Web Scrapping sobre Amazon
Una proposta d'obtenció de dades

Carlos A. García

November 10, 2018

Assignatura:	M2.951-Tipologia i cicle de vida de les dades
Professora associada:	Laia Subirats Maté

Figure 1: Imatge obtinguda mitjançant Web Scrapping

1 Dataset

El conjunt de dades generat es correspon amb les característiques d'una sèrie de productes a la web Amazon. El mecanisme d'obtenció del Dataset és el següent:

1. **Determinem les paraules clau a partir de les quals volem obtindre informació dels seus productes**. En el nostre cas hem cercat per marques comercials associades a la fotografia: Nikon, Canon, Leica i Fujifilm. Això és configurable i podem adaptar el Dataset a qualsevol concepte.
2. **Executem el codi proporcionat al projecte**
3. **Es genera un fitxer (CSV) amb les principals caracterítiques dels productes obtinguts**
4. **Depurar les dades.** Així com es proporcionen les dades no són tractables. S'han de depurar i netejar (que el preu sigui un numèric, que la valoració sigui numèrica, eliminar possibles duplicats, etc.). Aquest punt queda fora de l'abast del codi proporcionat
5. **El fitxer generat es pot emmagatzemar a una base de dades.** Aquest punt queda fora de l'abast del codi proporcionat

Així, el més important d'aquest exemple no és el Dataset en sí, sino que es pot adaptar a qualsevol producte, concepte, marca, etc.
Les principals caracterítiques del Dataset obtingut són:

1. 9 columnes
2. 1037 files que es corresponen amb 4 conceptes clau

2 Contexte

La monitorització de productes a Amazon pot ser molt útil per a diferents funcionalitats:

1. **Seguiment del preu a una web de descomptes.** Es pot fer una aplicació que avisi a l'usuari si un determinat producte baixa de preu
2. **Informació comercial.** Una tenda que vengui els mateixos productes pot fer un seguiment de les seves característiques a Amazon
3. **Completar informació.** Si tenim un Dataset amb productes, el podem completar amb la tècnica de Web Scrapping

3 Contingut

Els continguts obtinguts són els propis de qualsevol producte a una cerca a Amazon.

Figure 2: Imatge obtinguda mitjançant Web Scrapping

Obtenim les dades:

1. **imgName.** Nom del fitxer de l'imatge
2. **code.** Codi intern del producte per a Amazon. El podem fer servir de clau única del producte
3. **description.** Descripció del producte. El nom forma part de la descripció
4. **price.** Preu del producte
5. **image.** URL original de l'imatge; es descarrega en local i es desa a la carpeta img
6. **puntuation.** Valoració per part dels clients del producte (número de estrelles)
7. **altImage.** Text visible quan es situa el ratolí sobre l'imatge. Sembla que es correspon amb la descripció del producte
8. **stock.** Indicador de si hi ha stock limitat
9. **date.** Data de la captura. Permetrà monitoritzar evolucions de preu

Exemple de dades obtingudes:

imgName	51gYVi9XqDL._AC_US218_.jpg
code	B01C672QQ0
description	Nikon CoolPix B700 - Cámara Digital de 20.3 megapíxeles (Zoom Opt. 60 x, Full HD ...
price	EUR 459,00
image	https://images-eu.ssl-images-amazon.com/images/I/51gYVi9XqDL._AC_US218_.jpg
puntuation	3,8 de un máximo de 5 estrellas
altImage	Nikon CoolPix B700 - Cámara Digital de 20.3 megapíxeles (Zoom Opt. 60 x, Full HD ...
stock	Sólo hay 3 en stock. Cómpralo cuanto antes.
date	2018-10-31 08:33:41.330821

4 Agraïments

Les dades han estat obtingudes de la web www.amazon.es. Per a tal fi, s'ha fet servir l'IDE Intellij, el llenguatge de programació Python i tècniques de Web Scrapping

5 Inspiració

L'idea inicial va sorgir de forma casual, comprovant que les cerques a Amazon es poden parametritzar via URL. Els dos principals camps són:

1. **field-keywords.** Concepte pel que se fa la cerca
2. **page.** Ordre de la pàgina. Qualsevol cerca pot generar 1 o més pàgines navegables

A partir d'aquest punt inicial, vaig pensar que seria interessant monitoritzar l'evolució de les característiques dels productes; d'aquesta forma podem avaluar si un preu baixa o puja, si la valoració del producte varia, etc. Com es pot veure, no sóc el primer que ha arribat a una solució d'aquest tipus [1].

6 Llicència

La llicència triada és "GNU Lesser General Public License v3.0". Aquesta llicència permet:

1. **Commercial use.** Qualsevol empresa o particular pot fer servir el codi per a fer una obtenir una solució comercial
2. **Modification.** Qualsevol té permís de modificar el codi
3. **Distribution.** Qualsevol pot distribuir lliurement el codi
4. **Patent use.** Qualsevol pot fer servir les possibles patents que es puguin derivar del codi
5. **Private use.** Es permet fer un us privat del codi

La llicència **NO** permet:

1. **Liability.** L'autor no es fa responsable dels efectes que es puguin produir com a resultat de fer servir el codi
2. **Warranty.** No es proporciona garantia d'us

References

[1] *https://elpais.com/tecnologia/2017/09/21/actualidad/1506009655_-602120.html*