

Tipologia i cicle de les dades

Pràctica 1: Com podem capturar les dades de la web?

<https://github.com/Jidorr/SteamScraping>

1. Context

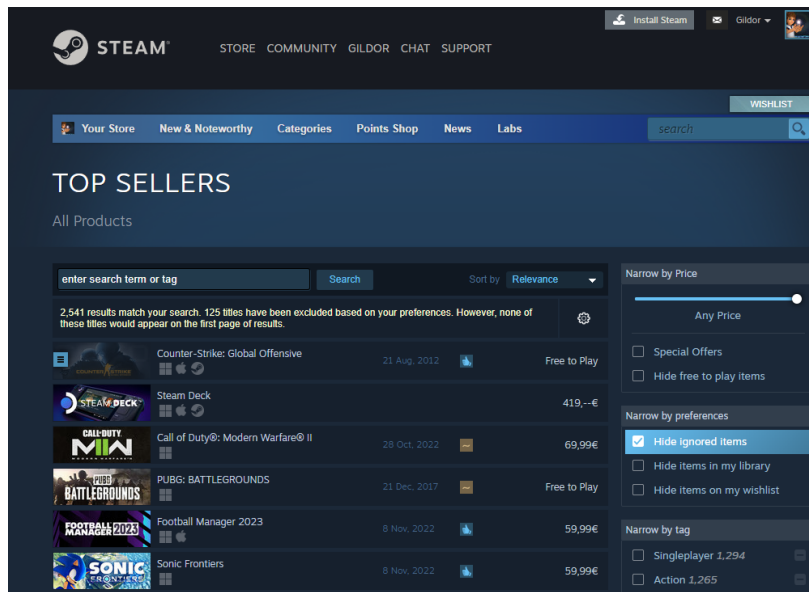
Per la realització d'aquesta pràctica elaborarem un cas pràctic orientat a identificar dades rellevants per a un projecte analític utilitzant tècniques d'extracció de dades.

Actualment, existeixen milers de llocs webs amb potencial per realitzar-hi extracció de dades, i fins i tot algunes d'elles disposen d'eines per ajudar-nos a fer-ho, com ara APIs públiques. Nosaltres, però, utilitzarem mètodes més genèrics que poden ser utilitzats en pràcticament qualsevol web que ho permeti.

Després d'un temps de reflexió, hem decidit fer scraping a la web de Steam (<https://store.steampowered.com>), un lloc web dedicat principalment a la distribució digital de videojocs.



Dins aquesta pàgina, trobem multitud d'informació sobre cadascun dels jocs existents a la plataforma; títol, categories a les qual pertany, preu, descomptes, popularitat, reviews, etc. El potencial analític és bastant elevat i podríem crear diferents projectes utilitzant les dades de la web. Pel nostre cas d'estudi, hem decidit centrar-nos en les ofertes existents dins els "top sellers". L'objectiu és tenir un dataset que ens mostri les millors ofertes dels millors jocs de la plataforma. Aquestes ofertes són rotatives, per tant aquest script és reutilitzable cada cert temps, ja que ens proporcionarà resultats diferents. La pàgina de top sellers és la següent: <https://store.steampowered.com/search/?filter=topsellers>.



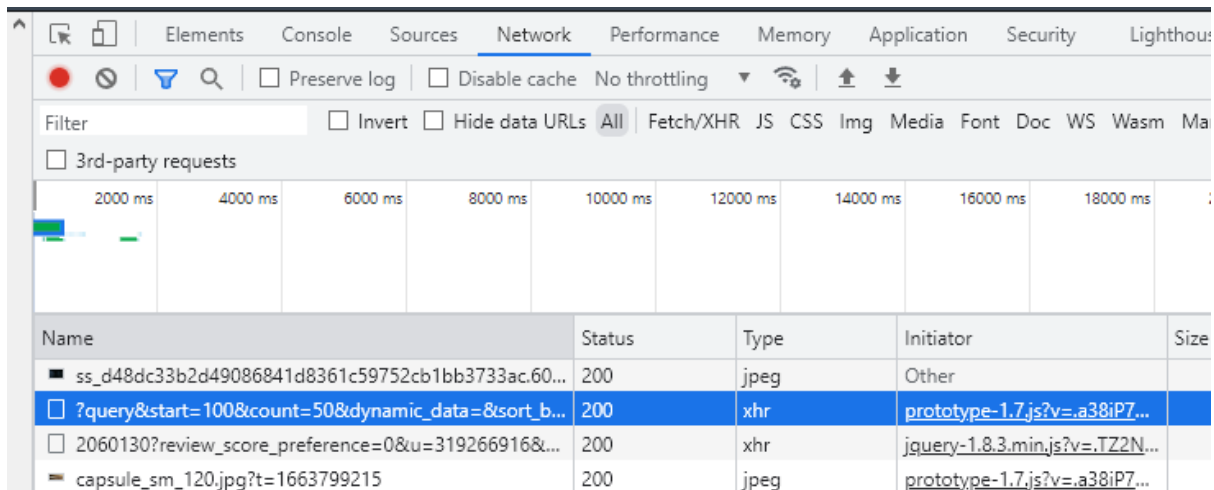
Veiem que dins aquesta mateixa pàgina trobem la majoria de l'informació necessària per a obtenir el dataset objectiu. Disposem del nom del joc i del preu. Quan el joc en qüestió té algun tipus de descompte també apareix a la casella del preu:



