Web Scraping: meneame.net

XAVIER JORDÀ MURRIA

ANNA SERENA LATRE

Màster Data Science UOC

ABRIL 2019

índex

PREGUNTA 1.

Context. En quin context s'ha recol·lectat la informació. Explicar perquè el lloc web proporciona aquesta informació.

La proposta que fem és la de rastreig de la website mename.net:

Mename.net (2019) Menéame peta. Menéame Comunicacions, SL. Simó Ballester 9 bajos 07011 Palma. Disponible a:

www.Meneame.net (2019)

Menéame es una xarxa social amb caràcter lúdic pensada perquè la comunitat d'usuaris puguin compartir enllaços referents a notícies i articles d'actualitat així com la publicació de notes expressant opinions, experiències i punts de vista.

Exposem el context de recol.lecció en dues fases, una primera orientada a la recerca d'informació per a la documentació de la website i l'altra fase orientada a l'aplicació d'una de les tècniques de rastreig de websites proposades en un procés de cerca selectiva i avançada per a la recollida de dades.

- **FASE 1: RECERCA documentació de la website Meneame.net.**

Informació de la website per tal de fonamentar l'actuació en un procés documentat amb garanties d'èxit.

a) **Fitxer robots.txt.**
b) **Mapa web Sitemap.**
c) **Profunditat de la website.**
d) **Tecnologia website.**

a) Fitxer robots.txt.

En el context situem algunes dades d'interès com la localització dels fitxers robots.txt de Meneame.net, a través del motor de cerca:

www.meneame.net/robots.txt

Constatem que la website en els fitxers robots.txt meneame.net dóna a conèixer quins directoris i pàgines deshabilita per als developers:

```
User-agent: *
Disallow: /search
Disallow: /between
Disallow: /login
Disallow: /shakeit.php
Disallow: /index.php
Disallow: /profile.php
Disallow: /between.php
Disallow: /login.php
Disallow: /submit.php
Disallow: /trackback.php
Disallow: /editlink.php
Disallow: /backend/
Disallow: /api/
Disallow: /index.php
Disallow: /comments_rss2.php
Disallow: /rss2.php?
Disallow: /javascript:
Disallow: /comments_rss2.php Disallow: /link_bookmark.php
Disallow: /search.php
Sitemap: http://www.meneame.net/sitemap
User-agent: Mediapartners-Google
Disallow:
```

Trobem que la pàgina Meneame.net inhabilita sis fitxers i catorze directoris a l'accés dels usuaris developers.

b) Mapa web Sitemap.

El mapa web Sitemap de meneame.net ens refereix a dos directoris l'statics i el last que són llocs de la website que visitarem per tal d'accedir a la informació rellevant de la pàgina.

https://www.meneame.net/sitemap

Vegem-ho:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<sitemapindex xmlns="https://www.sitemaps.org/schemas/sitemap/0.9"> <sitemap>
<loc>https://www.meneame.net/sitemap?statics</loc> </sitemap> <sitemap>
<loc>https://www.meneame.net/sitemap?last</loc> </sitemap> </sitemapindex>
```

En la direcció statics que ens proporciona el mapa web Sitemap se'ns refereix a un total d'onze directoris relacionats amb la website, i que aporten informació de caràcter estructura de la web en relació les pàgines en les quals podem navegar en la web.

- https://www.meneame.net/sitemap?statics

```
<urlset xmlns="https://www.sitemaps.org/schemas/sitemap/0.9">
<url>
<loc>https://www.meneame.net/</loc>
<priority>1.0</priority>
</url>
<url>
<loc>https://www.meneame.net/queue</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/sneak</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/notame/</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/cloud</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/popular</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/top_commented</loc>
<priority>0.8</priority>
```

