Pràctica 1 de Tipologia i cicle de vida de les dades

Aquesta és la documentació de la PRA1 de Tipologia i cicle de vida de les dades del màster de Ciència de Dades de la UOC.

Estudiants:

- Zenon Perise Alia
- Adrià Vidal de Palol

Semestre: Primavera 20/21

Context

La importància dels medis de comunicació a la societat actual és indiscutible. L'anomenat "quart poder" té la capacitat de, a part d'informar sobre l'actualitat, de formar, canviar i manipular l'opinió pública. Cada medi de comunicació té la seva línia editorial, que fa que es posi l'accent a diferents aspectes d'una mateixa notícia. En aquest context es fa necessari el contrast de noticies de diferents medis per tal d'esbrinar el que realment ha passat. D'aquest necessitat i amb l'ajut de les noves tecnologies, sorgeixen els anomenats agregadors de notícies. Aquests sistemes permeten agrupar notícies de diferents medis i presentar-les de forma conjunta per tal d'obtenir una visió general de la realitat.

Per aquest motiu, hem plantejat la possibilitat de fer un estudi sobre les notícies publicades durant un període de temps en un agregador de notícies per tal de veure quins esdeveniments o personatges han tingut una certa rellevància durant un temps en el passat.

Com a font de dades hem escollit la web https://www.meneame.net. Aquesta web és un agregador de noticies, és a dir, recull noticies d'altres medis. La particularitat d'aquesta pàgina és que la decisió sobre quines noticies apareixen a portada és pressa per la comunitat de la pàgina gracies a un sistema de votacions. Per tant, les notícies que apareixen en portada tenen una certa rellevància social o interès per la major part d'usuaris.

Títol del dataset

Anàlisi de les notícies publicades a la portada de Meneame.net del Maig 2020 fins Abril 2021

Descripció

El dataset recull les notícies que han arribat a la portada de Meneame.

Com que la pàgina de Meneame és un agregador de notícies, el primer camp ens mostra la URL de la notícia pròpiament dita, és a dir, és un enllac a una altra web, ja sigui la versió digital d'un diari o un blog o qualsevol altre tipus de pàgina. A continuació trobem els dos camps principals de la notícia, que contenen la notícia pròpiament dita. Aquests són el títol i el resum de la notícia. Es tracta d'un text escrit normalment en castellà.

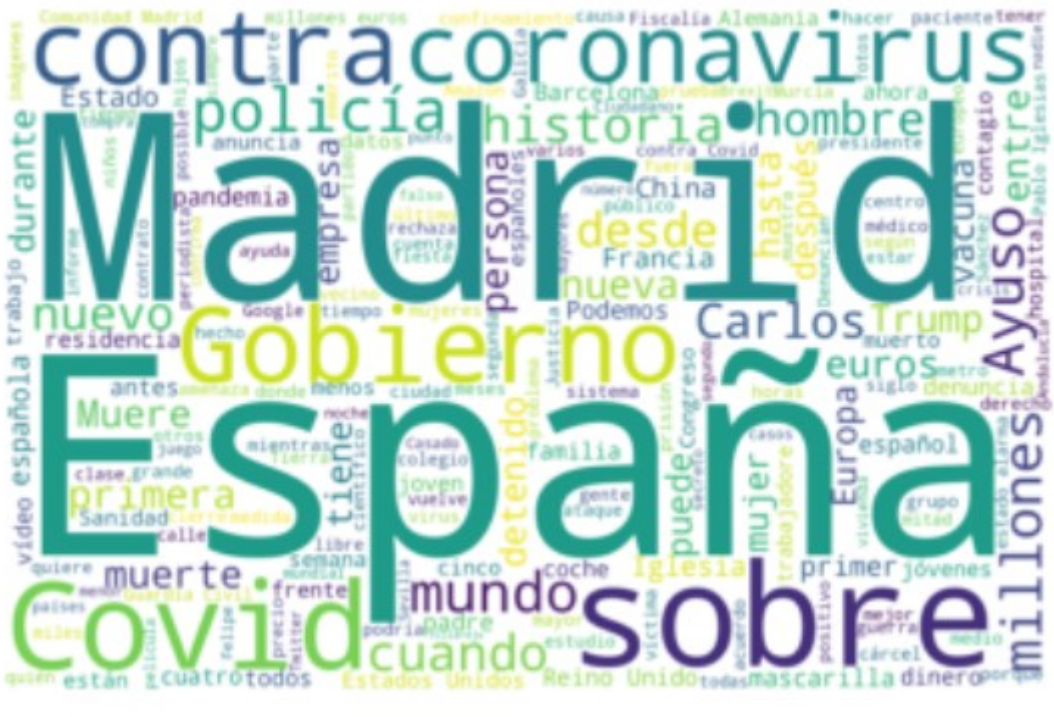
Per tal de posar les notícies en context temporal disposem de dues dates que enregistra Meneame. Aquestes són, per un costat, la data i hora en la que la notícia ha estat enregistrada a la pàgina. És a dir, el

