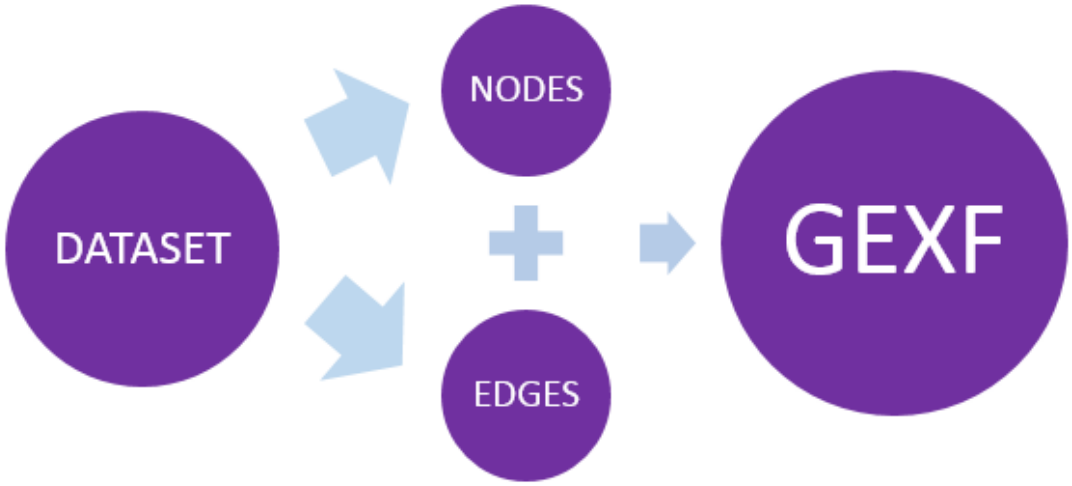
INTRODUCCIÓN:



Un archivo de datos para representar una red se divide en dos conjuntos de datos relacionados, uno corresponde a los nodos y el otro a las aristas.

Por tanto disponemos de cuatro BBDD:

1) Dataset: BBDD base estructurada por posts
2) NODES: BBDD estructurada por identidades
3) EDGES: BBDD estructurada por pares de identidades relacionadas entre sí que forman cada arista.
4) GEXF: BBDD bipartida que compila los datos de los nodos y aristas en un formato (GEXF) compatible con programas de análisis de redes (ej; Gephi)

1 - DATASET:

Nombre del archivo: DATASET Twitter-23-26-Mar-2014-MotoGP-Qatar

Descripción del contenido: Captura de datos del 23 al 26 de marzo. Contiene todo el ruido social generado en Twitter durante la carrera de Qatar Moto GP 2014 y los tres días posteriores.

Número total de variables: 19

Variables más relevantes: Source, Target, Body

Muestra de datos: (Ver tabla)

ID	PARENT-	Source	Mentions	Target	NAME Source	BODY	PUBDATE	URLs coma separated	Type TW-RT-MT
438372	437950	Motorsport79	box_repsol,marcmarquez93	box_repsol	Motorsport	rt @box_repsol: great photo of @marcmarq	23/03/2014 0:01	http://bit.ly/1hq6p7x,http://t.co/ctxgptimw0	RT
438373	437950	tash_bandicoot	box_repsol,marcmarquez93	box_repsol	tashh	rt @box_repsol: great photo of @marcmarq	23/03/2014 0:00	http://bit.ly/1hq6p7x,http://t.co/ctxgptimw0	RT
438394	430279	hermidoctor46	marcmarquez93	marcmarquez93	Hermi	rt @marcmarquez93: pole position muy esp	23/03/2014 0:05	0	RT