Com a dada extra, també obtindrem les categories de cadascun dels jocs, que no es troben en aquesta pàgina. És una dada que pot resultar molt atractiva quan realitzem l'anàlisi del dataset, ja que ens permetrà comparar descomptes entre categories, veure quin és el joc més barat de cadascuna, etc.

La pàgina de top sellers és d'scroll infinit, això significa que per molt que baixem a baix de tot de la pàgina, no pararem de carregar nous elements. Això pot arribar a ser un problema, ja que a priori ens limita els elements que podem scrapejar de manera

directa. Hem realitzat un petit anàlisi del lloc web per veure com es comporta a l'hora de fer les requests per obtenir nous elements. Amb l'ajuda de l'eina d'inspeccionar pàgina descobrim que cada vegada que es carrega una nova pàgina, apareix una nova ordre xhr amb la query que s'envia al servidor d'Steam per obtenir aquesta nova pàgina.



Name	Status	Type	Initiator	Size
ss_d48dc33b2d49086841d8361c59752cb1bb3733ac.60...	200	jpeg	Other	
?query&start=100&count=50&dynamic_data=&sort_b...	200	xhr	prototype-1.7.js?v=a38iP7...	
2060130?review_score_preference=0&u=319266916&...	200	xhr	jquery-1.8.3.min.js?v=TZ2N...	
capsule_sm_120.jpg?t=1663799215	200	jpeg	prototype-1.7.js?v=a38iP7...	

Si analitzem la URL que retorna aquesta request, veiem com funciona la crida interna i com la podem modificar:

https://store.steampowered.com/search/results/?query&start=100&count=50&dynamic_data=&sort_by=_ASC&snr=1_7_7_7000_7&filter=topsellers&infinite=1

La part interessant és la marcada en vermell, que diu al servidor en quin element ha de començar i quants elements ha de retornar. En el cas concret d'aquesta URL, és la crida de la segona pàgina de l'scroll infinit, ja que quan obrim la pàgina tenim start = 0. Canviant aquest paràmetre (en increments de 50) podem obtenir tants elements com desitgem.

L'script creat té com a paràmetre el número de pàgines que volem scrapejar (cada cop que es carrega una nova pàgina obtenim 50 nous elements).

2. Títol

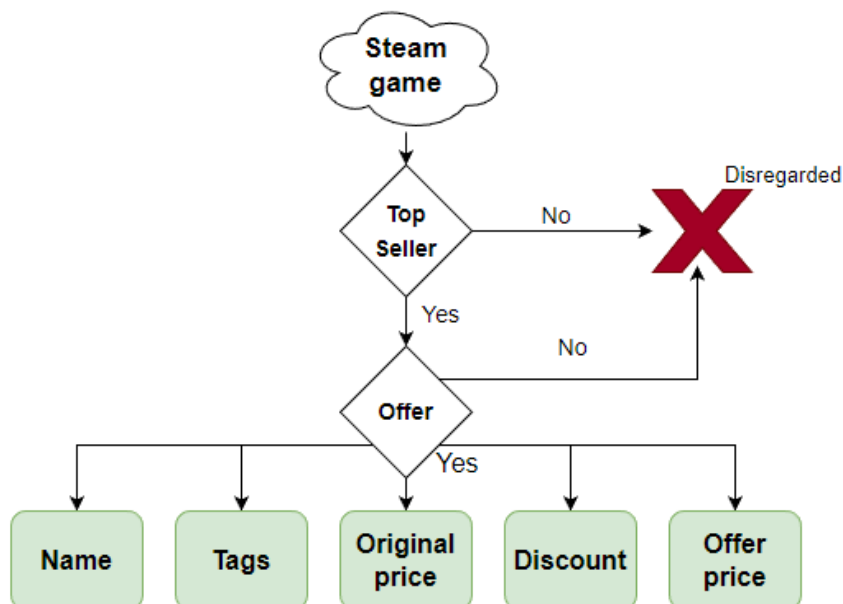
Ja que el que ens interessa és obtenir els descomptes dels millors jocs d'Steam, el títol escollit pel dataset resultant és "Top Steam Offers".

