Pràctica 1 - M2.951_20221

Gerard Ramos Gambús: gambusrg@uoc.edu
Jaume Verges Mont: jvergesmo@uoc.edu
22/11/2022

Contenido

Context

Per realitzar la pràctica de web scraping s'ha triat una web molt comuna entre la gent apassionada als videojocs. Aquesta web és Instant Gaming.
El que es vol aconseguir és poder fer una extracció d'una gran part dels jocs d'aquesta web tenint en compte els propis títols dels jocs, preus, descomptes, desenvolupadors i la data de llançament.
Per poder realitzar aquesta extracció s'han tingut vàries consideracions prèvies. En un principi es va intentar fer una extracció sobre la web de PcComponentes, però hi havia bastantes barreres que va resultar complicat passar, com ara la captcha o els bloquejos de l'user agent. En un primer cas el bloqueig de l'user agents es va resoldre, però la captcha va ser més complicat, per tant, es va triar una web similar la qual no disposés de tal barrera. D'altra banda, analitzant l'arxiu robots.txt de cada una de les webs, es pot veure que el de PcComponentes conté moltes més restriccions sobre user agents que el de Instant Gaming.
Observant el sitemap del lloc web, es veu que la web triada disposa de menys enllaços a l'hora de fer una navegació autònoma per tant es va decidir directament fer l'extracció sobre els jocs.

La idea de fer el web scraping sobre aquesta web consisteix a accedir a l'apartat de tendències, on es mostren tots els jocs dels que es disposa, ordenats per tendència. Un cop es passi aquest punt es recorre joc per joc i pàgina per pàgina, extraient la informació pertinent de cada joc.
Instant gaming: https://www.instant-gaming.com/es/

Títol

El títol del dataset és **products**, ja que a la web a part de jocs també hi ha productes que complementen els jocs.

