Tipologia i cicle de la vida de les dades

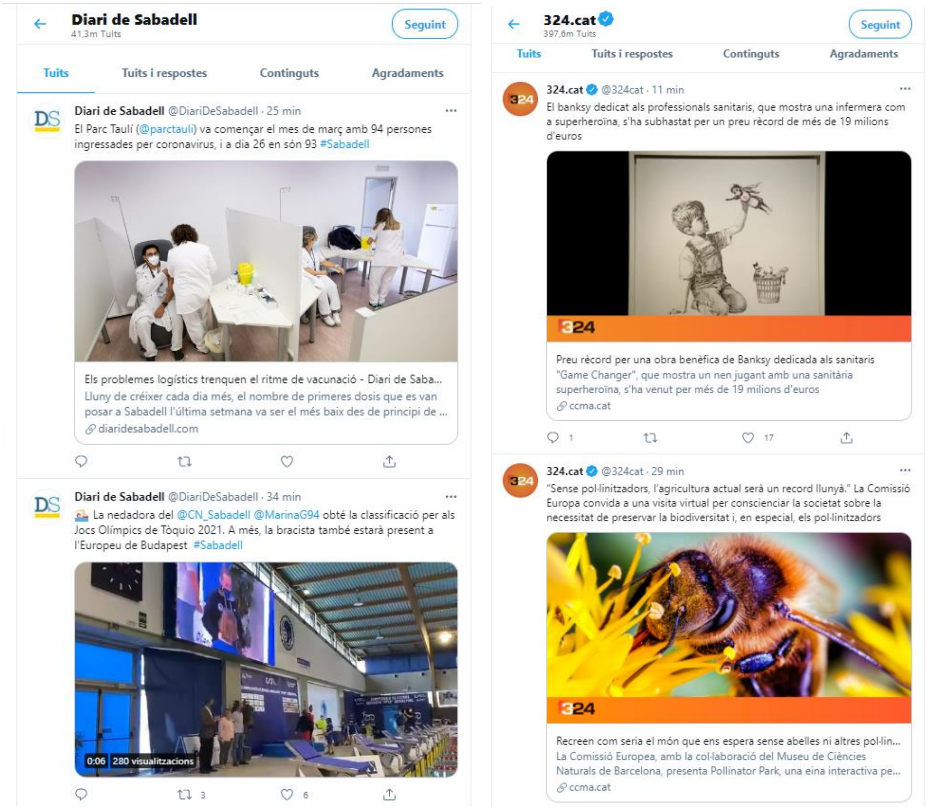
Pràctica 1: Web scraping

Context:

El context d'aquesta pràctica està centrat en estudiar com els mitjans de comunicació utilitzen les xarxes socials, en aquest cas, Twitter. En aquesta pràctica s'ha utilitzat la API de Twitter per consultar els tweets que publica un mitjà de comunicació, en el meu cas, el Diari de Sabadell.

Per situar-nos i per tal de presentar la idea de què volem obtenir amb aquesta pràctica, he realitzat dues captures a l'atzar de dos mitjans de comunicació. Un d'ells el mitjà de comunicació que estudiarem (el Diari de Sabadell) i per altra banda el 324 de la corporació Catalana de Mitjans Audiovisuals.

En les captures que presento a continuació veiem que els dos mitjans utilitzen la mateixa tàctica. Presenten el tweet amb un "resum" i adjunten un link de la seva pàgina web.



Realment és lògic utilitzar aquest mètode ja que molts mitjans de comunicació guanyen diners de la publicitat que tenen allotjada a la seva pàgina web. Si els mitjans només presentessin la notícia o el titular sense adjuntar la pàgina web deixarien de fer negoci.

Davant d'aquesta casuística em plantejo diferents preguntes.

- Els mitjans de comunicació saben utilitzar correctament les xarxes socials?
- Quants cops ens "saturen" de sobre informació només perquè entrem al seu link?

Amb aquesta pràctica vull presentar un dataset capaç de reflectir la "sobra" informació que ens presenten els mitjans de comunicació a les diferents xarxes socials. El dataset serà capaç de respondre les següents preguntes.

- Quants cops es fa un tweet presentant una notícia (url) presentada anteriorment?
- Quants likes aconsegueix en total la notícia? Quin és el número de likes major i menor de likes que ha aconseguit una notícia per separat?

