Universitat Oberta de Catalunya

TIPOLOGIA I CICLE DE VIDA DE LES DADES

Pràctica 1 – Web Scraping [Zacatrus]

Eduard Ruiz Solé
11/11/2019

INDEX

1. CONTEXT

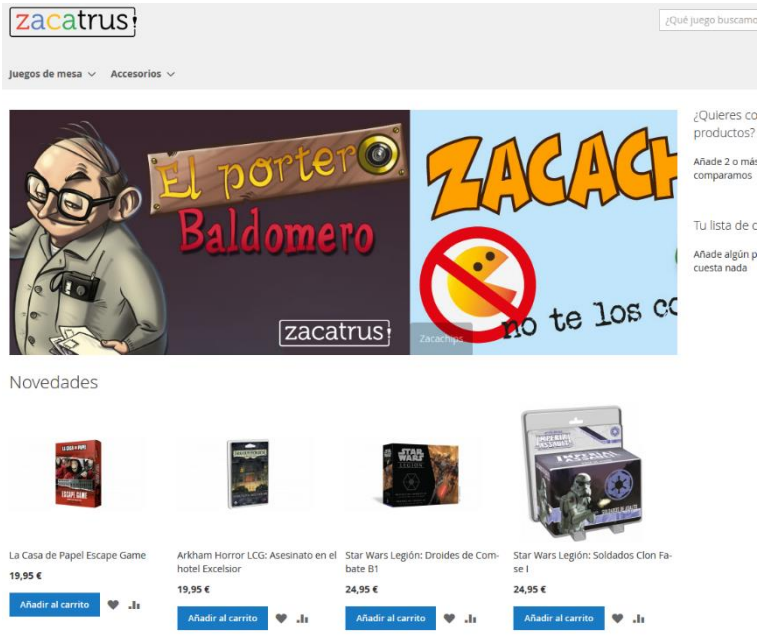
Explicar en quin context s'ha recol·lectat la informació. Explicar perquè el lloc web triat proporciona aquesta informació.

L'us del web scraping està ben lligat amb la necesitat d'adquirir un conjunt de dades que compleixin o satisfacin uns parametres definits. Per tant, es crearà un algoritme d'extracció d'informació per ajudar a un individu o a tot un col·lectiu. En aquest cas, s'ha pensat que com ja arriba l'època de nadal, amb els típics menjars familiars i els regals sota l'arbre i el tió, no estaria gens malament crear un algoritme que facilités aquell moment de dubte, el qual ens trobem en una botiga de joguines amb 4 jocs a les mans i un malestar general per no saber quin triar. Doncs bé, gràcies a aquest projecte es pot prevenir aquesta situació i desplaçar-nos cap a la botiga de joguines amb els deures fets i per tant, els jocs ben triats, fet que no comportarà cap sorpresa adicional.

De les múltiples pàgines web de venda de jocs, s'ha escollit Zacatrus, ja que no hi ha una limitació evident en l'ús de la tècnica de web scraping, no té una estructura molt complexa i és senzilla a l'hora d'utilitzar-la. De fet, abans d'escollir la web i la idea d'aquest projecte, centrat en una pàgina de jocs, un parell de dies abans vaig estar tafanejant per començar a comprar alguns dels regals de nadal. Aquest fet, va fer que vista una necessitat, només fes falta la creació d'una solució i s'ha pogut complir gràcies a aquesta pràctica.

L'objectiu de realitzar web scraping en aquesta pàgina consisteix en trobar jocs de taula de diferents categories i obtenir aquells que s'hagin vengut més. Aquesta és una manera d'estalviar temps de filtratge dins la pàgina web i arribar a trobar el joc més adient, a partir d'una llista amb totes les informacions necessàries.

Com podem observar, el lloc web és molt simple i bàsic, fet que fa que la majoria de productes no mostrin necessàriament de bones i primeres, aquelles dades que podrien ser d'especial interès, com per exemple : dificultat, descripció i jugadors. Cal esmentar que totes aquestes dades estan contingudes en el lloc web i és per això que ha estat l'escollit.



Captura de pantalla de la pàgina web Zacatrus.es

Un bon ús d'aquests recursos ens permetran obtenir aquells jocs més venuts i amb una clara tendència en la temporada actual, fet que ajudarà a escollir els regals més adients per a cadascun dels votres fills, cosins, amics i nebots.

2. DEFINIR UN TÍTOL PEL DATASET

Triar un títol que sigui descriptiu.

