



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Escola Tècnica
Superior d'Enginyeria
Informàtica

SOCIAL WEB BEHAVIOUR

PRÁCTICA 2. TIPIFICACIÓN DE USUARIOS

SOCIAL WEB BEHAVIOUR

CRISTINA I. FONT
DCADHA
crifonju@upv.es



1. Introducción

Compartir información utilizando la Web 2.0 y sus aplicaciones se ha convertido en algo común en los últimos años. El uso de redes sociales como Twitter, Facebook o Instagram para compartir gustos y aficiones, reivindicar o apoyar causas, comentar o discutir sobre temas concretos, movilizar a la gente o simplemente pasar el tiempo están a la orden del día.

Cada movimiento que un usuario realiza en la web lo define y sitúa en un espectro, indicando sus gustos e intereses.

Dependiendo de la plataforma o modo de análisis en la que se estudie al usuario los indicadores se encontrarán en diferentes lugares. Por ejemplo, si se analiza a los usuarios en Twitter se puede contabilizar:

- Número de seguidores (followers)
- Número de cuentas seguidas (followins)
- Número de tweets
- Número de likes
- Número de retweets
- Uso de hashtags

Pero, además, se puede realizar un análisis relacionado con:

- Tipo de vocabulario utilizado
- Número de palabras por tweet
- Si se realizan enlaces externos, tipos de enlaces, cantidad, etc.
- Si se utilizan imágenes, cantidad, tipología, etc.
- Método de comunicación (vía gifs, emoticonos, hashtags, etc.)
- Edad
- Género
- Localización

Poder conocer los grupos o clústers que se generan en función del comportamiento permite obtener una información valiosa para las distintas aplicaciones que existen de la misma. Por ejemplo, para el uso de sistemas de recomendación de contenidos o usuarios. Estos sistemas se utilizan en cualquier tipo de entorno, desde las propias plataformas para conectar nuevos usuarios, asistir en temas o mostrar información al usuario. Como en plataformas de compra para mostrar ítems similares a lo visto o comprado. O en plataformas de visualización como Netflix o YouTube en las que recomiendan títulos en función del contenido visualizado.

Todos estos análisis se realizan agrupando a los usuarios en diferentes conjuntos, en función de la categoría que se analice. Para ello se aplican algoritmos o formulas estadísticas, que emplean uno o más criterios de modo que se creen diferentes grupúsculos. Este será el tipo de análisis en el que se centre la presente práctica.

2. Objetivos de la práctica

- Analizar el comportamiento de usuarios de una plataforma o red social
- Conocer la aplicabilidad del análisis del comportamiento de los usuarios usando datos reales bajo una temática concreta
- Aprender la utilización de técnicas de clustering, a partir del uso de K-Means

3. Desarrollo de la práctica

Para esta práctica se va a descargar información de los usuarios relacionados con un hashtag, palabra clave o seguidores de una cuenta. Para ello, se descargará la información de estos utilizando alguno de los métodos propuestos a continuación.

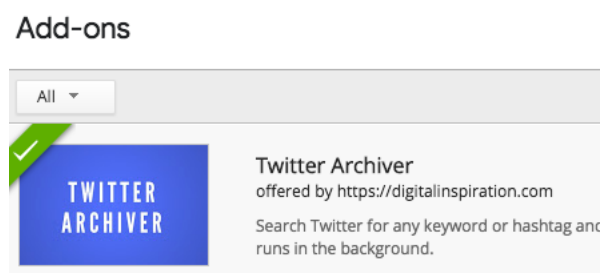
Se recomienda para esta práctica, generar una cuenta secundaria a modo de cuenta de desarrollador que permita realizar la conexión a las diferentes aplicaciones sin la interferencia con la cuenta del alumno.

Primera parte. Extracción de datos

Método 1. Twitter Archiver

Twitter Archiver es un addon que se instala en Google Sheets y permite la descarga de diferentes tipos de información relacionada con los usuarios de Twitter, Tweets, etc.

La instalación es sencilla en Google Sheets se debe acceder desde el menú superior a Add-ons > Get Add-ons y ahí buscar Twitter Archiver y añadirlo a la hoja.