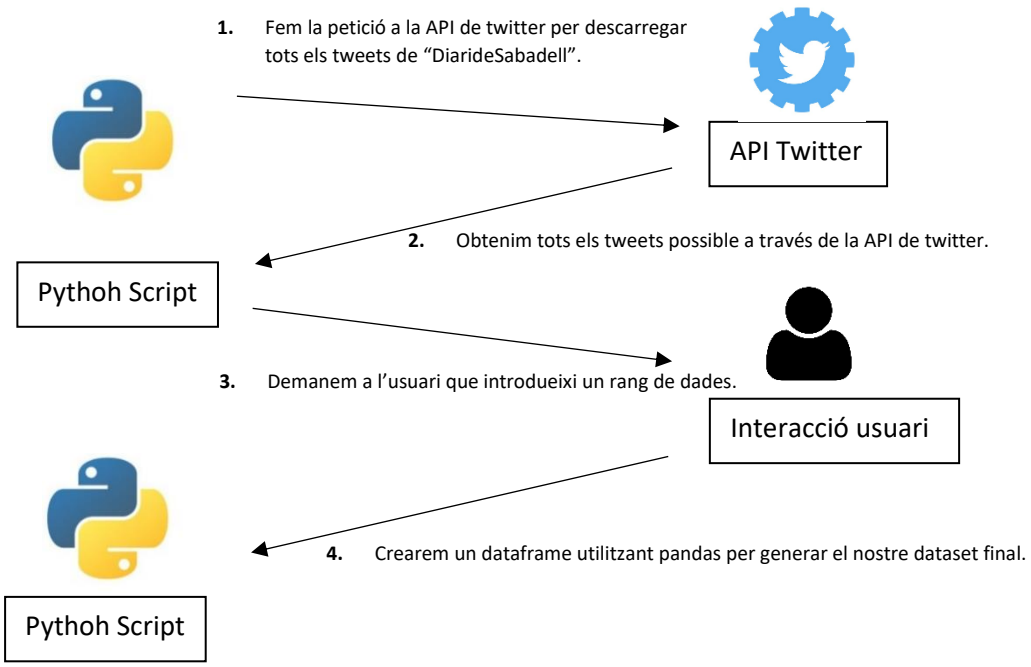
Dataset

Nom: repetibilitat_diari_de_sabadell

Descripció: En el dataset podem observar

Representació gràfica:

Presentar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.



Contingut dataset:

El script proporcionat en l'entrega d'aquesta pràctica crea dos fitxers .csv. El primer fitxer s'anomena "DiariDeSabadell_tweets.csv" i conté tots els tweets descarregats de Twitter utilitzant la API. El segon fitxer és el dataset que volíem aconseguir en aquesta pràctica, el fitxer s'anomena "repetibilitat_diari_de_sabadell.csv".

El dataset "repetibilitat_diari_de_sabadell.csv" conté un total de 8 columnes per tal de veure la informació processada.

A continuació detallarem el contingut de cada columna i el significat d'ells mateixos.

El període de temps de les dades és el període de temps introduït per l'usuari en l'execució del script. En l'exemple d'aquesta entrega hem decidit utilitzar el rang de dates entre el 01/04/2021 fins el 09/04/2021.

Exemple execució:

```
Siusplau introdueix una data inicial (format AAAA-MM-DD):
2021-04-01
You entered 2021-04-01
Siusplau introdueix una data final (format AAAA-MM-DD):
2021-04-09
You entered 2021-04-09
Dates a consultar:
['2021-04-01', '2021-04-02', '2021-04-03', '2021-04-04', '2021-04-05', '2021-04-06', '202
1-04-07', '2021-04-08', '2021-04-09']
```

Una vegada hem executat el script obtindrem els dos fitxers.

