

Scraping a meneame.net

Exercici de scraping PRA1

Autor: Antonio Nogueras

1. Context

Amb l'aparició d'internet han proliferat els diaris electrònics, blogs, webs de vídeos com youtube, xarxes socials, ... Tot això crea un munt d'informació i diferents punts per a recollir-la. Per aquesta raó van néixer els agregadors de notícies, un lloc on es recullen els diferents punts d'informació linkant a les notícies, vídeos, fotos, opinions, ...

En aquest context va néixer meneame.net, un agregador espanyol creat per Ricardo Galli o en es recull informació global de tot el món, i especialment d'Espanya i països de parla hispànica.

A meneame.net els usuaris envien «notícies» per a ser publicades. Un cop arriben a la web, aquestes notícies comencen a rebre vots, positius, negatius, comentaris i d'altres puntuacions que fan que la notícia pugui arribar a la portada o sigui descartada.

La motivació per fer scraping en aquesta web, és recollir les mètriques que afecten a les diferents notícies que ja hi són a portada, i poder entendre una mica com es relacionen aquestes mètriques i si podem treure algun patró que ens indiqui perquè arriben a portada.

Hem treballat programant amb Python a l'entorn Jupyter Notebook. Meneame.net guarda la portada i portades antigues, amb el que només amb una execució podem treure informació de la portada i versions anteriors, el que fa que només amb una execució puguem obtenir el que necessitem.

2. Títol

Scraping i anàlisi de les mètriques generades pels usuaris a meneame.net

3. Descripció del dataset

Titular:

Enunciat de la notícia a meneame.net.

Data creació:

Data en la que l'usuari va enviar la notícia a meneame.net.

Web:

URL de la web de la notícia.

Usuari:

Usuari que va enviar la notícia.

Meneos:

Clics "meneando" la noticia, una manera d'indicar que la noticia t'agrada i que vols que pugi a portada.

Clics:

Clics per visualitzar la notícia.

Comentarios:

Comentarios a la noticia.

Vots Positius:

Vots positius a la noticia.

Vots Anònims:

Vots anònims a la noticia.

Vots Negatius:

Vots negatius a la noticia.

Sub:

Temàtica (a meneame s'anomena sub) de la noticia.

Karma:

Càlcul que fa la pagina web sobre l'estat de "salut" de la noticia, un karma alt fa que pugi a portada, un karma baix fa que surti de portada.

4. Representació gràfica.

100
meneos

menéalo

1397 clics

Militar demanda a 2 policías por abuso de autoridad en EEUU

por **candonga1** a **noticieros.televisa.com** 06:20 publicado

En Estados Unidos, un teniente del Ejército de 27 años, demandó ante un tribunal federal a dos oficiales de la policía de Windsor, Virginia por rociarle gas pimienta en la cara y amenazarlo.

