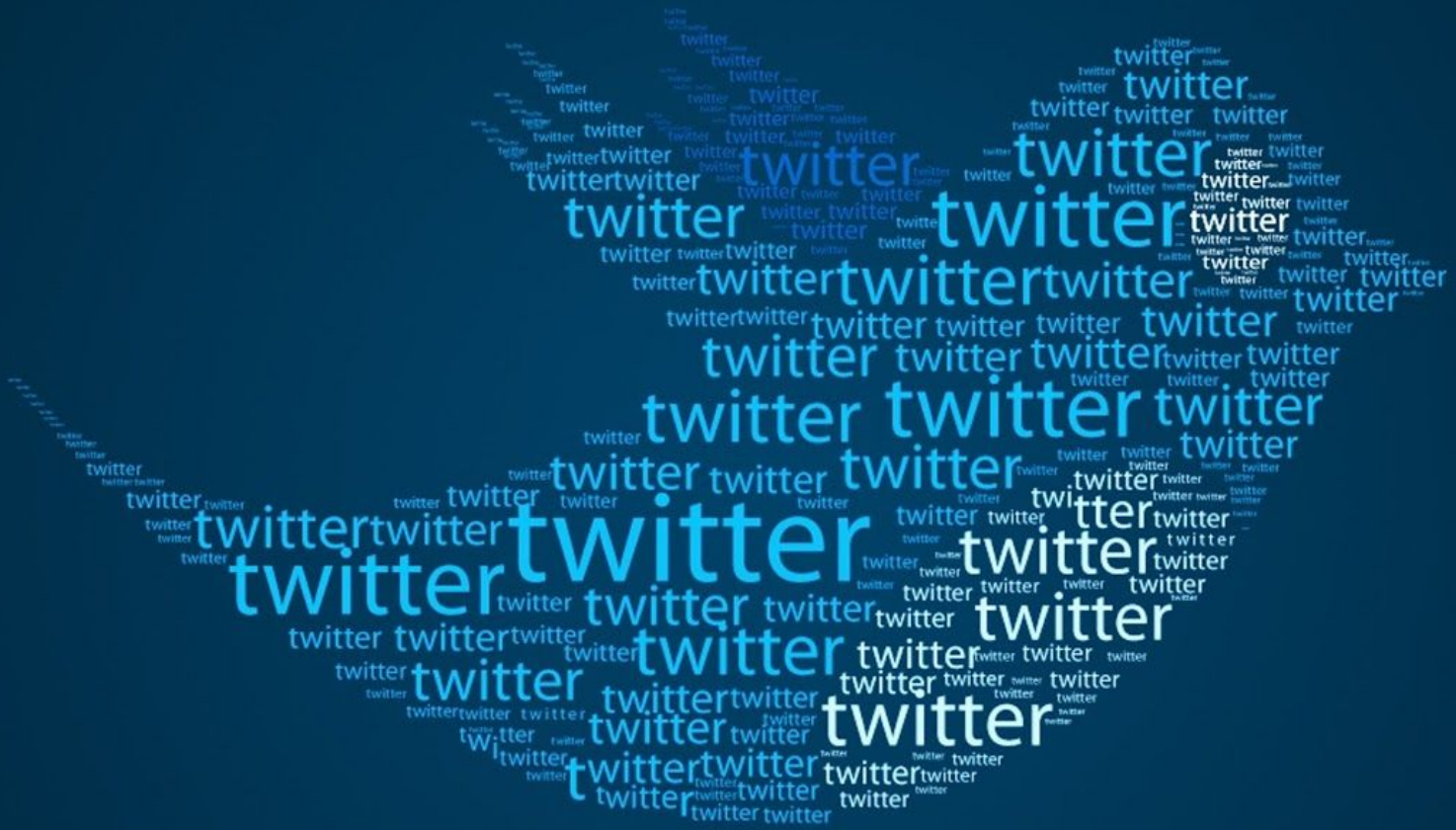
Applied Twitter Analytics



Grup 43
7 de gener del 2017
QT 2017/2018

Ariadna de Arriba
David Soldevila
Marc Casellas

Resum:

Objectius:

Estudiar si la interacció dels usuaris mitjançant likes i retuits de Twitter depèn de si el dia és laborable o no.

Mètodes:

Utilitzem un compte de Twitter amb el que hem seguit aproximadament 200 persones de diversos àmbits i usant un script de Phyton per obtenir el nombre de retuits i likes que cada usuari ha obtingut en les seves darreres 20 publicacions.

