Tipologia i cicle de vida de les dades

Pràctica 1

Web scraping

Bonnín Hernández, Joan
Dolz Fernández, José Luís

Índex

1. Títol del dataset **2**

2. Subtítol del dataset **2**

3. Imatge **2**

4. Context **2**

5. Contingut **2**

6. Agraïments **3**

7. Inspiració **3**

8. Llicència **3**

9. Codi **3**

10. Dataset **3**

1. Títol del dataset

Animals en adopció.

2. Subtítol del dataset

Estat dels animals en adopció de protectores que fan servir el CMS Bambú. Inclou registres d'animals d'arreu de l'estat espanyol, a més de l'històric d'adopcions.

3. Imatge

Imatge sota llicència CC0 Creative Commons. Font

4. Context

El conjunt de dades recull la informació dels animals en adopció d'aquelles protectores que fan servir el sistema de gestió de continguts conegut com a Bambú. Aquest sistema de gestió de continguts permet als diferents voluntaris de les protectores difondre les fitxes i els estats dels animals que tenen a càrrec de forma total gratuïta.

El conjunt de dades és una recopilació de la informació que es pot trobar als perfils dels animals en adopció, com pot ser el seu nom, estat de salut, descripció de la seva personalitat, edat, etc.

5. Contingut

Aquests són els camps que componen el dataset:

- **domain:** (cadena) El domini http des d'on s'han carregat les dades. Coincidirà amb la protectora que gestiona l'animal.
- **scraped_at:** (data i hora) Metadada en forma de marca temporal per identificar quan es va produir la consulta de la informació.
- **id**: (enter) Número identificador unívoc de l'animal dins de la protectora
- **name**: (cadena) Nom de l'animal.
- **image**: (cadena) Adreça http amb la imatge principal de l'animal.
- **status**: (cadena) Situació de l'animal: en adopció, adoptat, perdut, etc.
- **urgency**: (booleà) Indica si el cas és urgent.
- **special_case**: (booleà) Indica si es tracta d'un cas especial, habitualment per malalties.
- **class**: (cadena) Espècie a la qual pertany l'animal: gat, gos, lloro, cavall, etc.
- **since**: (data) Data en la qual es va donar d'alta el perfil de l'animal a la web de la protectora.
- **gender**: (cadena) Sexe de l'animal.
- **age**: (cadena) Edat de l'animal, en text lliure. Pot incloure tant els anys, com anys i mesos.
- **birthday**: (data) Data de naixement (exacta o aproximada) de l'animal.
- **race**: (cadena) Raça de l'animal dintre de la seva espècie.
- **size**: (cadena) Mida de l'animal (petit, mitjà, gran) dintre dels rangs de la seva espècie.
- **weitgh**: (cadena) Pes de l'animal.
- **chip**: (cadena) cadena indicant si l'animal és portador de xip. En alguns casos s'especifica el nombre identificador del mateix.
- **situation**: (cadena) Estat pel que fa al procés d'adopció en el que es troba l'animal.
- **location**: (cadena) Ciutat on es troba actualment l'animal.
- **health**: (cadena) Informe sobre l'estat de salut indicant possibles malalties, estat de les vacunes i si porta el xip identificador.
- **description**: (cadena) Historial de l'animal i altres dades importants.

Les dades carregades es corresponen als animals en adopció, en cerca o adoptats fins al dia 15/04/2018 a la vesprada. La càrrega de dades s'ha realitzat de forma espaiada i s'ha realitzat fins a la matinada del 16/04/2018.

Els registres amb més antiguitat dins els sistemes de les protectores daten, llevat d'alguns errors, de 2008, així que la temporalitat del dataset es situa entre aquest punt i l'actualitat.

6. Agraïments

Volem agrair a Helena Jimenez Heidenreich per haver creat el sistema de gestió de continguts (CMS) Bambú que permet a les protectores difondre l'adopció dels animals.

Agrair també les tasques de tots els voluntaris de les protectores per acollir, cuidar i cercar una nova llar de tots els animals que es troben.

Per últim, però no més important, agrair a ScrapingHub i tots els contribuïdors que han fet possible la creació del marc de treball Scrapy fins a arribar a la versió 1.5, que hem fet servir en aquesta pràctica.

