Autor: Carlos Pérez Martín
Autor: Òscar Fernández Castro

Tipologia i cicle de vida de les dades.

Pràctica 1: Web scraping

1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Dintre del mercat digital de videojocs, hi ha una gran oferta, tant en varietat de jocs com de plataformes on comprar els videojocs. Després de fer una mica d'investigació, es pot veure l'enorme diferència de preus entre les diferents plataformes de venda, i com molt sovint els preus fluctuen bastant sense saber clar el motiu.

Per això s'ha decidit que partint d'una de les pàgines més famoses i utilitzades per a la venda de videojocs digitals (https://www.instant-gaming.com/es/), analitzar les seves dades de manera periòdica per tal de crear un dataset que contingui la informació dels preus i els diferents factors que poden influir en les fluctuacions.

D'aquesta manera un usuari que vulgui fer una compra, podrà tenir tota la informació necessària per obtenir el preu que consideri òptim.

2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

El nom escollit per al dataset és: GamesDataPrices.

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).

El dataset tracta de mostrar l'evolució temporal dels videojocs que es venen digitalment en la web de instant-gaming (plataforma on-line de venda). Es tracta de mostrar l'evolució temporal dels preus i quins factors influeixen en aquests preus. Des de les diferents plataformes de distribució digitals que hi ha (steam, Uplay. etc), passant pel temps que fa que va sortir el videojoc i la valoració dels usuaris del producte.

Amb el dataset s'espera veure informació del preu dels productes, els possibles descomptes que se li apliquen, així com d'altres aspectes que afecten de manera significativa als preus, com ara la data de publicació o quan es va extreure la informació. Com a aspecte limitant tenim, que atributs significatius com ara el nombre de vendes en un determinat interval, no ha sigut possible extreure'l, ja que no és accessible.

Al dataset resultant ja se li realitza certa neteja de dades (format de dates, eliminació del tipus de moneda, etc). De totes maneres pot ser que requereixi certa neteja, ja que alguns jocs es publiquen a la pàgina quan encara no tenen un preu fix, una data de publicació determinada, o els usuaris no els hi han posat nota.

4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment.

Crec que la imatge inferior representa d'una manera gràfica el que es vol aconseguir amb el nostre dataset. Els videojocs són un passatemps àmpliament utilitzat en la societat, però que pot arribar a tenir un cost elevat.

El que es vol, és donar la màxima informació possible a l'usuari, perquè pugui estalviar el màxim en les seves compres.

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

Per raons de temps obvies, el dataset només compte un petit interval de temps capturat, però la idea darrera del dataset és calcular-lo de manera periòdica per poder veure l'evolució dels preus.

Atributs del dataset:

- name: Nom del producte.
- platform: Plataforma a la qual pertany el videojoc. Des de diferents plataformes de distribució digital de jocs d'ordinador (steam, Origin etc) a jocs de videoconsola, etc.
- original_price: Preu de compra estàndard, sense aplicar el descompte (en euros).
- discounted_price: preu de venda al públic un cop aplicat el descompte (en euros).
- discount: Percentatge de descompte entre el preu estàndard i el preu de compra.
- release_date: Data de sortida del videojoc.
- users_rating: Qualificació dels usuaris (sobre un total de 100 punts).
- category: Tipus d'objecte comprat (GAME, DLC)
- extraction_date: Data d'extracció del dataset

Per a recollir les dades s'ha utilitzat un programa de web scrapping desenvolupat en python, que es dedica a navegar per la pàgina web d'instant-gaming recopilant tota la informació necessària. A partir de la pàgina de la pàgina principal de cerca, accedeix als diferents enllaços que ès trobar per treure la informació individual dels diferents productes.

Amb els camps capturats i l'evolució dels mateixos s'espera poder donar la màxima informació possible als usuaris.

Cada dataset obtingut, és a dir, cada arxiu .csv conté un interval de temps molt petit, i la validesa d'aquestes dades és únicament d'un dia, ja que els preus varien constantment. Per una altra part, un cop obtinguda la informació de cada dia, durant tot un any per exemple, la validesa d'aquestes dades pot ser molt llarga, ja que aquestes dades permeten obtenir una evolució temporal dels preus mitjançant un tractament posterior de les dades. Mitjançant un tractament posterior de tot el conjunt de dades, es pot arribar a saber en quin moment un joc apareix nou a la pàgina, o en quin moment desapareix, es a dir, el temps que aquest joc es troba a la venda. Mitjançant l'anàlisi es podria arribar a predir, per un tipus de joc concret, quina podria ser l'evolució en el preu d'un joc nou, aprofitant tota la informació recollida.

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

La pàgina web de instant gaming pertany a Instant Gaming Ltd i es troba ubicada en Hong Kong.

Porta operant des de l'any 2005, i s'ha fet bastant famosa per oferir uns preus inferiors a la competència. Segons ells mateixos diuen, poden oferir aquests preus gràcies a ser jocs en format electrònic, amb la qual cosa retallen moltes despeses en emmagatzematge i logística, i que a més adquireixen els jocs en grans quantitats.

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

El conjunt de dades vol ajudar als usuaris a realitzar les seves compres de videojocs digitals amb la màxima informació possible. Es pretén mostra l'evolució temporal dels preus i quin són els criteris que influeixen en aquesta evolució.

Per exemple, quina reducció de preu pot tenir un determinat videojoc, que fa sortir a la venda fa 30 dies per a la plataforma steam? Si espero dos mesos més, quan valdrà el videojoc?

Baixen de preu igual de ràpid els videojocs que tenen valoracions molt positives, i els que tenen valoracions més dolentes? Quines plataformes tenen preus de sortida més alts? Quines fan les millors ofertes després de 3 mesos?

Totes aquestes preguntes i algunes més, són les que es pretenen contestar amb la creació d'aquest dataset.

Exemples similars al nostre dataset es poden trobar als enllaços 'www.steamprices.com/eu/tracker/' i 'https://camelcamelcamel.com'. En el cas de steamprices és molt similar al nostre dataset, en quant es basa en l'evolució temporal dels preus dels videojocs, però en aquest cas sobre la plataforma de venda steam. En el segon cas també es tracta d'un tracker de preus però no centrat únicament en videojocs, sinó en els diferents productes que es poden trobar a la venda en la pàgina d'Amazon.

En els dos casos, l'usuari pot seguir els evolutius dels preus d'uns determinats productes i intentar predir el seu comportament. Amb el nostre projecte, es podria complementar la informació d'aquests dos, afegint la que generem d'instant-gaming.

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 per al nostre dataset és la CC BY-NC-SA 4.0
S'ha escollit aquesta llicència perquè permet compartir el material creat i adaptar-lo i transformar-lo en cas necessari. Sempre que es reconegui l'autoria del mateix i no sigui utilitzat amb finalitats comercials.

Un dels grans avantatges d'aquest tipus de llicències és que obliga al fet que la nova obra resultant, ha de tenir la mateixa llicencia que l'original. D'aquesta manera es fomenta la transmissió del coneixement, especialment amb finalitats acadèmiques.

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

El codi en python es pot trobar dintre de la carpeta scraper del GitHub.

10. Dataset. Presentar el dataset en format CSV

Els datasets generats estan dintre de la carpeta Datasets del GitHub.

Taula de contribucions:

Contribucions	Signa
Recerca prèvia	CPM, OFC
Redacció de les respostes	CPM, OFC
Desenvolupament del codi	CPM, OFC