Universidad Popular del Cesar

Diplomado para Acceder a Grado
Módulo Ciencia de Datos

Análisis de datos de Redes Sociales.

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

Análisis de datos de Redes Sociales

Extracción, procesamiento y visualización

- Obtención de datos a través de API (Twitter, Facebook)
- Procesamiento de datos 1
- Procesamiento de datos 2
- Análisis de redes sociales

La ciencia de los datos y el científico

El porqué del análisis de datos

"El Científico de Datos es un profesional que combinando conocimientos de matemáticas, estadística y programación, se encarga de analizar los grandes volúmenes de datos."

La Ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva.

Objetivo

Al término de esta sesión debéis de:

- *Tener conocimiento de cómo acceder y extraer información de las API de TW y FB*
- *Utilizar herramientas para la consolidación de datos como Open Refine*
- *Saber utilizar una hoja de cálculo para calcular indicadores y generar archivos de análisis*
- *Realizar un análisis de redes con Gephi detectando usuarios relevantes y sub-comunidades.*

Obtención de datos de la API

Twitter y Facebook

- Extraer datos de la API de Twitter
  - Definir una app
  - Uso de R
  - Extracción de datos
- Extraer datos de la API de Facebook
  - Uso de Netvizz
  - Extracción de datos

Objetivo

Al finalizar este módulo deberéis tener un conocimiento básico del uso de R y RStudio para poder extraer datos de perfiles de la API de Twitter

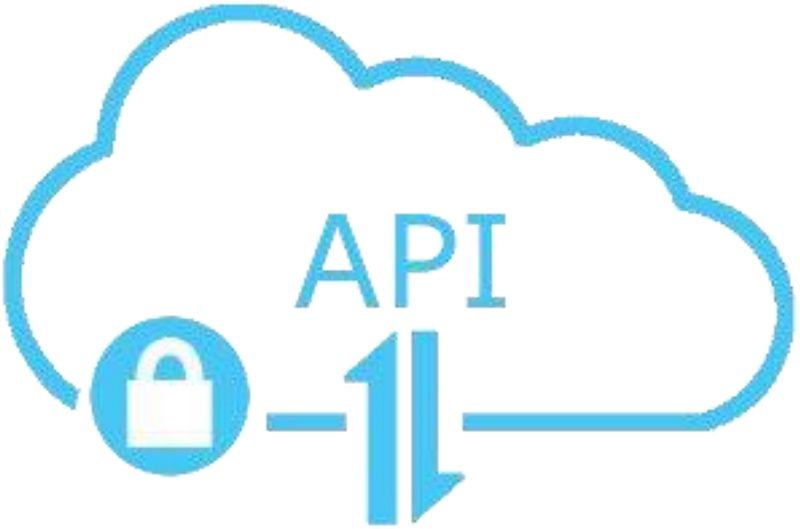
¿Qué experiencia tienes en programación?

Obtención de datos de la API

Twitter

Métodos de acceso a la API

A la API de Twitter se puede acceder de diferentes maneras. La opción más habitual es mediante una aplicación, o mediante un lenguaje de programación como R, Python, Java...



En nuestro caso lo haremos utilizando R y RStudio junto con el paquete TwitteR

Obtención de datos de la API

Twitter

R

R es un lenguaje y un entorno para la informática estadística y los gráficos.
R proporciona una amplia variedad de modelos estadísticos y técnicas gráficas, y es muy extensible mediante paquetes.

R está disponible como Software Libre bajo los términos de la GNU General Public License de la Fundación de Software Libre en forma de código fuente.

Obtención de datos de la API

Twitter

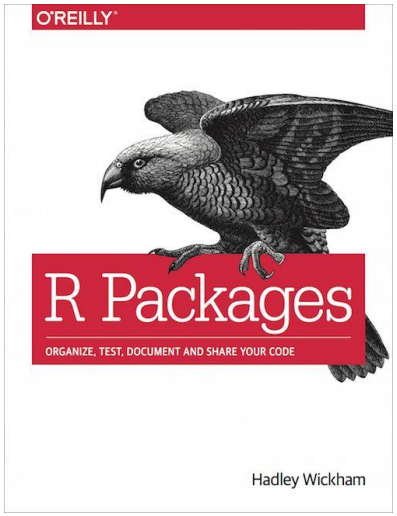
RStudio

RStudio es un entorno de desarrollo integrado (IDE) para R (lenguaje de programación) .Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux).