Una vez añadido, se le debe otorgar permisos de acceso a Twitter utilizando una cuenta de usuario.

Con los permisos otorgados, se puede generar la regla que permita extraer la información necesaria para realizar el estudio.

Update Twitter Rule

×

All of these words	<input type="text"/>	This exact phrase	<input type="text"/>
Any of these words	<input type="text"/>	None of these words	<input type="text"/>
These #hashtags	<input type="text"/>	Written in	Any Language ⌵
Near This Place	<input type="text" value="Introduce una ubicación"/>	Advanced Rules	<input type="text" value="-filter:retweets -filter:repl"/>

People

To these accounts	<input type="text"/>	Mentioning accounts	<input type="text"/>
From these accounts	<input type="text"/>		

Twitter Search Query: -filter:retweets -filter:replies

Update Search Rule

Upgrade to Premium

Cancel

Una vez creada la regla, bastará con esperar a que el programa realice la extracción. Al tratarse de una cuenta gratuita, únicamente recoge información cada hora, lo que supone una limitación.

Método 2. Python & Tweepy

En el equipo debe estar instalado [Python](https://www.python.org/) (se aconseja usar la versión 3) (<https://www.python.org/>). Así mismo, se debe instalar Tweepy (<http://www.tweepy.org/>). Una forma sencilla de hacerlo es usando Pip install o Easy Install

```
pip install tweepy
```

Una vez instalado es necesario obtener acceso a la API de Twitter. Para ello se debe acceder a la página de desarrollo de Twitter (<https://developer.twitter.com/en/apps>) y crear una nueva App. Para ello se debe cumplimentar:

- Nombre
- Descripción
- URL
- Diferentes URLs de políticas de privacidad, términos de uso, etc.
- Una breve descripción del uso de la App

Una vez realizado este paso, se creará un acceso a la información de la aplicación desde el que se podrá crear y copiar las Claves y Tokens para realizar la conexión desde los diferentes Scripts que se desee utilizar. En el Anexo I se puede encontrar un script de ejemplo para usar. Así mismo, la información sobre el uso de la API de Twitter se puede encontrar en: <https://developer.twitter.com/en/docs/basics/getting-started>

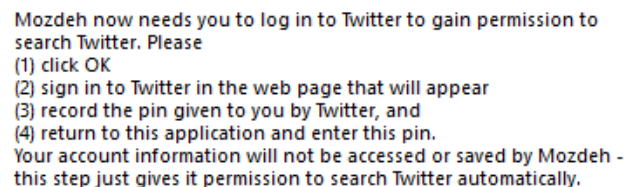
Método 3. Mozdeh

La herramienta Mozdeh (<http://mozdeh.wlv.ac.uk/index.html>) desarrollada por el grupo de investigación [Statistical Cybermetrics Research Group](#) de la Universidad de Wolverhampton, permite descargar tweets en función de hashtags, nombres de usuario o palabras clave.

Para utilizarla basta con ejecutarla, crear un proyecto y seleccionar la carpeta en la que se guardará. Una vez se arranca el proyecto, en la pestaña Tweets se introduce el término por el que se desea descargar los tweets.

Se debe introducir una por línea, sin caracteres de separación. Se selecciona el idioma y se ejecuta la búsqueda pulsando "Search Twitter Once".

Tras esto, se deberá permitir el acceso a Twitter desde la cuenta de usuario, una vez otorgado, se deberá introducir el código que ofrece la página en la herramienta, siguiendo las instrucciones que muestra la Figura 2.



Mozdeh now needs you to log in to Twitter to gain permission to search Twitter. Please
(1) click OK
(2) sign in to Twitter in the web page that will appear
(3) record the pin given to you by Twitter, and
(4) return to this application and enter this pin.
Your account information will not be accessed or saved by Mozdeh - this step just gives it permission to search Twitter automatically.

OK

Figura 1: Instrucciones acceso a Twitter

Una vez realizada la configuración, la aplicación preguntará si se desea eliminar los duplicados de los resultados, Figura 3, la respuesta recomendada es NO. Aparecerá una segunda ventana preguntando de nuevo, la respuesta recomendada volverá a ser NO. Esto puede cambiar en otro tipo de búsqueda o proyecto.