```
<url>
<loc>https://www.meneame.net/top_comments</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/top_users</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/legal</loc>
<priority>0.8</priority>
</url>
<url>
<loc>https://www.meneame.net/faq-es.php</loc>
<priority>0.8</priority>
</url>
</urlset>
```

Els directoris de pàgines vinculats a la web Meneame.net ens poden ser de gran utilitat perquè aporten informació rellevant del funcionament de la website.

Queue: notícies cua d'espera per tal que siguin publicades en la pàgina central o bé eliminades.

https://www.meneame.net/queue

Sneak: streaming source de l'activitat de la website.

https://www.meneame.net/sneak

Notame: streaming de les notes personalitzades dels usuaris registrats.

https://www.meneame.net/notame

Cloud: gràfic amb les paraules de més notorietat en les publicacions.

https://www.meneame.net/cloud

Popular: llista de les notícies amb més èxit, amb màxima notorietat en meneos, clics i vots.

https://www.meneame.net/popular

Top_commented: llista de les notícies amb més comentaris.

https://www.meneame.net/top_commented

Top_comments: llista dels comentaris més votats.

https://www.meneame.net/top_comments

Top_users: estadístiques dels usuaris més rellevants en quant a activitat karma, notícies, noticíes publicades, comentaris i vots.

https://www.meneame.net/top_users

Legal: informació legal i condicions d'ús.

https://www.meneame.net/legal

Faq: preguntes freqüents que es fan els usuaris que participen en la xarxa.

https://www.meneame.net/faq-es.php

En la direcció statics que ens proporciona el mapa web Sitemap se'ns exposa links que refereixen les notícies exposades en la pàgina central de la website que ja no són actives.

Vegem-ho:

- https://www.meneame.net/sitemap?lasts

c) Grandària de la website (Google).

La website meneame.net denega l'accés a la grandària de la website.

www.meneame.net/site

Dóna error: not found error 404, meneame peta te habíamos avisado.

En una cerca avançada al crawler de Google (https://www.google.com/advanced_search) obtenim un total de 176.000.000 cerques vinculants ja que meneame.net genera gran quantitat de links compartits amb mitjans de comunicació online i altres xarxes socials a mans dels usuaris.

https://www.google.com/search?as_q=meneame.net&as_epq=&as_oq=&as_eq=&as_nlo=&as_nhi=&lr=&cr=&as_qdr=all&as_sitesearch=&as_occt=any&safe=images&as_filetype=&as_rights=

d) Tecnologia utilitzada en la website .

La tecnologia usada en la construcció de la web meneame.net, és una informació que aconseguim amb l'ajuda de la llibreria builtwith.

```
pip install builtwith

import builtwith

print(builtwith.builtwith('https://www.meneame.net'))
```

```
{'web-servers': ['Nginx'],

'advertising-networks': ['DoubleClick for Publishers (DFP)'], 'font-scripts': ['Font Awesome', 'Google Font API'], 'javascript-graphics': ['Javascript Infovis Toolkit'], 'javascript-frameworks': ['RequireJS',
```

```
'Twitter typeahead.js', 'jQuery'],
'web-frameworks': ['Twitter Bootstrap']}
```

En la identificació de la tecnologia present en la website observem com meneame.net presta col.laboració amb empreses publicitàries i consta en el registre 'advertising-networks': ['DoubleClick for Publishers (DFP)']. Els fonaments de la xarxa social van orientats en última instància a sustentar aquesta activitat lucrativa:

CIF:B57466229
Registro Mercantil de Illes Balears: Tomo 2263, Libro 0, Folio 181, Hoja PM-57684, Inscripción 1.

També s'identifica el servidor web, així com l'empresa que dona suport a la infraestructura de la website 'Google Font API', i altres tecnologies usades com el llenguatge java i el software per a consultes 'jQuery'.

- **FASE 2: CRAWLING – Rastreig de la website Meneame.net.**

Estratègia de cerca selectiva de paràmetres automatitzada amb els codis python segons l'estructura html de la website, donat el cas que volem obtenir la informació publicada en la pàgina central de la web com a informació actualitzada i en clau qualitativa per a una primera aproximació a la comprensió del funcionament de la website. Els passos de la cerca selectiva de paràmetres automatitzada són:

a) **Entorn i llibreries**
b) **Iteració selectiva**
c) **Cerca avançada selectiva**

a) Entorn i llibreries:

Instal.lació de l'entorn i descàrrega de les llibreries Python per a una cerca i recol.lecció de dades automatitzada mitjançant la tècnica Web Scraping.

Entorn :