7. Inspiració

Ens han inspirat diferents amics i coneguts que són voluntaris i que empren Bambú. Es queixaven de poder pujar els casos al CMS, però que no hi havia manera de poder descarregar a un altre format (Excel o base de dades) tots els animals que tenen o han tingut. Per això hem pensat que mitjançant aquesta pràctica podran obtenir allò que volien en un format gairebé universal com un fitxer separat per comes.

D'aquesta manera, es poden respondre a preguntes com quants animals estan actualment en adopció en cada protectora; quants són de cada espècie; el nombre de casos urgents i/o especials; edat mitjana per espècie, etc. A més, en un futur es pot contrastar -quan un animal és adoptat- quant de temps es triga a adoptar un animal d'ençà que s'insereix a la pàgina, quines són les característiques que fan que uns s'adoptin més ràpidament, races que més s'adopten, etc.

Una altra idea motivadora per obtenir aquest dataset és la possibilitat de crear un agregador de fàcil accés on qualsevol usuari pugui consultar l'estat de diferents animals en adopció. Això simplificaria els processos, ja que no seria necessari consultar directament les fonts de diferents protectores.

8. Llicència

La llicència escollida per lliurar el codi del rastrejador i el conjunt de dades és CC BY-NC-SA 4.0. Això vol dir que es pot compartir i redistribuir en qualsevol medi o format, i que es pot adaptar i transformar, sempre sota els següents termes:

○ S'ha de donar el crèdit adequat als autors de l'obra proporcionat un enllaç directe a la llicència i indicar si s'han fet canvis respecte a l'original, però sense suggerir que els autors hi donen suport.

○ El contingut del codi i del conjunt de dades no pot ser emprat per motius comercials.

○ Compartir igual: si es barreja, transforma o es construeix sobre aquesta obra, s'ha de distribuir les noves contribucions sota la mateixa llicència.

Hem escollit aquesta llicència perquè altres persones també en puguin aprendre com fer servir un rastrejador amb el marc de treball Scrapy i utilitzar-ho si es troben en el mateix cas de les protectores amb les quals Bambú col·labora. Per altra banda, no volem que es faci un ús comercial de les dades que són proporcionades per protectores que fan un gran esforç per mantenir els animals i trobar-los una nova llar, a més del temps i els diners que els mateixos voluntaris dediquen de forma tan altruista.

9. Codi

El codi implementat per carregar el *dataset* està implementat en python, més concretament sobre el *framework* Scrapy. El codi està disponible al repositori de l'entrega, concretament al directori scranimal.

Ens agradaria fer algunes consideracions sobre la configuració i execució del projecte *scranimal*:

- **user-agent:** Definim un user-agent per tal que se'ns identifiqui correctament i s'entengui la intenció acadèmica del projecte.
- ***Caché* de peticions:** *Cachejam* les peticions satisfactòries durant una hora per evitar peticions excessives als sistemes accedits.
- ***Logging*:** En lloc de mostrar excessiva informació per pantalla durant l'execució, s'enregistren tots els missatges a l'arxiu *scrapy.log*.
- **Freqüència entre peticions:** Esperem tres segons entre successives peticions al mateix domini, evitant així que se'ns consideri usuaris abusius i saturar el sistema a atacar.
- **Execució de l'aranya:** Per simplificar l'execució de l'aranya hem creat un senzill *script* que elimina datasets anteriors al mateix directori i volca noves dades. En qualsevol cas, una manera senzilla d'executar l'eina és, des de línia de comanda, executar **scrapy crawl adopting -o <nom_dataset>.csv -t csv** al directori pare del codi: /scranimal.

10. Dataset

El dataset s'exporta en format CSV, en codificació UTF-8. El separador usat són comes i només s'usen cometes com a delimitador si el contingut ho requereix. A més a més, la nostra eina permet exportació amb diferents formats mitjançant paràmetres simples de l'execució.

El dataset exportat es pot consultar a l'arrel del repositori, i l'arxiu es pot consultar en cru a la següent adreça:

https://raw.githubusercontent.com/JoanBonnin/scraping-tipologia-dades/master/dataset.csv?token=AFXTb_qnwZ3XGRPdewm8iSpYSfz9Y2rBks5a3g5VwA%3D%3D