El primer fitxer podem veure tots els tweets descarregats, en aquest cas hem decidit quedar-nos només amb el ID del tweet, el dia de creació del tweet, el número de likes i de retweets, la URL publicada en el tweet i finalment el contingut de tot el tweet. Aquest fitxer veurem totes les dates possibles.

```
ID;Day_Created;Created;Favorites;Retweets;URL;fullTweet
1380907122708332545;2021-04-10;2021-04-10 15:34:48;0;0;;⚽ Comença al Municipal d'Arraona els juvenils del @CEMercantil i @UESantsoficial Dóna inici la segona fase d
1380905913138171907;2021-04-10;2021-04-10 15:30:00;0;0;https://www.diaridesabadell.com/2021/04/10/sabadell-demana-570-milions-a-la-ue/;Com Barcelona es va obrir al m
1380902138004480001;2021-04-10;2021-04-10 15:15:00;2;2;https://www.diaridesabadell.com/2021/04/10/la-policia-demana-collaboracio-per-trobar-lamo-duna-gossa-perduda/;
1380898365118971907;2021-04-10;2021-04-10 15:00:00;0;0;https://www.diaridesabadell.com/2021/04/10/les-treballadores-del-sad-segueixen-en-peu-de-guerra/;Unes cinquant
1380894588143988743;2021-04-10;2021-04-10 14:45:00;0;0;https://www.diaridesabadell.com/2021/04/10/la-vicepresidenta-del-govern-espanyol-critica-linterrogatori-de-la-
1380890814792953858;2021-04-10;2021-04-10 14:30:00;1;0;https://www.diaridesabadell.com/2021/04/10/homenatge-a-la-mort-del-periodista-xavier-vinader/;Homenatge a la m
1380887038753435650;2021-04-10;2021-04-10 14:15:00;1;0;https://www.diaridesabadell.com/2021/04/10/el-govern-treballa-perque-la-cultura-sigui-una-excepcio-al-confinam
1380883266807414789;2021-04-10;2021-04-10 14:00:01;0;0;https://www.diaridesabadell.com/2021/04/10/xavier-marcet-els-fons-de-la-ue-no-han-de-servir-per-a-ocurrencies/
1380879488536416256;2021-04-10;2021-04-10 13:45:00;0;0;https://www.diaridesabadell.com/2021/04/10/la-favs-recela-de-la-promesa-de-salut-de-reobrir-el-consultori-medi
1380875713704697857;2021-04-10;2021-04-10 13:30:00;1;0;https://www.diaridesabadell.com/2021/04/10/la-majoria-dabusos-i-agressions-sexuals-passen-dins-de-casa/;Des de
1380868171998105603;2021-04-10;2021-04-10 13:00:02;0;0;https://www.diaridesabadell.com/2021/04/10/sabadell-injecta-mes-diners-a-les-escoles-per-a-material-escolar-pe
1380865744964046851;2021-04-10;2021-04-10 12:50:23;0;0;https://www.diaridesabadell.com/2021/04/10/les-treballadores-del-sad-segueixen-en-peu-de-guerra/;⚫ Les trebal
1380860617578348547;2021-04-10;2021-04-10 12:30:01;6;0;https://www.diaridesabadell.com/2021/04/10/sabadell-demana-570-milions-a-la-ue/;El Castell de Can Feu, l'entor
```

Aquest fitxer només l'utilitzarem per fer un processat de la informació per generar el nostre dataset final.

El nostre dataset final tindrà la següent forma:

	A	B	C	D	E	F	G	H
1	URL	Total_likes	Tweet_max_likes	Tweet_min_likes	Total_retweets	Tweet_max_retweets	Tweet_min_retweets	Total_tweets_mateixa_URL
2	http://diaridesabadell.com/2021/04/08/aterra-a-sabadell-el-circ-raluy-la-vacuna-contra-el-pessimisme/	4	4	4	0	0	0	1
3	https://www.diaridesabadell.com/2021/03/30/maxim-docupacions-illegals-de-lultima-decada-a-sabadell/	0	0	0	4	4	4	1
4	https://www.diaridesabadell.com/2021/03/31/alberg-temporal-sensellar-sabadell/	0	0	0	1	1	0	2
5	https://www.diaridesabadell.com/2021/03/31/el-dijous-arriba-algun-nuvol-a-sabadell/	0	0	0	0	0	0	1
6	https://www.diaridesabadell.com/2021/03/31/els-professors-de-lescola-de-musica-de-barbera-denuncien-inseguretat-laboral/	0	0	0	0	0	0	1
7	https://www.diaridesabadell.com/2021/03/31/lavantguarda-es-medica-i-cientifica/	1	1	1	0	0	0	1
8	https://www.diaridesabadell.com/2021/03/31/mes-vorera-a-la-carretera-de-barcelona/	9	7	0	1	1	0	3
9	https://www.diaridesabadell.com/2021/03/31/obituari-de-sabadell-de-l1-dabril-de-2021/	0	0	0	0	0	0	1
10	https://www.diaridesabadell.com/2021/04/01/a-lhospital-un-jove-de-24-anys-despres-dun-incendi-a-casa-seva/	1	1	0	0	0	0	2
11	https://www.diaridesabadell.com/2021/04/01/detingut-un-home-per-intentar-colar-receptes-falsificades-en-una-farmacia/	0	0	0	0	0	0	1
12	https://www.diaridesabadell.com/2021/04/01/el-personatge-la-mascareta-i-la-platja-son-incompatibles/	1	1	1	0	0	0	1
13	https://www.diaridesabadell.com/2021/04/01/el-postureig-a-dalt-la-mola-fa-enrabiar-els-bombers/	30	16	6	16	9	3	3
14	https://www.diaridesabadell.com/2021/04/01/el-sabadell-recupera-la-millor-versio-de-stoichkov/	4	4	4	0	0	0	1
15	https://www.diaridesabadell.com/2021/04/01/judici-violacio-multiple-sabadell/	9	9	0	3	2	0	3
16	https://www.diaridesabadell.com/2021/04/01/la-generalitat-licita-el-projecte-duna-rotonda-a-la-b-140-entre-sabadell-i-barbera/	1	1	0	0	0	0	2
17	https://www.diaridesabadell.com/2021/04/01/obituari-de-sabadell-del-2-dabril-de-2021/	0	0	0	0	0	0	2
18	https://www.diaridesabadell.com/2021/04/01/ocupacions-conflictives/	10	10	10	9	9	9	1