Obtención de datos de la API

Twitter



Paquete TwitteR

TwitteR es un paquete R que proporciona acceso a la API de Twitter. La mayoría de la funcionalidad de la API se admiten, con un sesgo hacia las llamadas API que son más útiles en el análisis de datos en lugar de la interacción diaria.

Ejercicio

Se compone de las siguientes partes:

- Definir una app en Twitter
- Descargar e instalar R y RStudio
- Descargar el archivo twitter-info-perfiles.R
- Completar con los datos de la API
- Seleccionar un usuario con pocos seguidores y ejecutar el script de seguidores
- Exportar los datos a un fichero csv

Obtención de datos de la API

Facebook

Netvizz

Es una aplicación que se encuentra en Facebook.

Netvizz es una herramienta que extrae datos de diferentes secciones de Facebook, en particular grupos y páginas. Las salidas de archivos se pueden analizar fácilmente en el software estándar.

Obtención de datos de la API

Facebook

Módulo de datos de página

Este módulo obtiene mensajes (especifique el último n o un intervalo de fechas) en una página y crea una serie de archivos:

- Un archivo tabular (tsv) que lista una serie de métricas para cada publicación.
- Un archivo tabular (tsv) que lista las estadísticas básicas por día para el período cubierto por las publicaciones seleccionadas.
- Un archivo tabular (tsv) que enumera el número de usuarios de página por país (sólo para los 45 principales países).
- Un archivo tabular (tsv) que contiene el texto de los comentarios de los usuarios (usuarios anónimos).
- Un archivo de gráfico bipartito (gdf) que muestra mensajes, usuarios (anónimos) y conexiones entre los dos. Un usuario está conectado a un puesto si ha comentado o reaccionado en él.

Ejercicio

- Seleccionaremos una página, buscaremos su ID y la introducimos en el formulario
- Recuperar los post publicados entre el 1 y el 7 de marzo.
- Pedir todos los datos, no sólo los de la página, también los de los usuarios.
- Comprobar qué tipos de archivo ha generado.

Procesamiento de datos 1

Open Refine

- Procesamiento de datos
- Instalación de Open Refine
- Uso de GREL para extracción de:
  - Hashtags
  - Url
  - Menciones

Objetivo

Al finalizar este módulo deberéis tener un conocimiento básico del uso de Open Refine para consolidar extracción de datos.

¿Encuentras siempre datasets en perfectas condiciones para analizar?

Procesamiento de datos 1

Open Refine

OpenRefine (antes Google Refine) es una poderosa herramienta para trabajar con datos desordenados: limpiarlo; Transformándolo de un formato a otro; Y ampliarlo con servicios web y datos externos.

Desde el 2 de octubre de 2012, Google no está apoyando activamente este proyecto, que ahora se ha renombrado a OpenRefine. El desarrollo de proyectos, la documentación y la promoción están ahora plenamente apoyados por voluntarios.

url: http://openrefine.org/

Procesamiento de datos 1

Open Refine

En nuestro caso tenemos un dataset de tweets, con el que vamos a hacer lo siguiente:

- Extraer url utilizadas
- Extraer hashtags utilizados
- Extraer menciones utilizadas

Con estos datos adicionales vamos a poder calcular determinar lo más popular:

- Lo más compartido
- De lo que más se ha hablado
- A quién se ha mencionado más

El fichero resultante nos servirá para calcular diferentes archivos relacionales para utilizarlos en Gephi

Procesamiento de datos 1

Open Refine

Instalación

Nos bajaremos e instalaremos Open Refine
url: http://openrefine.org/download.html

Ejecución

Al ejecutar Open Refine se abre una pestaña en el navegador, que es dónde se trabaja.

Datos: Descargar y abrir

Utilizaremos el fichero tweets-opendata.csv

Procesamiento de datos 1

Open Refine

Extracción de URLs

1. Abrir Open Refine
2. Crear un nuevo proyecto importando el csv con los tweets
3. Crear una nueva columna basada en “text” llamada “links”
4. Edit column
5. Add column based on this column
6. Introducir el nombre de la nueva columna
7. Seleccionar como lenguaje GREL
8. En expression introducir lo siguiente:
   filter(split(value, " "),v,startsWith(v,"http")).join("||")

Ejercicio

- **Repetir la misma operación para extraer menciones y hashtags del dataset**
- **Exportar a csv el archivo resultante.**

Antes de proceder con los pasos del punto anterior para la extracción de usuarios (del 3 al 8 adaptando nombres de columnas), hay que homogeneizar el texto. Es decir, convertir todo el texto a minúsculas.