Resultats:

La mediana i la mitjana de likes i retuits en un dia laborable ha estat, respectivament, de 26.00 i 678.69 i les de un dia no laborable, de 25.5 i 1060.4.

Amb una confiança del 95% **no podem rebutjar la hipòtesis nul·la:** que el nombre de likes i retuits sigui el mateix en funció del tipus de dia (feiner o no feiner). Ja que:

- p-valor = 0.2254 > α = 0.05
- |z| = 0.7541 < $z_{1-\alpha/2}$ = 1.959964
- L'interval de confiança inclou el 0: [-880.8525, 117.3555].

Discussió:

Com a conclusió, **no rebutgem que la interacció dels usuaris en funció del tipus de dia sigui la mateixa**, ja que no hi ha evidència per declarar que el fet de ser un dia festiu/cap de setmana o un dia laboral afecti al nombre de *retuits* i *likes*.

Introducció:

Les xarxes socials són el nou portal d'informació: milions de persones poden difondre i compartir missatges, opinar i discutir sobre aquests. Però sobretot Twitter ha destacat i debutat com un nou mitjà de comunicació: més instantani i plural que la premsa tradicional (diaris, televisió...). Però realment, com es comporta la gent a la xarxa? Quan està més activa? I, per tant, quins dies podem arribar a més gent?

Objectiu:

Saber si els usuaris de Twitter interactuen igual amb les publicacions a través dels likes i retuits que donen depenent de si estem en un dia laborable o no (festius i caps de setmana).

Mètode:

Obtenció de les dades:

Per fer l'estudi, hem creat un compte de Twitter amb el que hem seguit aproximadament 200 persones de diversos àmbits (portals de notícies, polítics...). També hem creat un script amb phyton que ens guardava, cada cop que l'executàvem i per a cada un dels usuaris que seguíem, la ID única d'alguns dels 20 últims tuits publicats per aquest, la data de publicació i els retuits i m'agrada que havien obtingut, en un fitxer. A partir d'aquí, hem recollit les dades en un excel, hem sumat retuits i likes i hem separat les dades segons si era un dia laborable o un cap de setmana/festiu.

Exemple:

ID	Data	Rt	Likes	Suma (Rt + Likes)	Feiner (0) / No feiner (1)
940878171867631000	Wed Dec 13 9:26:51 2017	12769	109532	122301	0

Taula 1. Taula a mode d'exemple (veure dades completes a l'excel de l'annex).

Variables:

- SL : retuits i likes en dies laborables.
- SF: retuits i likes en dies no laborables (festius o caps de setmana).

Anàlisis:

Premisses:

- Mostra aleatòria simple independent.
- n_{SL}, n_{SF} >= 100

Hipòtesis:

H_0: $\mu_{SL} = \mu_{SF}$
H_1: $\mu_{SL} \neq \mu_{SF}$

Càlcul de l'estadístic:

$$\hat{z} = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

*On S_1 = SL i S_2 = SF

Distribució sota l' H_0:

$\hat{Z} \sim N(0,1)$

Estadístic:

Rebutjarem la hipòtesis nul·la (H_0) amb una confiança del 95% si es compleix el següent:

- p-valor < α = 0.05
- |z| > $z_{1-\alpha/2}$ = 1.959964
- L'interval de confiança inclou el 0.

Resultats:

Estadístic:

```
> nSL = length(datalaboral$Suma)
> nSF = length(datafestiu$Suma)
> n = nSL + nSF
> ySF = mean(datafestiu$Suma)
> ySL = mean(datalaboral$Suma)
> s2SL = sum((datalaboral$Suma-ySL)^2)/(nSL-1)
> s2SF = sum((datafestiu$Suma-ySF)^2)/(nSF-1)
> s2 = ((nSL-1)*s2SL + (nSF-1)*s2SF)/(nSL + nSF - 2)
> z = ((ySL-ySF)/sqrt((s2SL/nSL)+(s2SF/nSF))); z
[1] -0.7547695
```

Càlculs 1. Càlcul de l'estadístic amb R (script complet a l'annex)

z = -0.7547695

P-valor:

```
> pvalor = pnorm(z); pvalor
[1] 0.2251936
```

Càlculs 2. Càlcul del p-valor amb R (script complet a l'annex)

