Tipologia i cicle de vida de les dades

PRA1

Alumnes: David Navarro Brugal
Antoni Llussà Sala

Índex

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

La informació que s'ha recol·lectat, està relacionada amb jocs de la consola "Super Nintendo" de la marca Nintendo. també anomenada "Super Famicom". Aquesta consola va sortir al mercat a la dècada dels anys 90. Un dels motius perquè s'ha escollit recollir informació sobre els jocs de la consola, és per la nostàlgia que ens porta als membres del grup.

L'objectiu que ens havíem posat era buscar alguna font d'informació que ens portés el màxim d'informació dels jocs i que el catàleg fos gran. S'han buscat diferents base de dades de jocs de la "Super Nintendo":

[1] "Super Nintendo (SNES) Games Database - SNES Games & ROM Info | superfamicom.org." [Online]. Available: https://superfamicom.org/.

[2] "SNES DATABASE." [Online]. Available: https://jensma.de/snes/.

[3] "Super Nintendo Entertainment System Games - LaunchBox Games Database." [Online]. Available: https://gamesdb.launchbox-app.com/platforms/games/53.

[4] "Video Game Charts, Game Sales, Top Sellers, Game Data - VGChartz." [Online]. Available: https://www.vgchartz.com/gamedb/games.php?console=SNES.

[5] "All game videos - Games Database." [Online]. Available: https://www.gamesdatabase.org/all_videos.

La pàgina web que ens facilita la informació més extensa i interessant és la [1] ens aporta informació com el títol original, títol USA, el productor, any, gènere, número de sèrie, descripció, staff tècnic, referències, screenshots, vídeos de youtube, informació de les roms, etc.

S'ha de tenir en compte que és una BBDD que és el hobby del responsable de la web, i és una pàgina web cooperativa, i es va actualitzant constantment. No tots els jocs hi ha la mateixa informació, aquest aspecte, ens ha semblat interessant perquè pensem que posa complexitat alhora de generar el data set.

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

Informació general i de les roms dels jocs de la Super Nintendo.

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).

Llista de jocs de “Super Nintendo” indicant la seva informació general i de les ROMS per tal de saber a on van sortir a la venda.

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

Figura 1- Imatge on apareixen jocs de la super nintendo

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

En el dataset si pot trobar la següent informació:

- Title: Títol.
- Original: Títol original.
- Date: Data de publicació.
- Genre: Gènere.
- Producer: Qui el va produir.
- Serial: Número de sèrie del joc.
- Ruby: Desconeixença, ja que hi ha la informació amb japonès.
- Description: Descripció.
- References: Referències.
- Availability: Disponibilitat, a on es pot comprar.
- Final Thoughts: Pensaments finals.
- Glitches: Valoració personals de qui ha posat la informació.
- The story: Història del joc.
- Ratings: Percentatge de valoració
    - Fun
    - Graphics
    - Lastability
    - Playability
    - Sound
    - Total
- Roms Info: Informació especifica de la ROM.
    - CRC32
    - Common Filenames
    - Country
    - Internal CRC
    - Internal Title
    - MD5
    - ROM Bank
    - ROM Size
    - ROM Speed
    - ROM Type
    - SHA-1
    - SHA-256
    - SRAM Size
    - Revision
- Staff Info: Equip de persones que han desenvolupat el joc.
    - Assistant Producer
    - Director
    - ENIX staff
    - Graphic designer

- Music composar
- Producer
- Programmer
- Publisher
- Scenario
- Special Thanks

A continuació es mostra un exemple de model entitat relació:

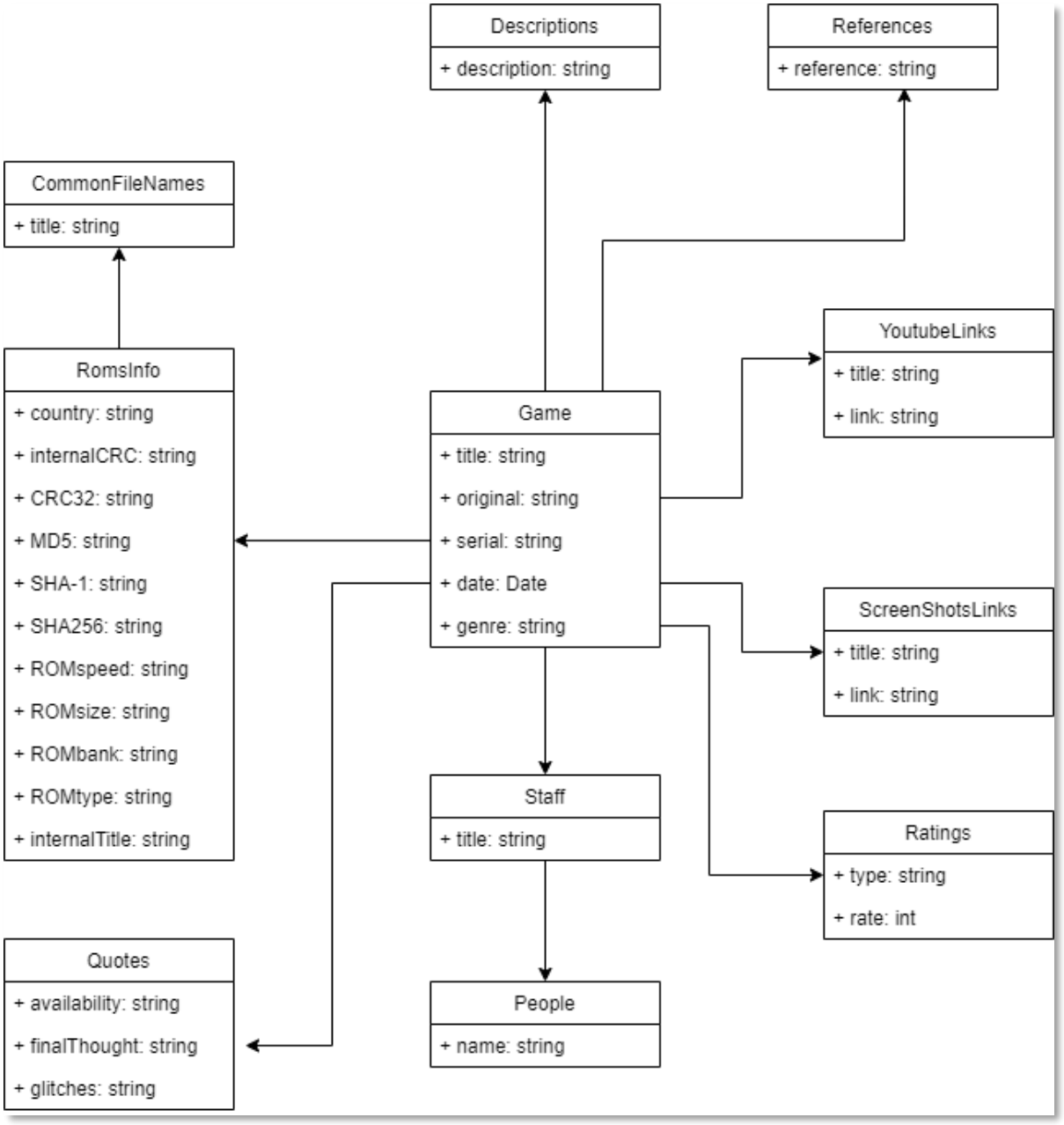

Figura 2 - Model relacional del dataset

La informació d'aquest dataset, és informació històrica dels jocs de la consola "Super Nintendo", la informació data en la dècada dels anys 90, que és l'època que va estar al mercat aquesta consola. En el peu de pàgina es pot veure que la pàgina web està en funcionament des del 2002 fins 2020. La actualització és mensual, per tant el catàleg va en augment.
La informació s'ha extret a la primera setmana d'abril del 2020.

El dataset, s'ha extret mitjançant tècniques de webscraping. Utilitzant el llenguatge Python i la llibreria especifica BeautifulSoup per fer webscraping. S'ha hagut d'analitzar l'estructura html, per poder extreure la informació. Els passos que s'han seguit han estat:

- Localitzar la URL del dataset i extreure els enllaços de totes les pàgines que hi ha jocs.
- Per cada pàgina de llistat de jocs, s'ha extret els enllaços dels jocs.
- Recórrer tots els enllaços dels jocs i extreure'n la informació rellevant.
- S'han descarregat les imatges "screenshots" dels jocs, les imatges quan es descarreguen s'han categoritzat per carpetes amb el títol del joc.

