Tipologia i Cicle de Vida de les dades

PRAC 1 – Web Scrapping

Revissió del Document

| Ver. No. | Ver. Data | Preparat per | Revissat per | Aprovat per | Secció Afectada i Resum del Canvi |
| --- | --- | --- | --- | --- | --- |
| 0 | 18/10 | Jose A.Montero |  |  | Draft Inicial |
| 1 | 28/10 | Mireia Mora |  |  | Entrega parcial |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Índice

[1. Introducció i contexte 4](#_Toc54816175)

[2. Descripció del Dataset 6](#_Toc54816176)

[3. Llicencia, codi i registre del dataset 8](#_Toc54816177)

[4. Referencies 12](#_Toc54816178)

[5. Contribucions 13](#_Toc54816179)

Introducció i contexte

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes en una web. Per a la seva realitzacio, s'han de complir una serie de punts.

**Context. Explicar en quin context s'ha recol・lectat la informacio. Explicar per que el lloc web triat proporciona aquesta informació.**

Pretenem fer un estudi de preus en el sector del ticketing, on hi podem trobar força diversitat segons la tipología del events i també dificultats en l’extracció de la información en funció dels controls que faci cadascun dels portals.

Entre els més destacats en l’àmbit estatal podem destacar els següents:

* **Koobin** : per ser una plataforma que integra molta diversitat de events (concerts, teatres, events esportius, etc)
* **Janto :** sobretot per events de teatre
* **Expertus :** pel mon del fútbol i events esportius
* **Clorian :** per museus i events no numerats (sense assignació de seient) en general

i a més grans canals de venda com:

* **El Corte Inglés**
* **TicketMaster**
* **Entradas.com**

La idea és analitzar un parell d’aquests portals i poder extreure informació de preus d’alguns dels events més representatius que tenen a la venta. Fent un primer análisis ens trobem que TicketMaster bloqueja la lectura de les seves planes fent servir Request desde Python, per tant ja és una opció que queda descartada.

No obstant, amb Koobin ens trobem patrons de cerca relativamente senzill per a obtenir preus

**Koobin**

**Madame Butterfly**

<https://operaoviedo.koobin.com/butterfly>

**Inicial per a buscar sessions**

Recinto : <meta content="Ópera de Oviedo" property="og:site\_name"/>

Evento : <meta content="Madama butterfly" property="og:title"/>

Lista de sesiones:

<div class="seleccioSessio">

<div class="titoldiasessio">

Lunes, 9 de noviembre de 2020

</div>

<div class="titolhorasessio">

Sesión 20:00 h

</div>

**Link**

<a class="**buttonBig**" href="https://operaoviedo.koobin.com/index.php?action=PU\_evento&amp;Ev\_id=422">

* El patró és

<https://operaoviedo.koobin.com/index.php?action=PU_evento&amp;Ev_id=>

**Madame Butterfly 09/11 20:00 🡪 patró per a preus**

Recinto : <meta content="Ópera de Oviedo" property="og:site\_name"/>

Evento : Title

<meta content="Madama butterfly" property="og:title"/>

Sesión : <span class="eventoTaronja">

lunes, 9 de noviembre de 2020, 20:00 h

</span>

Zona Preu : class="areaNom" (bucle por class="seleccioarea corner")

Preu : <div class="areaPreu">

Precio:

<strong class="Taronja nowrap">

**16,00 €**

</strong>

</div>

* Pel cas de Koobin es veu un patró relativamente sencill per a obtenir la informació via BeautifulSoup, la qüestió és que aquest patró és per a un event en concret perque cadascú tindrà un portal diferent, caldrà fer un webscrapper per cadascú (per exemple la información pel Mago Pop ens arrivarà en un format diferent).
* Hem de pensar a fer funcions en Python per fer el scrapping en cada cas i després desde el programa principal cridar-les convenientment

Descripció del Dataset

1. **Definir un titol pel dataset. Triar un titol que sigui descriptiu.**

Tenim en compte la temàtica escollida podriem agafar el següent títol : “**Comparador de ticketing”**

1. **Descripcio del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (es necessari que aquesta descripció tingui sentit amb el titol triat).**

El dataset no requereix un número de variables molt extens. El objectiu principal és poder comparar preus per les motivacions finals que pugui tenir. Si hem de crear una descripció de cara als possibles usuaris aquesta podría ser:

Aquest dataset conté informació de preus per diferents esdeveniments culturals futurs que estan essent oferts per diverses plataformes de ticketing. Mitjançant la seva consulta l’usuari podrà escollir quina és la millor opció per a la compra de tickets per a un esdeveniment concret o en general veure quines ofertes hi han pels propers dies.

