

Respostes preguntes Pràctica 1

Components del grup: Sara Quesada Gil

Pregunta 1: Context

El lloc web triat és: https://pokemon.fandom.com/es/wiki/Lista_de_Pok%C3%A9mon

Avui dia, qui no ha escoltat el nom Pokémon. No es parla d'altra cosa, especialment des de fa un parell d'anys. Si veus un grup de gent amb els caps en la pantalla del mòbil, has de saber que són jugadors de Pokémon. Si veus una estampida de persones cridant en un centre comercial amb els telèfons mòbils, has de saber que són jugadors de Pokémon, i un llarg etcètera.

I no és d'estranyar, si donem un cop d'ull al desorbitat nivell d'ingressos que ha generat el món Pokémon, veurem que guarda relació amb la gran quantitat de persones que dedica part del seu temps lliure a gaudir de les prestacions dels jocs.

Per mostrar una part de l'impacte del fenomen Pokémon, a continuació mostro una imatge resum extreta de l'enllaç <https://sensortower.com/blog/pokemon-go-revenue-year-two>, on es mostra que la facturació obtinguda per l'aplicació mòbil Pokémon Go és de 2 milions de dòlars diaris, aconseguint 1,8 bilions de dòlars durant els primers dos anys de creació de l'aplicació.

Pokémon GO Revenue Hits \$1.8 Billion on Its Two Year Launch Anniversary



Released two years ago today, Niantic's [Pokémon GO](#) became an overnight sensation. While the fervor has subsided in the 24 months since that game-changing day, players worldwide continue to spend more than \$2 million per day in their quest to catch those original Pokémon and the dozens that have been added since launch. This has led to the game reaching \$1.8 billion in player spending, according to the latest Sensor Tower [Store Intelligence](#) estimates, having [crossed the \\$1 billion milestone](#) in January 2017.

Així mateix, el següent enllaç mostra el creixement sostingut que ha tingut el fenomen Pokémon Go, i la previsió de tancar l'any 2019 amb una facturació acumulada de 3 bilions de dòlars: <https://variety.com/2019/gaming/news/pokemon-go-global-revenue-2018-1203098512/>

Per tot l'esmentat, trobo més que interessant analitzar què és el fenomen Pokémon, què són els Pokémon, quines característiques els han fet tan especials, pel que aprofitaré aquestes qüestions per desenvolupar aquesta pràctica.

El context en què s'ha recol·lectat la informació és un entorn amb una estructura HTML mitjanament complexa a causa de la gran quantitat de dades que hi han però, especialment, degut a la distribució d'aquestes, ja que algunes es trobaven en subpàgines de la principal. Així mateix, i després d'estudiar molt a fons l'estructura de la web, els materials facilitats per l'assignatura i un munt de pàgines per anar formant-me en la tècnica de Web scraping, m'ha semblat interessant adonar-me de què potser la pàgina en qüestió estava construïda per evitar que li fessin web scraping, ja que no era massa senzill obtenir automàticament la informació (per exemple, l'id i la classe del div que contenia dades de cada generació de Pokémon era la mateixa, i a més el text estava repartit en diversos paràgrafs i en diferents tipus de la subpàgina).

El lloc web proporciona aquesta informació per informar al gran volum de persones interessades en el joc de Pokémon (i derivats) sobre dades d'origen d'aquests, habilitats de joc, ... Considero que aquesta és una de les moltíssimes pàgines que han detectat el gran benefici que han obtingut infinitat d'empreses gràcies al món Pokémon (especialment Niantic, creadora de l'aplicació mòbil Pokemon Go, o altres aplicacions i pàgines que contenen informació del tema).

Pregunta 2: Definir un títol pel data set

El títol que s'ha definit pel dataset del programa ha sigut '*pokemon_atributs_coleccio*', ja que ho considero bastant descriptiu per la temàtica que ho caracteritza.

Pregunta 3: Descripció del data set

El conjunt de dades extret són 6 atributs de tipus llista que emmagatzemen les 807 varietats de Pokémon que hi han, dades necessàries pel públic objectiu del joc, que finalment es recullen en un dataset que s'exporta a un document amb extensió .csv. Els atributs seleccionats

són l'id, el nom, el dos tipus del Pokémon (pot tenir un o dos) recollits en les variables tipus1 i tipus2 respectivament, la generació a la qual correspon cada Pokémon, emmagatzemats en la variable generació, i la descripció, la qual s'obté mitjançant un link que conté la informació de cada Pokémon i el qual es visita un a un per obtenir la descripció d'aquest.