P-valor = 0.2251936

Interval de confiança (95%):

```
> conf = 0.95
> alpha = 1 - conf
> ic1 = (ySL-ySF)-qnorm(1-(alpha/2))*sqrt(s2/n)
> ic2 = (ySL-ySF)+qnorm(1-(alpha/2))*sqrt(s2/n)
> IC = c(ic1,ic2); IC
[1] -880.8525  117.3555
```

Càlculs 3. Càlcul de l'interval de confiança amb R (script complet a l'annex)

Interval de confiança: `[-880.8525, 117.3555]`

Resum de les dades obtingudes (summary):

Dies laborals (suma de retuits i likes)					
Mínim	**1r quantil**	**Mediana**	**Mitjana**	**3r quantil**	**Màxim**
0.00	8.00	26.00	678.69	96.75	122301.00

Taula 2. Resum de les dades de dies laborals (veure càlculs al script de R a l'annex)

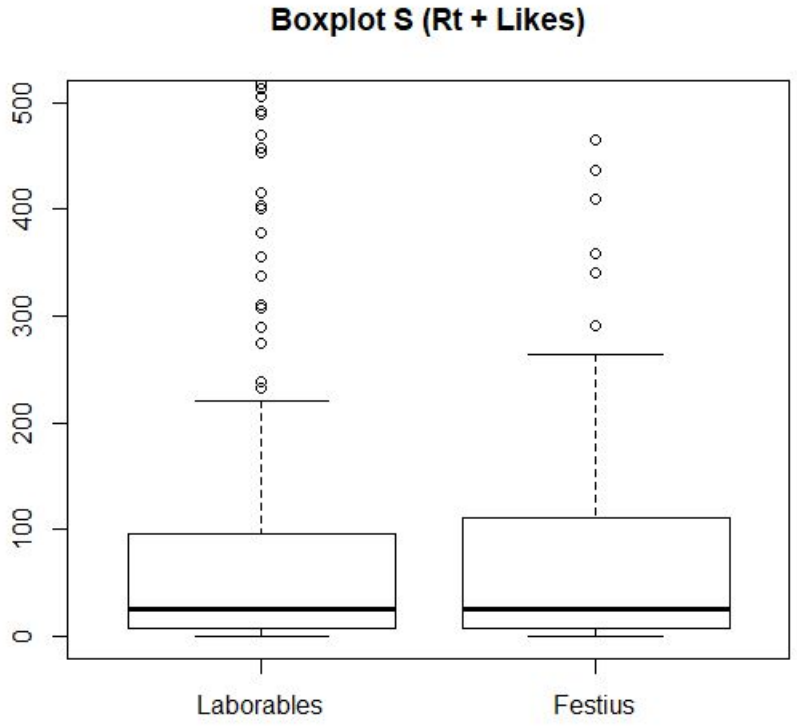
Dies festius/ cap de setmana (suma de retuits i likes)					
Mínim	**1r quantil**	**Mediana**	**Mitjana**	**3r quantil**	**Màxim**
0.00	8.00	25.50	1060.40	110.50	70258.00

Taula 3. Resum de les dades de dies festius / caps de setmana (veure càlculs al script de R a l'annex)

Boxplot de SL i SF:

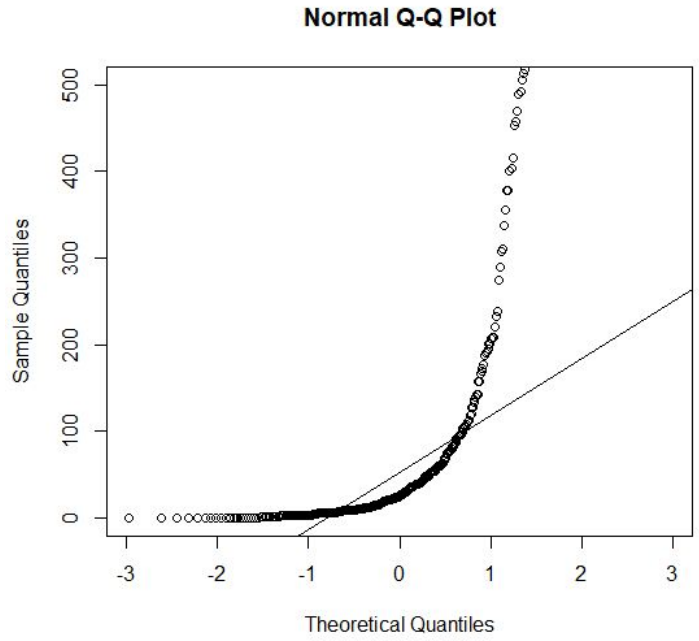
Abans d'entrar a estudiar el gràfic definirem el concepte de "*tweet star*":
És aquella persona que compta amb una gran fama i repercussió dins la xarxa social. Aquest tipus d'usuaris no els podem tractar com usuaris comuns.

