PRACTICA 01-Web Scraping

Luis Jerez Rincon

13 de noviembre, 2017

Table of Contents

Práctica 1 - Tipología y Ciclo de vida de los datos. Web Scraping

Objetivo (Transcripción del enunciado):

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en la web, indicando características requeridas para este práctica.

Los ficheros R, Rmd, Rproj, README, LICENSE, Wiki y demás están el repositorio GitHUb preparado para esta práctica:

https://github.com/ogorodriguez/web-scraping-practica-01-uoc.git

Mi cuenta en GitHub es ogorodriguez, la cual abría hace tiempo con un pseudónimo.

1. Título del dataset.

Poned un título que sea descriptivo.

El dataset lleva por título: "How Soon is Now: Morrissey On Tour. Estadísticas de conciertos, lugares y fechas"

2. Subtítulo del dataset.

Agregad una descripción ágil de vuestro conjunto de datos para el subtítulo.

El dataset present información sobre los conciertos y apariciones frente a público del cantante de Stephen Patrick Morrissey (born 22 May 1959), antiguo cantante de The Smiths. El set presenta los eventos pasados y actuales ofrecidos desde la API setlist.fm.

El cantante está actualmente de gira promocional de su último disco estrenado en Noviembre 2017, y se ha mantenido de gira en varios países a lo largo de los años desde que inició su carrera en solitario en 1988. Por tal razón es una oportunidad para sus fans, entre los que me incluyo, de verlo en escena, de ir a los sitios donde tenemos que ir a verle, conocer a alguien interesante; por lo que nos preguntamos: How soon is Now?

3. Imagen.

La imagen que identifica al set, y que encontraréis también en la cuenta GitHUb es:

Morrissey-Concert-Berlin

Autoría: By Alexander from Berlin, Germany (Morrissey #4) [CC BY 2.0 (http://creativecommons.org/licenses/by/2.0)], via Wikimedia Commons

4. Contexto

¿Cuál es la materia del conjuto de datos?

Estos datos se pueden categorizar dentro del renglón Entretenimiento u ocio. El mundo del espectáculo genera grandes movimientos de personas, materiales y financieros alrededor de las producciones o eventos ya sean de mayor o menor escala. Se buscan formas de aprovechar las herramientas de la ciencia de datos/big data/data mining, etc. para construir escenarios nuevos en donde competir y lograr ofertas atractivas (Lippell 2016). Entre ellos están los agregadores: páginas como rottentomatoes, metacritic, y las redes sociales, que gracias al aporte de usuarios a través de sus perfiles, comentarios e interacciones son un elemento clave para dicho sector, al tener por fin ese parte del conocimiento que posee el usuariro final.

5. Contenido

¿Qué campos incluye? ¿Cuál es el período de tiempo de los datos y cómo se han recogid?

El set de datos presenta los siguientes campos:

a. **id-setlist**: campo importante para el contenido extraído de setlist.fm Identifica la lista de canciones cantandas en el concierto
b. **tour**: nombre de la gira si la presentación ha sido parte de una serie de conciertos
c. **fecha-evento**: La fecha en que se celebró la presentación
d. **Id-artista**: es el código identificativo del artista facilitado por MusicBrainz
e. **artista**
f. **sala-lugar**: nombre con que se conoce la sala de espectáculos en el momento de celebrarse la presentación
g. **cuidad**: nombre de la cuidad del evento
h. **pais-código**: código de dos caracteres identificando el país.
i. **pais-nombre**: nombre extenso del país en donde se celebró el evento
j. **setlist**: listado de canciones del concierto
k. **mes** : Número del mes en que se celebró el evento mn. **año** : Año en que se celebró el evento.

Para más detalles sobre este punto ver el apartado *Observaciones*

6. Agradecimientos.

¿Quién es el propietario del conjunto de datos? Incluid citas de investigacion o análisis posteriores.

El propietario del conjunto de datos es el site: sitelist.fm. El site engloba principalmente la lista de canciones que un cantante o grupo haya presentado en un concierto, recital, festival, etc. Es una págian que lleva cierto tiempo y que ha ido desplegando su API para fines no-comercial, o comerciales previa comunicación con los propietarios.

En su política de permisos resaltan que es necesario hacer mención de la propiedad de los datos con los que se trabaja desde sus wikis incluyendo el snippet que pego a continuación:

```
Source: <a href="https://www.setlist.fm/>concert setlists on setlist.fm
```

ya sea en una web o bien dentro de un documento como éste. Como el objetivo de este trabajo no es comercial no es necesario avisar a setlist.fm.

Este trabajo estaré público y visible desde Github por lo que también se satisface un requerimiento de facil acceso por parte de los buscadores web.