1. **Representacio grafica. Presentar una imatge o esquema que identifiqui el**

**dataset visualment.**

1. **Contingut. Explicar els camps que inclou el dataset, el periode de temps de les**

**dades i com s'ha recollit.**

Un dataset que contingui preus pot estar força acotat a presentar la següent informació:

**< Ticketera, Tipo\_Evento, Recinte, Event, Sessió, Zona Preu, Preu, Data\_Registre >**

On passem a detallar cada camp:

**Ticketera** : es el portal d’on hem adquirit les dades (Koobin, Janto, etc)

**Tipo\_Evento** : representa la taxonomía, o tipología de l’event que s’està tractant, és a dir, la vertical/sector dintre del mon del ticketing. Pot ser Opera, Teatre, Museus, Futbol, Esports, Concerts, etc

**Recinte :** el lloc físic on es desenvolupen els events, per exemple el Teatre Tívoli.

**Event :** cadascun dels espectacles que s’organitzen dintre d’un recinte per a la difussió i gaudiment de la cultura. Per exemple, el Musical de La Jaula de Las Locas

**Sessió :** cadascuna de les representacions d’un event en el temps, és a dir, podríem dir que és la programació dels events, normalment haven-t’hi varies session per un event. Per exemple, del Musical de La Jaula de Las Locas, la sessió del dia 22 de octubre a les 22:00 hores.

**Zona de Preu :** es la divissió de l’aforament que es fa en els recintes per a escollir el preu en el moment de la compra del ticket. Acostuma a estar delimitada per l’arquitectura del recinte o bé per restriccions d’aforament com les existents actualment derivades de la pandèmia. Exemples podrien ser la Platea Esquerra, Platea Dreta, Palcos VIP, etc

**Preu :** import monetari del que costa un ticket a la zona de preu escollida.

**Data\_Registre** : moment en que es fa l’extracció de la informació.

Una instancia del dataset seguint els atributs considerats seria

Koobin, Opera, Opera de Oviedo, Madama butterfly, Lunes 9 de noviembre 2020 – 20:00h, Palco de Platea, 181.00, 18/10/2020 17:52

**En quan al periode de temps de dades** és important considerar l’ús que es poden fer de les mateixes :

1. **Ús personal com a agenda:** Extreure tota la informació com a ús personal, pot servir per tenir la informació a l’abast en un temps molt reduït, si el que es vol és extreure tota la informació possible d’un esdeveniment estalviant el temps d’haver de navegar per la web. D’aquest mode, l’estracció de les dades en un fitxer ens permetrà tenir la informació a l’abast per decidir quina serà la millor opció si es vol atendre a un esdeveniment concret.
2. **Ús empresarial com a estudi de mercat o competència :**
3. **Agraiments. Presentar el propietari del conjunt de dades. Es necessari incloure**

**cites de recerca o analisis anteriors (si n'hi ha).**

1. **Inspiracio. Explicar per que es interessant aquest conjunt de dades i quines**

**preguntes es pretenen respondre.**

Llicencia, codi i registre del dataset

1. **Llicencia. Seleccionar una d'aquestes llicencies pel dataset resultant i explicar el motiu de la seva seleccio:**