```
pip install requests
pip install beautifulsoup4
pip install news-please
```

Llibreries:

```
# PYTHON LYBRARIES FOR AN AUTOMATED WEB SCRAPING DATA COLLECTION
from bs4 import BeautifulSoup
import requests
from News import News
```

Requests és una llibreria software opensource de Python que ens possibilita la realització de peticions de consulta i la conseqüent transferència de dades de la resposta del servidor.

BeautifulSoup és una llibreria software opensource de Python amb eines per al rastreig i processament de dades en cerques html selectives en websites.

News és una llibreria software opensource de Python per a l'extracció de textos ja siguin notícies, articles, ... capaç de diferenciar l'estructura d'hyperlinks i habilitar el rascleig en opcions de filtre inserida en les tècniques pròpies del WebScraping.

b) Iteració selectiva:

Iterem en l'estructura html de la pàgina meneame.net, accedint a l'opció d'Inspecció amb el botó dret del ratolí per identificar certs paràmetres estructurals d'interès en relació a la cerca avançada de la classe.

En la inspecció de l'estructura html de la website trobem aquests paràmetres d'interès que són:

'a', 'div', 'span', 'h2'.

c) Cerca avançada:

En la cerca avançada selectiva en l'estructura html de la pàgina web, ens interessa la inspecció dels links de les notícies en portada per tal d'identificar la classe i els tokens en l'estructura dels links habilitats per a procedir amb les indicacions pertinents a la funció de cerca beautifulsoup.

Trobem que aquestes classificacions i tokens d'interès són:

'news-summary', 'clics', 'href', 'h2' 'text', 'news-content', 'news-deatils-data-up', 'votes', 'votes-down', 'votes-up', 'wideonly votes anonymous', 'tool sub-name', 'news-submitted', 'tool-subname', 'karma', 'comments'.

Partim de la notícia com a nucli d'interès objecte d'estudi i recol.lecció de dades.

PREGUNTA 2.

Definir un títol per el dataset. Triar un títol que sigui descriptiu.

Dataset: meneame_news.csv

El títol del dataset és *'meneame_news'* en referència a la llista notícies d'actualitat presentades en la pàgina central de la web meneame.net i que són les més rellevants i exitoses en el moment de la recollida de dades segons l'algorisme karma i a proposta de la comunitat d'usuaris.