53

47

0


K 378

actualidad

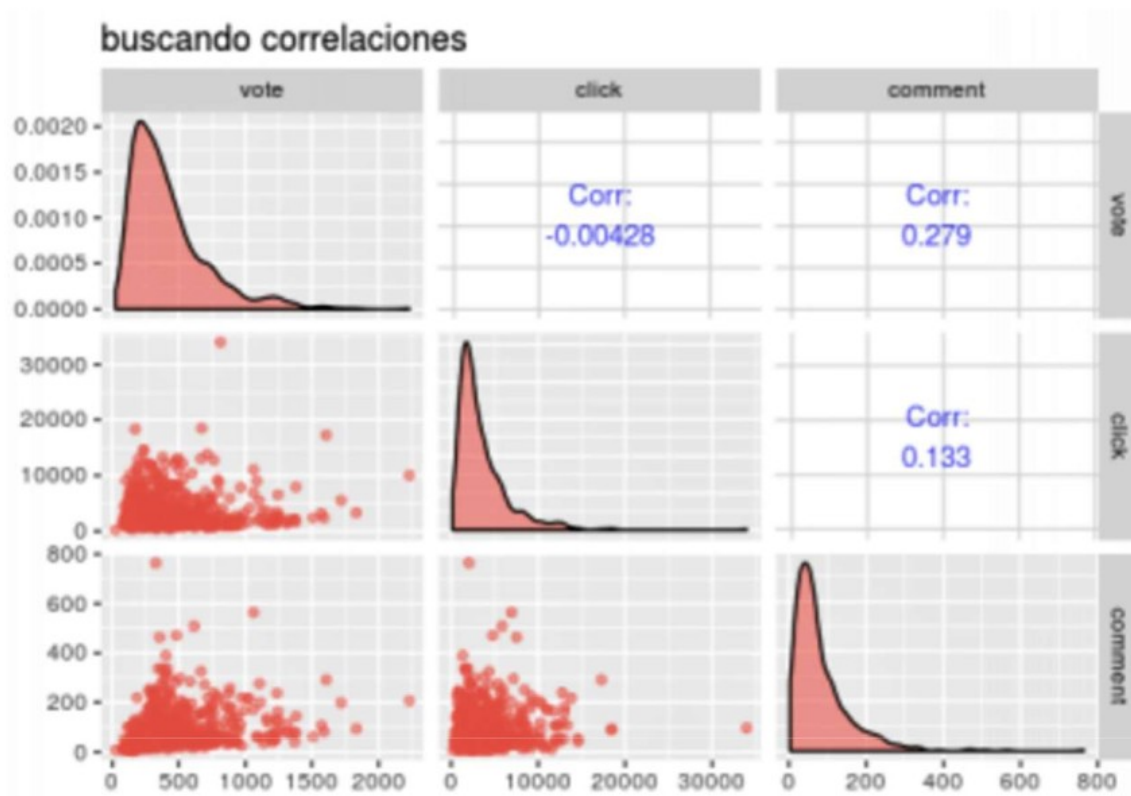
Reportar pro

33 comentarios

compartir



Aquí podem veure una captura de pantalla d'una notícia de meneame.net, amb algunes de les mètriques que recollim.



Gràfiques amb la correlació d'algunes de les variables.

5. Contingut

Els camps del dataset ja s'han descrit al punt 3.

Les dades les podíem treure de la portada directament (meneame.net) o de la cua de notícies noves (meneame.net/queue).

Al final hem decidit fer-ho de la portada directament ja que el que ens interessava era veure dades de notícies a portada, a més el fet que la portada reculli un històric ens permet recollir bastanta informació.

En el nostre cas amb una sola execució recollim la portada i 9 versions anteriors, recollint notícies enviades en diferents dies i diferent horari.

6. Agraïments

El treball de scraping s'ha fet per a llevar a terme una pràctica del màster de Ciència de Dades. Agraïm als propietaris de meneame.net les dades estretes del seu agregador, que es limiten a aquest àmbit, no es fa amb cap fi diferent que no sigui el d'aprendre a fer scraping d'una pàgina web.

Respecte a la font d'inspiració o cerca d'antecedents, després de decidir sobre que volíem fer scraping i amb quin objectiu, buscant a Google vam trobar un estudi fet abans:

<https://wiki.montera34.com/taller-web-scraping-hirakilabs/meneame-titulares>

aquest estudi utilitzava una llibreria urllib2 de Python 2 ja desapareguda.

7. Inspiració

Meneame.net és una web molt popular i un bon punt per a informar-se.

La web, per intentar evitar un esbiaix a l'hora de pujar notícies a portada, i per evitar que això ho faci una persona o un grup de persones, utilitza un algorisme que es basa en els vots, clicks, comentaris, ..., en resum en les dades que estem recollint.

Dins de la pròpia web sempre ha estat una mica polèmic aquest algorisme, a vegades arriben a portada notícies sense gaires comentaris, hi han notícies amb molts clicks i vots positius que són tombades per uns pocs vots negatius, ...

L'objectiu final d'aquest scraping era recollir aquestes mètriques i treure comparacions que ens permetessin intentar entendre una mica més com funciona aquest algorisme.

8. Llicència

Released Under CC0: Public Domain License

Creative Commons és una organització sense ànim de lucre que permet publicar les seves obres creatives sota una llicència que dona més flexibilitat als drets d'autor reservats. De fet és la llicència més utilitzada quan vols publicar a la web la teva obra i poder-la compartir.

9. Codi

https://github.com/JoniRotten/UOC_web_scraping

10. Dataset

<https://zenodo.org/record/4678236#.YHQkoXUzY5k>