PRACTICA 1- TIPOLOGIA Y CICLO DE VIDA DE LOS DATOS

Autores: Gabriel peso Bañuelos, Jon Ortiz Abalia

Fecha: 15 abril 2019

CONTENIDO

TÍTULO: "Muchos más espectáculos de los que imaginabas"

CONTEXTO

Actualmente existen múltiples portales y empresas de venta de entradas *online* de espectáculos culturales que pueden variar en cuanto a ámbito geográfico, tipos de espectáculos o granularidad de las características de los mismos y, lo que es más importante, ninguno de ellos presenta una oferta completa de espectáculos a nivel nacional. Por ejemplo, la obra "*Top girls*" (10 abr-21 abr) representada en el Teatro Valle-Inclán (Madrid) está ofertada en la página web del Instituto nacional de las Artes Escénicas y de la Música (INAEM) (1) y no así, en el buscador de Atrápalo (2) o de Ticketmaster (3), por citar algunos de los más importantes.

Existen múltiples causas para esta disparidad en la información ofrecida, entre ellas las condiciones económicas que cada una de estas plataformas ponen a los autores de las obras o la exclusividad del propio establecimiento a la hora de vender las entradas de sus funciones.

A nivel autonómico y, más concretamente, en el Pais Vasco, existe "Kulturklik" (4), portal dedicado a la Cultura lanzado por el Gobierno Vasco que sí que ofrece un listado muy completo de los espectáculos representados en la comunidad autónoma.

El presente proyecto se plantea como solución a la falta de un repositorio completo de espectáculos a nivel nacional. Para ello, se quiere sentar las bases para articular un mecanismo de *Web Scraping* y poder así integrar en un único listado todos los registros de espectáculos y eventos culturales que se ofrecen diariamente al público a través, tanto de las distintas plataformas de venta *on-line* como de las propias páginas web de las propias salas de representación.

Esto permitiría tener el listado más completo de espectáculos culturales a nivel nacional, tanto por el número de registros como del nivel de detalle de cada uno de ellos pudiendo servir como fuente de referencia para la consulta y la realización de proyectos de minería de datos con fines muy diferentes: análisis estadístico, estudios de mercado, etc.

Como prueba de concepto, el presente trabajo se ha desarrollado única y exclusivamente contra un buscador de eventos culturales muy utilizado como es "Atrápalo" (2), pero se ha adaptado el código para facilitar la adición de otras web relevantes del sector llegado el caso, como pueden ser: Entradas.com (5), Ticketmaster (3) etc.

DESCRIPCIÓN

El objetivo es extraer toda la información relacionada con los espectáculos culturales disponible en la siguiente página web de Atrapalo (2):

https://www.atrapalo.com/entradas/

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.

Cada uno de los espectáculos constituirá un registro de información (o fila) en el *dataset* e irá acompañado de los datos asociados disponibles en la página web como son el título y tipo de espectáculo, si es o no novedad, las fechas, el lugar y la localidad en que se representa, la duración, el idioma, el tipo de público al que va dirigido, el precio, la puntuación, el número de opiniones y la valoración hecha por los usuarios. A esta información extraída de la página añadiremos, además, algunos datos complementarios o metadatos como son la fecha de carga de los datos y el nombre de la página web que ha servido de fuente de los datos. Para más detalle sobre los atributos se puede consultar el apartado de "Contenido" de este mismo documento.

La idea es generar un primer volcado de la información en un *dataset* inicial y a continuación, realizar cargas nocturnas parciales e incrementales de la información con periodicidad diaria extrayendo cada día los datos referentes sólo a algunos de los tipos de espectáculos (llamados "Secciones" en el código desarrollado).Estos vendrán escogidos mediante el algoritmo implementado en la función del código '*secciones_de_hoy()*' en la que a cada día de la semana se le asocian sólo unos espectáculos en función del lugar que ocupan en la ordenación de la lista de secciones totales. De esta forma, reduciríamos el riesgo de sobrecargar el servidor asegurando al mismo tiempo que todas las secciones se actualizan una vez por semana lo que es suficiente para el tipo de información, poco cambiante, de que se trata.

El código desarrollado, por tanto, **guarda** los espectáculos antiguos, **actualiza** los existentes y **añade** los nuevos con lo cual mantenemos un listado con toda la información asociada totalmente actualizada desde el mismo momento que se empiecen a realizar las descargas diarias. Esto es importante ya que campos como "Descuento", "Puntuación", "Valoración" u "Opiniones" pueden sufrir cambios frecuentes.

Hemos comentado que el presente proyecto es una prueba concepto en la que se ha descargado el contenido de un sólo *site* pero que sería fácilmente extrapolable a otros llegado el caso. Por ello, hemos añadido la función *Sacar_atributos_otras_webs(espectaculo)* en donde habría que incluir el código específico para extraer los atributos de una nueva web.