Procesamiento de datos 2

Google Sheets

- Cálculo de indicadores de comunidad
  - Ratio de comunidad
  - Actividad media diaria
- Generación de archivos de grafo
  - Menciones, quién ha usado qué hashtags y quién ha compartido qué urls

Objetivo

Al finalizar este módulo deberéis tener un conocimiento básico del uso de Google Sheets generando indicadores básicos de comunidad para Twitter y generar archivos relacionales para Gephi.

¿Calculas indicadores propios? ¿O sólo utilizas los proporcionados?

Procesamiento de datos 2

Google Sheets

- Antes de proceder al análisis de datos comprobar la localización de la página archivo > configuración hoja de cálculo > locale
- Es importante que la hoja de cálculo tenga la localización de Estados Unidos.
- Hay diferencia en el carácter que define el decimal en los números.
- En EEUU es un „ mientras que en el sistema no anglosajón el decimal es una „
- La mayoría de aplicaciones / apis trabajan por defecto con el . para definir los decimales.

Procesamiento de datos 2

Google Sheets

Ratio de comunidad =seguidos / seguidores

Nos va a servir para poder tener un primer criterio para determinar la calidad de un usuario.

- Ratio > 1 Si la ratio es superior a uno significa que sigue a más usuarios que le siguen a él.
- Ratio < 1 Si la ratio es inferior a uno, significa que le siguen más usuarios que los que él sigue.

Procesamiento de datos 2

Google Sheets

Para la **actividad diaria** calcularemos la **media de tweets por día** para determinar:

- Si es un bot: Un usuario con una media de tweets diaria muy elevada (por encima de 25 tweets diarios de media hay que empezar a desconfiar) es muy probable que sea una cuenta automatizada.
- Si es un usuario que sólo utiliza twitter para informarse, para leer. No para participar. Estos usuarios tendrán una media muy baja. Podemos empezar a considerarlos como tales cuando la media es inferior a 1

Ejercicio

- Para realizar estos cálculos utilizaremos el archivo de seguidores generado en el primer módulo con los seguidores de un perfil de Twitter.
- Abrirlo en un archivo nuevo de Google Sheets
- Generar una columna llamada Ratio Comunidad y generar la fórmula para dividir los usuarios seguidos entre los seguidores.

Ejercicio

Para calcular la actividad media diaria de los usuarios:

1. Convetir la columna “created” en “fecha” y “hora” mediante un split.
2. Generar una columna nueva con la fecha actual.
3. Obtener en una nueva columna el total de días que hace que está activo.
4. Calcular la actividad media diaria.