PREGUNTA 3.

Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat):

meneame_news recull informació de les notícies publicades en primera pàgina amb caràcter central en la website. Hem procedit a una cerca de la informació a través dels paràmetres en l'estructura html de la website que hem hagut d'inspeccionar amb profunditat.

En la cerca selectiva hem pres la notícia com a entitat i nucli en el qual es focalitza la nostra cerca, destriant dotze atributs del camp notícia en pàgina central que presentem a continuació:

- clics: visites dels usuaris per a l'obtenció de la reproducció de la notícia segons la font d'informació.
- meneos: és a promoció efectiva de les notícies no publicades a mans dels usuaris registrats o anònims, en una llista d'espera amb ordre descendent segons la data de publicació i que seran publicades en la pàgina central segons els vots i la seva ponderació en un sumatori global que és el karma de la notícia.
- contentSumary: text que resumeix la notícia.
- title: titular de la notícia en format text.
- titleRef: link que és el directori que ens refereix a la notícia.
- category: classificació de la notícia segons la temàtica.
- votesUp: vots d'aprovació de la notícia que poden emetre els usuaris amb un karma mínim, i el valor d'aquest vot positiu val 6*karma personal.
- votesDown: vots de desaprovació que poden emetre els usuaris registrats amb un karma mínim, i el valor d'aquest vot negatiu val 6*karma personal.
- votesAnonymous: els vots anònims d'usuaris no registrats valen 6.
- newsPaper: s'identifica la font d'informació de la qual prové la noticía publicada en la llista d'espera o en la pàgina central.
- karma: és la valoració de la notícia en la suma de tots els resultats globals de vots que s'hi refereixen segons els usuaris amb les respectives ponderacions. És un algorisme complex que trobem explicitat en la website.
- comments: els usuaris amb un karma >17 en la categoria d'especials, o bé els usuaris admin, blogger i god estan habilitats per a la realització de comentaris.

