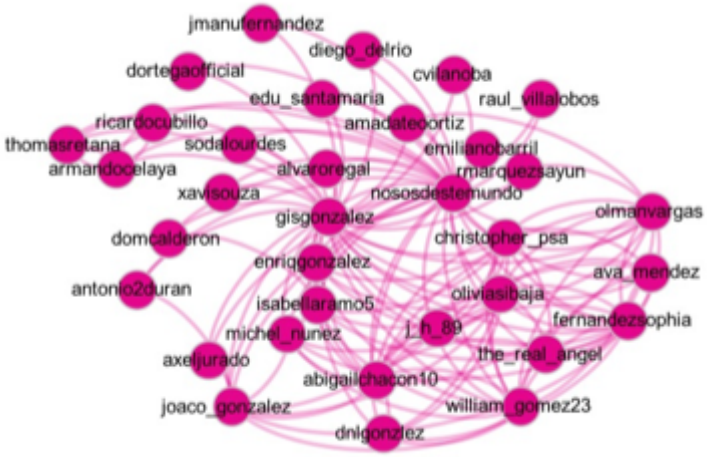
Extracto del informe:

La punta del día 7 de enero en Twitter se debe en parte a que 86 usuarios publicaron la

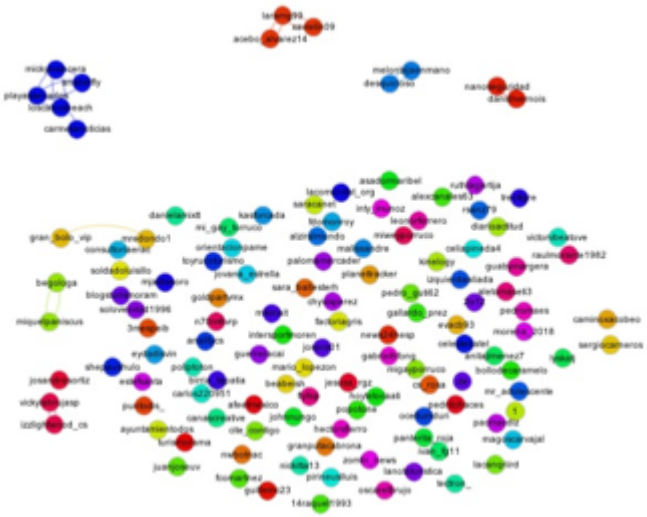
noticia publicada en El País "El Ayuntamiento de Castellón redactará un nuevo PGOU".

Estos 86 usuarios quedan fuera de la comunidad principal. Hecho extraño ya que es difícil que tantos usuarios "espontáneos" publiquen a la vez el mismo tweet. Se dividen en dos grandes grupos.

En el primero están todos conectados, además hay otra coincidencia. Todos, como se ve en la captura, al inicio de sus tweets introducen un hashtag con sus iniciales. Además, por supuesto, de publicar todos los mismo a la vez.



En el segundo grupo, ninguno está conectado pero tienen como elemento común que también publican todos los mismo a la vez. Hecho que demuestra que también pertenecen a una red de bots.



Acuerdo Senado Argentina

En el proyecto Senado, para la noticia sobre la presentación del acuerdo sobre la expropiación de

YPF a Repsol en el Senado Argentino, entre los días 10 y 13 de marzo de 2013 se detecta gran cantidad de bots. Están en la pestaña “acuerdo senado argentina”.

En este caso en el listado puede haber algún falso positivo. Para “identificarlos” se calculó la media de post por día y se incluyó en la lista todos aquellos que tienen un volumen de publicación superior a 30 tweets diarios. Las comprobaciones aleatorias que he hecho han sido positivas siempre…

Repsol crea un comité consultivo con sus accionistas minoritarios

En el proyecto Comité, entre 1 de marzo y el 3 de abril, para la noticia “Repsol crea un comité consultivo con sus accionistas minoritarios” se detecta un volumen inusual de publicación. La lista de usuarios que publicaron referencias a la noticia están en la pestaña “comité consultivo”. Puede haber algún falso positivo. Las comprobaciones aleatorias que he hecho han sido positivas siempre…

Rato, consejero de la inmobiliaria de La Caixa

Proyecto Rato, entre 1 de marzo y el 3 de abril. La noticia Rato, consejero de La Caixa, genera volumen inusual. Se detectan bots. La lista está en la pestaña “Rato - La Caixa”. Puede haber algún falso positivo. Las comprobaciones aleatorias que he hecho han sido positivas siempre…