Segmentación de Usuarios de Twitter

Agustín Capello
FaMAF - UNC

Problema

Segmentación (o división en clases), de usuarios de twitter en base al texto de sus tweets o sus retweets.

Datos

Se obtuvieron tweets/retweets entre las 18 a 00 hs del día 22 de noviembre con hashtags relacionados a las elecciones/ballottage en Argentina

A partir de estos tweets, se obtuvo toda la información relacionada a usuarios.

El objetivo es lograr descubrir inclinaciones políticas de usuarios, como así también grupos de usuarios relacionados por la misma temática.

Approach de la solución

Se tomó como un problema capaz de ser solucionado con machine learning, usando el algoritmo no supervisado de clustering K-Means.

Se vectorizaron los usuarios, cada uno como vector de features.

Los features utilizados fueron bag of words, bag of hashtags (ambos en base al texto de los tweets de un usuario) y relación con otros usuarios por retweets

Cada "feature" es una característica propia que posee cada usuario.

Algunos procesos que mejoraron la segmentación

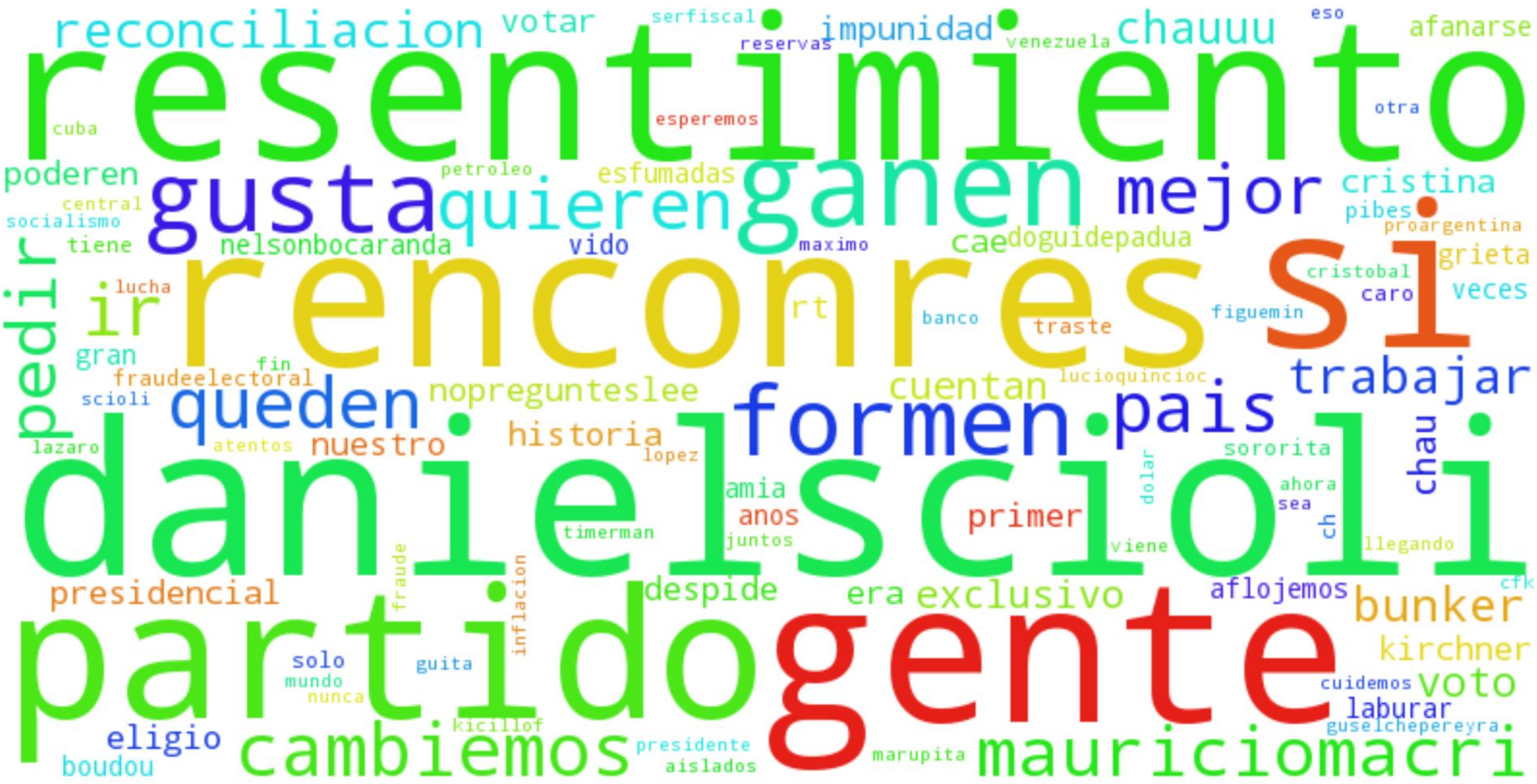
- Tokenizar/normalizar los tweets de todos los usuarios (en minúsculas, sin acentos, etc.).
- Quitar "function words" (artículos, preposiciones, pronombres, símbolos, etc.) de los tokens principales de cada usuario.
- Quitar palabras/tokens poco significativos y que se repiten mucho (ejm: argentina, balotaje2015, etc.).
- Dividir los usuarios en una mayor cantidad de clusters: con siete u ocho clusters se visualizaron mejor las diferencias que con tres o cuatro.
- Usar sólo features relevantes, y quitar algunos como la ubicación del usuario, su nombre o si está verificado.
- Utilización de diferentes métodos para evaluar la clasificación y analizar los resultados:
    a. Palabras más repetidas de cada cluster, y cantidad de usuarios que la usaron.
    b. Examinar los usuarios más relevantes de cada cluster (más cercanos a los centroides), y sus tweets.
    c. Examinar los features más representativos de cada cluster.
    d. Word-clouds de cada cluster.