Procesamiento de datos 2

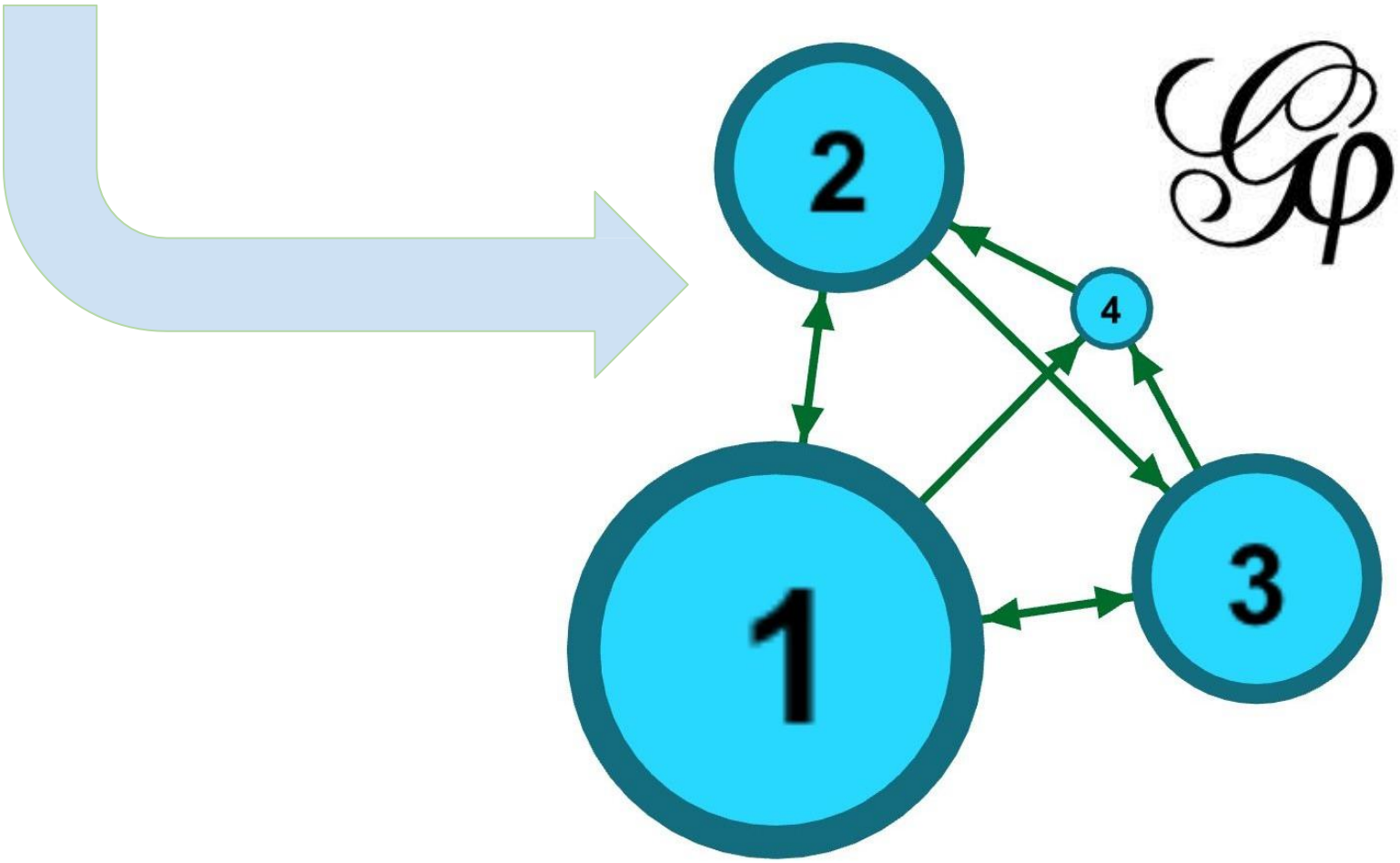
Google Sheets

Generación de archivos relacionales

Utilizaremos Google Sheets para generar archivos relacionales para poder utilizarlos en Gephi:

- Archivo de grafo de menciones para poder determinar a los usuarios clave y determinar sub-comunidades.
- Archivo de grafo de url para ver qué usuarios las han compartido
- Archivo de grafo de hashtags para determinar cómo se han utilizado, y cuales han tenido más relevancia.

	A	B	C	D
1	1	2	3	4
2	2	1	3	
3	4	2		
4	3	4	1	
5				



Ejercicio

1. Importar el archivo resultante del proceso con Open Refine en un archivo nuevo de Google Sheets
2. Generar los siguientes archivos de grafo:
   a. Menciones entre usuarios
   b. Usuarios > hashtags
   c. Usuarios > url
3. Exportar los archivos de grafo cada uno a un archivo .csv

*Antes de exportar, recordar borrar las cabeceras de las columnas, sino al abrirlo en **Gephi**, éste considerará que los nombres de las columnas también son nodos relacionados.*

Análisis de Redes Sociales

- Conceptos Básicos sobre Teoría de Grafos y el análisis de Redes
- Uso de Gephi para el cálculo de medidas de centralidad y detección de subcomunidades

Objetivo

El objetivo de este módulo es tener el conocimiento básico de uso de Gephi para procesar un archivo de grafo, poder determinar los nodos relevantes en base a medidas de centralidad y el cálculo de subcomunidades.

¿Qué podemos analizar con #ARS?

Procesamiento de datos 2

Conceptos básicos análisis de redes

Análisis de Redes

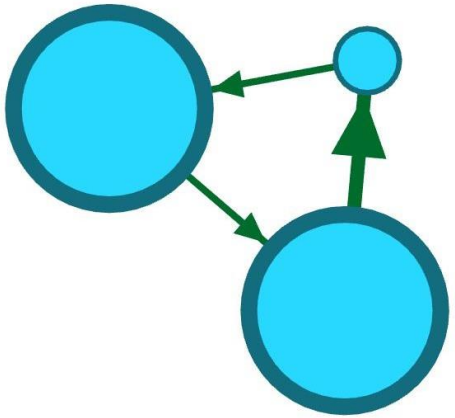
- Es el área encargada de analizar las redes mediante la teoría de redes
- Las redes pueden ser de diversos tipos: social, transporte, biológica, internet, información, etc.
- El estudio se centra en la asociación y medida de las relaciones y flujos entre las personas, grupos, organizaciones, computadoras, sitios web...
- Los nodos en la red, en este caso, son personas y grupos mientras que los enlaces muestran relaciones o flujos entre los nodos.