Para más información leer la información en el portal de setlist.fm Terms of Service

Previamente se han hecho desarrollos en Python tomando como fuente de datos los de la API de setlist.fm. Haciendo una búsqueda en GitHub aparecen varios desarollos en Python, Jave, etc. No aparecen aún desarrollos hechos directamente desde R.

Uno de los que destaco es el de Fabio Lamanna TheConcertsTracker.py que recoge el histórico de conciertos de setlist.fm. Debido a mi escaso conocimiento en Python, no he podido reproducir ni aprovechar el desarrollo. Sin embargo, sí me hizo ver las posibilidades que tiene el entorno de datos que posee setlist.fm en la materia, ya no de la granularidad que lleva una canción en sí, si no todas ellas en su conjunto y como van de la mano de una representación en vivo del artista. El desarrollo también permite hacer un rastreo de las setlist.

7. Inspiración

¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder a la comunidad?

A parte del tema personal sobre mi preferencia y gustos musicales, el set de datos me parece interesante en parte por lo mencionado en el apartado **4. Contexto**, y en parte por lo que supone para un usuario, digamos común, poder él mismo o ella misma, con una formación básica en gestión de datos y en entornos como R y Python, tener una visión muy diferente de la vida de su grupo o artista favorito, sin depender en cierta forma del producto hiper procesado de la presa, y otros medios. Un trabajo profundo, muy cercano a sus gustos, puede dar pie a la creación de valor inesperado de la mano de apps para móviles o de la creación de set de datos nacidos mediante la exploración. Estos sets pueden servir para que otros continúen descubriendo más conocimiento o creando otras aplicaciones.

Aprovecho este punto para ofrecer algunas observaciones sobre mi experiencia a lo largo de esta Práctica 01 y que creo ayudarán a entender el trabajo realizado recogiendo la información para crear el set de datos **How soon is Now**.

Observaciones:

```
a. Debo admitir que el set de datos no quedó como me hubiese gustado que
quedara.  La idea era poder incluso asociar las canciones de cada setlist, es
decir las canciones de manera individual con cada evento.  Mi limitación en
```

```
el uso del lenguaje JSON, y cómo este es interpretado por R mediante la
librería jsonlite presentaban una curva de aprendizaje bastante empinada.  En
todo caso, sí se pueden asociar las setlists en su conjunto mendiante su
código.

b. Las setlists quedan en una columnna y con format bastant complejo anidando
información según la estructura de setlist.fm.  Una trabajo quizá a través de
XML y sus librerías en R pudieran acercar un poco más la identificación de
las setlists con el resto, pero intuyo que sería un set separado que luego
habría que unificar a través de join, etc.

c. Si bien el paquete jsonlite presenta ciertas dificultados, sí ha podido
ofrecer una conversión, desde mi punto de vista, fluida hacia lo que es un
data frame.  Los datos se preprocesan hasta lograr dos variable calculadas
para identificar el tiempo.

d. La API de setlist.fm, si bien requiere su obtención a través de un correo
electrónico, obtener los informes con respuestas Json o XML no requiere
incluir el token correspondientes.  Version 1.0.
```

Con estas observaciones hechas procedo a las preguntas que me han surgido o que podría surgirle a alguien que mire este conjunto de datos.

- ¿En cuánto ha disminuido/aumentado/fluctuado la frecuencia de giras a lo largo de los años?
- ¿Se puede determinar el número de cancelaciones a lo largo del tiempo a través de la variable info?
- ¿Es extrapolable este set para hacer un análisis similar para otro artista?
- Preguntas tipo EDA: Año de mayor número de eventos, el mínimo de eventos.
- ¿Cambia el número de eventos tipo Tour a lo largo del tiempo?
- ¿Se puede predecir el patrón de países a visitar (o de cuidades (variable no incluida))
- ¿Se pueden vincular o hacer cross-reference con otros datasets como los de MusicBrainz, o Wikipedia (información financiera de los conciertos)

8. Licencia.

Seleccionad una de las licensias mostradas y justificad la elección.

Transcribo lo indicado en el fichero LICENCE CC0 en el repositorio GitHUb.

La licencia seleccionada es: Released Under CC0: Public Domain License

Create LICENSE CC0

He seleccionado esta licencia porque se adapta a la naturaleza académica de este trabajo. Además ofrece la flexibilidad a todo aquel que quiera expandir las acciones iniciadas aquí y que incluyo en los apartados correspondientes. La práctica me ha regalado más preguntas que respuestas, lo que me motiva a seguir indagando.

