



Universitat Oberta
de Catalunya

Tipologia i Cicle de Vida de les dades PRAC 1 – Web Scrapping

Revisió 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	04/11	Jose A.Montero /Mireia Mora			Ampliació informació del document en general
2	07/11	Jose A.Montero / Mireia Mora			Petits canvis del document en general

Índice

1. Introducció i contexte.....	4
2. Descripció del Dataset	5
3. Llicència, codi i registre del dataset.....	9
4. Implementació.....	13
5. Referències.....	15
6. Contribucions.....	15

1. Introducció i contexte

Qui no ha passat una tarda en una agència de viatges planejant les seves properes vacances, o ha anat a les guixetes d'un teatre per comprar les entrades per aquella obra de cap de setmana, o al punt d'informació d'un gran centre comercial i fer cua per comprar aquella entrada de concert tan desitjada.... Fa molts anys que la dinàmica habitual de compra online es va anar imposant a la societat, i mica en mica el fet de perdre un temps fora de casa (molts cops temps molt ampli) fent aquestes gestions per comprar cultura va anar donant pas a un sistema més còmode per les persones a base de clics des del sofà, des de l'ordinador, des del mòbil...

Avui dia, això es pot fer a cop de clic a través de plataformes del sector del ticketing en qualsevol sector d'oci, cultura, esports, etc...

La idea d'aquesta pràctica és realitzar un estudi de preus en el sector del ticketing on hi podem trobar força diversitat segons la tipologia del esdeveniments, i també dificultats en l'extracció de la informació en funció dels controls que faci cadascun dels portals.

Decidim realitzar la pràctica en Python tot i no tenir cap tipus de coneixements previs en aquest llenguatge, sigui així un repte i una bona manera de començar desde zero en aquest llenguatge de programació. Mitjançant els fitxers de documentació, així com les cerques fetes (veure punt 5 del document "Referencies"), es comença a planificar les diferents fases i objectiu de la pràctica.

L'objectiu d'aquesta activitat ha estat la creació d'un dataset a partir de les dades contingudes en tres plataformes diferents, i d'aquesta manera, extreure informació de preus d'alguns dels esdeveniments més representatius que tenen a la venda. Per començar, es va definir un primer llistat de les plataformes més destacades:

- **Koobin** : plataforma que integra molta diversitat de events (concerts, teatres, events esportius, etc)
- **Janto** : esdeveniments de teatre sobretot.
- **Vivaticket** : plataforma italiana d'esdeveniments de teatre, música, òpera i ballet, esport, art, altres...
- **Expertus** : món del fútbol i esdeveniments esportius
- **Clorian** : museus i esdeveniments no numerats (sense assignació de seient) en general

i a més grans canals de venda com:

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

Després de fer un primer anàlisis a TicketMaster, trobem que aquesta plataforma bloqueja la lectura de les seves planes fent servir Request desde Python, i per tant, aquesta opció queda descartada.

Es decideix extreure dades de les plataformes : **Koobin, Janto i Vivaticket**.

No obstant, amb la plataforma Koobin ens trobem patrons de cerca relativament senzills per a obtenir preus, on a través de la biblioteca de Python BeautifulSoup, ens trobem el següent codi HTML per poder extreure la informació desitjada:



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

→ El patró és

https://operaoviedo.koobin.com/index.php?action=PU_evento&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 : lunes, 9 de noviembre de 2020, 20:00 h

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

Preu : <div class="areaPreu"> Precio:

<strong class="Taronja nowrap">

16,00 €

</div>

2. Descripció del Dataset

L'objectiu principal de la informació recollida al dataset és poder llistar els preus dels diferents esdeveniments tenint en compte les sessions i les diferents zones de preus. per tal de comparar-los amb les diferents motivacions que es puguin tenir (personals, empresarials....).

Si hem de crear una descripció de cara als possibles usuaris aquesta podria 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.

Tenim en compte la temàtica escollida el títol escollit del dataset: “**Comparador de ticketing**”.

Aquest dataset no requereix un número de variables molt extens, es defineixen 8 variables que detallem a continuació:

['Ticketera', 'Tipus Event', 'Recinte', 'Event', 'Sessió', 'Zona Preu', 'Preu', 'Data Registre']

Ticketera : és el portal d'on hem adquirit les dades (Koobin, Janto, Vivaticket)

Tipo_Evento : representa la taxonomia, o tipologia de l'event que s'està tractant, és a dir, la vertical/sector dintre del món del ticketing. Pot ser Opera, Teatre, Museu, Futbol, Esports, Concerts, etc

Recinte : Lloc físic on es desenvolupa l'esdeveniment, per exemple el Teatre Tívoli.