3. Descripció / contingut del dataset

A partir de les dades de la pàgina, s'han extret els articles als que se'ls havia aplicat algun tipus de descompte. Aquesta informació s'ha emmagatzemat en un dataframe de cinc variables:

- Name **str**: nom del videojoc
- Tags **list**: nom dels tres tags principals del videojoc
- Original price **float**: preu inicial (sense descompte)
- Discount **string**: descompte aplicat (en %)
- Offer price **float**: preu final (després del descompte)

4. Representació gràfica



5. Propietari

Les dades obtingudes s'han extret de la pàgina web de Steam. Per tant, el propietari del conjunt de dades resultant és Valve Corporation, que és l'empresa que ha desenvolupat la plataforma.

Aquest anàlisi no s'ha basat pròpiament en cap d'anterior, però sí que s'han trobat anàlisis similars que poden justificar el que es presenta en aquesta memòria.

Per entrar en matèria, ens trobem amb un article publicat a la pàgina web de Hardzone per Jose Luís Sanz sota el títol de *Nadie quiere pagar por jugar, algo que deja claro el Top de más jugados en Steam*¹. En aquesta notícia, s'analitzen els jocs més venuts i jugats de la pàgina de top sellers de Steam i s'arriba a la conclusió que un percentatge important dels que es troben a les primeres posicions són gratuïts. Aquesta dada suggereix que el preu és una variable de gran importància per molts jugadors.

D'altra banda, també es troben articles com el de *Steam Summer Sale 2022 - Game Predictions and Discounts*², publicat a la web de CCL, que, basant-se en campanyes de rebaixes anteriors i en l'historial de preus dels jocs, intenta predir els articles als que se'ls aplicarà algun tipus de descompte i amb quin percentatge a la campanya de rebaixes de l'estiu de 2022. El resultat és una llista que conté aquestes dades. El fet de realitzar un anàlisi com aquest també contribueix a pensar que els descomptes juguen un paper important en la presa de decisió.

En termes de conjunts de dades, s'han trobat nombrosos datasets publicats a la web de Statista que recullen i analitzen diferents aspectes econòmics, com ara els ingressos generats a Steam³ o els videojocs que més còpies han venut⁴. En aquests casos, les conclusions afecten de manera directa a l'empresa, però la base dels anàlisis i les dades que es fan servir provenen de la mateixa font que estem treballant.

En conclusió, tots aquests exemples mostren la rellevància de les variables escollides, tant per usuaris com per la pròpia empresa. Així mateix, justifiquen que es reuneixin les millors ofertes de Steam en un sol dataset.

En relació als principis ètics i legals, s'ha escollit una font d'informació que és lliure i no conté dades confidencials. Per tant, no té restriccions d'ús, d'explotació ni de publicació, assegurant una correcta adhesió als principis mencionats.

6. Inspiració

Quan un usuari té la intenció de adquirir un videojoc, a banda d'avaluar si coincidirà amb les seves preferències en funció dels 'tags' que tingui assignats, es fixa en el preu. Si aquest és molt elevat, normalment els descomptes juguen un paper molt rellevant en prendre la decisió final.

Per això, s'ha creat un conjunt de dades que reuneix les variables que més informació proporcionen i en les que més es fixen els clients de Steam a l'hora d'adquirir un article.

Com s'ha mencionat a l'apartat anterior, no s'ha trobat una font que reculli les mateixes dades amb el mateix objectiu. Per aquesta raó i perquè ens ha semblat un dataset rellevant i molt útil, hem decidit que seria un bon punt de partida crear-lo.

7. Llicència

La llicència del dataset és 'Released Under CC0: Public Domain License'. La informació és lliure i de ús comercial i no conté dades personals dels usuaris.



8. Codi

El codi es troba en un únic script anomenat `steamScraper.py`, situat dins la carpeta `/source` del projecte.

El primer pas és l'importació de llibreries. Pel nostre cas d'estudi hem utilitzat les següents:

- `Requests`: ens ajuda a realitzar les peticions HTML de manera senzilla.
- `BeautifulSoup`: ens permet parsejar les pàgines HTML obtingudes amb l'anterior llibreria.
- `Pandas`: utilitzada per la creació del dataset final.
- `sys`: ens permet realitzar operacions de sistema dins l'script.
- `time`: ens permet comptar el temps d'execució de l'script.
- `date`: utilitzada per obtenir la data en la qual s'executa l'script.

