Scraping d'un agregador de notícies

Exercici de scraping PRA1

Autors: Antonio Nogueras i Alexandre Casanovas

Context

Els medis tradicionals com els periòdics eren breus, per motius de cost, i el seu nombre era limitat. Amb l'arribada dels medis a la digitalització, la quantitat de dades en els medis i el nombre de medis ha crescut de forma molt intensa, i han aparegut sistemes de resumir les notícies de diferents medis en un únic punt: un agregador.

Aquests agregadors solen oferir, a més, altres serveis, com puntuació de les notícies agregades i de la persona que ha aportat la notícia, o comentaris. Un típic agregador important és Reddit, i una versió en castellà, Menéame, va ser creada per l'informàtic Ricardo Galli, de la Universitat de les Illes Balears: https://www.meneame.net/m/Preg%C3%BAntame/soy-ricardo-galli-socio-fundador-programador-ex-administrador

i altres.

L'objectiu del scraping és analitzar les relacions entre paràmetres que apareixen en un agregador com el nombre de comentaris, els vots positius o vots negatius.

Ens hem centrat en un agregador, meneame.net, que proporciona un paràmetre que qualifica a l'usuari que aporta la notícia, que denominen karma, juntament amb uns altres com "meneos", vots positius, negatius i anònims, clics, una variable qualitativa que classifica el tema, l'usuari que aporta la notícia o la data d'aportació, entre altres.

Concretament volem comprovar si existeix alguna relació entre el karma i els altres paràmetres, per exemple, si més karma suposa més clics o més vots positius. El karma és com un carnet de punts: el karma inicial d'un usuari nou és baix, i pot pujar (o baixar) en funció de l'èxit de les aportacions.

La dificultat d'analitzar un agregador és que és un món canviant: clics, notícies, vots, meneos, ... canvien constantment, així que convé prendre una instantània el més àmplia possible, ja que en diferents temps compararem coses diferents. Especialment cal evitar notícies duplicades: en la cua d'entrada ('noves' en meneame.net) aquests duplicats són freqüents i s'insisteix en la mateixa notícia.

Hem treballat en Jupyter de Anaconda3 i Python amb les llibreries específiques request i Beautiful Soup. El que s'adjunta és el fitxer .ipynb i els fitxers d'eixida, en csv, per exemple.

Executant el .ipynb es pot llegir directament el dataframe que s'obté d'escanejar uns 250 notícies úniques i l'anàlisi de gràfiques i resultats que realitzem. Ens hem limitat a aqueixa quantitat per a no carregar la web amb una lectura successiva de moltes pàgines; això és només un exercici de scraping.

Als agregadors hi ha moltes entrades i canvien contínuament, de forma per trobar el que busquem suposa molt de temps i esforços. Si volem analitzar les puntuacions o el karma: un algorisme que valora les

aportacions de cada persona, en meneame.net seria quasi impossible perquè les dades varien contínuament; almenys, el 'clics'. El scraping ofereix una possibilitat de fer-ho.

Si volem analitzar dades sociològiques com que temes tenen més clics, vots positius o comentaris, cal així mateix fer scraping del agregador.

A més està el fenomen dels bots i de les fakes news. Medis com maldita.es o newtral.es es dediquen a verificar notícies per a la qual cosa han de fer un ús intens de la ciència de dades: per exemple, busquen l'origen real de les imatges d'unes fake news. Aquests fenòmens també apareixen als agregadors. El scraping ofereix així mateix la possibilitat de mesurar la incidència d'aquests fenòmens als agregadors.

Titol

"Scraping d'agregadors: el cas de meneame.net".

Descripció del data set

Al cas de meneame.net, s'obté 'Titular', 'URL' (l'enllaç), 'Usuari', 'Entrada', 'Meneos', 'Clics', 'Comentaris' (en nombre),'Vots Positius', 'Vots Anònims', 'Vots Negatius', 'Categoria','Karma' (puntuació de l'aportador), 'Data Submissió'.