Conclusiones - Resultados

El siguiente es un ejemplo de la evaluación de un modelo entrenado, es decir, el análisis de la clasificación.

En "Reporte.txt"

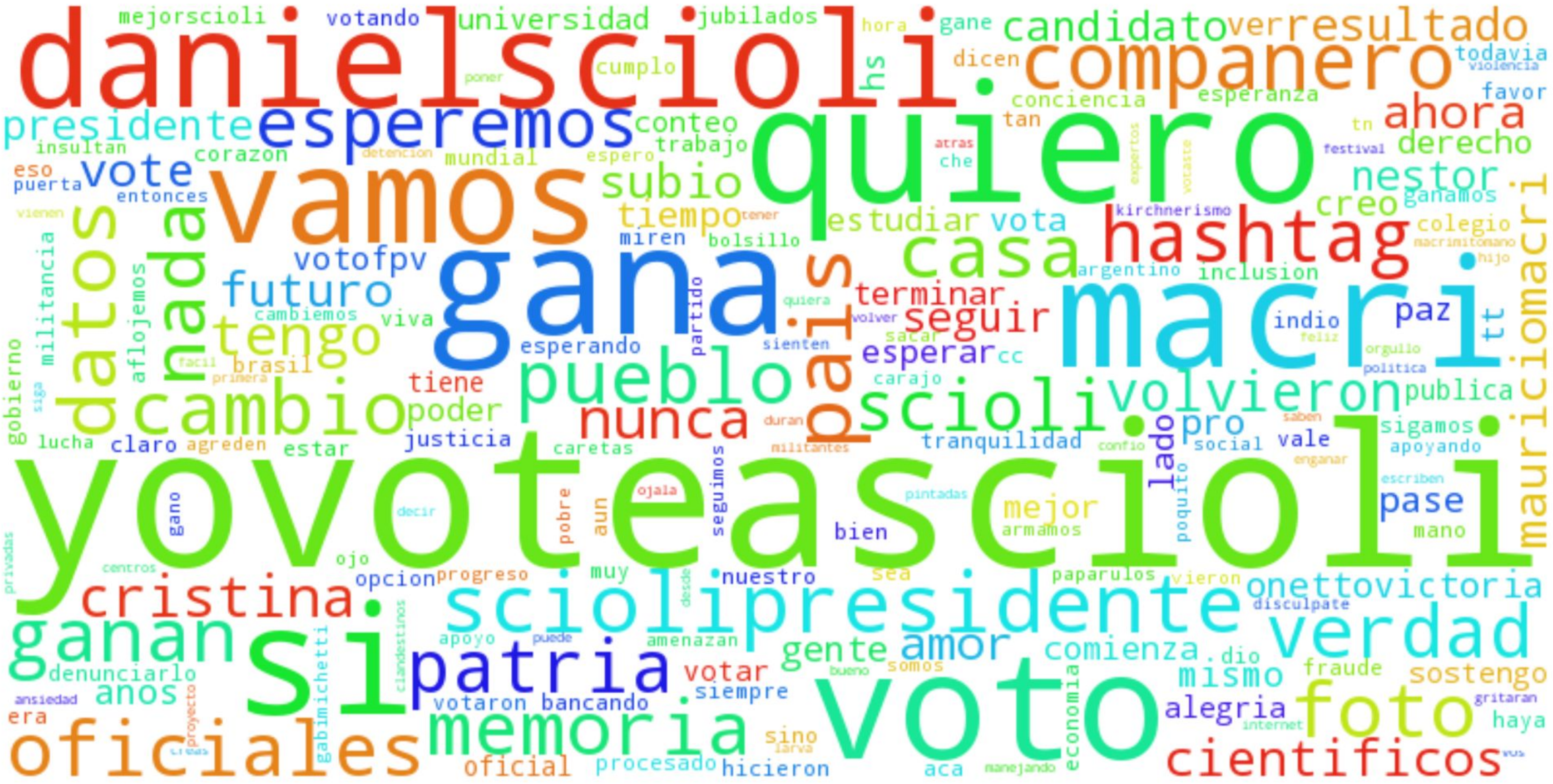

datooficialprovisorio
numeros
dentro
imbecil
triunfo
conferencia
claudio
telefecom
escucharlo
ceres
gt
danielscioli
escrutado
fpv
lamaxmaxima
propio
proximo
escrutadas
c5n
tendencia
mesa
elegistelt3
capital
vera
ensenada
indicaria
scioli
lider
deve
participacion
vantagem
em
puedas
acta
tv_publica
ahi
gano
pringles
ahora
cordoba
periodista
respetarlo
militante
presidenciales
presidente
rato
retiocba
mauriciomacri
eltreceoficial
gran
seria
cambiemos
sergiomassa
rosario
verdadero
espero
minutos
todonoticias
podran
soy
primeros
dar
pro
laburo
mauriciopresidente
eleger
rating
radiosuquia
governistas
loultimo
superior
americatv
local
falam
jbonifacio
cpugliese1968
votos
abrazo
laplata
macri
afano
escuela
mauricio
penalosa
transitocordoba
albertofernandez


reconciliacion
votar
serfiscal
reservas
impunidad
chauuu
venezuela
eso
afanarse
resentimiento
cuba
esperemos
otra
poderen
petroleo
esfumadas
gusta
quieren
ganen
mejor
cristina
central
socialismo
pibes
proargentina
tiene
nelsonbocaranda
vido
maximo
cae
doguidepadua
grieta
pedir
lucha
ir
renconres
si
cristobal
veces
rt
caro
banco
figuemin
traste
gran
fin
fraudeelectoral
lucioquincioc
trabajar
scioli
queden
nopregunteslee
cuentan
formen
pais
lazaro
atentos
nuestro
historia
lopez
chau
sororita
danielscioli
amia
anos
primer
dolar
ahora
sea
timerman
juntos
ch
viene
llegando
fraude
inflacion
presidencial
despide
era
exclusivo
aflojemos
cfk
bunker
partido
gente
kirchner
solo
guita
mundo
nunca
cuidemos
voto
laburar
kicillof
eligio
cambiemos
presidente
mauriciomacri
guselchepereyra
boudou
aislados
marupita