LINK	n1 Link	n1 Picture	PERSONAL-WEBSITE	COUNTRY	ALL-NICK-ACTIVITY-EVER	NICK-FOLLOWERS	FRIENDS-FOLLOWING-AUDIENCE	LOCATION
http://twitter.com/Motorsport79/sta	1	1	http://twitter.com/Motorsport79	es	20672	1537	1531	BOLZANO - ALTO ADIGE - ITALIA
http://twitter.com/tash_bandicoot/s	1	1	http://twitter.com/tash_bandicoot	es	6442	516	672	not public
http://twitter.com/hermidoctor46/st	0	0	http://twitter.com/hermidoctor46	es	15320	709	1300	Triana, Sevilla

Descripción detallada de variables:

- **ID:** Valor numérico N. Índice de posts identifica cada uno.
- **PARENT-SYS-ID:** Valor numérico o nulo (=“sin padre”). El valor es un ID que relaciona el post con aquel del cual depende por ser un retweet o una respuesta a un tweet
- **Source:** Nicks del usuario que emitió el post.
- **Mentions:** Nicks de usuarios de Twitter separados por comas presentes en el contenidos del mensaje. En definitiva son nicks de los usuarios retwitteados y/o mencionados.
- **Target:** Nick del usuario que es retwitteado
- **NAME Source:** Nombre del usuario de Twitter que emitió el post
- **BODY:** Contenido del mensaje
- **PUBDATE:** Fecha y hora de emisión del psot
- **URLs comma separated:** URLs separadas por comas existentes en el contenido del mensaje
- **Type TW-RT-MT:** Variable categorial. Valores:
    - **TW:** Tweet original
    - **MT:** Tweet que contiene una mención
    - **RT:** Tweet que es un retweet
- **LINK:** Link que da acceso al post de Twitter
- **n1 Link:** Variable categorial numérica. Valores:
    - **0:** (cero) El contenido del mensaje no contiene ningún link
    - **1:** (uno) El contenido del mensaje contiene algún link
- **n1 Picture:** Variable categorial numérica. Valores:
    - **0:** (cero) El contenido del mensaje no contiene ningún link que sea una foto
    - **1:** (uno) El contenido del mensaje contiene algún link que es una foto
- **PERSONAL-WEBSITE:** Link a la página personal indicada en el perfil de Twitter del usuario o en su defecto link al perfil de Twitter.
- **COUNTRY:** Código del país donde el usuario reside
- **ALL-NICK-ACTIVITY-EVER:** Número de tweets emitidos por el perfil desde que el usuario creó la cuenta en Twitter.
- **NICK-FOLLOWERS:** Número de seguidores
- **FRIENDS-FOLLOWING-AUDIENCE:** Número de perfiles que sigue
- **LOCATION:** Variable de Geolocalización. Si comienza por “ÜT”, se proporcionan las coordenadas exactas desde donde se emitió el tweet. En su defecto contiene la ubicación declarada por el perfil.

2 – CSV NODOS:

Nombre del archivo: TWITTER-RT-23-26-Mar-2014-MotoGP.gexf [Nodes] BBDD

Descripción del contenido: Contiene algunos de los datos del Dataset anterior pero estructurados en función del índice de nodos. También contiene variables nuevas fruto del análisis de los datos relacionados en formato red. Como nodos se representan aquellos usuarios que interaccionaron por retweet del 23 al 26 de marzo en relación a la carrera de Qatar MotoGP 2014.

Número total de variables: 20

Variables más relevantes: Id, Eigenvector Centrality, Modularity Class, Grado, Grado de entrada, Grado de Salida, FIRST PUBDATE, LAST PUBDATE,

Muestra de datos: (Ver tabla)

Id	Label	NAME	PERSONAL-WEBSITE	COUNTRY	LOCATION	ALL-NICK-ACTIVITY-EVER	NICK-FOLLOWERS	FRIENDS-FOLLOWING-AUDIENCE
motogp	motogp	MotoGP™	http://www.motogp.com	es	All around the world	17940	1036047	430
valeyellow46	valeyellow46	Valentino Rossi	http://www.valentinorossi	es	ITALIA	2737	2825957	335
lorenzo99	lorenzo99	Jorge Lorenzo	http://www.facebook.com	es	not public	8670	962465	479