Si busques jocs, aquí en tens.. i no pocs !

Sobretot en aquelles èpoques on els regals estan en boca de tothom, aquest projecte, extreu un llistat dels jocs més populars amb les característiques associades i d'interès. Gràcies al lloc web Zacatrus, obtenim una extensa llista pero no interminable de jocs de taula, ja que la idea és ajudar a triar el millor regal abans d'anar a la botiga, però no incrementar els dubtes sobre el joc a escollir.

El títol esmentat anteriorment juga amb la quantitat de jocs oferits, gràcies a l'aplicacio d'un algortime, ja que el resultat final obté un llistat de 144 registres, classificats en 6 categories diferents, com són: Juegos de cartas, Juegos de tablero, Juegos de Rol, etc ..

3. DESCRIPCIÓ DEL DATASET

Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

Les informacions extretes estan dividides en tres conjunts, el primer consta del nom del joc, la categoria i el preu, ja que aquestes són les dades de més importància pels consumidors. Per altra banda, tenim les informacions més concretes, com : els autors, la temàtica, el temps de joc, la dificultat, el nombre de jugadors, l'idioma i l'edat. Aquests valors detallen i fan més comprensible el primer conjunt esmentat. Per acabar, el tercer grup correspon a aquelles informacions més específiques i utilitzades quan ja es vol accedir al producte, com són la descripció del joc i la URL.

El dataset consta de les següents 12 variables :

NOMBRE
TIPO
AUTORES
TEMÁTICA
PRECIO
TIEMPO DE JUEGO
NÚMERO DE JUGADORES
DIFICULTAD
IDIOMA
DESCRIPCIÓN
EDAD
URL

Primer de tot es recopilen les dades : Nom, Tipus de joc (Categoria) i URL, i s'utilitza aquesta última (accedint-hi) per obtenir la resta de les dades.

A partir de l'URL recopilada anteriorment, s'extreuran les següents dades desitjades :

Autors, Temàtica, Preu, Temps de joc, Número de jugadors, Dificultat, Idioma, Descripció i Edat.

Draftosaurus

No hay opiniones sobre este artículo. ¿Nos das una?

19,95 € DISPONIBLE

¡Envío gratis! *Para todos los envíos a la península.

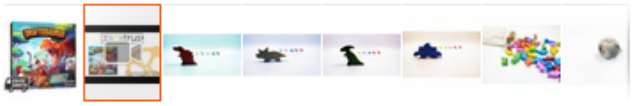
Conseguirás **59 Fichas** comprando esto ;)

Cantidad 1

Añadir al carrito

AÑADIR A TU LISTA DE DESEOS
AÑADIR PARA COMPARAR CORREO ELECTRÓNICO

Crea tu parque de atracciones de dinosaurios.



Autor	Antoine Bauza, Corentin Lebrat, Ludovic Maublanc, Théo Rivière
BGG	264055
Mecánica	Draft
Temática	Dinosaurios
Si buscas...	Fáciles, Familia, Para 2, Peques, Viaje
Edad	de 8 a 10 años, de 10 a 14 años, de 14 a 18 años, más de 18 años
Núm. jugadores	2, 3, 4, 5
Tiempo de juego	15 min
Medidas	200 x 200 x 55 mm
Complejidad	Medio
Editorial	Zacatrus
Contenido	• 5 Tableros de parque de dos lados • 60 Meeples de dinosaurio • 1 Dado de colocación • 1 Bolsa de tela • Reglamento
Idioma	Español
Dependencia del idioma	Nula (Sólo Instrucciones)
Envío gratis	Sí