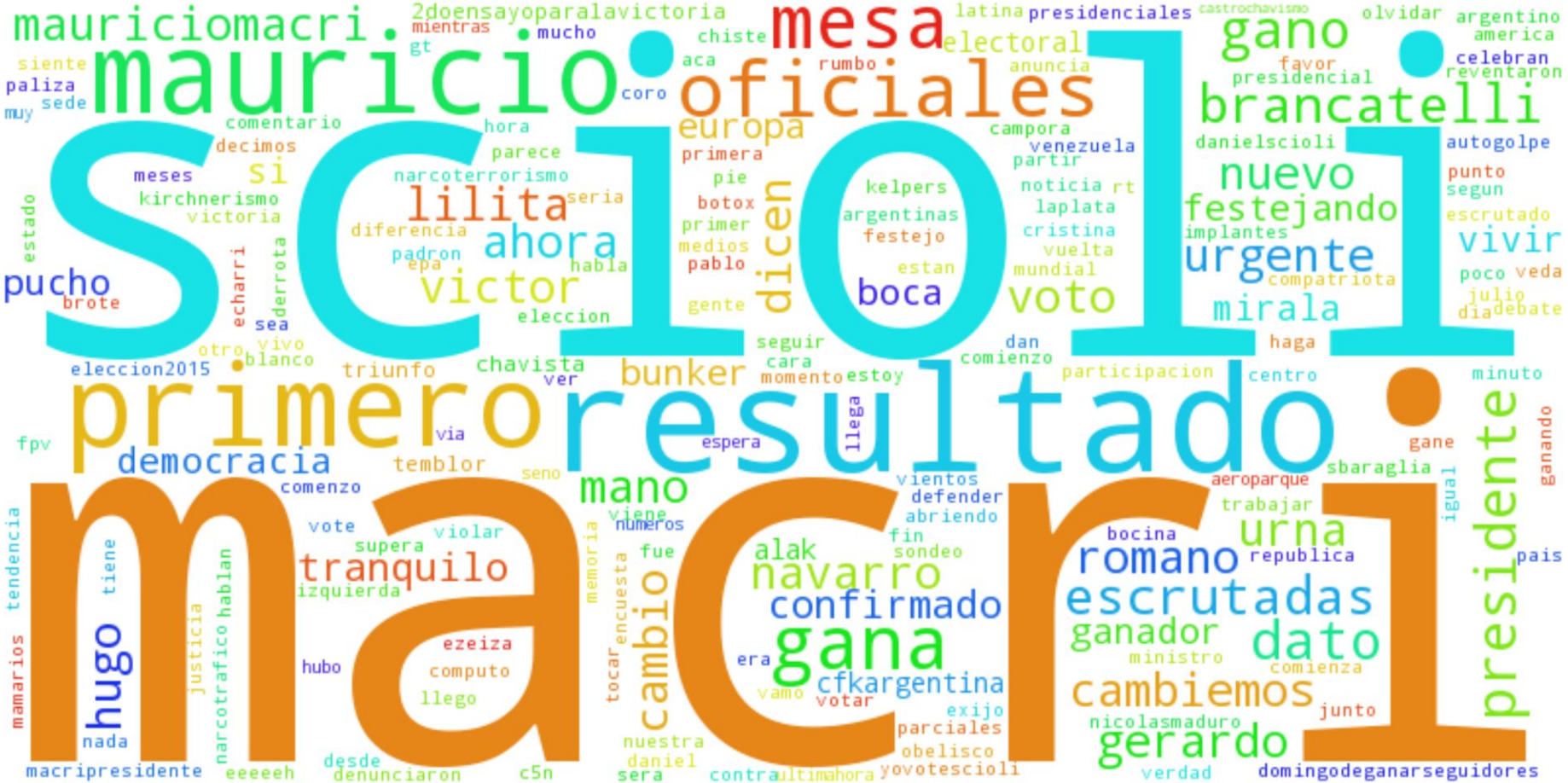

mauriciomacri
danielscioli
oficiales
dato
presidente
latinoamerica
primero
democracia
resultado
cambio
dia
voto
mesa
ano
historico
venezuela
si
nuevo
ahora
felicitaciones
presente
primer
pueblo
nisman
urna
lilita
acompanando
argentino
sanzernesto
desde
elisacarrio
victoria
participacion
escrutadas
anunciando
boletin
sola
socialismo
oficial
ganen
gana
pais
cambiemos
minuto
sera
vamo
gane
triunfo
esperanza
esperando
electoral
gusta
despues
partido
scioli
bueno
fin
viva
alegra
cfkargentina
felicidades
padron
yovoteamacri
falkland
politico
carajo
presidenciales
papelon
libre
bien
estan
siglo
comienza
formen
ojala
expulsado
estoy
indica
devolver
margarita
aqui
jubilo
generales
hermano
decoracion
estado
eso
dio
republica
diciembre
mariuvidal
solo
venezolano
gobierno
via
eres
alak
nuestro
ano
vuelva
wolffwaldo
votar
dime
muy
adios
casa
dara
c5n
saber
democrata
reacciono
interesa
eleccion2015
foto
saenzricardo
america
dire
hecho
urgente
golpe
historia
derecha
ver
tener
vota
justicia
indican
vzla
mm
dan
ramiro
punto
aire
bunker
alegria
estamos
tienen
hora
mucha
tiene
mencion
gente
medio
sirvio
ezequiel
cara
mira
hicieron
argentinas
tenemos
paz
julio
dijo
armen
cristina
quiero
boca
puede
rosada
entre
jose
junto
mejor
nacional
fue
momento
primera
stolbizer
6d
aplastante
espero
fernando
cerradas
liliantintori