Procesamiento de datos 2

Conceptos básicos análisis de redes

Nodos y Aristas

- **Nodo**: Representa cada uno de los elementos relacionados que vamos a analizar. Si analizamos una conversación de Twitter, los nodos son los usuarios.
- **Arista**: Representa la relación entre dos nodos. Por ejemplo en la conversación de Twitter, esta puede ser representada por RT o menciones entre usuarios. El grosor de las aristas nos mostrará la intensidad con la que se ha producido.

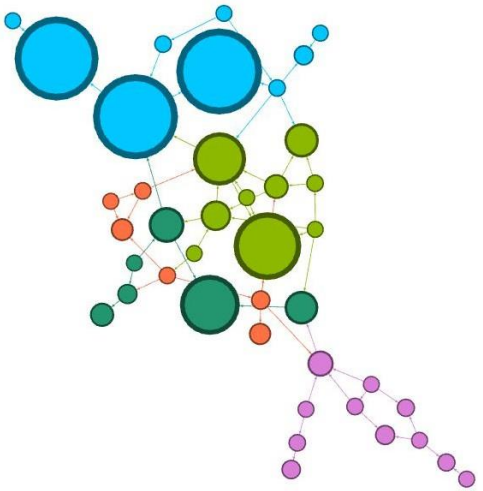


Procesamiento de datos 2

Conceptos básicos análisis de redes

Comunidades

- Saber cómo se han relacionado los nodos en el grafo analizado, nos puede aportar una visión clara de lo que ha sucedido.
- En función de la intensidad de las relaciones y de la cercanía con otros nodos se puede determinar que sub-comunidades se han generado.
- Mediante la visualización se puede determinar si una conversación ha estado muy polarizada.



Procesamiento de datos 2

Conceptos básicos análisis de redes

Tipos de grafo

- Por el **tipo de relación:** Encontramos dos tipos de grafo en función de si las relaciones son direccionales o no. Por lo tanto hay dirigidos y no dirigidos.
- Si hay **uno o más tipos de nodos:** En el caso de analizar sólo las menciones entre usuarios, tenemos un sólo tipo de nodos. En el caso de analizar cómo han utilizado los hashtags los usuarios, nos encontramos ante un grafo bipartito.

Procesamiento de datos 2

Conceptos básicos análisis de redes

Medidas de centralidad

Hay cuatro medidas de centralidad ampliamente utilizadas:

- La **centralidad de grado** («degree centrality»): Tiene en cuenta el número de relaciones entre los nodos.
- La cercanía («**closeness**») Tiene en cuenta la distancia de un nodo respecto del resto de los componentes del grafo.
- La **intermediación** («betweenness») Determina qué usuarios son esenciales en la unión entre diferentes grupos del grafo.
- La centralidad de vector propio («**eigenvector** centrality») Nos muestra qué nodos son los más relevantes.

Procesamiento de datos 2

Gephi

- Es software open-source de análisis de redes y visualización Permite analizar y visualizar archivos de grafo.
- Permite trabajar con diferentes tipos de archivos de grafo; .graphml, .gexf, .gdf, .gephi... E importar datos desde archivos .csv tanto para la generación de nodos, como la de aristas.
- Gephi ha sido utilizado en proyectos de académicos de investigación, periodismo y en otro lugares,
- Gephi es la herramienta que utilizaremos para realizar el análisis de redes.

Ejercicio

- Descargar Gephi e instalarlo
- Abrir Gephi y el csv generado en el módulo anterior con las menciones Archivo > Abrir
- Aplicar el layout Force Atlas 2
- En Estadísticas calcular la centralidad de vector propio y la modularidad
- Aplicar formato a los nodos. Para el tamaño utilizaremos la medida de centralidad Eigenvector. Para el color utilizaremos el valor Modularity Class
- Ir a previsualización
- Exportar archivo