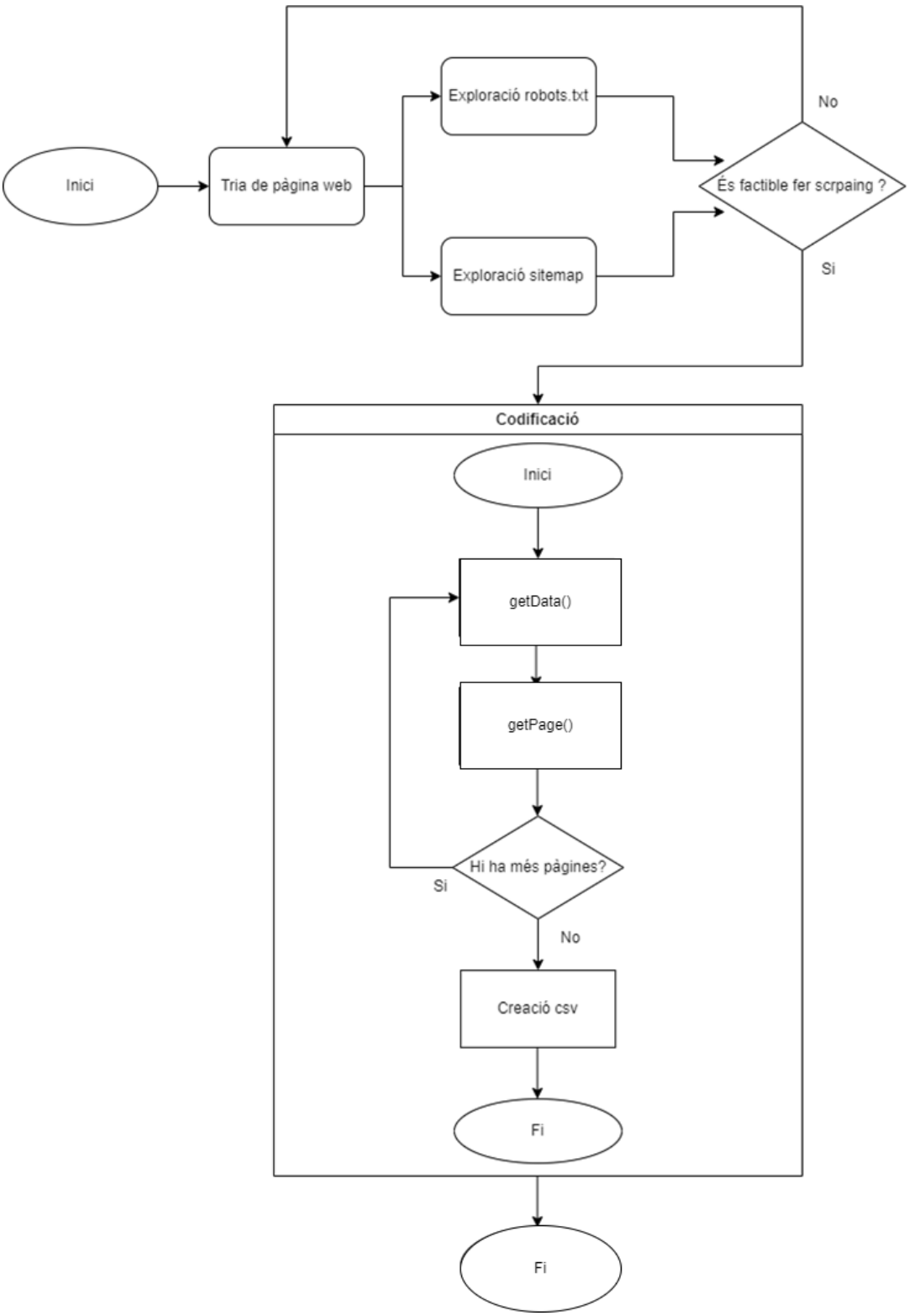
Descripció del data set

Com be he comentat a l'apartat 1, les extraccions que s'han realitzat han estat de jocs. Aquestes extraccions s'han emmagatzemat en llistes que posteriorment han estat introduïdes en un diccionari. Aquest, s'ha transformat en un data set.

Les dades que s'han extret son totes de tipus string, ja que en l'esquelet html estan emmagatzemades com a text.

El data set conté un conjunt ampli d'observacions. Concretament, a la web hi ha 60 observacions per pàgina i s'ha decidit fer una extracció de 100 pàgines de productes. Cada una de les observacions conté: Títol del producte, preu sense descompte, descompte, preu amb descompte aplicat, desenvolupador del producte, data de sortida del joc.

Representació gràfica



Contingut

Jocs:

- Títol del producte: Marvel's Spider-Man: Miles Morales
- Preu del producte sense descompte: 50€
- Descompte del producte: -31%
- Preu amb el descompte aplicat: 34.49€
- Desenvolupador del producte: Insomniac GamesNixxes Software
- Data de sortida del joc: 18 noviembre 2022

I també de complements de jocs:

- Títol del producte: FIFA 23: 2800 FUT Points
- Preu del producte sense descompte: 25€
- Descompte del producte: -20%
- Preu amb el descompte aplicat: 19.99€
- Desenvolupador del producte: EA Canada & EA Romania
- Data de sortida del joc: 1 octubre 2022

Com bé he comentat abans, les dades han estat extretes com a text, ja que a l'arbre html tot tipus de caràcter que es veu per pantalla està com a string.

Propietari

Domain Name: INSTANT-GAMING.COM
Registry Domain ID: 280093430_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.godaddy.com
Registrar URL: https://www.godaddy.com
Updated Date: 2020-12-02T00:12:49Z
Creation Date: 2005-12-09T15:26:16Z
Registrar Registration Expiration Date: 2025-12-02T06:59:59Z
Registrar: GoDaddy.com, LLC
Registrar IANA ID: 146
Registrar Abuse Contact Email: abuse@godaddy.com
Registrar Abuse Contact Phone: +1.4806242505
Domain Status: clientTransferProhibited https://icann.org/epp#clientTransferProhibited
Domain Status: clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited
Domain Status: clientRenewProhibited https://icann.org/epp#clientRenewProhibited
Domain Status: clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited
Registry Registrant ID: Not Available From Registry
Registrant Name: Registration Private
Registrant Organization: Domains By Proxy, LLC
Registrant Street: DomainsByProxy.com
Registrant Street: 2155 E Warner Rd
Registrant City: Tempe

