SENTIMENT ANALYSIS & CRAWLING CON SOFTWARE LIBRE

Christian López - @christianlrcalo - info@christianlr.es

//Codery_

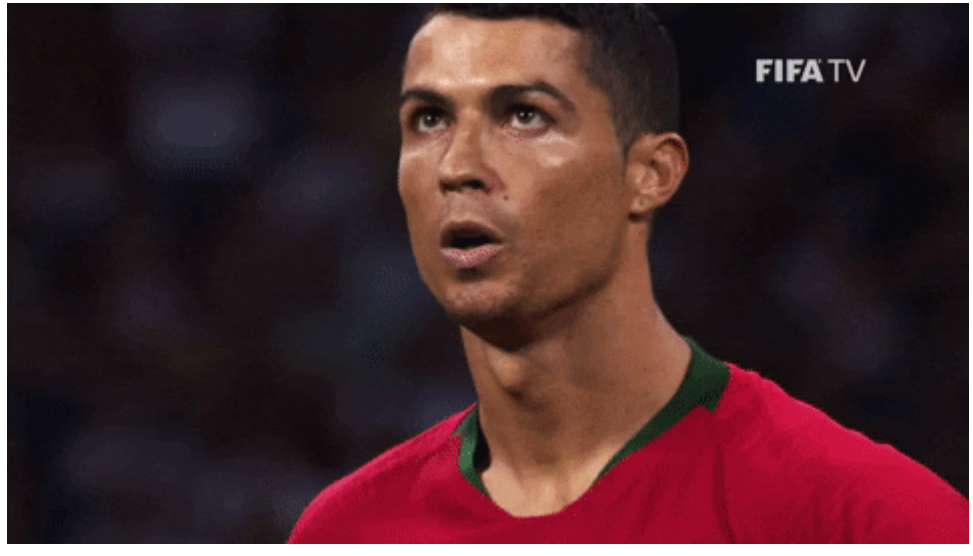
SOBRE CHRISTIAN

- Ingeniero Informático - USC
- Master of Applied Data Science - Michigan University
- Desarrollador Drupal en Codery_ desde 2017
- Devops - SysAdmin

CONTEXTO/ORIGEN PROYECTO

"Tenemos que montar algo para recuperar datos de las redes y que lo vean..."

... y tenemos sobre un mes!



BUENO, VAMOH A ELLO!

PRIMEROS PASOS

- Análisis - Diseño - Desarrollo - Pruebas => NOP!
- Java - Php - Python? => Python - Tweepy
- Facebook - Twitter - Instagram? => Twitter
- SQL - NoSQL? => PostgreSQL
- Entrega? => PgAdmin con consultas pre-cargadas

FASE1 - PRIMERA ENTREGA

Crawler de datos (cuenta gratuíta - límites)

Relación Tweet - Usuario

Búsqueda según hashtags y palabras clave (crontab)

Acceso a PgAdmin con consultas precargadas

FEEDBACK?

Quiero más!

Y podemos?, y podemos?, y si?

FASE2 - MÁS DIMENSIONES

Geolocalización de los datos

Datos macro y socioeconómicos

Datos meteorológicos

Datos de género

Configuración de búsqueda dinámica

Diferentes grupos de búsqueda

GIPD (Gráficas - Informes - Pdfs y Dashboard)

VENGA CHAO!

PROBLEMAS

- Ejecución estática con crontab
- Servidor por grupo de búsqueda => gestión de recuros (APIs)
- Relación/fusión de datos de todas las dimensiones
- Actualización versiones Tweepy - Api Twitter
- Generador informes + estadísticas + API + Dashboard
- Búsqueda/análisis de servicios (ubicación, geo, etc...)