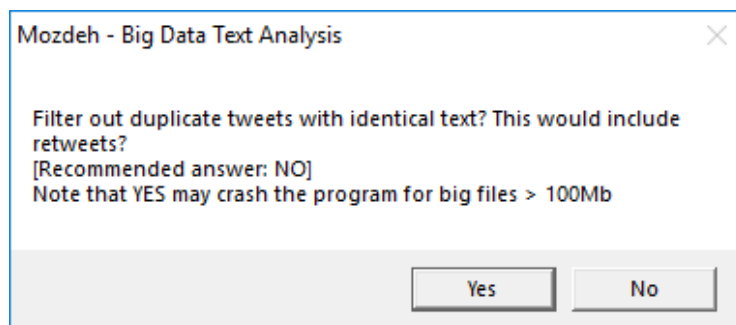


Figura 2: Pregunta de eliminación de resultados

Se puede cerrar la aplicación de análisis cuando se abra, ya que en esta práctica necesitamos los datos brutos.

En la carpeta donde hemos guardado el proyecto encontraremos una llamada Raw Data, dentro debe encontrarse un fichero .txt con los tweets.

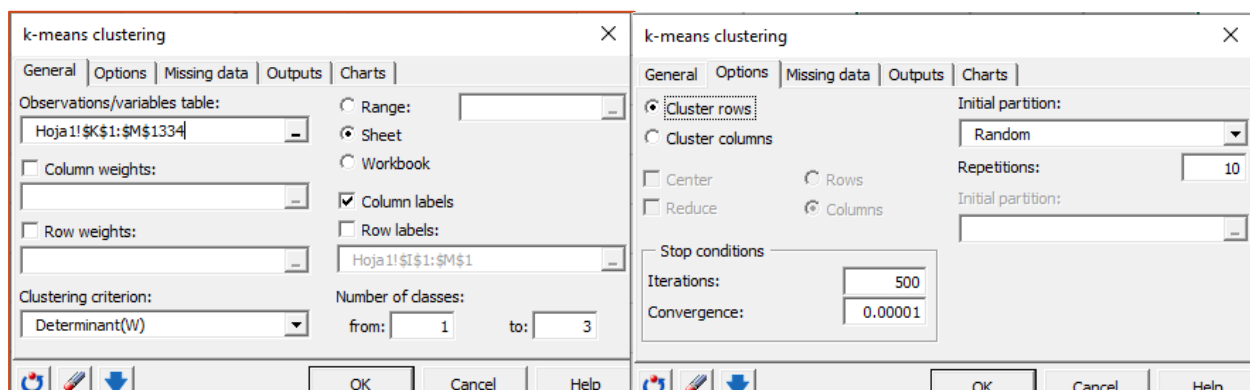
Para utilizarlos se puede abrir Excel, importar los datos y utilizar aquellos necesarios para la práctica.

Segunda parte. Análisis de datos con XLSTAT

Una vez extraída la información necesaria se pasará a analizar con un plugin para Excel, de modo que se puedan realizar diferentes análisis estadísticos.

Para ello se debe descargar desde la página del producto (<https://www.xlstat.com/es/>). El uso de la herramienta es muy sencillo:

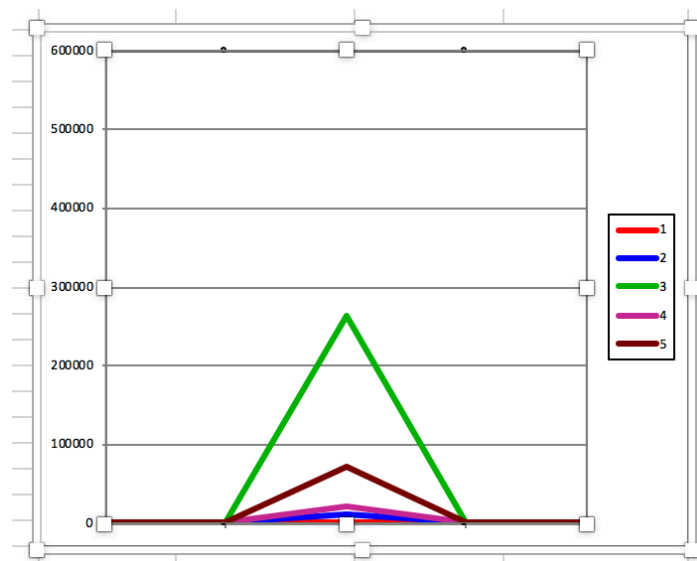
- Se abre Excel y se cargan los datos extraídos en la primera parte de la práctica.
- Una vez cargados, una primera aproximación al análisis de estos puede ser categorizar o tipificar a los usuarios en función de los Retweets, Favoritos, Followers, Follows y Listados.
- Para ello, se seleccionarán las columnas pertinentes, y se llevarán a una nueva hoja del libro de Excel, dejando la primera columna libre.
- En esta nueva hoja, se colocará en la primera columna los nombres de los usuarios copiados de la hoja anterior.
- Una vez los datos se encuentren preparados, se seleccionará en el apartado de **XLSTAT** la opción Modelado de datos (Data Modeling) en el menú superior -> K-Means Clustering.
- Cuando aparezca la ventana emergente, pasaremos a seleccionar los datos a analizar:
 - o Se debe seleccionar las columnas a analizar para complementar el campo **Observations/Variables Table**
 - o Se debe desmarcar la opción **Row Labels**
 - o Se debe indicar el **número de clases** en función de la cantidad de columnas seleccionadas (*from 1 to 3/4/5/n*)
 - o Por último, como configuración inicial se pasará a la pestaña Options y se seleccionará **Cluster Rows**



Tras pulsar OK comenzará el análisis, una vez termine, aparecerá una nueva hoja con los resultados llamada K-Means, en la que se podrá interpretar los diferentes resultados.

Los principales resultados ha tener en cuenta son:

- Class Centroids: indican los valores que conforman cada centroide. La observación promedio para cada clase analizada.
- Central Objects: indica la observación principal para cada clase junto con su valor.
- Results by class: indica el número de observaciones por clase, así como cada observación que incluye el cluster.
- Gráfica Profile: en la que se encuentran las clases generadas con la característica que define a la clase (las variables en las que es fuerte o débil).



Se aconseja realizar diversas ejecuciones con diferentes combinaciones de datos o Criterios de Clustering, cambiar la visualización de la gráfica por una de tipo radial, etc. para tratar de mejorar el análisis de los resultados.

4. Preguntas y entrega de la práctica

La práctica será entregada mediante Tareas en Poliformat por grupo. Deberá subirse un documento con un informe en el que se describa el método utilizado para la extracción de tweets (en caso de utilizar un script este deberá añadirse como anexo), así como capturas del análisis de datos realizado con XLSTATS.

Las capturas y explicaciones deben responder principalmente a las preguntas:

- Indicar el tamaño de los clusters
- Discutir las diferencias entre tamaños, a qué puede deberse
- Explicar la característica esencial de cada cluster en función de las variables

Fecha máxima de entrega: 2 de marzo 2021, 16:30h

Anexo I

Script de ejemplo para descarga de Tweets relacionados con un Hashtag. Se debe guardar el fichero con terminación .py

```
import tweepy
import csv
import pandas as pd

## Introducir las credenciales extraidas de Twitter Apps
consumer_key = "
consumer_secret = "
access_token = "
access_token_secret = "

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)

## Creación y escritura en el fichero CSV
csvFile = open('hashtag_tweets.csv', 'a')
csvWriter = csv.writer(csvFile)

## Extraccion del hashtag, indicando el lenguaje y la fecha de inicio. Maximo 9 dias
for tweet in tweepy.Cursor(api.search, q="#SBLIII", count=100,
                           lang="en",
                           since="2018-02-01").items():
    print (tweet.created_at, tweet.text)
    csvWriter.writerow([tweet.created_at, tweet.text.encode('utf-8')])
```