Detalles | Comentarios

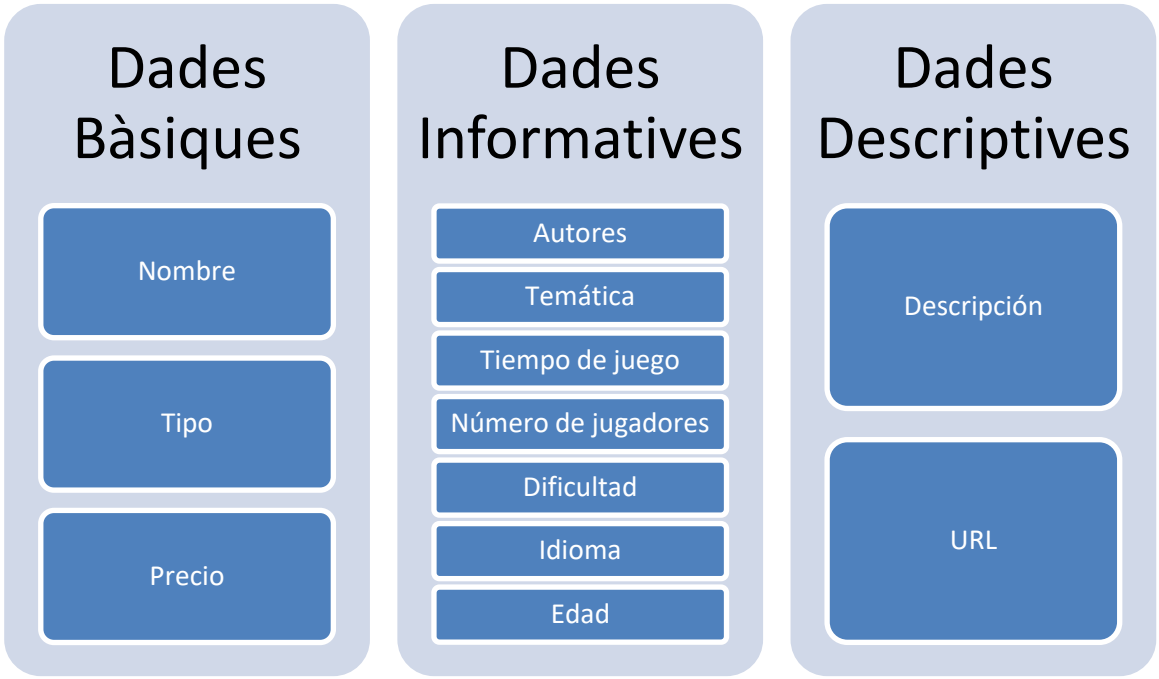
Administra tu parcela de dinosaurios en el parque temático que los científicos han conseguido crear una vez clonados estos saurios gigantes.

Escoge el dinosaurio apropiado en cada momento para obtener el mayor número de puntos posible y pasa el resto a tus rivales.

4. REPRESENTACIÓ GRÀFICA

Presentar una imatge o esquema que identifiqui el dataset visualment.

La representació gràfica permet mostrar la divisió dels tres grups de dades extretes, mitjançant la tècnica de web scraping. La subdivisió de conjunts es produeix per facilitar la decisió dels consumidors i aclarir els dubtes que es puguin tenir, ja que es consideren les dades bàsiques com aquelles que simbolitzen un primer nivell de convenciment. És a dir, si la informació proporcionada en les dades bàsiques t'atrau, l'usuari passarà a veure les dades informatives per tal de confirmar la seva elecció i si finalment, ha quedat algun dubte, l'usuari es recolzarà amb les dades descriptives per arribar a una conclusió definitiva. Doncs, com es pot observar, aquests conjunts formen una cadena de decisió prou sòlida.



A partir de les dades obtingudes en aquest projecte, un exercici de comparació entre pàgines de venta de jocs de taula, podria ser realitzat. No és el cas d'aquest algoritme pero es podria implementar d'una manera dinàmica.

5. CONTINGUT

Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

La informació s'obté instantàniament des de la pàgina web, a l'execució de l'algortime de scraping. Per tant, el programa extraurà les dades en el moment concret d'inici i pot variar si s'executa en un període temporal posterior. L'extracció s'efectua sobre els valors desitjats, que són localitzats posicionalment dins l'estructura HTML pertinent.

El dataset està format pels atributs de cadascuna de les variables següents :

- **Nombre** : Indica el títol del joc i és el valor més cercat i important pels consumidors.

  Exemple : Draftosaurus

- **Tipos**: Indica la categoria en el qual el joc és classificat.

  Categories : 'Jocs de Tauler', 'Jocs de Cartes', 'Jocs de Rol', 'Jocs de Wargamers', 'Jocs de Miniatures', 'Jocs de Dades'.

- **Precio** : Indica el valor monetari del joc i és una dada imprescindible pels usuaris.

  Exemple : 19,95 €

- **Temática** : Indica l'àrea d'influència del joc.

  Exemple : Dinosaurios.

- **Tiempo de juego** : Indica el temps per partida, del joc en questió.

  Exemple : 15 min

- **Número de jugadores** : Informa del nombre de jugadors permesos en una partida.

  Exemple : 2, 3, 4, 5

- **Dificultad** : Indica el nivell de complexitat del joc.

