Pràctica 1

Tipologia i cicle de vida de les dades

Abril 2023
Pau Bernabé
Sergi Crespi

1. Context

S'han avaluat diferents pàgines web de les quals es pogués obtenir dades mitjançant web scraping. A causa de l'interès dels integrants del grup de la pràctica pel món dels videojocs, es va plantejar orientar l'extracció de dades en aquest àmbit, i es van avaluar diferents pàgines, tals com Twitch o Steam. Al final, es va optar per a la popular plataforma de venda de videojocs com a objectiu de web scraping.

Aconseguir dades del líder del sector de venda de videojocs en format digital pot ser interessant per diversos motius, des d'un punt de vista de preguntes de negoci, conèixer en detall els preus, com van evolucionant, la quantitat de videojocs, poder detectar ofertes... així com des del punt de vista de si es treballa en una companyia rival del sector, tal com Epic Games, pot ser interessant per a fer anàlisi de mercats.

Tot i que Steam ofereix una API, l'objectiu d'aquest pot ser lleugerament diferent, més orientada cap a equips de màrqueting o donar servei a llocs web de tercers, per exemple. Així i tot, part de la informació que hi ha a la web de Steam pot no estar disponible en aquesta API, i pot ser interessant procedir amb tècniques de web scraping.

L'adreça del lloc web d'Steam és: https://store.steampowered.com/.
Concretament, se centrarà el web scraping al seu catàleg a: https://store.steampowered.com/search/?category1=998&page=1&ndl=1, on el paràmetre de la URL fa que només es mostrin els productes de la categoria videojoc.

2. Títol

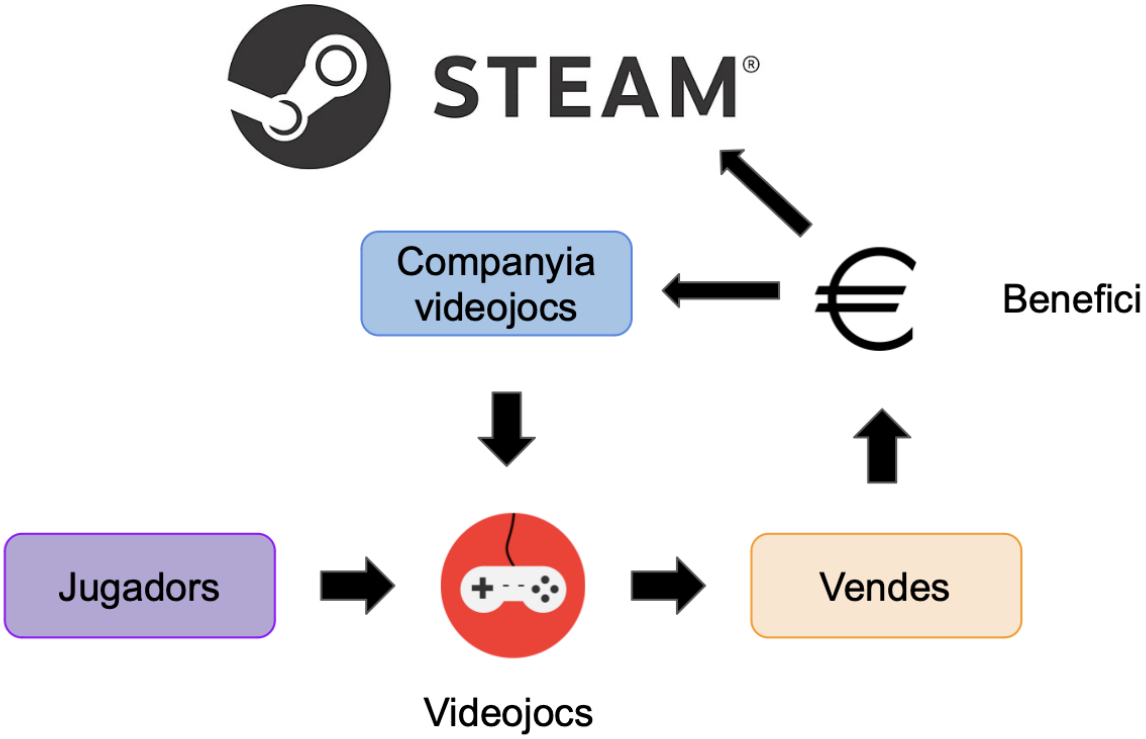
El títol del dataset és "steam_games". De cara a publicar-lo a Zenodo, podria tenir un títol com: Sample of Steam catalog.