Event : cadascun dels espectacles que s'organitzen dintre d'un recinte per a la difusió 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 sessions 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 : és la divisió 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 instància 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

De les dades extretes és important considerar l'ús que es poden fer de les mateixes diferenciant en quin àmbit es pot utilitzar aquest ús :

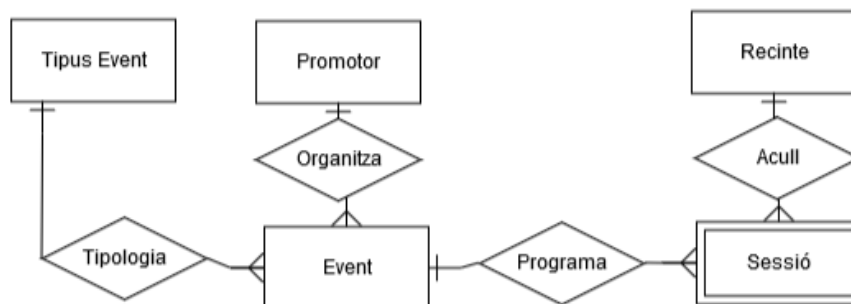
a) **Ús personal com a agenda:** Extreure tota la informació com a ús personal, pot servir per tenir la informació a l'abast en un temps molt reduït, si el que es vol és extreure tota la informació possible

d'un o varis esdeveniment estalviant el temps d'haver de navegar per la web. D'aquest mode, l'extracció 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 amb un estalvi de temps considerable.

- b) **Ús empresarial com a estudi de mercat o competència** : En aquest àmbit podria ser interessant fer un estudi de la competència per controlar quins preus són els que tenen en l'actualitat per cada esdeveniment. Com un dels camps del dataset ens marca la data i l'hora de l'extracció, podríem analitzar quan i en quin esdeveniment pot haver una modificació dels preus. Anant més enllà, també es podria analitzar quins esdeveniments podrien ser per exemple els més visitats si afegíssim més camps al dataset, o quines zones serien les més venudes abans, i per tant més demandades, per exemple si afegíssim el nombre de localitats venudes en cada zona.

Un ús conegut d'informació d'aquest tipus, i que podríem considerar poc ètic, és per part dels portal de revenda de tickets o de mercari secundari com Viagogo. Es coneixen pràctiques de web scrapping per part d'aquests portals cap als llocs oficials de venda d'entrades o directament cap als organitzadors d'event amb l'objectiu d'extreure preus i poder **implementar un Sistema de preus dinàmics** als seus portal de venda amb l'objectiu de generar més beneficis. Normalments aquests preus dinàmics es generaran en base a comparatives entres el preus actuals que tinguin els revendedors i el que conseqüeixen extreure dels portals oficials i probablement aplicant algun model de predicció.

Representacio grafica. Presentar una imatge o esquema que identifiqui el dataset visualment.



Agraïments. Presentar el propietari del conjunt de dades. Es necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Volem agrair particularment al portal **Koobin** i al seu CTO que ens van autoritzar a fer extraccions de manera controlada i amb finalitats merament acadèmiques. En aquest sentit van posar a disposició nostre el seu entorn de test per a poder extreure informació d'un evento Esportiu (partit de basket) ja en producció no hi tenen actualment res a la venta pel tancament de recintes degut a la pandèmia. Degut a la seva diversitat d'event i a la homogeneïtat en com estan construïdes les seves planes lo que ens ha facilitat molt la implementació dels scrappers, ha sigut el portal amb el que més hem treballat.

No ens ha estat possible contactar amb ningú de Janto ni Vivaticket per a avisar-los dels nostres propòsits, per tant no hem volgut ser massa intrusius amb aquests dos portals.

Hem intentat buscar cites i anàlisis anteriors d'aquestes plataformes a través de Google acadèmic o portals com a ResearchGate però per aquesta temàtica escollida de la pràctica, plenament empresarial, no hem aconseguit trobar cap article, ni publicació prèvia.

Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

En el món actual, on les presses i el ritme vertiginós del dia a dia ens porta a voler gaudir del nostre temps d'oci d'una manera ràpida i des del sofà, fa que el conjunt de dades escollit per fer aquesta pràctica sigui un objectiu àgil, resumit i amb la informació necessària a l'abast.