MTED	MTING	TWING	nº URLs	Eigenvector Centrality	Modularity Class	Grado de entrada	Grado de salida	Grado	FIRST PUBDATE	LAST PUBDATE
1049	56	26	40	0.9341255132966219	88	11109	7	11116	23/03/2014 21:34	26/03/2014 14:15
1296	4	1	10	0.6776367314463162	70	8067	7	8074	23/03/2014 00:08	26/03/2014 14:15
938	3	4	3	0.29291577939342617	94	3485	0	3485	23/03/2014 21:36	26/03/2014 14:15

Descripción detallada de variables:

- **Id:** Nick del usuario representado como nodo. La variable es un índice de nodos que identifica a cada uno.
- **Label:** En nuestra BBDD tiene igual valor que Id. Su valor es la etiqueta del nodo utilizada en la visualización del grafo.
- **NAME:** Nombre del usuario de Twitter
- **PERSONAL-WEBSITE:** Link a la página personal indicada en el perfil de Twitter del usuario o en su defecto link al perfil de Twitter.
- **COUNTRY:** Código del país donde el usuario reside
- **LOCATION:** Variable de Geolocalización. Si comienza por "ÜT", se proporcionan las coordenadas exactas desde donde se emitió el tweet. En su defecto contiene la ubicación declarada por el perfil.
- **ALL-NICK-ACTIVITY-EVER:** Número de tweets emitidos por el perfil desde que el usuario creó la cuenta en Twitter.
- **NICK-FOLLOWERS:** Número de seguidores
- **FRIENDS-FOLLOWING-AUDIENCE:** Número de perfiles que sigue
- **MTED:** Número de veces que dicho nodo ha sido mencionado en un tweet
- **MTING:** Número de menciones que este usuario ha realizado a otros
- **TWING:** Número de tweets originales que este usuario ha enviado
- **nº URL:** Número de URLs difundidas p
- **Eigenvector Centrality:** Métrica de análisis de redes. Valores entre 0-1. Esta métrica evalúa la importancia de un nodo dentro de la estructura de la red otorgando valores más altos a aquellos nodos conectados con nodos que tienen muchas conexiones. En una estructura de comunicación como esta indica la capacidad que tiene un usuario para que sus mensajes alcancen a la globalidad de la red.
- **Modularity Class:** Variable categorial numérica. Cada número es asignado a un módulo, un conjunto de nodos densamente interconectados si los comparamos un una red con conexiones aleatorias. Con esta variable
- **Grado de entrada:** Número de conexiones que se dirigen al nodo, es decir, retweets recibidos.
- **Grado de salida:** Número de conexiones que salen del nodo, es decir, número de retweets que el usuario realiza a otros.
- **Grado:** Número total de conexiones entrantes y salientes.

- **FIRST PUBDATE:** Fecha y hora de emisión del primer post capturado en el Dataset. En una visualización temporal esta variable es útil para hacer aparecer el nodo.
- **LAST PUBDATE:** Fecha y hora de emisión del último post capturado en el Dataset. En una visualización temporal esta variable es útil para hacer desaparecer el nodo.

3 – EDGES:

Nombre del archivo: TWITTER-RT-23-26-Mar-2014-MotoGP.gexf [Edges] BBDD

Descripción del contenido: Contiene algunos de los datos del Dataset pero estructurados en función de pares de nodos que interactúan por retweet formando aristas. También contiene variables nuevas fruto del análisis de los datos relacionados en formato red. Estos datos representan todos los retweets del 23 al 26 de marzo en relación a la carrera de Qatar MotoGP 2014.