dia i hora en el que un usuari ha decidit que una determinada notícia d'un medi extern pot ser rellevant per a la comunitat i l'ha enregistrar a Meneame. Aquesta és la data d'enviament. Després de cert temps, la comunitat ha pogut veure aquesta notícia i ha pogut votar si és rellevant o no. L'algorisme de Meneame decideix a partir d'aquests vots sí la notícia és rellevant i per tant, apareix en portada. Aquesta és la data de publicació i és la segona data que enregistrem.

Tal com hem dit, la decisió de publicació es pren en funció de les votacions que ha rebut una notícia determinada. Per tal de poder ponderar el pes d'una notícia determinada el nostre dataset també conté el número de vots que ha obtingut.

Per últim, no totes les noticies que s'envien a Meneame són d'actualitat pròpiament dites. També apareixen altres tipus com articles culturals o històrics. Per tal de diferenciar les notícies d'actualitat d'altre tipus de textos, disposem d'una categoria que és assignada per l'usuari que envia la notícia.

Representació Gràfica



Aquest WorkCloud ha estat generat amb el codi disponible a wordcloud.py utilitzant els títols de les notícies del dataset.

Contingut

El dataset conté les noticies de les 600 primeres pàgines de Meneame en el moment de l'extracció (8. Abril 2021) Això correspon a 15000 registres. Les dades obtingudes es troben totes a la portada de Meneame, és a dir, a la pàgina principal. No ha estat necessari carregar les pàgines de les notícies individualment. La pàgina principal a més està paginada, és a dir, al final de la pàgina trobem enllaços per accedir a la següent pàgina. Totes les pàgines són accessibles amb un paràmetre a la URL.

Els camps extrets són els següents:

- Pàgina: Número de pàgina de Meneame on apareix la notícia
- Títol: Títol de la noticia. Text curt.

- Resum: Resum de la notícia amb una llargada de 4 o 5 línies.
- URL: URL de la notícia original.
- Vots: Numero de vots que havia rebut la notícia en el moment de fer la recollida de dades.
- Data Enviament: Data d'enviament de la notícia a Meneame. El format és el número de segons des de mitjanit del 1 de Gener de 1970 a la zona horària UTC.
- Data Publicació: Data en la que la notícia va ser publicada a la portada de Meneame. El format és el número de segons des de mitjanit del 1 de Gener de 1970 a la zona horària UTC.
- Categoria: Categoria de l'article: Notícia d'actualitat, curiositat, ...

Exemple de dades al Dataset

```
0,https://www.20minutos.es/noticia/4648655/0/encuentran-en-egipto-bajo-la-
arena-una-gran-ciudad-perdida-de-3-000-anos/," Encuentran en Egipto el
'Ascenso de Atón', el descubrimiento más importante desde la tumba de
Tutankamón ","El Gobierno egipcio anunció este jueves el hallazgo bajo la
arena en la monumental Luxor de una gran ciudad de unos 3.000 años de
antigüedad que se hallaba perdida y que se encuentra en un buen estado de
conservación. Se trata del ""mayor asentamiento administrativo e industrial
de la era del Imperio Egipcio en la orilla occidental de Luxor"" y ""la
mayor ciudad jamás encontrada en Egipto"". La urbe recibió el nombre de
""El Ascenso de Atón"" y estuvo activa durante los reinados de faraones de
la Dinastía XVIII, como Amenhotep III o
Tutankamón.",135,1617889103,1617891003,actualidad
0,https://www.huffingtonpost.es/entry/santiago-abascal-vallecas-
policia_es_606ef3bdc5b6c70eccaeeb0d?due, La Policía señala a Santiago
Abascal por los disturbios en Vallecas ,"El líder de Vox es el señalado por
la Policía Nacional, según fuentes policiales citadas por el diario ABC:
“Si Abascal no hubiera hecho de policía tal vez no habría 21 agentes
heridos”. Este diario afirma que un agente pidió a Abascal que no cruzara
el cordón de seguridad y que éste “hizo caso omiso”. El líder
ultraderechista se bajó de la tribuna, donde había dado un breve discurso,
y se dirigió hacia los manifestantes. Según ABC, “se dedicó a contar pasos,
a tomar medidas imaginarias en el aire y a encararse con quienes tenía
enfrente”.",619,1617888733,1617890155,actualidad
```