El fet de tenir en un sol fitxer la informació de varis espectacles dels que, o bé es vol fer un ús personal, o bé un ús de caire empresarial, ens fa tenir en un sol click i sobretot un estalvi de temps important unes dades de les que poder fer un ús a posteriori.

Per tal de poder decidir en qualsevol dels àmbits mencionats (personal o empresarial), i tenint en compte l'objectiu marcat, ens podríem formular les següents preguntes:

- ¿ Puc aconseguir tota la informació d'un espectacle resumida per tal de decidir-me en l'elecció d'un dia per assistir-hi ?
- ¿ Puc detallar els preus dels espectacles de totes les zones i de totes les sessions programades ?
- ¿ Puc saber si s'han actualitzat els preus respecte a una extracció anterior ?
- ¿ Puc comparar la informació dels espectacles de varies plataformes amb varis fitxers d'extracció ?
- ¿ Puc utilitzar les dades generades per tal de poder filtrar informació posteriorment ?
- ¿ Està venent la competència a uns preus diferents als que jo estic oferint?

3. Llicència, codi i registre del dataset


Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

- ☐ 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, i quines característiques té cadascuna, així que a continuació exposem d'una manera resumida les característiques de cadascuna :


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

Sin derechos autorales



La persona que asoció una obra con este resumen ha **dedicado** la obra al dominio público, mediante la renuncia a todos sus derechos a la obra bajo las leyes de derechos autorales en todo el mundo, incluyendo todos los derechos conexos y afines, en la medida permitida por la ley.

Puede copiar, modificar, distribuir e interpretar la obra, incluso para propósitos comerciales, sin pedir permiso. Vea **Otra información** debajo.



Más información

- CC0 no afecta en ninguna forma los derechos de patentes o de marcas sobre la obra, ni derechos que otras personas puedan tener en la obra o en cómo la obra es usada, como [derechos de publicidad o privacidad](#).
- A menos que esté expresamente señalado, la persona que asoció una obra con este resumen no da garantías sobre la obra, y se exime de toda responsabilidad por los usos de la misma, en la medida permitida por la ley.
- Al usar o citar la obra, no debería insinuar [aprobación](#) de la autora o la afirmadora.

Font: creativecommons.org



➤ Released Under CC BY-NC-SA 4.0 License


Usted es libre de:


Compartir — copiar y redistribuir el material en cualquier medio o formato

Adaptar — remezclar, transformar y crear a partir del material

El licenciador no puede revocar estas libertades mientras cumpla con los términos de la licencia.

Bajo las condiciones siguientes:

 **Reconocimiento** — Debe [reconocer adecuadamente](#) la autoría, proporcionar un enlace a la licencia e [indicar si se han realizado cambios](#). Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.

 **NoComercial** — No puede utilizar el material para una [finalidad comercial](#).

 **CompartirIgual** — Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la [misma licencia que el original](#).

No hay restricciones adicionales — [No puede aplicar términos legales o medidas tecnológicas](#) que legalmente restrinjan realizar aquello que la licencia permite.

Font: creativecommons.org

➤ Released Under CC BY-SA 4.0 License

Eres libre de:

Compartir : copia y redistribuye el material en cualquier medio o formato.

Adaptarse : remezclar, transformar y construir sobre el material para cualquier propósito, incluso comercial.

El licenciente no puede revocar estas libertades siempre que siga los términos de la licencia.



Bajo los siguientes términos:



Atribución : debe otorgar [el crédito correspondiente](#), proporcionar un enlace a la licencia e [indicar si se realizaron cambios](#). Puede hacerlo de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciente lo respalda a usted o su uso.



ShareAlike : si remezcla, transforma o construye sobre el material, debe distribuir sus contribuciones bajo la [misma licencia](#) que el original.

Sin restricciones adicionales : no puede aplicar términos legales o [medidas tecnológicas](#) que restrinjan legalmente a otros de hacer cualquier cosa que permita la licencia.

Font: creativecommons.org

➤ Database released under Open Database License, individual content 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](https://opendatacommons.org)

Tenim present les pràctiques conegudes dels revenedors d'entrades optariem per la llicència de tipus **Released Under CC BY-NC-SA 4.0 License**, per a protegir el seu ús de caràcter comercial i restringint el seu ús inicial a caràcter personal. Expliquem els paràmetres que porten a aquesta decisió

1. **Atribució (BY)**: per utilitzar una obra a qualsevol tipus de mitjà és imprescindible citar a l'autor de forma explícita.
2. **NoComercial (NC)**: pots utilitzar la obra per generar un altra, sempre i quan no vagis a aconseguir diners amb ella. Molt comú a fotos de Flickr.
3. **CompartirIgual (SA)**: es pot utilitzar la obra per crear una altra, sempre i quan aquesta es publiqui amb la mateixa llicència que la obra original.