Recull 10 pàgines web, que suposen 250 entrades no repetides.

```
menDF= pd.read_excel(excel)
# Comprovem el
menDF.head()
```

run cell, select below

Out[129]:

	Titular	URL	Usuari	Entrada	Meneos	Clics	Comentaris	Vots Positius	Vots Anònims	Vots Negatius	Categoria	Karma	Data Submissió
0	Comunicado Sociedad Española de Trombosis y H...	https://www.seth.es/index.php/noticias/noticia...	Inutil	Desde la Sociedad Española de Trombosis y Hemo...	38	363	8	30	8	5	actualidad	449	11-04-21 23:59
1	El explorador muere en el hielo	https://www.revistamercurio.es/2021/04/11/el-e...	Tieso	15 de enero en la Antártida. Pleno verano con ...	24	575	1	20	4	0	cultura	440	11-04-21 12:38
2	Cuando el banco se convierte en juez y 'conde...	https://www.elconfidencial.com/amp/espana/2021...	Octaviano	Una llamada, una carta y, de la noche a la mañ...	57	249	12	35	22	0	actualidad	434	11-04-21 21:40
3	La Comunidad de Madrid suspende la actividad ...	https://www.europapress.es/madrid/noticia-comu...	unmundofeliz	El Gobierno regional ha explicado en un comuni...	172	278	13	103	69	0	Rescates	428	11-04-21 20:26
4	13 películas imprescindibles de ciencia-ficci...	https://www.xataka.com/cine-y-tv/13-peliculas-...	ChanVader	Hay un tópico asociado férreamente al género d...	77	3155	41	54	23	11	actualidad	527	11-04-21 22:22

In [130]:
```
menDF.tail()
```

Out[130]:

	Titular	URL	Usuari	Entrada	Meneos	Clics	Comentaris	Vots Positius	Vots Anònims	Vots Negatius	Categoria	Karma	Data Submissió
245	La historia de la imagen que plasma la	https://www.xataka.com/otros/historia-imagen-q...	NubisMusic	Que las hazañas y momentos históricos	247	5695	65	114	133	4	cultura	440	06-04-21 14:01

Contingut

Apareix les dades de 10 pàgines de l'agregador del fil principal de https://www.meneame (hi ha així mateix la cua d'entrada https://www.meneame.net/queue que porta menys informació perquè hi ha moltes notícies repetides).

Finalment hem triat el fil principal de https://www.meneame al llarg de 10 pàgines, uns 250 agregats, perquè en fer scraping de més pàgines (hem provat 30) ens censuren i no tenim resposta. Podíem gastar un time delay, però considerem que no canvia l'exemple de scraping, i hi ha el perill de què, entretant, canvien clics, meneos, dades.... entre pàgines. A més, no té sentit comparar el nombre de clics de notícies al llarg d'una setmana amb les d'ahir, amb duració d'un dia.

Hem comprovat que la fulla de pàgina zero de meneame.net és una selecció de 25 agregats dels agregats acceptats al fil principal de meneame. Per evitar repeticions, comencem a la pàgina 1 fins a la 11.

Al cas de meneame les dades són seleccions de la cua d'entrada basada en meneos, clics, vots positius,... etc. i l'eixida estàndard és un fitxer format .csv, amb delimitador '.', i en el seu cas (versió 2) en format Excel .xlsx. Internament l'estructura bàsica és el datagrama.

El scraping que fem és una mena d'instantània que podem repetir en el temps i agregar als scrapings anteriors. Els datagrames poden encadenar-se via append. El nom de cada ú dels fitxers d'eixida inclou la data per conèixer el moment del scraping.

Hem treballat en Jupyter d'Anaconda3 i Python amb les llibreries específiques request i Beautiful Soup. El que s'adjunta és el fitxer .ipynb i els fitxers d'eixida, en csv, o .xlsx, en el seu cas.

El dataFrame permet fer gràfiques i anàlisis variades. Executant el .ipynb es pot llegir directament aquest dataFrame que s'obté d'escanejar unes 250 notícies úniques i l'anàlisi de gràfiques i resultats que realitzem. Ens hem limitat a aqueixa quantitat per a no carregar la web amb una lectura successiva de moltes pàgines; això és només un exercici de scraping.

El karma és un paràmetre característic de meneame.net. Ens interessava descobrir alguna relació del karma amb els altres paràmetres. Concloem, entre altres coses, que el valor de karma de les notícies del fil principal no sembla dependre dels altres paràmetres, ja que ha de ser obtingut probablement per un període extraordinàriament llarg d'aportacions; el karma mitjà d'aquests aportadors pot estar en un valor de 350.

Hi ha una versió 2: meneame_scrapingV2.ipynb per al cas de fallades en Windows 10. Produeix a més un fitxer Excel de les dades.