3. Descripció del Dataset

Tal i com es dóna a entendre al títol, es tracta d'una mostra del catàleg de jocs d'steam obtingut a data del 11/04/2023. Es tracta, doncs d'una foto de part del catàleg de jocs a aquesta data, i el dataset no ha passat preprocessament de les dades, o de neteja, que sigui rellevant, així doncs pots haver-hi dades en formats, o presentar alguna dada que no estigui present, ja que no hi era en el moment de l'extracció (per exemple, jocs que no tenen valor de "review" ja que encara no s'han publicat o no tenen puntuació, jocs sense "developers" ja que tampoc estava present a l'hora de l'extracció de la informació). A destacar també que alguns camps estan format per un array d'strings.
A apartats posteriors es donaran més detalls del camps i del seu contingut, però es destaca que l'arxiu generat és un arxiu en format CSV per a facilitar el seu posterior tractament, amb el separador ";".

4. Representació gràfica

Esquema que representa el dataset i el projecte escollit:



5. Contingut

El contingut del set de dades és una foto estàtica del dia 11/04/2023, i està format pels següents camps:

Columna	Tipus	Descripció
game_id	integer	Identificador del joc
title	string	Nom del joc
price	float	Preu final del joc
discount	integer	Descompte del joc, en cas d'haver-ne

review_score	integer	Puntuació del joc, en tant per cent
number_reviews	integer	Nombre de reviews
released	date	Data de publicació del joc
platforms	array	Noms de les plataformes suportades pel joc
developers	array	Noms dels desenvolupadors del joc
genres	array	Categories descriptives del joc

6. Propietari

De cara a saber el propietari de l'adreça web https://store.steampowered.com s'ha executat el següent codi:

```
import whois
w = whois.whois("https://store.steampowered.com")
print(w)
```

I, la sortida resultant de l'execució de l'script és:


```
{
  "domain_name": "STEAMPOWERED.COM",
  "registrar": "Network Solutions, LLC",
  "whois_server": "whois.networksolutions.com",
  "referral_url": null,
  "updated_date": [
    "2022-06-10 08:11:38",
    "2022-08-20 05:56:15"
  ],
  "creation_date": "2000-08-09 16:03:56",
  "expiration_date": "2024-08-09 16:03:56",
  "name_servers": [
    "A1-164.AKAM.NET",
    "A11-67.AKAM.NET",
    "A24-64.AKAM.NET",
    "A26-65.AKAM.NET",
    "A8-66.AKAM.NET",
    "A9-67.AKAM.NET"
  ],
  "status": "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
  "emails": [
    "abuse@web.com",
    "itgext@valvesoftware.com",
    "domain.operations@web.com"
  ],
  "dnssec": "unsigned",
  "name": "Valve Corporation",
  "org": "Valve Corporation",
  "address": "10400 NE 4TH ST STE 1400",
  "city": "BELLEVUE",
  "state": "WA",
  "registrant_postal_code": "98004-5174",
  "country": "US"
}
```

Així doncs, es pot veure que el propietari de l'adreça web és Valve Corporation, amb seu a Bellevue, Washington, Estats Units.
Existeixen altres anàlisis anteriors que han obtingut informació de Steam, tals com:

- https://www.kaggle.com/datasets/trolukovich/steam-games-complete-dataset
- https://www.kaggle.com/datasets/nikdavis/steam-store-games
- https://www.kaggle.com/datasets/fronkongames/steam-games-dataset