Una vez descargados los datos, y previo a su volcado al fichero, se necesitaría hacer una limpieza a posteriori. Hemos identificado las siguientes acciones a realizar:

- Cambiar tipos de variable:
    - De "*string*" a "*number*": atributos "Opiniones" y "Duración"
    - De "*string*" a "*float*": atributos "Precio" y "Descuento"
- Estandarizar la unidad de medida de "Duración" a minutos.

- Eliminar los espacios en blanco antes y después de las cadenas de caracteres: atributos "Evento", "Título", "Ubicación", "Localidad", "Fechas", "Idioma", "Publico", "Valoración".

- Estandarizar los campos "Título" y "Ubicación" atendiendo a las mayúsculas y minúsculas (primera letra en mayúscula y resto en minúsculas). Esto nos permitiría identificar si un mismo espectáculo está representado en diferentes localizaciones u ofertado por distintas plataformas (en el momento en que se añadan más fuentes de datos en el campo "Web").

- Revisar la consistencia en cuanto al vocabulario controlado utilizado en los atributos "Localidad", "Idioma", "Publico" y "Valoración.

REPRESENTACIÓN GRÁFICA

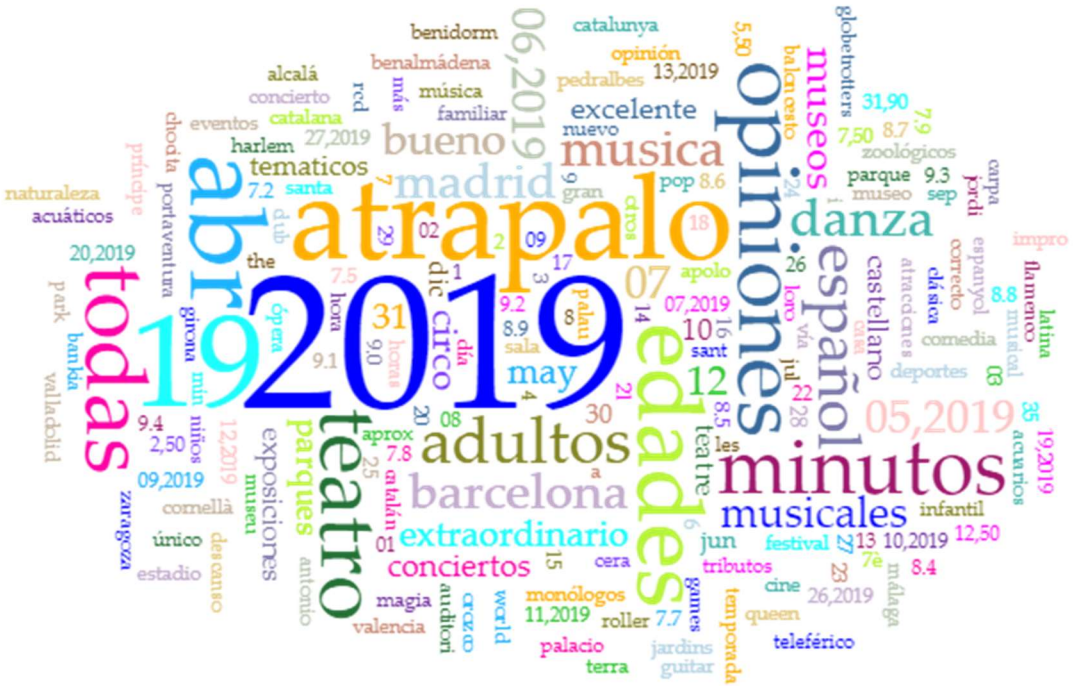

Figura 1. Nube de palabras generada con la herramienta "Cirrus" de Voyant-tools.

CONTENIDO

El *dataset* se compone de 1546 registros y 19 atributos. La carga del archivo se ha realizado el día 7 de abril de 2019.

Como paso previo y fundamental a la extracción de los datos se ha procedido a evaluar si era posible realizar ésta. Por un lado, se ha revisado el archivo *Robots.txt* en donde no se ha identificado ninguna prohibición en cuanto a "*user-agent*" ni a ficheros que pudiera afectar al presente proyecto.

Aparte de esta revisión inicial, en el código se incluye una comprobación explícita mediante la función '*licencia_robots()*' cada vez que se inicia la descarga por si el archivo *Robots.txt* hubiera podido haber sido modificado por el *site*.