Font : genbeta.com/herramientas/licencias-creative-commons-explicadas-para-dummies

4. Implementació

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

El codi es compon del següents fitxers:

- Koobin_scrapper_functions.py : fitxer de python on estan recollides les funcions per extreure i parsejar dades segons la tipologia d'events de qualsevol de les pàgines que es vulgui extreure informació. Es fa crida d'aquestes funcions desde els altres scripts python.
- Koobin_scrapper_prices.py : scrapper per a events de teatre, opera i basket (entorn de test de Koobin).
- Janto_scrapper_prices.py : scrapper per a esdeveniments musical, Magia i Familiar de Janto.
- Vivaticket_scrapper.py : scrapper per a events de opera del portal italià Vivaticket.

Cal comentar que ens trobem amb diferents casuístiques a les diferents webs que hem volgut fer scraping:

- **Koobin** : deixava accedir sense problema fent servir la llibreria **BeautifulSoup** i el patró de búsqueda a les seves pàgines és bastant comú
- **Janto** : ha calgut indicar el parametre headers a la crida request.get

```
v_headers = {  
    "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp, \n  
    /*;q=0.8",  
    "Accept-Encoding": "gzip, deflate, sdch, br",  
    "Accept-Language": "en-US,en;q=0.8",  
    "Cache-Control": "no-cache",  
    "dnt": "1",  
    "Pragma": "no-cache",  
    "Upgrade-Insecure-Requests": "1",  
    "User-Agent": user_agent_desktop}
```

i la crida :

```
page = requests.get(p_url, headers = v_headers)
```

- **VivaTickets** : inicialment vam obtenir una resposta que semblava indicar que ens detecten i ens encapsulen la petició

```
<html>  
<head>
```

```
<meta content="noindex,nofollow" name="robots"/>
<script
src="/_Incapsula_Resource?SWJIYLWA=5074a744e2e3d891814e9a2dace20bd4,719d34d31c8e3a6e6fffd425f7e
032f3">
</script>
<body>
</body></head></html>
```

i finalment hem pogut fer scrapping d'aquest portal fent servir Selenium e Xpath com a mètode de cerca a les planes HTML. Destacar que les lectures amb Selenium son més lentes que amb BeautifulSoup, aspecte a considerar si volem periodificar scrips o bé fer extraccions més complexes que les que hem tractat.

Per carregar el driver de Selenium hem fet

```
#Necessitem treballar amb Selenium per extreure d'aquest portal
DRIVER_PATH = "d:\\UOC_ML\\chromedriver.exe"
options = Options()
options.headless = True
options.add_argument("--window-size=1920,1200")
```

Per extreure el codi html d'una url concreta

```
#invoquem el driver per obtenir el html
driver = webdriver.Chrome(options=options, executable_path=DRIVER_PATH)
soup = driver.get(p_url)
```

I per fer cerques amb Xpath

```
v_event = driver.find_element_by_xpath("//h2[@class='__title -uppercase']")
```

5. References

Document	Descripció	Path
Python Pocket Reference	Referencia per Python	Editorial O'Reilly
Web Scrapping using Selenium and Python	Tutorial de com fer servir Selenium	https://www.scrapingbee.com/blog/selenium-python/
Practical XPath for Web Scrapping	Utilitzar XPath expressions per fer web scrapping	https://www.scrapingbee.com/blog/practical-xpath-for-web-scrapping/
Web amb exemples i explicacions ús BeautifulSoup	Exemple de l'ús de BeautifulSoup	https://jarroba.com/scraping-python-beautifulsoup-ejemplos/
Selenium with Python	Document sobre Selenium	https://selenium-python.readthedocs.io/
Llicències creative commons	Web sobre les llicències	https://creativecommons.org/licenses/
Llicències creative commons	Web sobre les llicències	https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.es

6. Contribucions

Contribució	Signa
Recerca prèvia	José Antonio Montero, Mireia Mora
Redacció de les respostes	José Antonio Montero, Mireia Mora
Desenvolupament de codi	José Antonio Montero, Mireia Mora