7. Inspiració

Aquest set de dades pot ser interessant d'analitzar ja que es pot tenir una radiografia del mercat dels videojocs d'ordinador, una indústria que mou mil·lions d'euros, i, si se seguissin obtenint fotos de diferents dies, es podria comprovar l'evolució d'aquest mercat. Tanmateix, pot ser interessant per a companyies de la competència, tals com Epic Games o d'altres, fer un seguiment del catàleg disponible del líder del mercat de cara a poder veure quotes de *market share* o poder definir estratègies de negoci adients.
Algunes preguntes que es poden respondre amba quest set de dades, són:

- Quants jocs hi ha disponibles per a comprar-se a Steam?
- Quin és el descompte més gran que ofereix Steam?
- Quants jocs amb descompte hi ha?
- Quina és la puntuació mitjana de valoracions de jocs de Steam?
- Quin és el top 10 de millor jocs, segons valoracions, de Steam?
- Quin és el top 10 de pitjors jocs, segons valoracions, de Steam?
- Quina és la distribució de jocs segons plataforma?
- Quina és la puntuació mitjana per plataforma?

Tot i que s'ha vist que ja hi ha datasets anteriors del catàleg de jocs de Steam, els que s'han trobat no permeten contestar totes les preguntes que s'exposen anteriorment, s'hauria de recórrer a diferents, i tots no tenen tota la informació o no està formatejada per a que sigui senzill contestar-les, s'ha produit el web scraping de manera que contestar aquestes preguntes sigui més senzill.

8. Llicència

La llicència que s'ha pensat fer servir és la CC BY-NC-SA ja que:

- BY: permet l'ús, sempre citant als autors.
- NC: sense que se'n permetin finalitats comercials.
- SA: possibles obres derivades estiguin sota la mateixa llicència.

D’aquesta manera, s’evita que es pogués fer un ús comercial de les dades extretes, tot i que les seves aplicacions comercials podrien ser limitades, s’escull aquesta llicència per a preservar *l’ownership* de les dades per part de la Valve Corporation.