En la selecció hem eludit la data de publicació, l'hora de publicació, la identificació de l'usuari que la publica, el compartiment de la publicació en altres xarxes socials com twitter i facebook, l'enviament per e-mail i els copy efectuats en la notícia.

En el cicle de vida de les notícies, aquestes són proposades per usuaris registrats i l'input es manté en la cua en un directori anàleg esperant la valoració de l'algorisme karma per tal d'entrar en aquesta pàgina central d'actualitat de meneame.net.

La valoració de l'usuari registrat obté un karma en valor de sis i en cas que publiqui notícies obté una valoració de 0 a vint segons la qualitat de les emissions en els resultats (vots positius sumen karma, vots negatius resten i vots rebuts sobre comentaris i notes, la inactivitat resta). En cas d'obtenir un karma > 17 s'obté la categoria d'usuari especial amb la facultat d'editar i descartar notícies de la pàgina central. L'usuari admin, blogger i god no es veuen afectats per la variació del karma i el seus status és permanent. Els anys d'antiguitat també són un premi per als usuaris de meneame.net.

També cal destacar que els vots estan habilitats en una vigència de 30 dies de vida de la notícia i que els usuaris poden ser deshabilitats en cas de procedir amb un comportament no desitjat.

PREGUNTA 4.

Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment.

Procedim a una exposició gràfica dels fluxos segons el cicle de vida de la notícia en portada i els seus atributs.

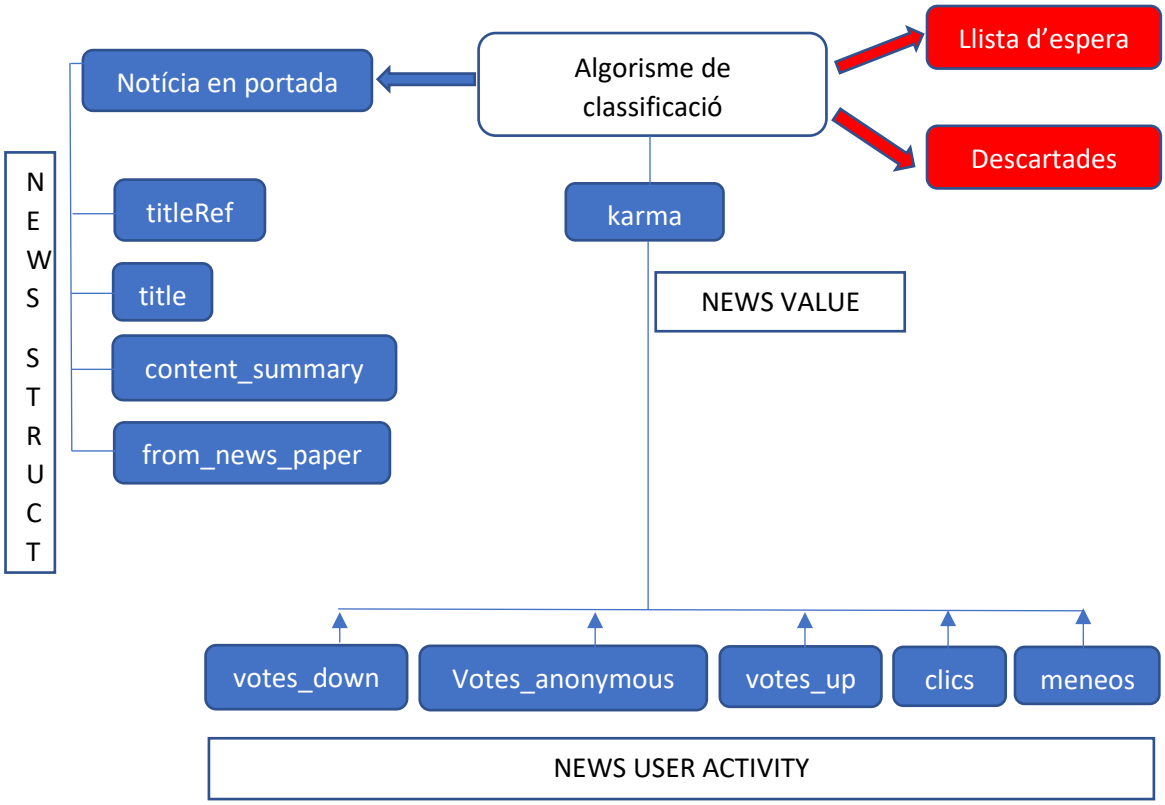

Figura 1. Representació gràfica dels fluxos en el cicle de vida de la notícia.

Els atributs de la notícia en portada venen representats en blau i estan agrupats segons la seva funcionalitat entorn al nucli notícia en portada: estructura, activitat i valor.

Els atributs referents a l'estructura són estables i tenen un caràcter permanent en el cicle de vida de la notícia en totes les seves fases : ja sigui en llista d'espera, en portada o bé descartades.