S'ha recopilat informació de 1868 joc amb un temps de 52 minuts.

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 pertany a Mattew Callis , de l'organització eludevisibility, de la ciutat Nashville de US.

"MATTHEWCALLIS (MATTHEW CALLIS)." [ONLINE]. AVAILABLE: HTTPS://GITHUB.COM/MATTHEWCALLIS.

La llista de webs que es van estudiar abans d'escollir el [1] són la següent llista:

[1] "SUPER NINTENDO (SNES) GAMES DATABASE - SNES GAMES & ROM INFO | SUPERFAMICOM.ORG." [ONLINE]. AVAILABLE: HTTPS://SUPERFAMICOM.ORG/.

[2] "SNES DATABASE." [ONLINE]. AVAILABLE: HTTPS://JENSMA.DE/SNES/.

[3] "SUPER NINTENDO ENTERTAINMENT SYSTEM GAMES - LAUNCHBOX GAMES DATABASE." [ONLINE]. AVAILABLE: HTTPS://GAMESDB.LAUNCHBOX-APP.COM/PLATFORMS/GAMES/53.

[4] "VIDEO GAME CHARTS, GAME SALES, TOP SELLERS, GAME DATA - VGCHARTZ." [ONLINE]. AVAILABLE: HTTPS://WWW.VGCHARTZ.COM/GAMEDB/GAMES.PHP?CONSOLE=SNES.

[5] "ALL GAME VIDEOS - GAMES DATABASE." [ONLINE]. AVAILABLE: HTTPS://WWW.GAMESDATABASE.ORG/ALL_VIDEOS.

Només la número [1] tenia informació general i especifica del dataset que ens varem proposar per recollir informació.

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

Aquest dataset és interessat perquè recull més de 1800 jocs de la SNES de les diferents regions que es venia la consola (USA, Japan, Germany, France, Europe), facilita informació de la gent d'staff que va realitzar els jocs, informació rellevant com el títol i títol original, l'any de publicació i el gènere.

Les preguntes que vol respondre:

- Quina regions estava disponible el joc.
- La data de publicació
- Com s'anomena el joc originàriament.
- Quin gènere és el joc.
- Quin es l'equip professional que el va crear? Scenario, Programmer, Graphic Designer, Music Composer, ENIX Staff, Assitant Producer, Director, Publisher
- Quina valoració té el joc en diferents aspectes. (Graphics, Sound, Playability, Lastability, Fun Factor i Total)

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 que pensem que s'escau millor al tipus de pràctica que s'està realitzant és:

Releases Under CC BY-SA 4.0 License:

Vostè és lliure de:

Compartir – copiar i redistribuir el material en qualsevol mitjà i format.
Adaptar – remesclar, transformar i construir a partir del material per qualsevol propòsit, incloent el comercial.

El llicenciador no pot revocar aquestes llibertats mentre vostè segueixi els termes de llicència.

Sota els següents termes:

Atribució – Vostè ha de donar crèdit de manera adequada, posar un enllaç a la llicència, i indicar si s'han realitzat canvis. Pot fer-ho en qualsevol forma raonable, però no de forma que suggereixi que vostè o el seu ús tingui el suport del llicenciador.

Compartir igual – Si remescla, transforma o crea a partir del material, ha de distribuir la seva contribució sota la mateixa llicència original.

No hi ha restriccions addicionals – No pot aplicar termes legals ni mesures tecnològiques que restringeixin legalment a altres o faci qualsevol us permès per la llicència.

Avisos:

No ha de complir amb la llicència per elements del material en el domini públic o quan el seu ús es permès per una excepció o limitació aplicable.

No es donen garanties. La llicència podria no donar tots els permisos que necessita per l'ús que en tingui previst. Per exemple, altres drets com publicitat, privacitat o drets morals poden limitar la forma en que utilizi el material.

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

S'adjunta el document: pra1.webscraping.ipynb

10. Dataset. Publicar el dataset en format CSV a Zenodo amb una xicoteta descripció.

La URL del Zenodo: https://zenodo.org/record/3746655#.Xo9mVMj7SUk

11. Lliurar. Presentar el treball amb el DOI del dataset a Github

La URL del Github és: https://github.com/tllussa/UOC-PRA1-M2.951-WebScraping-dnavarro30-tllusa

LA URL del DOI és: https://doi.org/10.5281/zenodo.3746888