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

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.

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 geor
2	Ivysaur	Tipo planta	Tipo veneno	Primera Generación	Ivysaur posee un color azulado más vivo que su preevolución Bulbasaur. Además, sus ojos ad
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 cre
4	Charmander	Tipo fuego	-	Primera Generación	Charmander es un Pokémon de tipo fuego introducido en la primera generación. Es uno de l
5	Charmeleon	Tipo fuego	-	Primera Generación	Charmeleon es un Pokémon de tipo fuego introducido en la primera generación. Es la evoluc
6	Charizard	Tipo fuego	Tipo volador	Primera Generación	Su nombre es una contracción de las palabras inglesas char (carbonizar, quemar, incinerar) y
7	Squirtle	Tipo agua	-	Primera Generación	Su nombre proviene de las palabras en inglés squirt (disparar un chorro de agua) y turtle (tor
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
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 (libr
13	Weedle	Tipo bicho	Tipo veneno	Primera Generación	Su nombre proviene de "wee", "pequeñito" en inglés (aunque esta palabra solo adopta este
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), en alu

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

A continuació es descriuen aquests 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.

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

Pregunta 7: Inspiració

El motiu pel qual he escollit aquesta temàtica ha sigut principalment perquè 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'ambit 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.

Pregunta 8: Llicència

La llicència escollida ha sigut GNU GENERAL PUBLIC LICENSE (Other, de l'enunciat de la pràctica), ja que és molt restrictiva en termes de difusió. He dubtat en escollir una llicència MIT que és menys restrictiva per motius de reutilització, però finalment m'he decantat per la GNU.

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.