Per veure-ho més clar, analitzem el boxplot de retuits i m'agrada distingint entre dies laborables i festius.

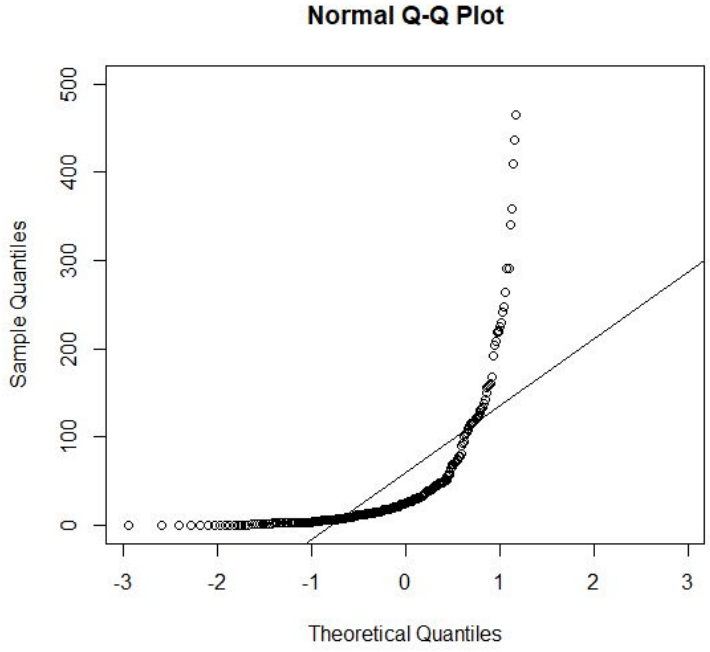


Gràfic 1. Boxplot de S en dies laborables i en dies festius / caps de setmana

En el boxplot, juntament amb la funció summary de R, podem observar les medianes, que són les que ens donen una estimació per la gran majoria d'usuaris, deixant a part els *tweet star* que són els valors extrems. Veiem que la diferència de les mitjanes és bastant significativa, però en les medianes, que ens proporcionen informació més real, la diferencia és mínima, concretament de 0,5 més en els dies laborables. Tot i això, la mitjana és més alta en els no laborables, degut als efectes dels tuits més virals, els quals no afecten a les interaccions de l'usuari mitjà. Així doncs, considerem la mitjana com un valor poc fiable.



Gràfic 2. QQnorm de SL (laborables)



Gràfic 3. QQnorm de SF (festius)

Els gràfics anteriors, gràfics 2 i 3, ens donen informació sobre la normalitat.
Es veu reflectida la normalitat segons la distribució dels punts sobra la recta: si aquests es situen sobre la línia, existeix normalitat; en cas contrari, no.
En el nostre cas, és trivial dir que no segueixen una distribució normal en cap dels dos. És lògic pensar que no ho segueixen, ja que la gran majoria de tuits tenen un nombre molt baix d'interaccions i només la gran minoria en tenen un gran nombre.

Discussions:

Conclusions:

No rebutgem la H_0, ja que:

$$\textbf{p-valor} = 0.2254 > \alpha = 0.05 \text{ o } |z| = 0.7541 < z_{1-\alpha/2} = 1.959964$$

Per tant, **no hi ha evidència per declarar que el fet de ser un dia festiu/cap de setmana o un dia laboral comporti diferències a la interacció dels usuaris** (donar retuits i likes).

Limitacions i treball futur:

La limitació ve donada per la generalització de la mostra. No hem agafat un perfil d'usuari (target) específic. Per fer un estudi el més realista possible s'hauria de classificar en diferents tipus d'usuaris segons els seus gustos, que es podrien saber segons el tipus de gent que segueixen. Seguidament fer els mateixos càlculs que hem fet, però aplicats als subgrups. En aquest cas seria possible que en certs grups, com comptes d'empreses i altres organitzacions, baixessin en festius. O per contra, que usuaris que es poguessin classificar en un grup d'oci tinguessin un alt nivell d'activitat els dies festius.