macri
eleccion2015
resultados
scioli
mesa
ybtuc
mauriciomacri
facucarrillo
twitter
mayor
arrasooooo
cordoba
agusffox
menciones
terminamos
momento
bacomuna2
mauricio
ernesto
cantidad
danielscioli
caripela
famosos
datos
voto
julio
ganaron
casi
carajo
formaron
mastercard
escrutadas
precio
tiene
existe
sanz
lilita
partido
demas
caroya
maria
villa
colonia
ahora
fpv
c5n
gano
oficiales
analizan
vamos
tercero
peluquita
diferencia
lugar
carrio2
hoymacripresidente
tomatelas
lozanoveroam
fiscales
quiero
digo
ja
entre
15min
estamos
tendra
plza
yoli
felices
muy
elisacarrio
diegobranca
laburar
votar
tanta
pais
anda
padron
asistido
acto
serfiscal
normalidad
elisacarrioarg
montero
festejar
presidente
ciudadanos
tenes
llevado
pena
algun
coincido
comparto
martayatz
alrededor
seria
centrodecomputos
dan
repiten
habla
destacar
negativas
ojoelcorreo
mayo
eleccion
escrutados
estan
amigui
leo
pondremos
ciudad
marcos
positivas
miccaucr
senor
alak
terminos
estaria
tras
prometio
concurrido
chau
campora
pasaporte
millones
total
proargentina
gusta
esperar
ademas
nada
militonto
mensajito
mientras
fidelioa
solesalva
seguira
cierre
vago
adentro
si

macripresidente
mauriciomacri
cambiemos
macri
presidente
yocambio
cambio
oficial
campora
pobre
desde
scioli
vamos
gano
nuevo
resultado
cfkargentina
cristina
yovoteamacri
llego
si
muchachos
cambiemosconmacri
hora
llora
siente
rt
electo
graciaslilita
democracia
via
pidiendo
dijo
chaucristina
bravooooooo
latenesadentro
mauricio
kumpa
olvide
chau
ahora
miedo
voto
kretina
mesa
paz
bueno
datos
pais
descansar
ganomacri
venezuela
maduro
muy
primeros
campana
muerto
podra
pueblo
debe
tan
momento
jujuy
america
da
corrupto
corrupcion
libertad
fin
acta
oficiales
americalatina
siempre
falta
antes
cara
mauriciopresidente
entera
paso
bonelli
asesinado
fiscal
gente
chauuuuu
danielscioli
nunca
gana
bunker
911rosario
politico
contra
toda
aire
hacer
buena
22nov
yoli
eso
colombia
favor
otra
provincia
argentino
nissman
huevos
yovotoamacri
salir
gabimichetti
vivaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
votoexterior
tanta


yovoteascioli
danielscioli
quiero
macri
gana
vamos
voto
si
sciolipresidente
companero
esperemos
presidente
datos
nada
cambio
tengo
pueblo
pais
scioli
volvieron
hashtag
casa
oficiales
memoria
patria
ganan
cristina
verdad
foto
cientificos
mauriciomacri
candidato
resultado
ahora
nestor
subio
universidad
jubilados
mejorscioli
votando
ver
vote
futuro
votofpv
tiempo
nunca
seguir
terminar
estudiar
vota
creo
derecho
onettovictoria
amor
gente
comienza
mismo
alegria
sostengo
anos
votar
denunciarlo
votaron bancando
hicieron
procesado
oficial
siempre
mejor
pase
paz
pro
publica
sigamos
tranquilidad
justicia
poder
claro
agreden
estar
lucha
conteo
trabajo
corazon
mundial
conciencia
esperanza
cumplo
dicen
gane
hora
todavia
favor
tn
tan
hs
insultan
eso
puerta
entonces
miren
bolsillo
kirchnerismo
ganamos
colegio
argentino
inclusion
indio
tt
esperar
cc
esperando
brasil
cambiemos
viva
sacar
carajo
tiene
lado
vale
social
apoyando
caretas
seguimos
ojala
pobre
aun
gano
bien
armamos
poquito
mano
ojo
opcion
progreso
muy
nuestro
sea
paparulos
vieron
disculpate
apoyo
amenazan
dio
fraude
somos
economia
sino
aca
haya
era
gabimichetti
militancia
aflojemos
gobierno
partido
sienten
volver
quiera
espero
che
festival
orgullo
politica
escriben
pintadas
decir
centros
privadas
clandestinos
ansiedad
proyecto
internet
gritaran
manejando
bueno
puede
desde
duran
militantes
confio
enganar
saben
hijo
atras
tener
facil
primera
siga
vienen
poner
detencion
violencia
macrimitomano
feliz
expertos
votaste
vos