Número total de variables: 17

Variables más relevantes: Source, Target, Body

Muestra de datos: (Ver tabla)

Source	Target	Type	Id	Label	Weight	n Type MT	n Contiene Foto	n Contiene URL1_Foto1	n Contiene URL1_FOTO0	n Contiene URL contenido
sararamos69	nickyhayden	Directed	15938	rt @nickyhayden: haha this one sent t	2.0	0	2	0	0	2
ornxbyaza_47	motogp	Directed	19595	rt @motogp: rt @marcmarquez93 spe	5.0	0	2	0	2	4
rizkitriana96	lorenzo99	Directed	26193	rt @lorenzo99: 5a posición! #keepfigh	1.0	0	1	0	0	1

URLs coma separated	FIRST PUBDATE	LAST PUBDATE	COUNTRY	LINK	NICK-FOLLOWERS
http://t.co/sy1sisxxbh	23/03/2014 00:00	23/03/2014 00:00	es	http://twitter.com/SaraRamos69/statuses/447508226490503169,http://twitter.com/SaraRam	204
http://t.co/tyx2tl6zyj	23/03/2014 00:00	25/03/2014 11:14	es	http://twitter.com/ornxbyaza_47/statuses/447508021829054464,http://twitter.com/ornxbya	187
http://t.co/xwaaocposy	23/03/2014 00:00	23/03/2014 00:00	es	http://twitter.com/rizkitriana96/statuses/447508031547248642	285

Descripción detallada de variables:

- **Source:** Nicks del usuario que emitió el post.
- **Target:** Nick del usuario que es retwitteado
- **Type:** En grafos donde las interacciones tienen un sentido, es decir, son dirigidos, su valor es “Directed”. En otros casos, como co-ocurrencia de palabras clave, su valor es “Undirected”
- **Id:** Número identificador de la arista. Variable necesaria para el programa de análisis de redes Gephi. Aunque tenga el mismo nombre que otras variables no está relacionado.
- **Label:** Contiene un tweet de ejemplo que intercambiaron los usuarios “Source” y “Target”. Dado que dos usuarios pueden interaccionar más de una vez, no todos los contenidos están en esta BBDD.
- **Weight:** Peso de la arista. Suma el número de interacciones que se producen entre dos usuarios.
- **n Type:** número de menciones a terceros usuarios en todos los tweets intercambiados entre el par de usuarios de la arista.
- **n Contiene Foto:** número de tweets con foto intercambiados entre el par de usuarios
- **n Contiene URL1_Foto1:** número de tweets con una foto y link en el mismo post intercambiados entre el par de usuarios

- **n Contiene URL1_FOTO0:** número de tweets que contienen un link pero ninguna foto intercambiados entre el par de usuarios.
- **n Contiene URL contenido:** número de tweets intercambiados entre el par de usuarios que contienen un link independientemente si este es a una foto o a otro contenido
- **NAME Source:** Nombre del usuario de Twitter que emitió el post
- **BODY:** Contenido del mensaje
- **PUBDATE:** Fecha y hora de emisión del psot
- **URLs comma separated:** URLs separadas por comas existentes en el contenido del mensaje ejemplo.
- **FIRST PUBDATE:** Fecha y hora de emisión del primer post intercambiado entre ambos usuarios capturado en el Dataset. En una visualización temporal esta variable es útil para hacer aparecer la arista.
- **LAST PUBDATE:** Fecha y hora de emisión del último post intercambiado entre ambos usuarios capturado en el Dataset. En una visualización temporal esta variable es útil para hacer desaparecer la arista.
- **COUNTRY:** Código del país donde el usuario reside
- **LINK:** Link que da acceso al post de Twitter
- **NICK-FOLLOWERS:** Número de seguidores de quién realiza el retweet, y por tanto número potencial de usuarios impactados por cada interacción.

4 – GEXF:

Nombre del archivo: TWITTER-RT-23-26-Mar-2014-MotoGP.gexf

Descripción del contenido: Archivo formato GEXF compatible con el programa de análisis de redes Gephi. Contiene las BBDD de NODES y EDGES descritas. Contiene calculadas algunas métricas como las de grado, Eigenvector Centralitya y Modularity Class. La distribución de los datos es aleatoria puesto que no está analizada.