El primer que fem és guardar el temps inicial d'execució del programa. Això ens permetrà restar-lo al temps final per tal de saber quants segons ha tardat a executar-se tot el programa.

A continuació, prenem el primer argument passat a l'execució del programa, que ha de ser un número no més gran a 10. Aquest número (`numPagines`) determinarà el número de pàgines que s'scrapejaran i posarà un límit a l'scroll infinit. Si no es passa cap valor com a argument, aquest és invàlid o bé més gran a 10, es prendrà com a default `numPagines = 2`.

A partir d'aquí, tot el programa es troba dins un loop, que és dependent del paràmetre `numPagines`. Per cada valor extra d'aquest paràmetre, obtindrem una nova pàgina HTML amb 50 resultats més.

El que fem a continuació és obtenir la pàgina HTML de la url utilitzant `requests`. Ja que aquesta pàgina es troba dins d'un json, hem d'agafar la clau que ens interessa (`results_html`) i parsejar el resultat mitjançant `BeautifulSoup`.

Mitjançant el mètode `.find_all` que ens proporciona `BeautifulSoup` obtenim una llista de tags HTML que conté tots els jocs de la pàgina. Inspeccionant la pàgina veiem que tan sols existeix un sol tipus de tag "a", per la qual cosa és fàcilment obtenible. Guardem aquesta llista en una variable anomenada `all_games`.

El proper pas és iterar amb un loop aquesta llista, ja que de cada joc volem prendre'n un seguit de dades. Per obtenir el nom i el preu simplement hem de buscar segons els tags que ho contenen. Utilitzem altre vegada el mètode `.find` de BeautifulSoup per fer-ho. Utilitzem un `try except` per quedar-nos només amb els jocs que presenten un descompte. Els que no en tenen, es troben dins un tag de diferent nom, per tant amb la sentència `continue` els ignorem. Separem el preu inicial i el preu d'oferta, així com el % de descompte, que es troba dins un altre tag d'html.

També volem prendre les 3 categories principals de cadascun dels jocs, però això no ho trobem dins d'aquesta mateixa pàgina. El que fem és obtenir la url de la pàgina de cada joc per també scrapejar-la i obtenir els tags que ens interessin. Utilitzem un `try except`, ja que hi ha jocs que no estan classificats i no tenen categoria. En aquests casos, simplement inserim "untagged" a la llista de categories.

Deixem passar 5 segons entre pàgina i pàgina del loop principal, per tal de no sobrecarregar el servidor i evitar un possible ban d'ip per realitzar masses peticions. Un cop tenim totes les dades que ens interessin, creem un diccionari que posteriorment convertim a un dataframe de pandas. Anomenem aquest resultat amb la data d'extracció, que ens pot ser útil per comparar diferents resultats en diferents períodes de temps. El csv resultant és guardat dins la carpeta `/dataset`.

Per acabar, fem un print del temps que ha tardat a executar tot el programa.

9. Dataset

El dataset final s'ha publicat a Zenodo amb DOI <https://doi.org/10.5281/zenodo.7311975>.

Es troba també dins la carpeta `/dataset` del projecte.

10. Video

https://drive.google.com/file/d/1CQrFtMRm-kp0oZ81-PPbbniwD2ul7Zp_/view?usp=share_link

Taula de contribucions

Contribucions	Signatura
Investigació prèvia	J.S.A., M.V.D.
Redacció de les respostes	J.S.A., M.V.D.
Desenvolupament del codi	J.S.A.
Participació al video	J.S.A., M.V.D.

Referències bibliogràfiques

1. Sanz, Jose Luís. (10 de novembre, 2022). *Nadie quiere pagar por jugar, algo que deja claro el Top de más jugados en Steam*. Hardzone. Consultat l'11 de novembre de 2022.
<https://hardzone.es/noticias/juegos/steam-top-mas-jugados/>
2. Byrne, Mark. (5 d'abril, 2022) *Steam Summer Sale 2022 - Game Predictions and Discounts*. CCL. Consultat l'11 de novembre de 2022.
<https://www.cclonline.com/article/2089/News/CCL-Gaming-PCs/Steam-Summer-Sale-2022-Game-Predictions-and-Discounts/>
3. Statista. (14 de setembre, 2022). *Revenue generated by game sales on Steam from 2020 to 2027 (in million U.S. dollars)*. Statista. Consultat l'11 de novembre de 2022.
<https://www.statista.com/statistics/547025/steam-game-sales-revenue/>.
4. Medium (galyonk.in). (4 d'abril, 2018). *Leading paid game titles on Steam in 2017, by number of units sold (in 1,000s)*. Statista. Consultat l'11 de novembre de 2022.
<https://www.statista.com/statistics/499390/leading-pc-steam-games-by-unit-sales/>.