Agraïments

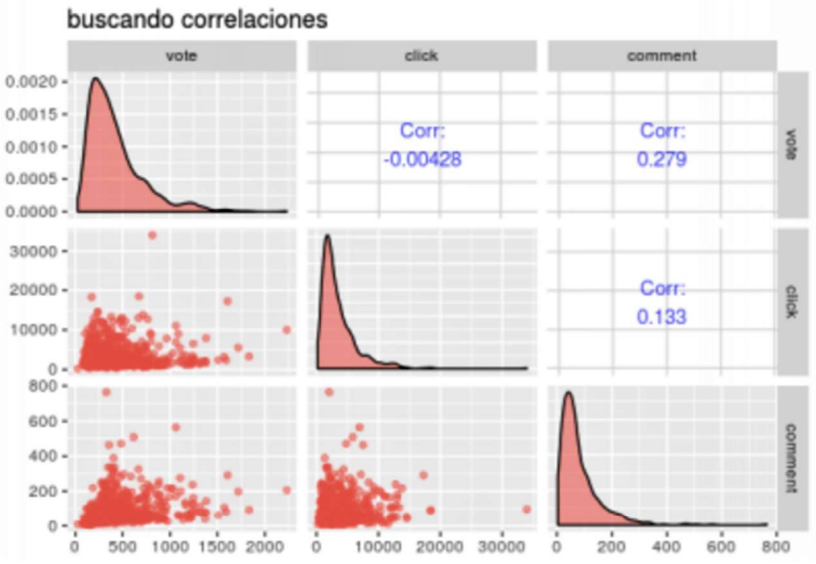
El treball de scraping s'ha fet per a llevar a terme una pràctica del màster de Ciència de Dades. Agraïm als propietaris de meneame.net les dades estretes del seu agregador, que es limiten a aquest àmbit. Les notícies en si no són seves, soles l'estructura i tractament dels enllaços als medis originals.

Respecte a la font d'inspiració o cerca d'antecedents, en cerca en Google trobarem un estudi de https://wiki.montera34.com/taller-web-scraping-hirikilabs/meneame-titulares

que gastava una llibreria urllib2 de Python 2 ja desapareguda, i l'interès era fer gràfiques entre paràmetres com, potser, karma i clics de meneame.net:

```
ggpairs(meneame,
        columns = c(2:4),
        title = "buscando correlaciones",
        # upper = list(continuous = "density"),
        aes(alpha = 0.1))
```

Esta es la visualicación que genera:



Podemos ver los diagramas de dispersión que comparan las tres variables numéricas y los coeficientes de correlación calculados.

taller-web-scraping-hirikilabs/meneame-titulares.txt · Última modificación: 2018/02/07 15:45 por numeroteca

Inspiració

Els motius que ens han portat a triar els dos casos estan ja anomenats en el punt 1 Context:

> "volem comprovar si existeix alguna relació entre el karma i els altres paràmetres, per exemple, si més karma suposa més clics o més vots positius...
>
> un agregador és que és un món canviant: clics, notícies, vots, meneos, ... canvien constantment, així que convé prendre una instantània el més àmplia possible ... Als agregadors hi ha moltes entrades i canvien contínuament, ... analitzar les puntuacions ... seria quasi impossible perquè les dades varien contínuament ... el scraping ofereix una possibilitat de fer-ho.
>
> "Si volem analitzar dades sociològiques ..., cal així mateix fer scraping del agregador".
>
> ... Aquests fenòmens (bots i fakes news) també apareixen als agregadors. El scraping ofereix ... la possibilitat de mesurar la incidència d'aquests fenòmens als agregadors.
>
> Hem triat aplicar l'exercici de scraping al agregador meneame.net com una activitat actual i com una forma d'analitzar les seves notícies."

Els agregadors tenen moltes visites en la web (meneame.net és la quarantena del món en la categoria 'news and media' (https://www.similarweb.com/website/meneame.net/) i hem vist que el scraping dels agregadors és un tema d'interès, i una idea per al punt de partida de la pràctica.

En llibres com el de Laszlo i en la xarxa podem veure com aplicar la llibreria Beautiful Soup.