9. Codi

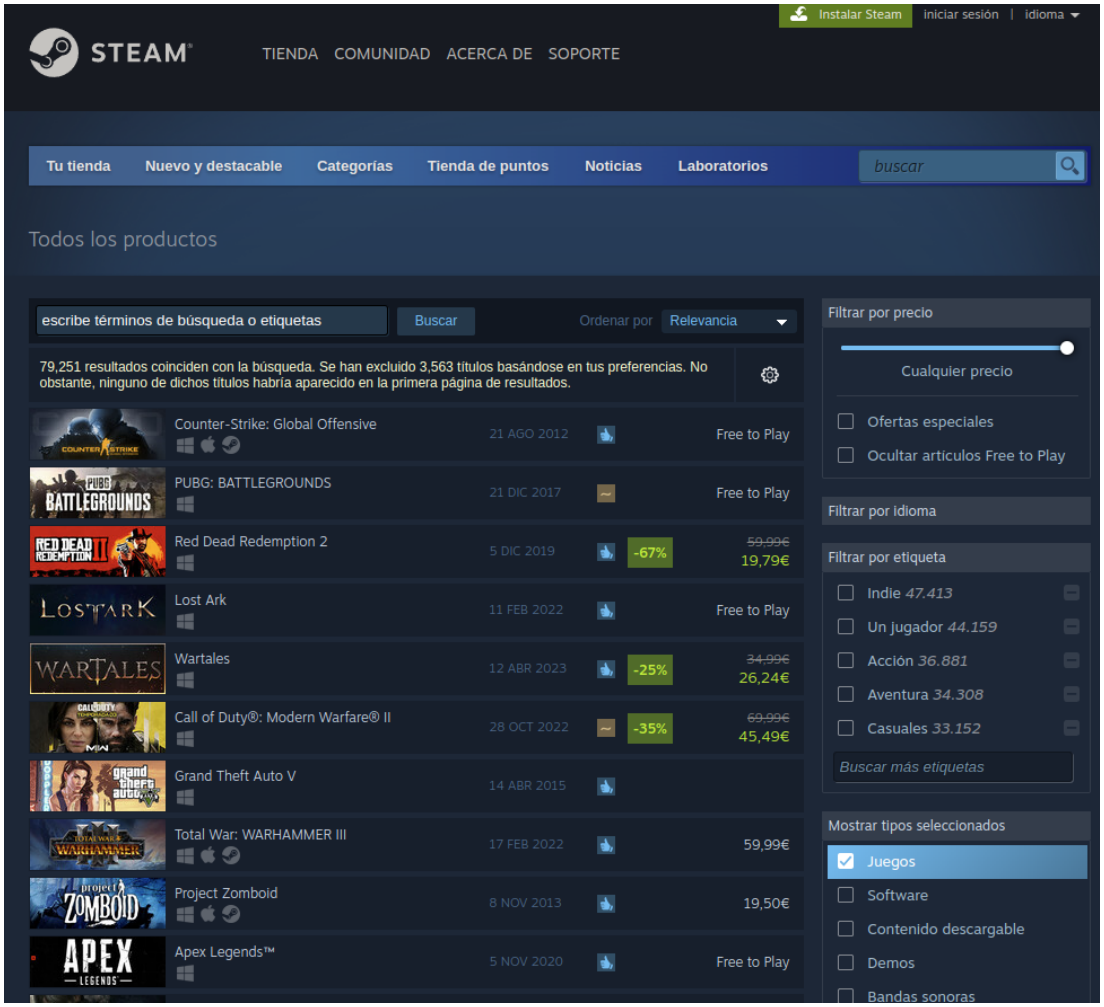
El codi es pot trobar al repositori: https://github.com/sercres/pra1_tipologia
Per a dur a terme la recol·lecció de les dades s’ha creat un *script* escrit amb Python. Aquest procés s’encarrega d’automatitzar la navegació dins de la pàgina web i de recolectar les dades requerides per la creació del conjunt de dades.

Per al desenvolupament del codi s’han emprat els següents moduls que inclou Python:

- Scrapy
- Selenium
- Llibreries de python nadiu

Recol·lecció de les dades amb Scrapy

Posant un cas d'exemple, en la següent figura es veu una de les pàgines d'on s'obtindran part de les dades de cadascun dels videojocs.



Per tal d'obtenir les dades pertinents a cada joc, cal inspeccionar la plantilla de la pàgina web, estructurada amb html. Per fer-ho, cal emprar les eines de desenvolupador que proporcionen la majoria dels navegadors actuals, en aquest cas s'utilitzarà Chrome. S'exemplifica amb el títol de cadascun dels jocs.

Seleccionant l'opció **Inspeccionar element** sobre el títol se'ns dirigeix cap a l'element corresponent del codi:

```
▼<a href="https://store.steampowered.com/app/730/CounterStrike_Global_Offensive/?snr=1_7_7_230_150_1" data-ds-appid="730" data-ds-itemkey="App_730" data-ds-tagids="[1663,1774,3859,3878,19,5711,5055]" data-ds-descids="[2,5]" data-ds-crtrids="[4]" onmouseover="GameHover( this, event, 'global_hover', {"type":"app","id":730,"public":1,"v6":1} );" onmouseout="HideGameHover( this, event, 'global_hover' )" class="search_result_row ds_collapse_flag " data-search-page="1" data-gpnav="item" data-ds-steam-deck-compat-handled="true">
 ▶<div class="col search_capsule">…</div>
 ▼<div class="responsive_search_name_combined">
  ▼<div class="col search_name ellipsis">
     <span class="title">Counter-Strike: Global Offensive</span> == $0
    ▶<div>…</div>
    </div>
    <div class="col search_released responsive_secondrow">21 AGO 2012</div>
   ▶<div class="col search_reviewscore responsive_secondrow">…</div>
   ▶<div class="col search_price_discount_combined responsive_secondrow" data-price-final="1419">…</div>
  </div>
  <div style="clear: left;"></div>
 </a>
```

Per tal que **Scrapy** pugui identificar l'element es fa mitjançant els **Selectors** d'Scrapy. Els Selectors s'anomenen així perquè seleccionen els elements de la plantilla html que s'especifiquen mitjançant l'**xpath** o les classes **css**. En el nostre projecte s'escull l'opció d'especificar les classes css per extreure els elements que les implementin.