Agraïments

La pàgina Meneame.net és un agregador de notícies en castellà creat al 2005 per Ricardo Galli, un professor de la Universitat de les Illes Balears (UIB). Meneame ofereix la possibilitat d'enviar notícies d'altres medis, votar-les i comentar-les. Les notícies s'ordenen segons la seva popularitat, mesurada a partir dels vots i número de comentaris. Les més populars apareixen a la portada de Meneame.

El fitxer robots.txt té el següent contingut:

```
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:
```

El nostre Scrapper obté la informació de les adreces https://www.meneame.net/?page=XX, on XX és el número de pàgina que volem obtenir. Al fitxer robots.txt no apareix aquesta URL, per tant no està explícitament prohibit accedir-hi de forma automàtica.

Pel que fa a estudis anteriors, hem trobat que s'han utilitzat medis digitals per a fer un estudi sobre els titulars de notícies, majoritàriament de llengua anglesa. També s'ha utilitzar Meneame per a extreure altres tipus d'informació. A continuació veiem els enllaços:

Analisis anteriors - Noticies de medis digitals

- Estudi dels titulars del New York Times de Gener a Juliol del 2020 utilitzant NLP (Natural Language Processing) https://towardsdatascience.com/headlines-articles-analysis-and-nlp-4013a66dbac
- Estudi dels titulars publicats a CNET https://towardsdatascience.com/analyzing-cnets-headlines-3f350bb97cd4
- Dataset amb els titulars de les noticies publicades per la ABC (Australian Broadcasting Corporation) durant un periode de 8 anys https://www.kaggle.com/therohk/million-headlines
- Dataset amb els titulars del ABC-News, un diari australià, durant un període que va del 2003 al 2016 https://www.kaggle.com/richel145/analysis-of-a-million-news-headlines

Analisis anteriors - Meneame

- Dataset que inclou tota la informació apareguda a la portada de Meneame des del 2005 fins al 2017 https://www.kaggle.com/mrverde/meneamenet-front-page-news
- Estudi que mostra com afecta la aparició de temes nous a una plataforma de discussió online, en aquest cas Meneame https://www.researchgate.net/figure/Scatter-plot-of-days-in-the-dataset-of-Meneame-2011-2015-Each-day-is-represented-by-a_fig2_318914420
- Dataset utilitzat a l'estudi del punt anterior https://zenodo.org/record/2536218

- Dataset que conté algunes noticies aparegudes a la portada de Meneame https://zenodo.org/record/4122059

Inspiració

Aquest dataset vol mostrar quins conceptes o personatges han tingut rellevància durant certs períodes de temps tenint en compte les notícies publicades a la premsa. Aquestes notícies, al estar publicades a Meneame, ja han passat un filtre de rellevància per part de la seva comunitat. Per tant, amb Meneame disposem d'una font de notícies rellevants ordenades cronològicament. Això ens permet fer diferents estudis

- Esdeveniments / Personatges / Conceptes
- populars o més freqüents en un determinat interval de temps
- ponderant amb el pes en vots
- Clustering de notícies que parlem de temes similars
- Classificació de notícies futures
- Estudi de paraules que apareixen associades. Per exemple (Fernando Simon + COVID)
- Veure sí esdeveniments que han ocorregut es reflexen a les notícies. Per exemple, Brexit.

Llicència

La llicència del contingut de Meneame es troba dins la distribució lliure amb la condició de citar i referenciar la autoria del propietari CC BY 3.0 ES. (Enllaç: https://creativecommons.org/licenses/by/3.0/es/) Segons el nostre criteri, seleccionem la llicència Released Under CC BY-NC-SA 4.0 License, on se'ns permet utilitzar el contingut per la mateixa finalitat que utilitza el propietari(SA -Compartir igual).

Dataset a Zenodo

Enllaç: https://doi.org/10.5281/zenodo.4677284

Contribucions

Contribució	***Signatura***
Recerca prèvia	zperise, avidaldepalol
Documentació	zperise, avidaldepalol
Codi	zperise, avidaldepalol