Guillem Vidal Pallarols
Lídia Bandrés Solé

Web Scraping - Pràctica 1

1. Context. Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Aquests darrers dies estem veient com l'economia mundial ha entrat dins un espiral d'inflació. Els preus de les matèries primeres no deixen de pujar i això esta provocant un desequilibri important en les economies i generant situacions de desproveïment. El cas més destacat de les últimes setmanes ha estat la falta de transportistes al Regne Unit que ha provocat una reducció de l'oferta de combustibles a tot el país. Des d'aquí, l'opinió pública s'ha posicionat dient que és una altra conseqüència del Brexit i que aquí no passarà. Tot i això volem investigar quina opinió té la gent d'aquesta possible relació.

Per dur a terme aquesta tasca hem escollit la xarxa social on podem obtenir més opinions de la població, Twitter. Aquest portal ens permet obtenir milers de comentaris sobre un tema en concret fent una simple cerca. És per això que hem decidit fer web scrape en aquest portal. D'aquesta manera aconseguirem un data set amb les opinions dels ciutadans durant un període concret. Un cop obtingudes les dades es podrien utilitzar per tal de conèixer els corrents d'opinió, detectar notícies falses o fer estudis sociològics.

2. Títol. Definir un títol que sigui descriptiu pel dataset.

Recull d'opinions a Twitter sobre els efectes del Brexit.

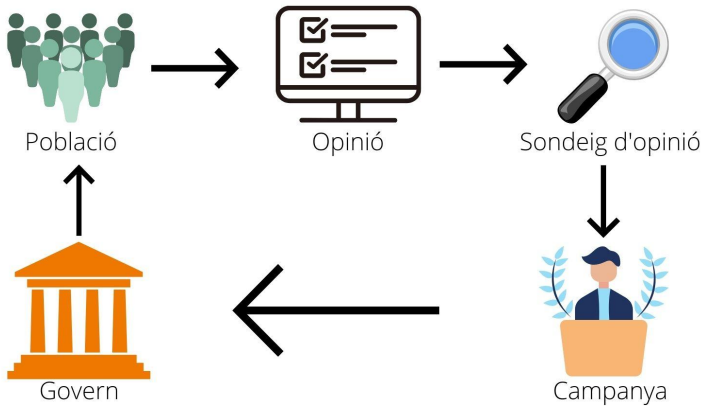
3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.

En aquest dataset trobarem un recull d'opinions per part d'usuaris de la xarxa social Twitter sobre el Brexit. D'aquesta forma trobar quina relació observa la població entre el Brexit i els esdeveniments del present. En diferents camps hi trobarem tant les opinions com la id i la data del tuit. Els tuits són comentaris trobats com a resultat de la cerca de la paraula "brexit" en el cercador i filtrant els 1000 primers resultats. Les 1000 entrades que tindrem seran les més recents a data d'extracció però com la data de la publicació del tweet sortirà al dataset, sempre es podrà contextualitzar.

4. Representació gràfica. Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.

La població expressa la seva opinió. Aquesta opinió s'utlitza per fer un sondeig d'opinió que s'utilitzarà a l'hora de planificar les campanyes electorals. Segons la campanya electoral,

s'escollirà un partit o un altre per governar i aquest govern serà el que determinarà l'opinió de la població.



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

Primer de tot analitzarem els diferents camps que inclou el dataset, en total són 8, sense tenir en compte el camp creat automàticament que informarà sobre el número de fila de cada un dels camps.

Date created: és una variable del tipus string. Representa el temps en format UTC quan es va crear el tweet.
Id: és una variable del tipus integer. És un identificador únic del tweet en qüestió.
Text: és una variable del tipus string. Emmagatzema el text que té el tweet en format UTF-8.
Truncated: és una variable del tipus boolean. Ens diu si la variable text ha estat truncada en diferents tweets.
Reply name: és una variable del tipus string, por ser null. Si el tweet en qüestió és una resposta a un altre tweet, aquesta variable ens donarà el nom de l'autor original del tweet al que s'ha respost.
Place: té una categoria place, pot ser null. Ens dirà si el tweet està associat a un lloc en concret.
Is quote: és una variable del tipus boolean. Ens dirà si el tweet amb el que estem tractant és en realitat una cita d'un altre.
Retweeted: és una variable del tipus boolean. Ens dirà si el tweet amb el que estem tractant és un retweet d'un altre tweet original.

La següent taula representa els diferents camps del dataset. Tots els camps fan referència a la informació trobada en cada un dels tweets

	Date created	Id	Text	Truncated	Reply name	Place	Is quote	Retweeted
1	2021-11-07 13:33:50 + 00:00	142335654	Brexit was not so bad	False			False	True

Un cop explicats els camps del dataset hem de saber el temps de les dades. Com podrem veure més endavant al codi, agafarem un total de 1000 objectes. Això farà que ens surtin 1000 entrades amb 1000 tweets diferents i seran els últims en haver estat publicats, per tant seran les dades més recents que tinguem en l'actualitat.

El procés de recollida és du a terme a través de l'API de twitter. Primer hem hagut de demanar accés i un cop ens han donat les claus per poder accedir-hi, hem ulitizat la llibreria Tweepy habilitada per poder extreure informació de l'API de Twitter. A partir d'aquí la feina consisteix en llegir i entendre el funcionament de la llibreria Tweepy per poder saber com extreure la informació que volem. Per exemple, tots els objectes que hi ha dins d'un tweet i quin codi hem d'escriure per poder-lo extreure.