Annexos:

Obtenció de dades (python):

get_id_list.py
Crea una llista de les ID's dels usuaris que segueixes.

```
import os
import json
from twitter import Api

CONSUMER_KEY = ‘’
CONSUMER_SECRET = ‘’
ACCESS_TOKEN = ‘’
ACCESS_TOKEN_SECRET = ‘’

api=Api(CONSUMER_KEY,CONSUMER_SECRET,ACCESS_TOKEN, ACCESS_TOKEN_SECRET)

with open("llista d’ids", 'w') as idlist:
      for line in api.GetFriendIDs(user_id = la teva id):
            idlist.write(str(line))
                  idlist.write('\n')
      idlist.close()
```

pe.py
Escriu amb una de cada set (aleatoriament) dels vint últims tuits de l'usuari. Escriu la data i el nombre de tuits en un arxiu.

```
import os
import json
import datetime
import random
from twitter import Api

CONSUMER_KEY = ‘’
CONSUMER_SECRET=’’
ACCESS_TOKEN =’’
ACCESS_TOKEN_SECRET = ‘’

api=Api(CONSUMER_KEY,CONSUMER_SECRET,ACCESS_TOKEN,ACCESS_TOKEN_SECRET)
```

```
def main():
      cont = 0
      i = 0
      with open('arxiu de dades', 'a') as output:
          with open('arxiu d'actualitzacions', 'a')as output_date:
                i+=1
                cont = 0
                output_date.write(str(datetime.datetime.now()))
                output_date.write(' ')
                print('begin at:', str(datetime.datetime.now()))
                with open('llista d'ids', 'r') as idlist:
                    for idline in idlist:
                        for line in api.GetUserTimeline(user_id=idline,
screen_name=None, since_id=None, max_id=None, count=None,
include_rts=True, trim_user=False, exclude_replies=False):
                            if random.randint(0, 255) % 7 == 0:
                                output.write("0")
                                output.write(str(line.id))
                                output.write('\t')
                                output.write(str(line.created_at))
                                output.write('\t')
                                output.write(str(line.retweet_count))
                                output.write('\t')
                                output.write(str(line.favorite_count))
                                output.write('\n')
                                cont += 1
                        print("end of user")
                print('end at:', str(datetime.datetime.now()))
                output_date.write(str(cont))
                output_date.write('\n')

if __name__ == '__main__':
      main()
```

Script R:

```
################################################
#         APPLIED TWITTER ANALYTICS            #
################################################

# DADES
data <- read.delim("clipboard",dec=",")                              #
Totes les dades (copiar de la pestanya "R" de l'excel "Dades.xlsx)
datalaboral = subset(data, Laboral.0...Cap.setmana.festiu.1. == 0)  #
```

```
Dades dies laborals
datafestiu = subset(data, Laboral.0...Cap.setmana.festiu.1. == 1)   #
Dades dies festius/cap de setmana


# CÀLCULS
nSL = length(datalaboral$Suma)
nSF = length(datafestiu$Suma)
n = nSL + nSF

ySL = mean(datalaboral$Suma)
ySF = mean(datafestiu$Suma)


s2SL = sum((datalaboral$Suma-ySL)^2)/(nSL-1)
s2SF = sum((datafestiu$Suma-ySF)^2)/(nSF-1)

s2 = ((nSL-1)*s2SL + (nSF-1)*s2SF)/(nSL + nSF - 2)

# Estadístic
z = ((ySL-ySF)/sqrt((s2SL/nSL)+(s2SF/nSF))); z

# P-valor
pvalor = pnorm(z); pvalor

# Interval de confiança
conf = 0.95
alpha = 1 - conf

ic1 = (ySL-ySF)-qnorm(1-(alpha/2))*sqrt(s2/n)
ic2 = (ySL-ySF)+qnorm(1-(alpha/2))*sqrt(s2/n)

IC = c(ic1,ic2); IC


# RESULTATS
summary(datalaboral$Suma)
summary(datafestiu$Suma)


# GRÁFIQUES

# Boxplot
boxplot(datalaboral$Suma, datafestiu$Suma, ylim = c(0,500), names =
c("Laborables", "Festius"), main = "Boxplot S (Rt + Likes)")

# QQnorm
prob = c(0,500)
qqnorm(datalaboral$Suma, prob)
qqline(datalaboral$Suma)
```

```
qqnorm(datafestiu$Suma, prob)
qqline(datafestiu$Suma)
```

Dades:

Estan adjuntes, en format digital, en un Excel.