Repositori de notícies

Aleix Martínez i Ignasi Vilarasau

15 d'abril de 2019

1 Introducció

Imaginem, per exemple, que una marca de roba vol publicitar-se en la pàgina web d'un diari per intentar guanyar nous clients. En aquest sentit, al diari li interessaria poder disposar d'un dataset propi creat amb tot la informació relacionada amb l'històric de notícies que han anat publicat durant un període de temps en la seva pàgina web i en funció dels resultats de l'estudi de mineria de dades poder oferir una proposta de pressupost a la marca de roba interessada en publicitar-se.

Hem recopilat informació relacionada amb les notícies publicades en la pàgina web del diari *El País* ([1]) per poder crear un dataframe com a respositori de dades on hi hagi suficient informació emmagatzemada, per a què, finalment puguem estudiar quines notícies acaben sent més vistes i perquè. I d'aquesta forma poder oferir-li aquest repositori i anàlisi al diari per a que pugui negociar amb propostes de pressupost més alt a marques que volen publicitar-se en la pàgina web del diari.

Figure 1: Publicitat en la pàgina web del diari *El País*

Per a la creació del dataset s'ha extret tot el contingut rellevant de la pàgina web del diari *El País* en diversos moments del dia. D'aquesta forma podem realitzar els anàl·lisis pertinents per obtenir informació valuosa de les noticies més vistes pels lectors del diari. Així el diari pot tenir una sòlida base sobre la qual poder negociar contractes amb companyies publicitàries poguent garantir una certa repercusió en funció d'on s'ubica la publicitat.

2 Avaluació inicial

Un cop vam haver definit el context d'estudi sobre el projecte que volíem dur a terme mitjançant la tècnica de *web scraping* vam començar el procés d'investigació de la pàgina web en qüestió, [1]: El primer pas en el procés d'avaluació de la pàgina web és comprovar l'arxiu *robots.txt*, [3], [2], ja que si anem a l'adreça web *www.elpais.com/robots.txt* podrem veure les restriccions associades a aquesta pàgina web. Podem observar doncs que en l'arxiu *robots.txt* permet l'accés a tots els *crawlers*, però no ens permeten accedir a les següents URL:

User-Agent: *
Disallow: /buscador/
Disallow: /m/buscador/
Disallow: /pruebas/
Disallow: /.well-known/amphtml/
Disallow: /t/
Disallow: /publicidad/
Disallow: /notificarelacionadas
Disallow: /*.swf$
Disallow: /eskupTSUpdate

A continuació, varem examinar a fons el format de la pàgina web, estudiant totes les etiquetes i identificant en quines s'hi identificaven els atributs que necessitàvem capturar mitjançant el *web scraping* per poder crear el nostre dataset diari. En aquest sentit, també vam estudiar la grandària i la tecnologia de la pàgina web. Vam veure que la grandària era correcta, és a dir que podíem aplicar les tècniques sense por a perdre molta eficiència del codi. A més a més, vam poder observar que la pàgina web en qüestió, [1], havia estat construida mitjançant *NGINX* (**input**: *builtwith.builtwith('http://www.elpais.com')*, **output**: *'web-servers': ['Nginx']*), un servidor HTTP *open-source* de gran rendiment i amb proxy de tipus invers. Per tant, al veure que era un servidor HTTP, vam veure que podríem utilitzar les llibreries bàsiques de python de *web scraping* sense cap problema.

Seguidament vam investigar el propietari de la web. Mitjançant la comanda *whois.whois('http://www.elpais.com')* vam poder veure que el propietari de la pàgina web és *Ascio Technologies, Inc.*, una empresa danesa de tecnologia que té registrades més de dos milions de dominis a internet i és la responsable de la prestació dels serveis d'aquests mateixos dominis.

Un cop llesta l'avaluació inicial, ja vam procedir a configurar el codi per a poder capturar tots els valors dels atributs necessaris per crear els datasets diari i el dataset final amb el contingut de notícies de diversos dies del diari *El País*.

3 Repositori final

El codi que hem creat ens generava un dataset a partir de totes les noticies que el diari mostrava en la seva pàgina principal. El proccés d'extracció es podia produïr més d'una vegada al dia ja que les noticies van canviant a mesura que passa el temps, i així podem analitzar totes i cadascuna de les noticies. Cada extracció ens generava un fitxer amb els mateixos camps, d'aquesta manera podiem anar afegint l'informació extreta en un repositori de notícies final, sobre el qual després podriem procedir a anal·litzar. En aquest repositori hi hem emmagatzemat tots els **atributs** que hem considerat necessaris per poder procedir més tard amb tots els datasets diaris recopilats:

1. **Extraction Date:** La data de la captura de les dades de la pàgina web del diari o dels arxius .html emmagatzemats.
2. **Extraction Hour:** L'hora en la qual s'ha produït la captura de dades de la pàgina web del diari o dels arxius .html emmagatzemats.
3. **Publication Date:** La data de publicació de l'article en el diari.
4. **Publication Hour:** L'hora de publicació de l'article en el diari.
5. **Title:** Títol de l'article publicat en el diari.
6. **Author:** Autor de l'article en qüestió.
7. **Location:** Ubicació del succés que explica la notícia de la que hem emmagatzemat l'hora i dia de publicació, el títol i l'autor.
8. **Num Comments:** Número de comentaris que té l'article realitzats per usuaris registrats en la pàgina web del diari.

9. **Photo Author:** Autor o propietari de la fotografia que acompanya l'article.
10. **Photo Text:** Text del peu de foto de la fotografia que acompanya l'article..
11. **Section:** Secció a la que pertany la notícia emmagatzemada.
12. **Subsection:** Subsecció a la que pertany la notícia emmagatzemada.

Un cop obtingut el dataset final, el repositori de notícies, hem hagut d'emmagatzemar-lo per poder seguir amb l'estudi que teníem planificat. Per poder emmagatzemar-lo hem hagut de tenir en compte quina mena de llicència era la més apropiada pel nostre dataset.

Tenint en compte que el nostre dataset ha estat extret a partir d'una pàgina web pública i els atributs que hem acabat extreient eren d'abast públic. En aquest sentit hem optat per la llicència ***Released Under CC BY-SA 4.0 License***, ja que com hem comentat, partim d'unes dades obtingudes en una web púbilca i que estan a l'abast de qualsevol persona, ja que no s'exclou cap mena de bot per intentar extreure les dades que nosaltres hem extret, per tant permetem que es facin adaptacions del nostre dataset final i permetem que se'n facin usos comercials, [4].

References

[1] Diario **El País**, edición online, *www.elpais.com*.

[2] *Lawson, R. (2015).* **Web Scraping with Python. Introduction to Web Scraping.**

[3] *Subirats, L., Calvo, M. (2019).* **Web Scraping**. Editorial UOC.

[4] **Creative Commons**. PO Box 1866, Mountain View, CA 94042: *https://creativecommons.org/choose/*.