  Niveles : Fácil, Medio i Difícil

- **Idioma** : Indica l'idioma disponible del joc.

  Exemple : Español

- **Edad** : Indica l'edat òptima per jugar al joc.

  Exemple : de 14 a 18 años, más de 18 años

- **Descripción** : Informa de les característiques més representatives del joc.

  Exemple : Administra tu parcela de dinosaurios en el parque temático que los científicos han conseguido crear una vez clonados estos saurios gigantes. Escoge el dinosaurio apropiado en cada momento para obtener el mayor número de puntos posible y pasa el resto a tus rivales.

- **URL** : Mostra l'enllaç per accedir directament al joc.

  Exemple : https://zacatrus.es/draftosaurus.html

6. AGRAÏMENTS

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

Primer de tot, agraeixo públicament el portal de jocs Zacratus, ja que gràcies a ell s'han pogut obtenir les dades desitjades, mitjançant un algortime de web scraping. Les informacions obtingudes tenen una única finalitat educativa i acadèmica, i els processos emprats són exclusivament desenvolupats per la millora de coneixement en la tècnica d'extracció de dades.

Zacatrus ofereix un accés lliure de les dades amb finalitat no comercial. Tot i això, s'ha verificat en tot moment el fitxer robots.txt, per no causar danys en la política del lloc web.

Els recursos utilitzats són els mateixos que han estat entregats pel professorat dins l'assigantura del màster i estan detallats en l'apartat (11) d'aquest document.

7. INSPIRACIÓ

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

Actualment els clients busquen eines tecnològiques de comparació i/o extracció que els permetin optimitzar temps i diners. Una petita contribució en aquesta causa, ha estat crear una eina que permeti recuperar una llista amb les informacions més important de jocs de taula. L'objectiu de tot plegat és la recopilació de manera directa d'aquelles informacions útils pels consumidors i la rapidesa de les mateixes, sense la necessitat de navegar entre diferents pàgines per trobar aquelles dades que faran sentir segur a qualsevol consumidor de que la compra que vol realitzar serà l'apropiada.

Com ja s'ha comentat anteriorment, aquestes dades podrien formar part d'una llarga llista comparativa amb altres pàgines web de venda de jocs de taula, amb el fi de trobar aquells productes desitjats a un millor preu. Actualment estan de moda, les pàgines com *Idealo*, les quals et transmeten les diferents empreses comercialitzadores d'un producte escollit, amb la comparativa de preus i valoracions.

8. 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

La llicència escollida és la Released Under **CC BY-NC-SA 4.0** License, ja que s'ajusta a les nostres necesitats i voluntats.

A nivell informatiu, cal recordar que :

Les llicències **CC** (Creative Commons) permeten uns drets d'autor sense restriccions, tot i aixo, no pot existir cap tipus d'intencionalitat comercial ni realització de modificacions. Activitats no lucratives.

El valor **BY** permet la utilització de les dades per copiar, transmetre i utilitzar tant de manera visual com a l'origen de nous treballs, si es dóna constància de l'autor.

La categoria **SA** permet que les informacions s'utilitzin a l'origen de nous treballs, però aquests han de tenir la mateixa llicència que el treball original.

NC Permet copiar, transmetre, utilitzar i realitzar treballs que derivin d'aquest, però amb una finalitat no comercial.

L'última versió de CC Correspon a la **4.0**, i aquesta s'utilitza en gran mesura per les jurisdiccions.

Per tant, a grans trets, aquesta llicència permet compartir, copiar i redistribuir el dataset, com també poder modificar-lo. Per poder-ho fer, es precisa : Agraïr el propietari de les dades, referenciar amb un enllaç les dades d'origen i no usar les dades per a un ús comercial i lucratiu.

9. CODI

Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

El codi es presenta en un format .py, dins el següent repertori GitHub :

https://github.com/thedu7/WebScrapingUOC/code

Dividit en 2 fitxers :

- */code/main.py* : Document que inicia el procés de Web Scraping.
- */code/function.py* : Document que conté totes les funcions necessàries per realitzar l'extracció de les dades desitjades.

10. DATASET

Presentar el dataset en format CSV.

El dataset es presenta en un format .csv, dins el següent repertori GitHub :

https://github.com/thedu7/WebScrapingUOC/csv

Anomenat : *Juegos-Zacatrus.csv*

11. COMPONENTS DEL GRUP I REFERÈNCIES

La pràctica ha estat realitzada individualment per l'**Eduard Ruiz Solé**.

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.