Llicència

És difícil conèixer ben bé les diferències entre les diverses llicències que són bàsicament un tema legal, però triem CC0 Public Domain License per al GitHub:

> *"CC0 helps ... creators a way to waive all their copyright and related rights in their works to the fullest extent allowed by law. CC0 is a universal instrument that is not adapted to the laws of any particular legal jurisdiction, similar to many open source software licenses ... provides the best and most complete alternative for contributing a work to the public domain given the many complex and diverse copyright and database systems around the world."*

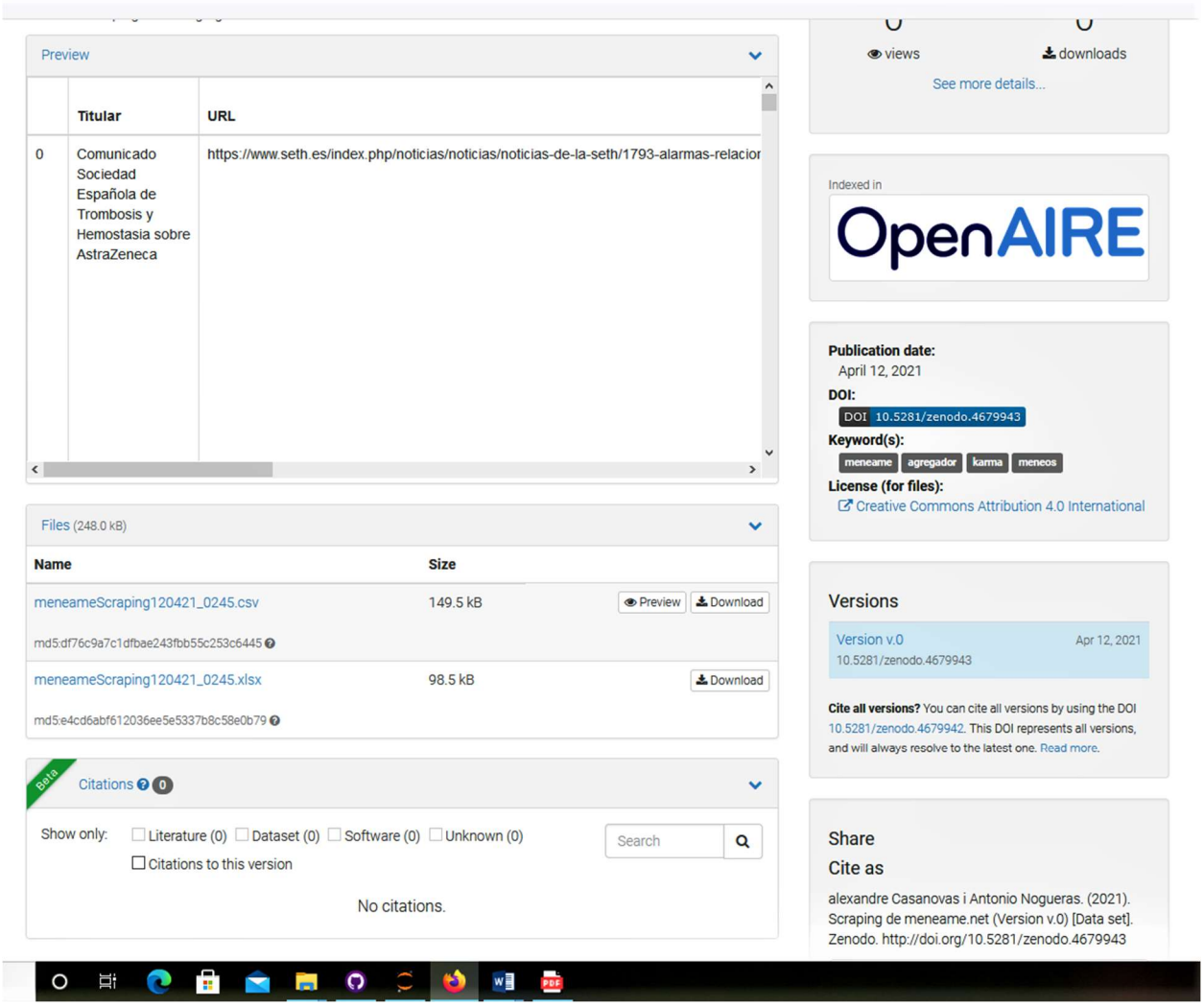
que seria també l'elecció si fora una base pública, si bé depèn dels drets legals dels agregadors estudiats. Com a pràctica, no necessita una llicència, la UOC es reserva els drets de còpia dels treballs presentats.

Codi

Presentem el codi en fitxers .ipynb de Jupyter en Python que s'adjunten, així com el fitxer .csv d'eixida (i en el seu cas, el .xlsx d'Excel).

Hi ha una versió 2: meneame_scrapingV2.ipynb per al cas de fallades en Windows 10. Produeix a més un fitxer Excel de les dades.

Dataset



SIGNAT PER ORDRE ALFABÈTIC:

Contribucions	Signa
Recerca prèvia	AC, AN
Redacció de les respostes	AC, AN
Desenvolupament codi	AC, AN