oficiales
resultado
mauriciomacri
dato
primero
democracia
escrutado
danielscioli
scioli
campora
gane
cambiemos
presidente
esperando
voto
cambio
estampitas
gano
quieren
militopresidente
electoral
momento
san
alak
bunker
sea
miedo
diferencia
fiscales
participacion
ojala
acabaron
mientras
seguir
nacional
aquello
bien
justicia
cara
centrodecomputos
daniel
llego
parte
militante
tanto
padron
mesa
tenemos
atencion
gente
nuevo
periodista
f5
manana
espero
concepto
sciolipresidente
c5n
fpv
vivo
ganen
importante
cfkargentina
gana
atentaron
quiero
nunca
saber
ver
desde
pais
parece
estan
minuto
pobreza
fin
propuesta
atullio
odio
historico
comienza
ultimo
mucho
hacer
dijimos
doy
futuro
salio
nacio
votar
diegobranca
crecer
siente
cristina
padre
vader
mayo
mejor
sobres
mauricio
otra
pusieron
dame
julio
lilita
junto
nombre
partido
yovoteamacri
genio
sanguches
estamos
termina
tanta
votaron
dia
politica
salame
felicitamos
reaparecio
final
continuidad
extrano
ganador
venezuela
votaste
trabajar
responde
afuera
escuela
estaran
argentino
acta
hubo
favor
ahora
perdidos
linda
yaaaaaaaaaaaaaa
disponibles
computo
darth
tenes
habla
alegria
deportista
festejo
chau
prensa
ano
politico
palabra
si
cayetano
vamo
brancatelli
dio
muchachos
caos
feliz


scioli
macri
mauricio
resultado
primero
presidente
mauriciomacri
mesa
oficiales
gano
brancatelli
nuevo
festejando
urgente
vivir
lilita
ahora
victor
pucho
boca
voto
mirala
dicen
europa
electoral
democracia
tranquilo
mano
cambio
navarro
confirmado
gana
cfkargentina
alak
romano
escrutadas
ganador
dato
cambiemos
gerardo
urna
hugo
bunker
chavista
triunfo
temblor
eleccion2015
macripresidente
2doensayoparalavictoria
mucho
chiste
latina
presidenciales
castrochavismo
olvidar
argentino
america
celebran
reventaron
favor
presidencial
anuncia
rumbo
aca
coro
siente
paliza
sede
muy
comentario
hora
decimos
parece
si
meses
kirchnerismo
victoria
narcoterrorismo
seria
diferencia
padron
epa
habla
primera
pie
botox
primer
medios
pablo
kelpers
argentinas
festejo
estan
noticia
rt
laplata
cristina
vuelta
mundial
campora
venezuela
partir
danielscioli
autogolpe
punto
segun
escrutado
implantes
compatriota
poco
veda
julio
dia
debate
estado
echarri
derrota
brote
sea
otro
vivo
blanco
eleccion
gente
seguir
cara
momento
estoy
dan
comienzo
haga
participacion
centro
minuto
ver
via
espera
llega
fpv
comenzo
vote
violar
supera
izquierda
ezeiza
computo
llego
hubo
tiene
nada
justicia
narcotrafico
hablan
mamarios
tendencia
eeeeeh
desde
denunciaron
c5n
sera
contra
ultimahora
yovotescioli
verdad
domingodeganarseguidores
nuestra
daniel
obelisco
parciales
exijo
votar
vamo
era
tocar
encuesta
memoria
numeros
viene
vientos
defender
abriendo
fin
sondeo
fue
aeroparque
trabajar
bocina
republica
sbaraglia
gane
ganando
igual
pais
ministro
comienza
junto
nicolasmaduro