ENTREGA/SOLUCIÓN

Geolocalización -> locationIQ

Datos macro -> La propia entidad

Meteorología -> weatherapi

Género -> genderApi.io

Búsqueda dinámicas -> ficheros yaml

Grupos de búsqueda -> varias máquinas (de 1 a 4)

Aumento sustancia de datos -> Sistema interno de generaición de stadísticas

GIPD (Gráficas - Informes - Pdfs y Dashboard) ->

Wkhtmltopodf + VueJs + PHP + SlimAPI

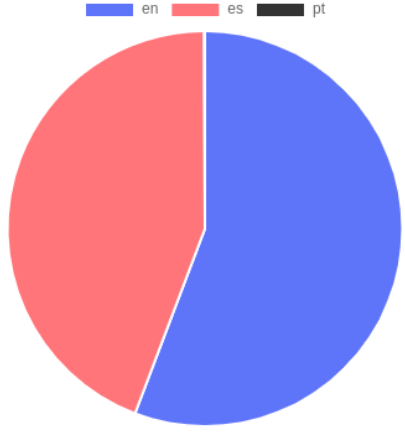
101894730	29038
TWEETS	AVERAGE LAST WEEK

18042650	8839
USERS	AVERAGE LAST WEEK

22462455	201
GENDERS	AVERAGE LAST WEEK

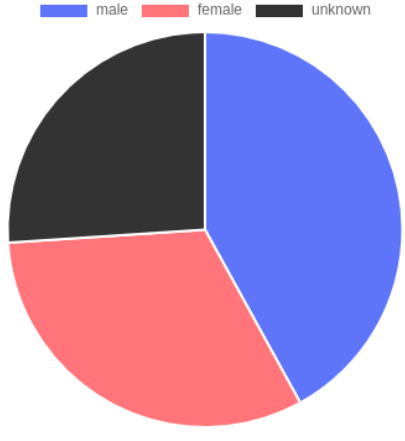
% of languages

% of top3 languages stored



% of genders

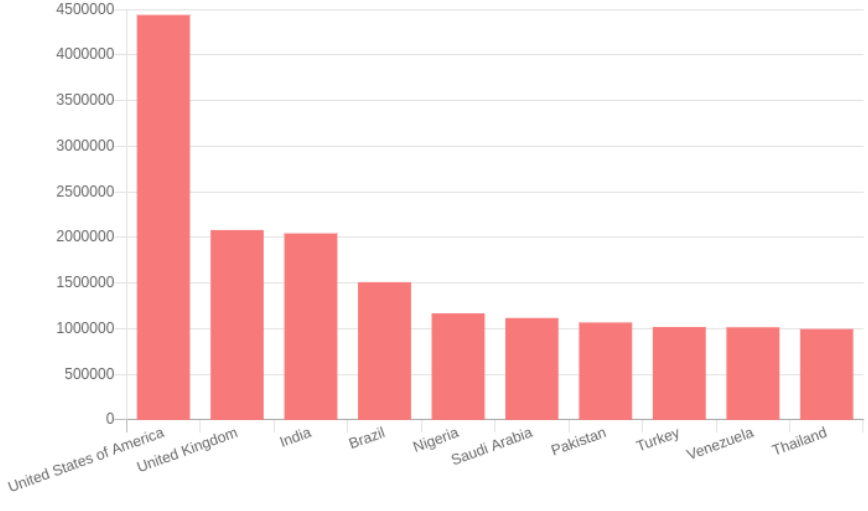
% of genders stored



Languages	User places	Keywords	Hashtags
0	12510135	18	27

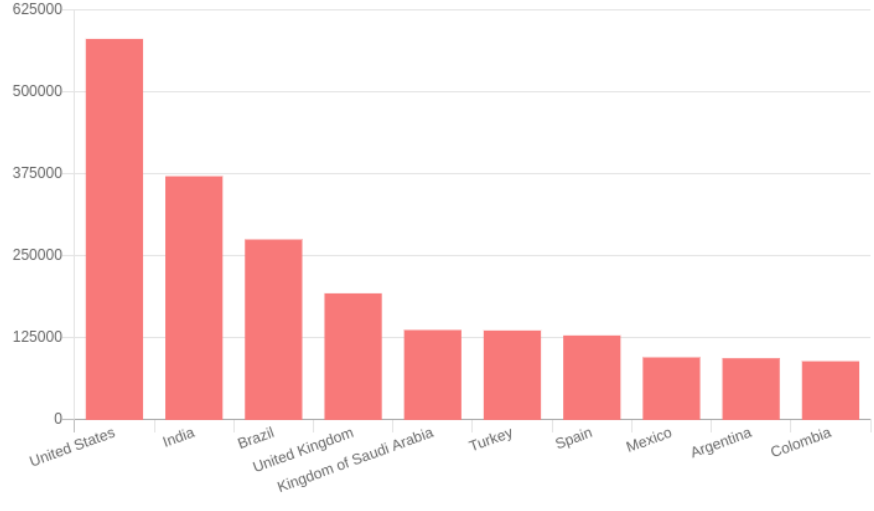
Top 10 user places

Top 10 of tweets by user location



Top 10 tweet places

Top 10 of tweets by tweet location



FASE3 - SENTIMENTANALYSIS

"Quiero saber el sentimiento o preocupación sobre un X en este dataset"