Els atributs relatius a l'activitat amb caràcter quantitatiu reben uns inputs provinents de l'activitat dels usuaris de la website, de manera que van patint modificacions constantment. Els inputs o modificacions en l'activitat generen uns fluxos que van a la variable karma o el que és l'algorisme de valoració de la notícia segons criteris comentats anteriorment.

Observem com l'atribut karma ocupa una posició central que també té connexió amb l'algorisme de classificació de les notícies segons el seu karma són dos filtres capaços de gestionar les notícies en la llista d'espera, en portada o bé descartades, camps que no existeixen en el dataset perquè són propis de l'administrador i hem volgut posar de relleu per tal d'observar la importància dels algorismes karma i de classificació en la gestió de notícies al llarg del seu cicle de vida en la website.

PREGUNTA 5.

Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

Presentem un esquema del disseny conceptual, lògic i físic de l'estructura relacional del dataset en el seu entorn immediat de manera simplificada, en identificació del camp New_list com a dades públiques de la notícia que posicionem en una taula perifèrica donat que la dimensió karma és el camp que ocupa un lloc central en l'estructura relacional:

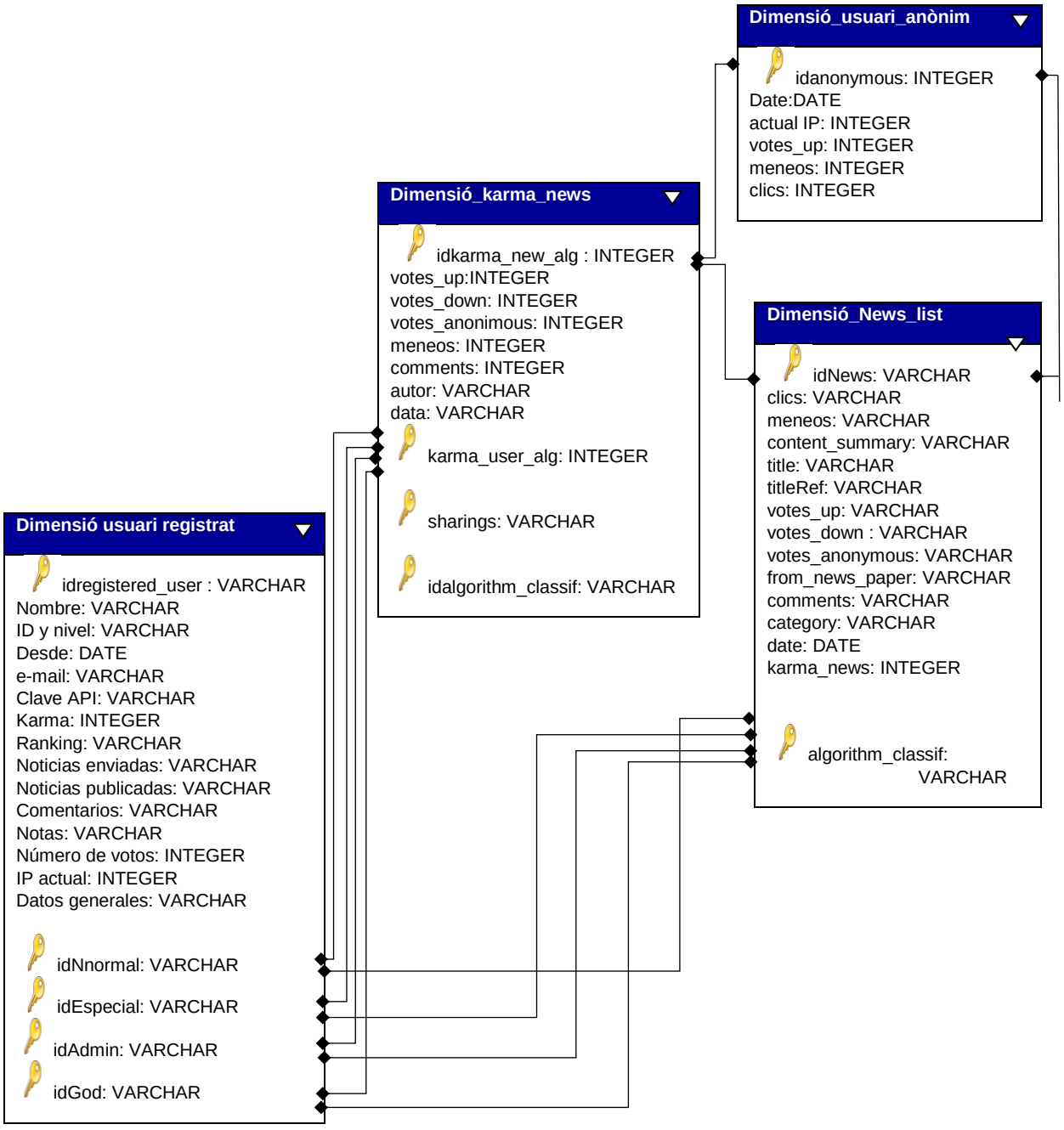

Figura 2. Esquema de l'estructura relacional del camp de la notícia publicada en pàgina central (Font: elaboració pròpia).

És rellevant destacar l'algorisme karma que procedeix a una valoració de la notícia en un pla individualitzat, però comentar que a nivell d'usuari també existeix un karma individualitzat en funció de l'activitat de l'usuari i les valoracions que rep, la seva antiguitat així com del seu estatus en funció de la classificació de l'usuari en normal, especial, admin i god, que tot plegat el que fa es ponderar el seu vot en funció de la seva vàlua tant si s'emet en positiu com negatiu.

Vegeu: https://www.meneame.net/values

Les dades que recollim responen a un cicle de vida de la notícia en portada que expira en el termini màxim de trenta dies. El temps d'execució i recollida de les dades és inferior als cinc segons.

PREGUNTA 6.

Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Utilitzem la llibreria whois de python per a la presentació del conjunt de dades:

```
Pip3 install python-whois

import whois

print(whois.whois('https://www.meneame.net'))
```

La funció whois ens mostrarà el propietari de la web en el paràmetre "org.":


```
{
  "domain_name": [
    "MENEAME.NET",
    "meneame.net"
  ],
  "registrar": "CSL Computer Service Langenbach GmbH d/b/a joker.co
m",
  "whois_server": "whois.joker.com",
  "referral_url": null,
  "updated_date": [
    "2018-07-20 18:28:23",
    "2018-07-20 18:28:13"
  ],
  "creation_date": [
    "2005-11-30 12:47:58",
    "2005-11-30 12:47:59"
  ],
  "expiration_date": "2020-11-30 12:47:58",
  "name_servers": [
    "NS-1116.AWSDNS-11.ORG",
    "NS-172.AWSDNS-21.COM",
    "NS-1982.AWSDNS-55.CO.UK",
    "NS-799.AWSDNS-35.NET",
    "ns-1982.awsdns-55.co.uk",
    "ns-799.awsdns-35.net",
```

```
    "ns-172.awsdns-21.com",
    "ns-1116.awsdns-11.org"
  ],
  "status": "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
  "emails": "abuse@joker.com",
  "dnssec": "unsigned",
  "name": null,
  "org": "Meneame Comunicacions, SL",
  "address": null,
  "city": null,
  "state": "Balears",
  "zipcode": null,
  "country": "ES"
}
```


El propietari de la web i dels continguts que els usuaris publiquen és la companyia balear Meneame Comunications, SL, i segons la wikipedia.org (2019) a mans dels seus fundadors Ricardo Galli i Benjamí Villoslada junt amb el seu soci capitalista Martín Varsavsky a parts iguals. Hem de dir però que en relació al contingut de les notícies ens hem de referir als mitjans de comunicació responsables de la seva publicació.

PREGUNTA 7.

Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Perquè és interessant el conjunt de dades ?

El conjunt de dades que hem seleccionat en la web meneame.net és d'especial interès per a entendre el funcionament de la pàgina des de la perspectiva d'usuari i la seva experiència, donat que la notícia es converteix en l'objecte lúdic dels usuaris i com a tal pateix uns cicles de vida immersos en unes regulacions com a normes de joc.

El conjunt de dades vinculats amb la notícia és interessant perquè la valoració dels paràmetres individualitzats i contrastats en el conjunt de totes les notícies, és l'algorisme que gestiona la pàgina central de la web, de manera que són els usuaris amb la seva participació activa en l'objecte notícia els que acaben determinant quines són les notícies en portada així com els que marquen els índexs d'activitat en la website.

Quines preguntes es pretenen respondre ?

En aquest cas la recol.lecció del conjunt de dades en base al nucli notícia en portada és molt vàlid com a peça angular per a l'exposició de la complexitat relacional que existeix entre els camps i les dades en el funcionament de la web, donat que la gestió de la pàgina es duu mitjançant algorismes que valoren la notícia en base a l'activitat dels mateixos usuaris en els paràmetres d'actuació. Aquests algorismes, el de valoració de la notícia que alhora exigeix una valoració dels usuaris sota el nom de karma, i el de classificació de les notícies en portada segons el seu karma en una visió de conjunt del total de les notícies, no són de caràcter públic de

manera que no hi tenim accés i nosaltres hem procedit amb una aproximació en una versió simplificada del que podrien ser les dimensions i les relacions en funció de les indicacions de funcionament del algorisme que es donen en la website.

Les preguntes que ens hem fet entorn al funcionament i gestió de les dades posen en evidencia que el funcionament de la xarxa social meneame.net està pensat per a protegir els interessos dels usuaris registrats en el desenvolupament de la seva activitat a nivell de grup, i per a aquests efectes les normes en funcionament que regulen la valoració del karma i classificació de les notícies si que queden ben especificats en la website, de manera que només els usuaris registrats poden publicar notícies, fer comentaris, emetre vots negatius i en funció del seu estatus poden suprimir notícies i publicar notes. En aquesta lògica es desprèn que els vots negatius dels usuaris registrats tenen més pes.

Per altra banda la website realment és un negoci el funcionament del qual se sustenta en les premisses de maximització de les ràtios de participació per a una revalorització de les quotes publicitàries, per tant en una visió de negoci a llarg termini de manera que els algorismes en última instància defensen la consecució d'aquests propòsits en unes regles que no queden explicades amb prou claredat en relació a les potestats especials d'usuaris registrats com els admin i god, i que els càlculs algorísmics no són explicitats.

PREGUNTA 8.

Llicència.

La llicència utilitzada en la website meneame.net per qüestions de contingut explicitat en la website és la CC BY 3.0 ES:

https://creativecommons.org/licenses/by/3.0/es/

El tema 3.0 apareix en associació a aspectes de contingut i les sigles ES per reconeixement de la legislació en el país d'actuació de la website, de manera que el codi al qual ens refereixen seria en primer de la llista que proposeu:

> Released Under CC0: Public Domain License.

En altres paraules en relació a la llicència de continguts de text, en la website es reconeix qualsevol persona amb dret d'ús dels continguts publicats en la website per a les finalitats que estimi més oportunes i convenients sempre i quan no s'atribueixin continguts falsos als usuaris de la xarxa social meneame.net.

PREGUNTA 9.

Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python.