**○ Released Under CC0: Public Domain License**

**○ Released Under CC BY-NC-SA 4.0 License**

**○ Released Under CC BY-SA 4.0 License**

**○ Database released under Open Database License, individual contents**

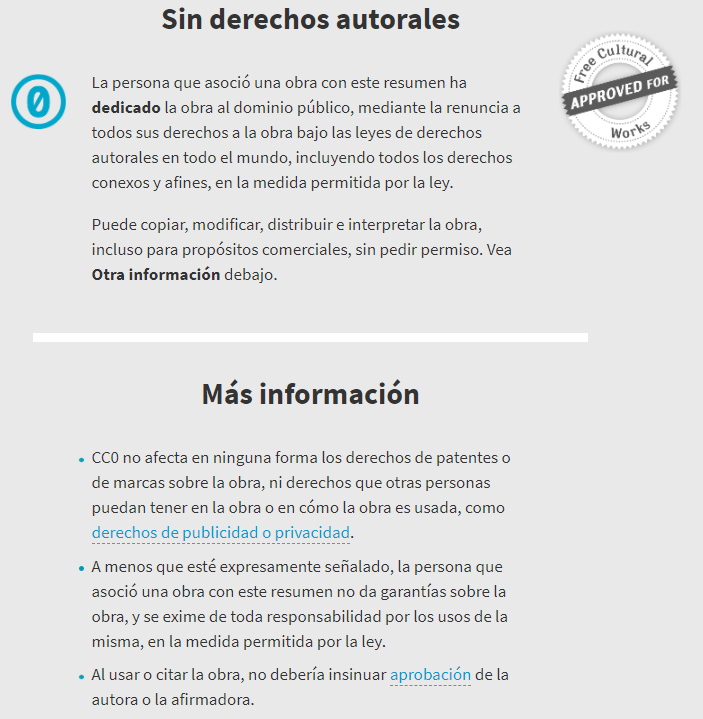
**under Database Contents License**

**○ Other (specified above)**

**○ Unknown License**

Abans de seleccionar una de les llicències esmentades anteriorment, caldria detallar bé a què correspon cada llicència, quines característiques té cadascuna :

**○ Released Under CC0: Public Domain License**



**Font: creativecommons.org**

**○ Released Under CC BY-NC-SA 4.0 License**



**Font: creativecommons.org**

**○ Released Under CC BY-SA 4.0 License**



**○ Database released under Open Database License, individual contents**

**under Database Contents License**

Estas libre:

* *Compartir* : copiar, distribuir y utilizar la base de datos.
* *Crear* : Producir obras a partir de la base de datos.
* *Adaptar* : modificar, transformar y construir sobre la base de datos.

Siempre que usted:

* *Atributo* : Debe atribuir cualquier uso público de la base de datos, o trabajos producidos a partir de la base de datos, de la manera especificada en la ODbL. Para cualquier uso o redistribución de la base de datos, o de los trabajos producidos a partir de ella, debe dejar en claro a los demás la licencia de la base de datos y mantener intactos los avisos en la base de datos original.
* *Share-Alike* : Si utiliza públicamente cualquier versión adaptada de esta base de datos, o trabajos producidos a partir de una base de datos adaptada, también debe ofrecer esa base de datos adaptada bajo la ODbL.
* *Mantener abierto* : si redistribuye la base de datos, o una versión adaptada de ella, entonces puede usar medidas tecnológicas que restrinjan el trabajo (como DRM) siempre que también redistribuya una versión sin tales medidas.

**Font: Opendatacommons.org**

**○ Other (specified above)**

**○ Unknown License**

1. **Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en**

**Python o, alternativament, en R.**

<veure fitxer adjunts>

Hem treballar amb el IDE PyCharm per a desenvolupar els nostres scripts python.

El codi es compossa del següents fitxers:

* Koobin\_scrapper\_functions.py : amb funcions per extrereu i parsejar dades segons la tipologia d’events
* Koobin\_scrapper\_prices.py : scrapper per a events de teatre, opera I basket (entorn de test de Koobin)
* Koobin\_scrapper\_noa.py : scrapper per a events de teatre de Koobin
* Janto\_scrapper\_musical\_jaula.py : scrapper per a events musicals de Janto

Cal comentar que ens trobem amb diferentes casuístiques a les diferentes web que hem volgut fer scraping:

* Koobin : deixava accedir sense problema i el patró de búsqueda a les seves planes és bastant comú
* Janto : ha calgut indicar el parametre a la crida request.get
* VivaTickets : obtenim una resposta que sembla indicar que ens detecten i ens encapsulen la petició

<html>

<head>

<meta content="noindex,nofollow" name="robots"/>

<script src="/\_Incapsula\_Resource?SWJIYLWA=5074a744e2e3d891814e9a2dace20bd4,719d34d31c8e3a6e6fffd425f7e032f3">

</script>

<body>

</body></head></html>

1. **Dataset. Publicar el dataset en format CSV a Zenodo (obtencio del DOI) amb**

**una breu descripció.**

Referencies

|  |  |  |
| --- | --- | --- |
| **Document** | **Descripció** | **Path** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Contribucions

|  |  |
| --- | --- |
| **Contribució** | **Signa** |
| Recerca prèvia |  |
| Redacció de les respostes |  |
| Desenvolupament de codi |  |
|  |  |
|  |  |