Registrant State/Province: Arizona
Registrant Postal Code: 85284
Registrant Country: US
Registrant Phone: +1.4806242599
Registrant Phone Ext:
Registrant Fax: +1.4806242598
Registrant Fax Ext:
Registrant Email: Select Contact Domain Holder link at
https://www.godaddy.com/whois/results.aspx?domain=INSTANT-GAMING.COM
Registry Admin ID: Not Available From Registry
Admin Name: Registration Private
Admin Organization: Domains By Proxy, LLC
Admin Street: DomainsByProxy.com
Admin Street: 2155 E Warner Rd
Admin City: Tempe
Admin State/Province: Arizona
Admin Postal Code: 85284
Admin Country: US
Admin Phone: +1.4806242599
Admin Phone Ext:
Admin Fax: +1.4806242598
Admin Fax Ext:
Admin Email: Select Contact Domain Holder link at
https://www.godaddy.com/whois/results.aspx?domain=INSTANT-GAMING.COM
Registry Tech ID: Not Available From Registry
Tech Name: Registration Private
Tech Organization: Domains By Proxy, LLC
Tech Street: DomainsByProxy.com
Tech Street: 2155 E Warner Rd
Tech City: Tempe
Tech State/Province: Arizona
Tech Postal Code: 85284
Tech Country: US
Tech Phone: +1.4806242599
Tech Phone Ext:
Tech Fax: +1.4806242598
Tech Fax Ext:
Tech Email: Select Contact Domain Holder link at
https://www.godaddy.com/whois/results.aspx?domain=INSTANT-GAMING.COM
Name Server: BOB.NS.CLOUDFLARE.COM
Name Server: GAIL.NS.CLOUDFLARE.COM
DNSSEC: signedDelegation

Com es pot veure moltes de les dades son privades.

Per no trencar la llei ni els principis ètics, s'han realitzat un seguit de comprovacions:

1. Si la pàgina web disposa de l'arxiu robots.txt, que en aquest cas sí en disposa https://www.instant-gaming.com/robots.txt i s'ha comprovat si hi ha exclusió o no de robots.
   Aquest arxiu està dient amb el "Disallow: " on no podem accedir amb el nostre user agent. En aquest prohibeix accedir a llocs privats com seria la pestanya de la compra, pagament, usuari, les meves comandes, etc.
   Per exemple:

- User-agent: *
- Disallow: /?q=
- Disallow: /es/pagos-*
- Disallow: /user/*
- Disallow: /*/my-credits/

2. S'ha comprovat que la pàgina web no demani cap mena de terme ni condició per a poder fer web scrpaing, per tant, es tracta de dades públiques que a priori es pot fer scraping sobre elles.
3. L'accés que es realitza es sempre a espais no protegits.
4. Les dades extretes tenen un fi únicament didàctic.

Inspiració

Avui dia el mon digital ho és tot, i els preus només fan que pujar. La societat s'ha tornat molt consumista i cada vegada gent més jove està entrant al món dels videojocs.
Amb aquest conjunt de dades es podrà comprovar l'evolució dels preus dels jocs any rere any, quins son els tipus/categories de jocs més comprats, sobre quines plataformes es juga més. A més, aquest conjunt de dades permet ampliar-se fàcilment ja sigui incloent les ressenyes(estrelles) que té cada joc, o les valoracions que aporten els usuaris, els requeriments de l'ordinador que es necessiten per poder fer córrer un joc, etc.
Per sintetitzar la informació, el dataset pretén respondre:

- Com evoluciona el preu d'un joc al llarg d'un any?
- Quines plataformes son les més jugades?
- Quins jocs son els més comprats?
- Quina classe de jocs son els que agraden més?

En anàlisis anterior com el del enllaç de Kaggle podeu trobar una quantitat d'informació referenciada extraordinaria. S'ha estimat que aquestes dades són excessivament àmplies i centrades en aspectes tècnics del joc. En canvi el dataset proposat aporta dades centrades en els aspectes comercials que defineixen l'èxit i l'impacte, positiu o negatiu. També s'hi inclou el cost econòmic i la seva evolució per a avaluar el poder de compra de l'usuari final.

Llicència

Les llicències que se li poden aplicar al dataset son:

- **Public-domain-equivalent license**: El seu ús és per fer que qualsevol individu pugui utilitzar les dades amb drets d'autor sense condicions. No es requereix cap permís per a dades de domini públic, per exemple un copyright caducat.
- **CC BY-NC-SA 4.0 License**: Permet a altres usuaris barrejar i reconstruir el nostre treball de manera no comercial i no haver de llicenciar-ho.
- **CC BY-SA 4.0 License**: Permet a altres usuaris barrejar i reconstruir el nostre treball de manera comercial, però sí ho han de llicenciar. Es podria comparar amb les llicencies de open source.
- **Open Database License, individual contents under Database Contents License**: Permet als usuaris compartir les dades amb llibertat i snese haver de patir pels drets d'autor, i a partir de les dades que han recollit afegir-les a les bases de dades.