```
# PYTHON CODE FOR AN AUTOMATED WEB SCRAPING DATA COLLECTION
from bs4 import BeautifulSoup
import requests
from News import News
from FileWriterNewsCSV import FileWriterNewsCSV
```

```
class Scraper:

    def __init__(self):
        self.url = "https://www.meneame.net"

    def __download_html(self, url:str):
        response = requests.get(url)
        html = response.content
        return html

    def scrape(self):

        print("Web Scraping of planes crashes data from " + self.url + "...")

       # Download HTML:
        html = self.__download_html(self.url)
        soup = BeautifulSoup(html, 'html.parser')
        rest = soup.find_all("div", {"class": "news-summary"})
        news_list = []

        for link in rest:
            clics = link.find("div", {"class":"clics"}).get_text().split( )[0]
            print("clicks: {}".format(clics))

            meneos = link.find("div", {"class":"votes"}).get_text().split( )[0]
            print("meneos: {}".format(meneos))

            contentSumary = link.find("div", {"class":"news-content"}).get_text()
            print("contentSumary: {}".format(contentSumary))

            try:
                title = link.find('h2')
                a = title.find('a')
                titleRef = a['href']
                print("TitleRef: {}".format(titleRef))
                title = title.text
                print("Title: {}".format(title))
            except AttributeError:
                continue

            news_details = link.find("div", {"class":"news-details-data-up"})
            votes_up = news_details.find("span", {"class":"votes-up"})
            votes_up = votes_up.find('strong').text
```

```
            print("votes_up: {}".format(votes_up))

            votes_down = news_details.find("span", {"class":"votes-down"})
            votes_down = votes_down.find('strong').text
            print("votes_down: {}".format(votes_down))

            votes_anonymous = news_details.find("span", {"class":"wideonly votes-
                               anonymous"})
            votes_anonymous = votes_anonymous.find('strong').text
            print("votes_anonymous: {}".format(votes_anonymous))

            news_submitted = link.find("div", {"class":"news-submitted"})
            from_news_paper = news_submitted.select_one("span").text
            print("from_news_paper: {}".format(from_news_paper))

            karma = link.find("span", {"class":"karma"}).text.split( )[1]
            print("karma: {}".format(karma))

            category = link.find("span", {"class":"tool sub-name"}).text
            print("category: {}".format(category))

            comments = link.find("a", {"class":"comments"}).text.split( )[0]
            print("comments: {}".format(comments))

            news = News(clics,
                        meneos,
                        contentSumary,
                        title,
                        titleRef,
                        votes_up,
                        votes_down,
                        votes_anonymous,
                        from_news_paper,
                        karma,
                        category,
                        comments)

            news_list.append(news)

        print("newsSize:{}".format(len(news_list)))
        writer = FileWriterNewsCSV()
        writer.persistNews(news_list)
```

PREGUNTA 10.

Dataset. Presentar el Dataset en format csv.

Dataset *meneame_news*.csv :

NOM	MIDA	DATA	HORA	TIPUS
meneame_news	16KB	20/03/2019	19:37 h	Arxiu CSV

Característiques:

25x12 : recull de 25 notícies en portada i dotze atributs per cadascuna d'elles.

Llibreries Python utilitzades per el formateig de les dades a .csv:

```
from FileWriterNewsCSV import FileWriterNewsCSV
```

FileWriterNewsCSV és una llibreria software opensource de Python per tal de formatejar una base de dades o dataframe en .csv.

Codi Python per al formateig de les dades:

```
writer = FileWriterNewsCSV()
writer.persistNews(news_list)
```

Podeu trobar el dataset .csv, documents i fitxers en les wikis:

Xavier Jordà Murria: https://github.com/XavierJordaMurria/WebScraping

Anna Serena Latre: https://github.com/AnnaSerenaLatre/PAC1_WEB_SCRAPING

ANNEXOS

Taula de contribucions al treball:

Contribucions	Signa
Recerca prèvia	X.J.M, A.S.L.
Redacció de les respostes	X.J.M, A.S.L.
Desenvolupament del codi	X.J.M, A.S.L