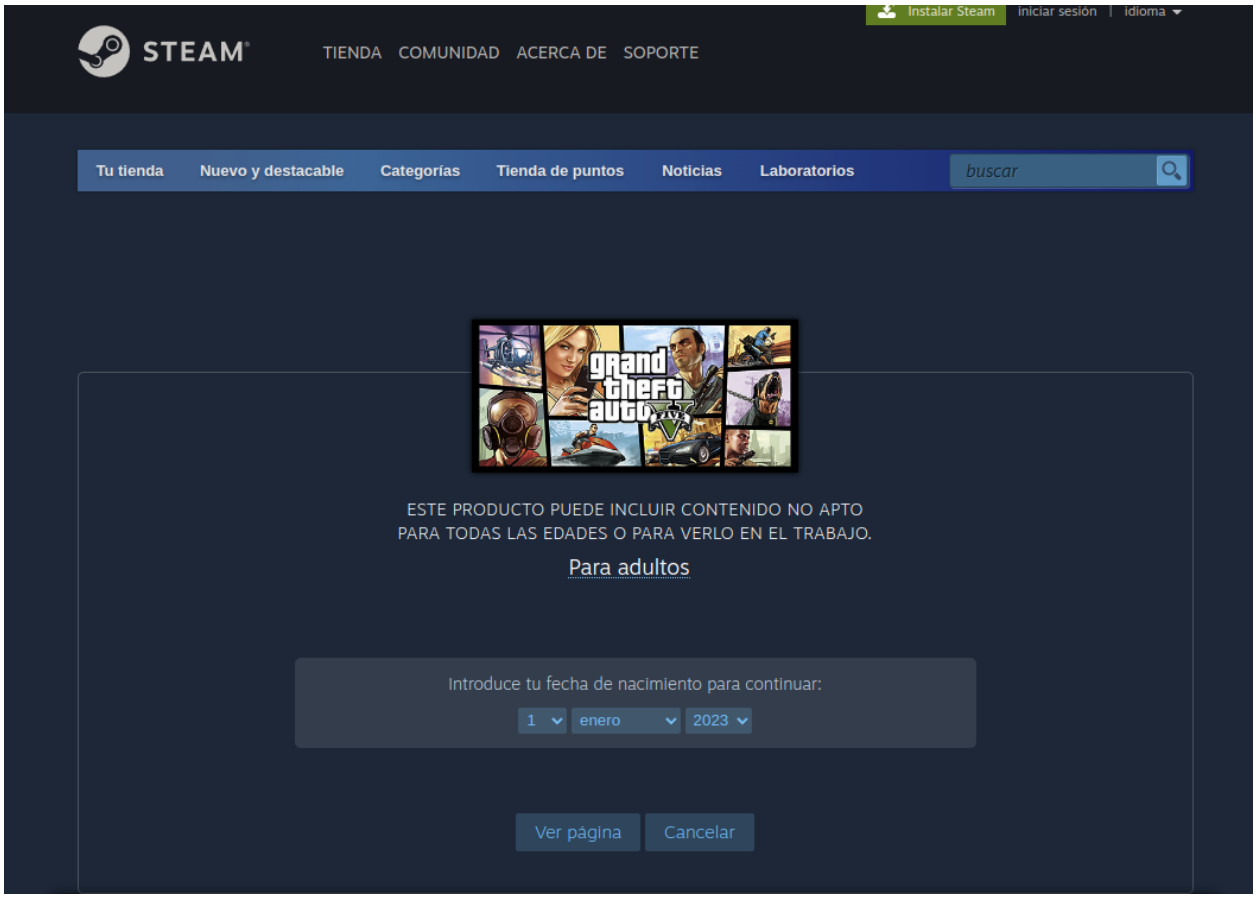
Com es pot observar, tenim la línia de color blau el que representa l'element del títol del joc en qüestió, que implementa la classe **title**. Per filar més prim i que la cerca sigui més acurada, s'inclou la classe de l'element pare, en aquest cas **search_result_row**.
La classe **search_result_row** és comuna a tots els elements de la llista, és a dir, cadascun dels elements que el selector trobi amb aquesta classe representarà un videojoc. Per aquesta raó, es pot iterar per tots els elements que es trobin amb aquesta classe i extreure'n el títol mencionat anteriorment.

```
for game in response.css('.search_result_row'):
            title = game.css('.title::text').get().strip()
```