Mitjançant aquest sistema, i aprofitant que tindrem el contingut de la subpàgina, també s'obté l'id. Aquesta informació es guarda en l'atribut descripcio. Cal tenir en compte que per la gran quantitat de dades que cal gestionar per l'obtenció d'aquest atribut (concretament emmagatzemar 807 cops el contingut de la subpàgina corresponent a cada diferent Pokémon), únicament he incorporat al codi Python el procés per a la primera generació de Pokémon (151), ja que fent-lo pels 807 el programa trigava entre 3 i 4 minuts en finalitzar, d'aquesta manera triga entre 30 i 40 segons. Per omplir la informació de l'id i de la descripció de la resta de generacions, s'ha optat per omplir l'id en funció d'un comptador, que respecta al 100% l'ordre i numeració original i un ' - ' per omplir el camp de descripció.

Un exemple d'una fila o element del dataset seria el següent:

1	Bulbasaur	Tipo planta	Tipo veneno	Primera Generación	Bulbasaur es un Pokémon cuadrúpedo de color ve
---	-----------	-------------	-------------	--------------------	--

On el nombre 1 representa l'atribut id, "Bulbasaur" representa el nom del Pokémon, "Tipo Planta" representa l'atribut tipus1, "Tipo veneno" represent l'atribut tipus2, "Primera Generación" fa referència a l'atribut generacio i "Bulbasaur es un..." representa el camp descripcio. El conjunt d'aquestes dades representa perfectament l'objecte d'estudi Pokémon.

És important mencionar que el contingut d'aquest dataset no necessitarà neteja a posteriori i les seves dades són 100% consistents, aquestes no hauran de variar mai. L'únic inconvenient seria que en un futur, quan el nombre de generacions vagi augmentant, haurem d'actualitzar el dataset amb les noves generacions. En aquest cas, únicament haurem d'aplicar el mateix patró que hem seguit, però augmentant el nombre d'elements. Igualment, encara que augmentessin el nombre de generacions, no afectaria al nostre dataset, però no estaria actualitzat.

Pregunta 4: Representació gràfica

Per a representar gràficament el conjunt de dades obtingut, he optat per mostrar la següent taula, que correspon al contingut del dataset exportat al fitxer .csv, al qual s'ha donat

format de codificació UTF-8 i s'ha separat per comes mitjançant l'opció d'obtenir dades des de text/csv que proporciona el mateix programa. Aquest procés s'explica en el fitxer README.md.

Id	Nombre	Tipo 1	Tipo 2	Generación	Descripción
1	Bulbasaur	Tipo planta	Tipo veneno	Primera Generación	Bulbasaur es un Pokémon cuadrúpedo de color verde y manchas más oscuras de formas geométricas.
2	Ivysaur	Tipo planta	Tipo veneno	Primera Generación	Ivysaur posee un color azulado más vivo que su preevolución Bulbasaur. Además, sus ojos adquieren un color rojo.
3	Venusaur	Tipo planta	Tipo veneno	Primera Generación	Su nombre es una combinación de las palabras Venus (una flor parecida a la planta que le crece) y saur.
4	Charmander	Tipo fuego	-	Primera Generación	Charmander es un Pokémon de tipo fuego introducido en la primera generación. Es uno de los tres Pokémon de tipo fuego introducidos en la primera generación.
5	Charmeleon	Tipo fuego	-	Primera Generación	Charmeleon es un Pokémon de tipo fuego introducido en la primera generación. Es la evolución de Charmander.
6	Charizard	Tipo fuego	Tipo volador	Primera Generación	Su nombre es una contracción de las palabras inglesas char (carbonizar, quemar, incinerar) y lizard (lagarto).
7	Squirtle	Tipo agua	-	Primera Generación	Su nombre proviene de las palabras en inglés squirt (disparar un chorro de agua) y turtle (tortuga).
8	Wartortle	Tipo agua	-	Primera Generación	Su nombre proviene de las palabras en inglés war (guerra) y turtle (tortuga).
9	Blastoise	Tipo agua	-	Primera Generación	Su nombre es una combinación de las palabras en inglés blast (explosión o ráfaga) y tortoise (tortuga).
10	Caterpie	Tipo bicho	-	Primera Generación	Su nombre deriva del inglés caterpillar (oruga, animal con el que comparte algunos rasgos).
11	Metapod	Tipo bicho	-	Primera Generación	Su nombre proviene de metamorfosis, proceso de transformación de la oruga en mariposa.
12	Butterfree	Tipo bicho	Tipo volador	Primera Generación	Su nombre es el resultado de la combinación de las palabras butterfly (mariposa) y free (libre).
13	Weedle	Tipo bicho	Tipo veneno	Primera Generación	Su nombre proviene de "wee", "pequeñito" en inglés (aunque esta palabra solo adopta este significado en el Reino Unido).
14	Kakuna	Tipo bicho	Tipo veneno	Primera Generación	Su nombre, tanto japonés como inglés, proviene del inglés cocoon (crisálida).
15	Beedrill	Tipo bicho	Tipo veneno	Primera Generación	Su nombre proviene de la unión de las palabras en inglés bee (abeja) y drill (taladro).

