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 | 04/11 | Mireia Mora /Jose A.Montero |  |  | Entrega final |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Índice

[1. Introducció i contexte 4](#_Toc53932283)

[2. Descripció del Dataset 6](#_Toc53932284)

[3. Llicencia, codi i registre del dataset 8](#_Toc53932285)

[4. Referencies 9](#_Toc53932286)

[5. Contribucions 9](#_Toc53932287)

Introducció i contexte

Qui no ha passat una tarda en una agencia 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 semana, o al punt d’informació d’un gran centre comercial i fer cua per comprar aquella entrada de concert tan dessitjada…. Fa molts anys que la dinámica 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 día, 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 práctica és realitzar un estudi de preus en el sector del ticketing on hi podem trobar força diversitat segons la tipología del esdeveniments, i també dificultats en l’extracció de la informació en funció dels controls que faci cadascun dels portals.

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 venta. 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 análisis 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.**

Afegir aquí info de robots.txt ?

Afegir aquí info de whois o no cal ?

Afegir casuística amb headers de Janto i Selenium de vivaticket (això millor explicar-ho a la part de codi)

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 Beautiful Soup, ens trobem el següent codi HTML per poder extreure la información 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**

<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>

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 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.*

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

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

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

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

**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, 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 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 :** 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 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

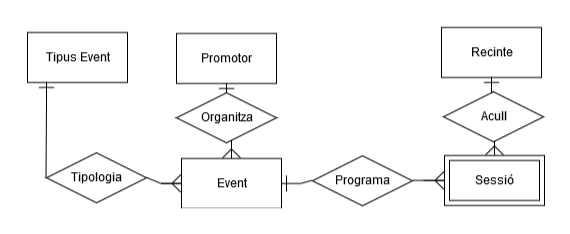
De les dades extretesés important considerar l’ús que es poden fer de les mateixes diferenciant en quin àmbit es pot utilitzar aquest ús :

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 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.
2. **Ús empresarial com a estudi de mercat o competència :** En aquest àmbit podria ser interessant fer un estudi de la competencia 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, podriem 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 afegissim més camps al dataset, o quines zones serien les més venudes abans, I per tant més demandades, per exemple si afegissim el nombre de localitats venudes en cada zona. (ampliar per exemple comparative de preus per ofertes ??? check coherencia info)

Un ús conegut de informació d’aquest tipus, i que podriem 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 revendors i el que consegueixen extreure dels portals oficials i probablement aplicant algun model de predicció.

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

|  |  |  |
| --- | --- | --- |
| **Interfaz de usuario gráfica  Descripción generada automáticamente** | **Logotipo, nombre de la empresa  Descripción generada automáticamente** |  |

****

1. **Agraiments. Presentar el propietari del conjunt de dades. Es necessari incloure**

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

Volem agraïr particulament 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 possar a disposició nostre el seu entorn de test per a poder extreure información d’un evento Esportiu (partit de basket) ja en en producción no hi tenen actualment res a la venta pel tancament de recintes degut a la pandemia. Degut a la seva diversitat d’event i a la homogeneitat en com están construides les seves planes lo que ens ha facilitat molt la implementación 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 nostre propòsits, per tant no hem volgut ser massa intrusius amb aquests dos portals.

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

**preguntes es pretenen respondre.**

En el món actual, on les presses i el ritme vertiginós del día a día 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 práctica sigui un objectiu ágil, 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 podriem formular les següents preguntes:

* Puc aconseguir tota la informació d’un espectacle resumida per tal de decidir-me en l’elecció d’un día per asistir-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 utilizar les dades generades per tal de poder filtrar información posteriorment ?
* ¿Està venent la competencia a uns preus diferents als que jo estic oferint?

Llicencia, codi i registre del dataset

**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**

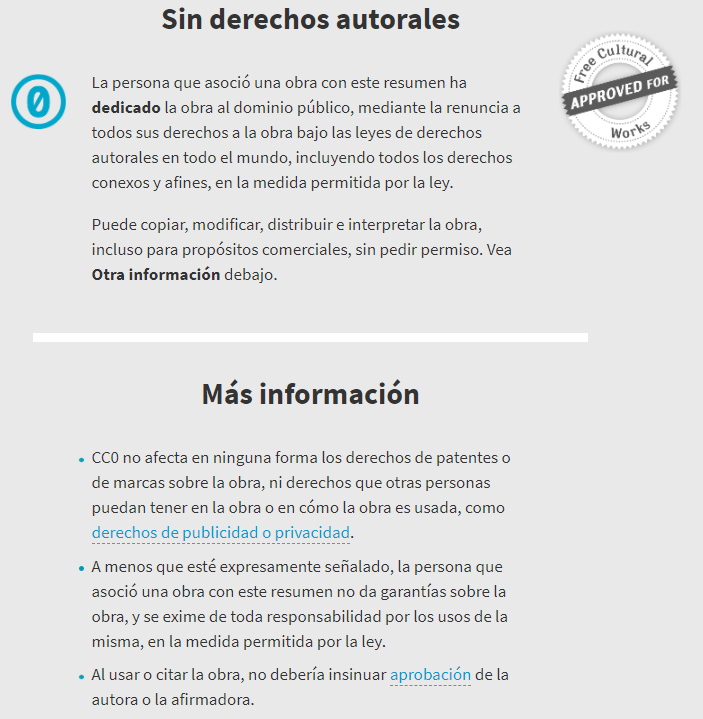
**under Database Contents License**

**○ Other (specified above)**

**○ Unknown License**

Abans de seleccionar una de les llicències esmentades anteriorment, caldria detallar bé a què correspon cada llicència, i quines característiques té cadascuna, així que a continuació exposem d’una manera resumida les característiques de 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 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**

* **Other (specified above)**
* **Unknown License**

Tenim present les pràctiques conegudes dels revenedors d’entrades optariem per la llicencia 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.

Implementació

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

El codi es compossa del següents fitxers:

* scrapper\_functions.py : fitxer de python on estan recollides les funcionsper extrereu 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 les altres pàgines,
* 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 diferentes casuístiques a les diferentes webs que hem volgut fer scraping:

* **Koobin** : deixava accedir sense problema fent server la llibreria **BeautifulSoup** i el patró de búsqueda a les seves planes 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,\  
 \*/\*;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=5074a744e2e3d891814e9a2dace20bd4,719d34d31c8e3a6e6fffd425f7e032f3">

</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']"**)

Referencies

|  |  |  |
| --- | --- | --- |
| **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-scraping/ |
| 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/ |

Contribucions

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