Utilització de Selenium

L'ús de Selenium ha sigut de manera auxiliar per algunes funcions que Scrapy no cobria, com per exemple omplir formularis fàcilment i simular accions amb els elements, com podria ser fer click a un botó. Aquest cas es dona quan el crawler vol accedir a la pàgina d'un joc restringit per

majors d'edat i cal que ompli un formulari per tal de verificar que no és menor. Un cop s'ha omplert, fa clic al botó de "Veure pàgina" i redirigeix a la pàgina del joc en qüestió.



Addicionalment, s'ha afegit Selenium per veure en directe l'extracció de dades del crawler, creiem que és una funcionalitat útil a l'hora de depurar el procés d'extracció de dades.

Flux de la recolecció de dades

El flux de la recolecció de les dades segueix un ordre.

1. El crawler es situa a la pàgina de cerca, on es troba una llista de 50 jocs aproximadament.
2. Seguidament, itera per la llista on per cada entrada obté les dades que es troben al mateix element, com pot ser el títol, preu, ressenyes, etc.
3. Un cop té aquesta informació d'un joc, entra dins la pàgina detallada del joc per extreure informació addicional. En aquest cas, es pot donar l'escenari de completar el formulari per la restricció d'edat.
4. Extreu la informació addicional del videjoc.

5. Es torna al pas 1. En el cas que no hi hagin mes elements (s'ha processat tota la pàgina) es passarà a la següent pàgina.

Observacions i dificultats

Realitzar un procés automàtic de recolecció de dades és a vegades complex, ja que, cal controlar múltiples aspectes que no es tenen en compte quan un usuari normal navega per una pàgina web. En aquest projecte s'han trobat alguns aspectes que ha calgut controlar:

- Gestió del temps de càrrega de la pàgina web: Al ser una pàgina que ven videojocs, conté una gran quantitat de contigut multimèdia com imatges i videos. Aquest fet causa un temps de càrrega major que en la majoria de llocs web. La solució ha sigut controlar-ho mitjançant esperes d'alguns segons entre pàgina i pàgina.
- Prevenció de possibles bloquejos al crawler: S'ha especificat el nom del User Agent per tal s'evitar aquest tipus de casos.
- Gestió de formularis: Com s'ha mencionat anteriorment, s'empra Selenium per gestionar els formularis. La raó és que Scrapy no proporciona les eines necessàries (o no les mes adequades) per gestionar aquest tipus de casuística.
- Gestió d'excepcions: Alguns elements no es troben presents i cal tractar aquestes casuístiques. En el cas que no es trobi un camp, es deixa buit.

10. Dataset

Link al dataset a Zenodo: https://doi.org/10.5281/zenodo.7832930

11. Vídeo

Link del vídeo:
https://drive.google.com/file/d/1DF-pr-6KsDPYpcjUgYr9JvKKvUlBPQrG/view?usp=share_link

Contribucions	Signatura
Investigació prèvia	P. B., S. C.
Redacció de les respostes	P. B., S. C.
Desenvolupament del codi	P. B., S. C.
Participació al vídeo	P. B., S. C.