**PRACTICA 1- TIPOLOGIA Y CICLO DE VIDA DE LOS DATOS**

**Autores:** Gabriel peso Bañuelos, Jon Ortiz Abalia

**Fecha:** abril 2019

**Guía de eventos culturales de ámbito nacional**

**Contexto**

Actualmente existen múltiples buscadores online de espectáculos culturales que pueden variar en cuanto a ámbito geográfico, tipos de espectáculos o granularidad de las características de los mismos.

La idea con este proyecto es, por un lado, integrar en un único listado los registros contenidos en los múltiples buscadores ofreciendo así un registro de referencia de eventos culturales y, por otro, aportar un repositorio de calidad para la realización de proyectos de minería de datos.

Como prueba de concepto, el presente trabajo tiene como objetivo el volcado de los datos de un buscador de eventos culturales muy utilizado como es “Atrápalo” (<https://www.atrapalo.com/entradas/home_nacional/>).

**Título**

Eventos\_culturales: “**eventos\_culturales.csv**”

Eventos: “**eventos.csv**”

Cultura: “**cultura.csv**”

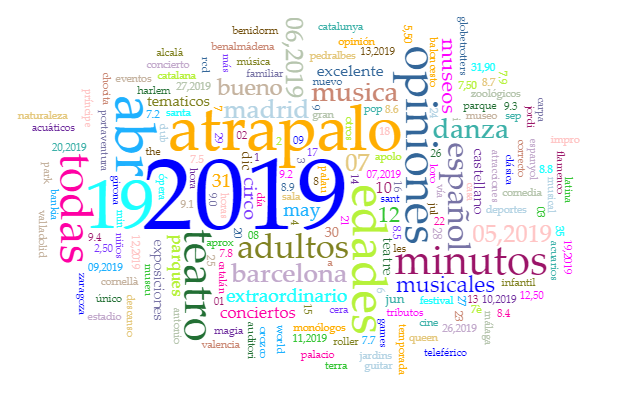
Espectaculos: “**espectaculos.csv”**

**Descripción**

El conjunto de datos se ha extraído de la siguiente página web de [Atrápalo](https://www.atrapalo.com/entradas/home_nacional/). Cada uno de los espectáculos constituye un registro de información (o fila) en el *dataset*. Dichos espectáculos abarcan todo el territorio nacional y se han extraído de todas las secciones disponibles en el buscador como son: Teatro y danza, Música, Musicales, Museos y exposiciones, Circo, Parques temáticos, Deportes, Cine, Conferencias y Ferias.

La idea es generar un primer volcado de la información en un *dataset* inicial y a continuación, realizar cargas incrementales de la información con periodicidad diaria.

**Representación gráfica**

****

**Figura 1.** Nube de palabras generada con la herramienta “Cirrus” de [Voyant-tools](https://voyant-tools.org/).

**Contenido**

El *dataset* se compone de XXXX registros y 19 atributos. La carga del archivo se ha realizado el día XX de abril de 2019 utilizando la técnica de *Web Scraping* y utilizando las librerías *Python requests* y *BeautifulSoup.*

Los campos que incluye el *dataset* son:

* **Captura** (*date*): fecha de la captura de los datos
* **Web** (*string*): nombre identificativo del buscador fuente de los datos
* **Evento** (*string*): tipo de espectáculo
* **Título** (*string*): título del espectáculo
* **Categoría** (*string*): subtipo de espectáculo
* **Novedad** (*string*): especifica si el espectáculo es una novedad
* **Ubicación** (*string*): lugar donde se realiza el espectáculo
* **Localidad** (*string*): localidad donde se realiza el espectáculo
* **Precio** (*float*): precio de la entrada para asistir al espectáculo (euros)
* **Descuento** (*float*): descuento ofertado en la entrada (euros)
* **Fechas** (*string*): rango de fechas en las que se representa el espectáculo
* **Fecha\_inicio** (*date*): fecha de inicio de la representación del espectáculo
* **Fecha\_fin** (*date*): fecha de fin de la representación del espectáculo
* **Duración** (*string*): duración del espectáculo
* **Idioma** (*string*): idioma en el que se representa el espectáculo
* **Público** (*string*): categorías de edad a los que va dirigido el espectáculo
* **Puntuación** (*float*): puntuación del espectáculo dada por los usuarios del buscador
* **Valoración** (*string*): valoración del espectáculo según la puntuación
* **Opiniones** (*integer*): número de opiniones del espectáculo realizadas por los usuarios del buscador

**Agradecimientos**

Los datos se han obtenido del buscador online “Atrápalo” propiedad de **Atrapalo.com** en la siguiente web: <https://www.atrapalo.com/entradas/home_nacional/>.

Entendemos que el uso que se hacen de los datos obtenidos de la página web no es comercial por lo que no es necesaria la autorización previa de **Atrapalo.com** según queda recogido en la sección “Aviso legal” de la página web.

**Inspiración**

Este repositorio de datos podría utilizarse para proyectos de minería de datos que abarquen tanto la estadística descriptiva como inferencial. Podríamos pensar en los siguientes:

* Cómputo del número de espectáculos tanto total como desglosado por: localidad, tipo/subtipo de espectáculo o tipo de público.
* Comparación del número y tipo/subtipo de espectáculos entre localidades.
* Generación de rankings de espectáculos por popularidad (puntuación, valoración o número de opiniones) o precio.
* Aplicación de modelos predictivos para:
  + Conocer qué variables predicen una valoración alta del espectáculo (en términos de puntuación, valoración o número de opiniones)

Además, el uso combinado del presente dataset junto con otros datos externos podría dar lugar a otros proyectos como:

* Junto con datos de recaudación se podrían aplicar modelos predictivos para conocer qué variables predicen mejor una alta recaudación.
* Junto con datos del Ministerio, de las consejerías o concejalías de Cultura se podría estudiar si existe una relación entre las subvenciones estatales, autonómicas o locales destinadas a la promoción de la cultura y la cantidad de espectáculos que se estrenan.

**Licencia**

La licencia tiene que respetar en todo momento la restricción de “no uso comercial” que tiene el propietario de los datos en origen (Atrapalo.com) con lo cual elegiríamos la licencia **CC BY-NC-SA 4.0 License.**

Esta licencia permite compartir (copiar y redistribuir) el *dataset* y trabajar sobre él para transformarlo a conveniencia.

Esta licencia conlleva:

* Dar crédito (agradecer) al propietario de los datos
* Proveer un link a la licencia e indicar qué cambios se han realizado.
* No se puede hacer un uso comercial de lo datos
* Aplicar a los nuevos datos (en el caso de ser modificados) la misma licencia que tienen los datos usados en origen.

**Recursos**

* Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
* Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
* Brody, H (2017). The Ultimate Guide to Web Scraping. Leanpub

**Tabla de contribuciones**

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | Gabriel Peso Bañuelos, Jon Ortiz Abalia |
| Redacción de las respuestas | Gabriel Peso Bañuelos, Jon Ortiz Abalia |
| Dsarrollo código | Gabriel Peso Bañuelos, Jon Ortiz Abalia |