Más detalles sobre la misma en la ruta del repo: web-scraping-practica-01-uoc/LICENSE , y aquí

9. Código

Adjuntar el código que ha ayudado a generar el dataset (en R)

```
# Unimos los objetos correspondientes y revisamos la clase
jsonLocsTotal <- rbind_pages(pagesT)

class(jsonLocsTotal)

## [1] "data.frame"
```

Ahora haremos la selección de las columnas que nos interesen y cambiamos su nombres por otros más amigables

```
dataSet <- jsonLocsTotal %>%
              select(id.setlist = setlists.setlist..id,
                     tour = setlists.setlist..tour,
                     fecha.evento = setlists.setlist..eventDate,
                     info = setlists.setlist.info,
                     id.artista = setlists.setlist.artist..mbid,
                     artista = setlists.setlist.artist..name,
                     lugar = setlists.setlist.venue..name,
                     ciudad = setlists.setlist.venue.city..name,
                     pais.codigo = setlists.setlist.venue.city.country..code,
                     pais.nombre = setlists.setlist.venue.city.country..name,
                     setlist = setlists.setlist.sets.set)

# Vemos la estructura
glimpse(dataSet)

## Observations: 1,002
## Variables: 11
## $ id.setlist   <chr> "53e09b3d", "33e0a859", "3be08898", "5be357c0", "...
## $ tour         <chr> "Low In High School", "Low In High School", "Low ...
## $ fecha.evento <chr> "11-11-2017", "10-11-2017", "05-11-2017", "04-11-...
## $ info         <chr> NA, NA, NA, NA, NA, NA, "Webcast live by Arte Con...
## $ id.artista   <chr> "013fa897-86db-41d3-8e9f-386c8a34f4e6", "013fa897...
## $ artista      <chr> "Morrissey", "Morrissey", "Morrissey", "Morrissey...
## $ lugar        <chr> "Hollywood Bowl", "Hollywood Bowl", "Vina Robles ...
## $ ciudad       <chr> "Hollywood", "Hollywood", "Paso Robles", "San Fra...
## $ pais.codigo  <chr> "US", "US", "US", "US", "US", "US", "DE", "FR", "...
## $ pais.nombre  <chr> "United States", "United States", "United States"...
## $ setlist      <list> [<c("You'll Be Gone, Alma Matters, When Last I S...
```

Agregamos variables calculadas para determinar el mes y el año. La variable fecha.evento está en formato chr. Hacemos también cambios a factor en tour, pais, ciudad, y lugar. El dato setlist lo paso a character para que permita guardar en csv.

```
dataSet <- dataSet %>%
  mutate(tour = as.factor(tour),
         fecha.evento = dmy(fecha.evento),
         ciudad = as.factor(ciudad),
         pais.codigo = as.factor(pais.codigo),
         lugar = as.factor(lugar),
         pais.nombre = as.factor(pais.nombre),
         setlist = as.character(setlist))

glimpse(dataSet)
```

```
## Observations: 1,002
## Variables: 11
## $ id.setlist   <chr> "53e09b3d", "33e0a859", "3be08898", "5be357c0", "...
## $ tour         <fctr> Low In High School, Low In High School, Low In H...
## $ fecha.evento <date> 2017-11-11, 2017-11-10, 2017-11-05, 2017-11-04, ...
## $ info         <chr> NA, NA, NA, NA, NA, NA, "Webcast live by Arte Con...
## $ id.artista   <chr> "013fa897-86db-41d3-8e9f-386c8a34f4e6", "013fa897...
## $ artista      <chr> "Morrissey", "Morrissey", "Morrissey", "Morrissey...
## $ lugar        <fctr> Hollywood Bowl, Hollywood Bowl, Vina Robles Amph...
## $ ciudad       <fctr> Hollywood, Hollywood, Paso Robles, San Francisco...
## $ pais.codigo  <fctr> US, US, US, US, US, US, DE, FR, GB, GB, US, US, ...
## $ pais.nombre  <fctr> United States, United States, United States, Uni...
## $ setlist      <chr> "list(song = list(list(`@name` = c(\"You'll Be Go...
```

Añadimos dos variables calculadas a partir del parámetro fecha.evento. El mes y el año.

```
dataSet <- dataSet %>%
  mutate(mes = month(fecha.evento)) %>%
  mutate(año = year(fecha.evento))  # en inglés p
```

Con esto el código ya estaría listo para analizar.

Guardamos el código resultante.

```
write.csv(dataSet, "dataSet.csv")
```

Observaciones finales.

Este código se puede reutilizar para cualquier otro artista simplemente cambiando el código correspondiente.

Una

Bibliografía y Referencias

Lippell, Helen. 2016. “Big Data in the Media and Entertainment Sectors.” In *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe,*

edited by José María Cavanillas, Edward Curry, and Wolfgang Wahlster, 245–59. Cham: Springer International Publishing. doi:10.1007/978-3-319-21569-3_14.