Així mateix, trobo molt representativa la següent imatge, on es mostra tot el que abasta el tema en qüestió, Pokémon, obtinguda del següent enllaç:

<http://edition.cnn.com/travel/gallery/pokemon-go-hong-kong/index.html?gallery=0>



Pregunta 5: Contingut

El contingut del dataset està compost pels següents 6 atributs:

- ids: Llista que emmagatzema el id oficial de cada Pokémon.
- noms: Llista que guarda el nom de cada Pokémon.

- tipus1: Llista que emmagatzema el tipus de cada Pokémon. Cada Pokémon pot tenir 1 o 2 tipus, però mínimament ha de tenir 1.
- tipus2: Llista que conté el tipus 2 de cada Pokémon. En cas que el Pokémon únicament correspongui a un tipus, aquest atribut figurarà amb el valor ' - '.
- generacio: Llista que guarda la generació de cada Pokémon, string obtingut de la llista de generacions proporcionada per la web, però no directament per la taula de cada generació.
- descripcions: Llista que emmagatzema les descripcions de cada Pokémon, obtinguda de cada subpàgina.

Per poder extreure aquestes dades de la web en qüestió https://pokemon.fandom.com/es/wiki/Lista_de_Pok%C3%A9mon, mitjançant la tècnica de web scraping, es va recórrer a l'ús de les següents llibreries:

- requests: Llibreria que permet fer peticions HTTP, FTP,...
- BeautifulSoup: Llibreria que permet analitzar i extreure dades de documents HTML.
- lxml: Llibreria que permet treballar amb documents XML.
- pandas: Paquet per poder fer ús de l'objecte dataframe.
- csv: Llibreria que permet treballar i manipular arxius .csv.

El llenguatge amb el qual es va desenvolupar el codi va ser amb Python 3.7.1 i l'IDE utilitzat va ser Visual Studio Code.

D'altra banda, cal recordar el mencionat al tercer apartat, que les dades del dataset són 100% consistents, amb el petit inconvenient que en un futur haurem d'actualitzar el dataset amb les noves generacions que apareguin, encara que de totes maneres, això no afectaria el nostre dataset, però aquest no estaria actualitzat.

Pregunta 6: Agraïments

El propietari de la pàgina web d'on s'han obtingut les dades del dataset objecte de la pràctica, és FANDOM (també conegut per Wikia), companya registrada en Delaware (EEUU) i CIF C2935209.

Mitjançant la recerca de les seves condicions d'ús analitzant l'enllaç <https://www.fandom.com/es/licensing-es>, s'ha determinat que la seva llicència és *Creative*

Commons, de la qual s'adjunta enllaç, i la qual és molt permissiva quant a difusió i reutilització de codi:

<https://creativecommons.org/licenses/by-sa/3.0/deed.es>

Després de realitzar cerques en Google, no he trobat cap referència a aquesta web.