En aquest cas el dataset se li atribueix la llicència **CC BY-NC-SA 4.0 License**, ja que es tracta de dades públiques les quals provenen d'una web amb copyright però es poden extreure amb la fi de no utilitzar-les per la comercialització.

Codi

A continuació es troba una descripció de descripció del codi emprat en aquesta pràctica:
El fil de codi principal empra dos mètodes:

- getData(): Obté les dades de cada un dels jocs de cada pàgina
- getPage(): Avança a la pàgina següent

El mètode principal del codi busca el rang de pàgines al web i per a cada una d'elles obté diferents informacions que es van afegint a una llista. Finalment, aquestes llistes són exportades a un arxiu .csv

Els problemes que hem tingut han estat a l'hora d'utilitzar agents web, que ens donaven bastants errors, però al final preparant els headers i utilitzant el web driver de mozzilla ho hem pogut resoldre.

Requisits:

anyio==3.6.2
argon2-cffi==21.3.0
argon2-cffi-bindings==21.2.0
asttokens==2.1.0
async-generator==1.10
attrs==22.1.0
backcall==0.2.0

beautifulsoup4==4.11.1
bleach==5.0.1
bs4==0.0.1
builtwith==1.3.4
certifi==2022.9.24
cffi==1.15.1
charset-normalizer==2.1.1
colorama==0.4.6
debugpy==1.6.3
decorator==5.1.1
defusedxml==0.7.1
distlib==0.3.6
entrypoints==0.4
executing==1.2.0
fastjsonschema==2.16.2
filelock==3.8.0
future==0.18.2
h11==0.14.0
idna==3.4
ipykernel==6.17.1
ipython==8.6.0
ipython-genutils==0.2.0
jedi==0.18.1
Jinja2==3.1.2
jsonschema==4.17.0
jupyter-server==1.23.1
jupyter_client==7.4.5
jupyter_core==5.0.0
jupyterlab-pygments==0.2.2
MarkupSafe==2.1.1
matplotlib-inline==0.1.6
mistune==2.0.4
nbclassic==0.4.8
nbclient==0.7.0
nbconvert==7.2.4
nbformat==5.7.0
nest-asyncio==1.5.6
notebook==6.5.2
notebook_shim==0.2.2
numpy==1.23.5
outcome==1.2.0
packaging==21.3
pandas==1.5.1
pandocfilters==1.5.0
parso==0.8.3

pickleshare==0.7.5
platformdirs==2.5.3
prometheus-client==0.15.0
prompt-toolkit==3.0.32
psutil==5.9.4
pure-eval==0.2.2
pycparser==2.21
Pygments==2.13.0
pyparsing==3.0.9
pyrsistent==0.19.2
PySocks==1.7.1
python-dateutil==2.8.2
python-dotenv==0.21.0
python-whois==0.8.0
pytz==2022.6
pywin32==305
pywinpty==2.0.9
pyzmq==24.0.1
requests==2.28.1
selenium==4.6.0
Send2Trash==1.8.0
sniffio==1.3.0
sortedcontainers==2.4.0
soupsieve==2.3.2.post1
stack-data==0.6.1
terminado==0.17.0
tinycss2==1.2.1
tornado==6.2
tqdm==4.64.1
traitlets==5.5.0
trio==0.22.0
trio-websocket==0.9.2
urllib3==1.26.12
virtualenv==20.16.6
wcwidth==0.2.5
webdriver-manager==3.8.5
webencodings==0.5.1
websocket-client==1.4.2
wsproto==1.2.0

Dataset

El dataset es pot trobar en el següent enllaç de Zenodo
https://doi.org/10.5281/zenodo.7349065

Video

Link vídeo:
https://drive.google.com/file/d/1GQvMcmxABcZrzWmIYdOkz9_mkn6YKET0/view?usp=share_link

Contribucions	Signatura
Investigació prèvia	Gerard Ramos
Redacció de les respotes	Gerard Ramos
Desenvolupament del codi	Gerard Ramos
Participació al video	Gerard Ramos