Por otro, se ha leído detenidamente el apartado "Aviso legal" de la página web en donde se dice que sólo se requiere autorización de Atrapalo.com en el caso de utilizar los datos para fines comerciales ("*…quedan expresamente prohibidas la reproducción, la distribución y la comunicación pública, incluida su modalidad de puesta a disposición, de la totalidad o parte de los contenidos de esta página web, con fines comerciales, en cualquier soporte y por cualquier medio técnico, sin la autorización de Atrapalo.com").*

La técnica utilizada en la extracción de datos ha sido *Web Scraping,* para la cual se ha hecho servir el lenguaje de programación Python y las librerías relacionadas con esta técnica como son *Python requests* y *BeautifulSoup.*

Python requests se ha utilizado para enviar la petición HTTP de descarga a la página Web y obtener una respuesta en formato HTML. Se ha tenido que cambiar el contenido del campo "*User-agent*" de la cabecera de la petición para poder evitar la respuesta de código 403 de prohibición de acceso.

En el código se ha introducido un mecanismo de reintentos espaciados en el tiempo, dentro de la función '*Sacar_espectaculos()*' por si hubiera problemas de desconexión durante la descarga o el propio *site* bloqueara nuestras peticiones.

BeautifulSoup se ha utilizado para transformar el documento HTML en una estructura anidada y poder identificar fácilmente las clases e identificadores de los distintos campos cuya información necesitamos extraer.

Así, de cada una de los "Tipos de espectáculo" listados en el buscador, se han extraído 17 campos a los cuales se han añadido 2 campos más creados por nosotros, a modo de metadatos, como la "Captura" y "Web".

A continuación, se detallan los campos que incluye el *dataset*:

- **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** (*string*): precio de la entrada para asistir al espectáculo (euros)

- **Descuento** (*string*): 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** (*string*): número de opiniones del espectáculo realizadas por los usuarios del buscador

La unicidad de cada espectáculo queda establecida por los siguientes atributos: "Título" y "Ubicación" de tal forma que si un espectáculo se descarga y éste ya existe en nuestra base de datos (listado existente) se procede a actualizar la información del mismo.

Este mecanismo asegura que la información existente en nuestro listado esté totalmente actualizada para todos los atributos, especialmente campos sujetos a muchos cambios como "Opiniones" y "Puntuación".

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 (6).

Además, se ha verificado el fichero *Robots.txt* para verificar que nuestra descarga no es contraria a la política del *site* y se ha reducido el número de espectáculos diarios a descargar para no sobrecargar a la operativa de la web.

Cabe mencionar que no hemos encontrado análisis similares realizado con este tipo de datos.

INSPIRACIÓN

Este repositorio de datos constituiría el listado más completo de eventos culturales representados en territorio nacional.

A modo de ejemplo del valor añadido que aportaría este proyecto estaría el uso que por parte del Ministerio de Cultura y Deporte podría hacer de los datos. Actualmente, el Ministerio en su anuario (7) de estadísticas culturales informa del número de representaciones de teatro,

danza y música, datos que a su vez obtiene del anuario (8) de las artes escénicas, musicales y audiovisuales de la SGAE (Sociedad General de Autores y Editores). Por lo tanto, en las estadísticas que maneja el Ministerio sólo figurarían las obras registradas en la SGAE, las cuales no representan la totalidad de las obras que se representan en España. En consecuencia, en cuanto a número de representaciones, la base de datos ideada en este proyecto podría dar valores más cercanos a la realidad que los ofrecidos por la SGAE.

Además, estos datos podrían utilizarse para proyectos de minería de datos que abarquen tanto la estadística descriptiva como inferencial y que podría tener muy diferentes fines. Podríamos pensar en los siguientes:

1) Estudios de mercado (para promotores)

- 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).

2) Análisis de situación (para organismos culturales como ministerios, ayuntamientos etc.)

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 (7), 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.

3) Propuestas y previsión (para el público en general)

Para el desarrollo de aplicaciones y propuestas de ocio por parte de empresas del sector.

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

REFERENCIAS

1. INAEM. [En línea] https://www.entradasinaem.es/.

2. Atrapalo. [En línea] https://www.atrapalo.com/entradas/.

3. *Ticketmaster.* [En línea] https://www.ticketmaster.es/?language=es-es.

4. *Kulturklik.* [En línea] http://www.kulturklik.euskadi.eus/inicio/.

5. *Entradas.com.* [En línea] https://www.entradas.com/.

6. *Atrapalo, Aviso Legal.* [En línea] https://www.atrapalo.com/common/aviso_legal/.

7. Deporte, Ministerio de Cultura y. *Anuario de Estadísticas Culturales.* [En línea] http://www.culturaydeporte.gob.es/servicios-al-ciudadano/estadisticas/cultura/mc/naec/portada.html.

8. SGAE, Sociedad General de Autores y Editores. *Anuarios de la SGAE.* [En línea] http://www.anuariossgae.com/home.html.

TABLA DE CONTRIBUCIONES

Contribuciones	Firma
Investigación previa	Gabriel Peso Bañuelos, Jon Ortiz Abalia
Redacción de las respuestas	Gabriel Peso Bañuelos, Jon Ortiz Abalia
Desarrollo código	Gabriel Peso Bañuelos, Jon Ortiz Abalia