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

En aquest exercici de web scraping hem decidit obtenir dades de twitter ja que permet obtenir fàcilment una API. Aquesta API l'ofereix la mateixa plataforma després de justificar els motius de l'extracció de dades. D'aquesta manera ens assegurem que les dades extretes no poden generar cap problema de drets d'autor o d'incompliment de termes i condicions.

Cercant anàlisis anteriors hem trobat un recull de treballs d'anàlisis similars fent web scrape a twitter.

- En un primer cas es tracta d'un exercici per tal d'obtenir les darreres opinions de l'ex president dels EEUU Doland Trump. Amb aquest estudi s'extreien els seus darrers 250 twits i s'analitzava quines eren les paraules que més utilitza. D'aquesta forma s'obté que les paraules "Biden", "fake" i "news" eren algunes de les paraules més utilitzades en el seu compte. D'aquesta forma es pot elaborar un estudi sobre quina és l'estratègia comunicativa de l'expresident.

  https://github.com/AlexTheAnalyst/PythonCode/blob/master/Twitter%20Scraper%20V8.ipynb

Guillem Vidal Pallarols
Lídia Bandrés Solé

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

Com hem explicat a l'inici, Twitter és la xarxa social on pots trobar més opinions per part de la ciutadania. El fet de poder filtrar sobre temes facilita molt l'enfocament que se li pot donar a l'estudi. En el nostre cas hem trobat interessant saber els corrents d'opinió sobre el brexit dels darrers dies ja que amb els problemes amb el combustible al Regne Unit, és un debat que ha tornat a aflorar. Amb el dataset que hem obtingut podem fer un estudi sobre les tendències d'opinió. Un exemple seria per fer-nos una idea de com sortirien els resultats d'un nou referèndum del Brexit actualment.

Comparat amb l'estudi mencionat anteriorment hi ha la diferència que no es té en compte un sol usuari sinó tots els usuaris que escriuen una paraula clau. Com que no busquem cap informació concreta no fem, encara, un anàlisi concret de les dades.

8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**
- **Unknown License**

La llicència escollida serà CC BY-NC 4.0, que realment no és cap de les especificades a l'enunciat. Volem que les persones que utilitzin el dataset ens donin crèdit ja que això dóna valor a la feina que hem fet, per això no ho fiquem amb una llicència oberta. Per altra banda, al sol·licitar l'accés a l'API de Twitter sempre s'ha especificat que els objectius per obtenir l'accés eren acadèmics. Per tant, no ens sembla lícit que les dades obtingudes es puguin fer servir per fins comercials. Tampoc tenim necessitat de restringir la llicència del resultat del treball que hagi utilitzat el nostre dataset, és per això que no s'utilitzarà la nomenclatura SA.

9. Codi. Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

#El primer pas serà importar les llibreries que necessitarem
import tweepy

Guillem Vidal Pallarols
Lídia Bandrés Solé

```
import pandas
import time

#A continuació inicialitzarem les claus donades per l'API de Twitter per poder accedir-hi
consumer_key = "4ikBRcdJVZJmcXF8WOsuQ0VY5"
consumer_secret = "ksmxgT77nUprvScPsPqb3bgy387cLkuTaiIMmqsKfsOcoA5Bbh"
access_token = "1455605536582512643-zJ9scn7mf8WYmNca2DB90cNLHYVC8P"
access_token_secret = "ZLuoquSYSjEyUVzKxDFlLoSkGG9588Ee7ZCftqCQZhoeM"

#El pas següent serà per poder obtenir accés a l'API
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)

#El text que buscarem serà la paraula clau "Brexit" i ho farem fins a 1000 iteracions
text_query = 'brexit'
count = 1000

try:
    # Crearem la query utilitzant les opcions que necessitem de l'API
    tweets = tweepy.Cursor(api.search_tweets, q=text_query).items(count)

    # Treurem la informació de cada un dels iterables dins la variable tweets
    tweets_list = [[tweet.created_at,
                    tweet.id,
                    tweet.text,
                    tweet.truncated,
                    tweet.in_reply_to_screen_name,
                    tweet.place,
                    tweet.is_quote_status,
                    tweet.retweeted] for tweet in tweets]

    # Creem un dataset a partir de tweets_list
    tweets_df = pandas.DataFrame(tweets_list)

except BaseException as e:
    print('failed on_status,', str(e))
    time.sleep(3)

#Per poder visualitzar i treballar amb el dataset, l'exportarem a csv
#Cal ficar el directori  d'on volem guardar el document
tweets_df.to_csv(r'C:\Users\lidia\OneDrive\Escriptori\MÀSTER UOC\Brexit_consequences.csv')
```

Guillem Vidal Pallarols
Lídia Bandrés Solé

10. Dataset. Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

A continuació trobarem l'enllaç al dataset publicat a Zenodo:

https://doi.org/10.5281/zenodo.5651923

Contribucions	Signatura
Investigació prèvia	LBS, GVP
Redacció de les respostes	LBS, GVP
Desenvolupament del codi	LBS, GVP

Bibliografia:

Per realitzar el codi s'ha utilitzat com a base el codi proporcionat per la pàgina web towardsdatascience:

https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1

Per poder entendre el funcionament de la llibreria Tweepy s'han consultat les següents pàgines web:

https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet
https://docs.tweepy.org/en/latest/api.html#API.search