Cal destacar que el contingut d'aquesta web, el qual està recollit en el dataset d'aquest projecte, és molt útil per dissenyar estratègies de joc per qualsevol de les edicions de Pokémon. Per exemple, si un consumidor d'algun joc o aplicacions de Pokémon vol fer un combat, necessitarà consultar el contingut de la web o el dataset per saber de quins tipus són els Pokémon contra els que vol lluitar i d'aquesta manera escollir el més apropiat dels del seu equip.

Pregunta 7: Inspiració

El motiu pel qual he escollit aquesta temàtica ha sigut principalment, com s'ha comentat a l'apartat de context, que em crida molt l'atenció la voràgine que ha causat el moviment Pokémon, especialment amb l'arribada de l'aplicació mòbil Pokémon Go, a més de per motius d'interés particulars. Penso que és increïble el nivell d'abastament al qual ha arribat, facturacions de molts dígits, publicitat constant, patrocinis i mecenatges per part de grans empreses,... Però no tan sols en l'àmbit econòmic, també ha afavorit les relacions personals.

D'altra banda, trobo que l'estructura per fer web scraping en una web tan densa, era un repte per ser el meu primer cop. He hagut de dedicar moltes hores per entendre tot el funcionament, fer proves, ... però personalment això em genera motivació.

Així mateix, considero que la selecció particular del conjunt de dades pot respondre pràcticament totes les preguntes que un usuari del joc es faria, com per exemple la generació a la qual correspon un Pokémon, el tipus què és, descripció bàsica,... A més, es recullen totes les generacions en un únic document, el que facilita la recerca de dades d'un element en particular.

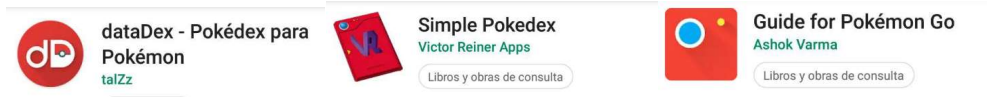
A continuació, menciono algunes de les pàgines que treballen amb dades similars a les que recull el dataset creat en aquest projecte:

https://www.wikidex.net/wiki/Lista_de_Pok%C3%A9mon

<https://www.pkparaiso.com/pokemon/lista-pokemon.php>

https://pokefanaticos.com/pokedex/pokedex_listar.php

I algunes aplicacions mòbils que segueixen el mateix patró:



Les pàgines i aplicacions mostrades, responen a les mateixes necessitats que les que respon el meu dataset, no obstant això, realitzant comparacions amb les pàgines, el valor afegit que representa el dataset d'aquest estudi respecte, per exemple el de la web “pokefanaticos”, és que el d'aquest estudi conté directament la generació a la qual correspon cada Pokémon, a més de mostrar tots els elements en un únic fitxer (en la pàgina de “pokefanaticos”, les dades es mostren de 20 en 20, d'un total de 807 elements).

Pregunta 8: Llicència

La llicència escollida pel dataset de la pràctica ha sigut una de les llicències Creative Commons, concretament la CC BY-SA 4.0 License pels següents motius:

- Les llicències Creative Commons, són una eina legal de caràcter gratuït que permet als usuaris tenir dret a reproduir l'obra i a distribuir-la sense càrrec.
- El beneficiari de la llicència té el dret de copiar, distribuir, exhibir i representar l'obra i fer obres derivades sempre que es reconegui i es citi l'obra de la forma especificada per l'autor d'aquesta.
- És de caràcter comercial, fet que motivarà la seva distribució (per aquest motiu no hem escollit l'opció No comercial (NC))
- El beneficiari de la llicència té el dret de distribuir obres derivades sota una llicència idèntica a la llicència que regula l'obra original.

D'altra banda, la llicència escollida pel projecte ha sigut MIT License, ja que no és gaire restrictiva en termes de difusió el qual potenciarà la distribució del projecte. Aquesta llicència ha sigut aplicada des del mateix repositori de GitHub i s'ha generat l'arxiu LICENSE per constatar-ho, el qual es troba dins del repositori.

Pregunta 9: Codi

(Fitxer codi font de nom: *Fitxer pokemon.py*)

Pregunta 10: Dataset

(Arxiu anomenat com: *Llista_Pokemon.csv*)

Contribucions	Signa
Recerca prèvia	S.Q.G.
Redacció de les respostes	S.Q.G.
Desenvolupament codi	S.Q.G.