El nostre dataset podrem veure l'agrupació de diferent tipus d'informacions. Per entendre com funciona, en el nostre cas s'ha decidit agafar només els tweets que s'ha publicat una URL. En aquest cas no ens quedarem amb el tweet però si que ens quedarem amb la URL. D'aquesta manera la nostre primera columna podrem veure totes les URLS publicades a twitter per un compte.

Les següents columnes es donen informació dels tweets que han tingut la URL.

URL -> URL publicada a twitter per DiariDeSabadell

Total_likes -> Número total de likes que ha obtingut la URL en els diferents tweets.

Tweet_max_likes -> Màxim número de likes que ha obtingut la URL per separat.

Tweet_min_likes -> Mínim número de likes que ha obtingut la URL per separat.

Total_retweets -> Número total de retweets que ha obtingut la URL en els diferents tweets.

Tweet_max_retweets -> Màxim número de retweets que ha obtingut la URL per separat.

Tweet_min_retweets -> Màxim número de retweets que ha obtingut la URL per separat.

Total_tweets_mateixa_URL -> Número de tweets en què s'ha publicat la URL.

Nota: En el dataset no podem veure la data de creació ja que estem agafant un rang de tweets i agrupant els tweets per la URL. Això significa que podem veure tweets/URLs de dades anteriors, per exemple, en el nostre cas veiem URLs del dia 30 i 31 de març quan aquestes dades no les hem seleccionat. Això és degut perquè el compte de twitter està republicant notícies d'altres dies.

Agraïments

En el nostre cas hem utilitzat la API de Twitter per descarregar els tweets d'un compte de twitter. Al no tractar amb cap web directament ni amb cap usuari (estem explotant informació que pertany a twitter) no hem d'agraïr a ningú l'accés a les dades. Només a la comunitat de python per crear les llibreries per accedir fàcilment a la API de Twitter.

De cara a citar anàlisis anteriors no n'he trobat, aquest anàlisis ve donat per la curiositat d'entendre si els mitjans de comunicació fan un ús abusiu de les xarxes socials. En aquest cas l'objectiu era veure quants tweets fan publicant el mateix, aprofitant aquest anàlisis ha estat fàcil treure'n més valors (likes, retweets...).

Nota: Realitzant l'anàlisis és important comentar que a vegades podem veure una URL publicada moltes vegades, els mitjans de comunicació a vegades ho fan per actualitzar una notícia o bé per exemple el seguiment d'un partit de futbol.

Inspiració

Com he comentat en l'apartat anterior l'inspiració ve donada de veure molts cops la mateixa notícia repetida en el TL de Twitter. En aquest sentit el dataset permet veure quin efecte té publicar la URL moltes vegades (sovint més likes i més retweets).

Llicència

En el meu cas la llicència entenc que forma part de Twitter però estic agafant informació dels seus tweets. Entenc que seria la següent:

Released Under CC0: Public Domain License

De totes maneres no n'estic segur i no he trobat informació relacionada.

Codi

El podem veure en el repositori de GitHub.

Dataset.

A més a més de publicar el dataset a Zenodo.

http://doi.org/10.5281/zenodo.4678511