Si puede ser tanto en inglés como en español

Y además, si puede ser, dividido en regiones/países

Y claro está, envíame todas las dimensiones

Tenéis 1 mes!, como mucho mes y medio que tenemos que publicar el estudio

NOOP!, ESTA VEZ NOP!

NO!, PORQUÉ?

- No hay nada formateado/limpio
- Máquinas de crawler sin conexión entre ellas
- Desconocimiento de los datos actuales. Las estadísitcas son básicas
- Gran dispersidad de datos (culturas, idiomas, contextos, etc...)
- Diccionarios/corpus orientados a felicidad (+/-)

ENTREGA

- Reunificación de todas las máquinas en una sola
- Reorganización de las ddbb a una misma
- Reconfiguración de los accesos a Twitter y métodos de búsqueda
- Limpieza de datos duplicados (ubicación, género, etc..)
- Mejora del generador de estadísticas

- Restricción de países a los que tengan más información
- Restricción a 2 idiomas. ES/EN
- Necesidad de Traducir ES a EN. OpenNMT
- Reorientacion de preocupación hacia felicidad por preocupación
- Apoyar el proceso mediante datos estadísticos

ENTREGA TÉCNICA

PostgresSQL de + 300GB (800M registros)

Creación de índices y varias vistas

Automatización en las estadísticas con systemD/servicios

Instalación de Jupyter Notebook

MODELOS/LEXICONS DE SENTIMENT ANALYSIS

VADER

LabMT

Afinn

Data: 24-10-2022

Totais

- Tweets: 180728089.00
- Tweets Rts: 2011476.00
- Tweets non Rts: 178716613
- Tweets onte: 332796.00
- Media de tweets última semana por día: 316259
- Tweets de contas verificadas: 3926304.00
- Tweets de contas verificadas: 3926304.00
- Tweets de contas non verificadas: 176665993.00
- Usuarios: 42362785.00
- Tweets xeoreferenciados: 3693619.00
- Tweets con localización de país: 3689945.00
- Tweets con localización de país e NON RT: 3689887.00

In [3]:
```
# Load models and instances of labMT
lang = 'spanish'
#labMT_dic, labMT_matrix, labMTwordList = emotionFileReader(stopval=0.0, lang=lang, returnVector=True)
labMT_dic, labMT_matrix, labMTwordList = emotionFileReader(stopval=1.0, lang=lang, returnVector=True)
```

In [4]:
```
# Testing main operations
print("--- EJEMPLOS CON EL DICCIONARIO DIRECTAMENTE ---")
print("Ejemplo con la palabra risa")
print(labMT_dic['risa'])
print("Ejemplo con la palabra mal")
print(labMT_dic['mal'])
```

```
--- EJEMPLOS CON EL DICCIONARIO DIRECTAMENTE ---
Ejemplo con la palabra risa
['61', '8.14', '0.9691', '5060', '531', '--', 'laughter\n']
Ejemplo con la palabra mal
['9749', '3.02', '1.6349', '678', '130', '438', 'wrong\n']
```

In [5]:
```
# More examples to see working modes
print("EJEMPLOS DE LAS PALABRAS CON MEJOR PUNTUACION")
print(labMT_matrix[0:5])
print(labMTwordList[0:5])
```

```
EJEMPLOS DE LAS PALABRAS CON MEJOR PUNTUACION
[8.68, 8.6, 8.48, 8.42, 8.42]
['amor', 'felicidad', 'felicidades', 'paz', 'jajajajajajajajaja']
```

In [6]:
```
# More examples to see working modes
print("EJEMPLOS DE LAS PALABRAS CON PEOR PUNTUACION")
print(labMT_matrix[-5:])
print(labMTwordList[-5:])
```

```
EJEMPLOS DE LAS PALABRAS CON PEOR PUNTUACION
[1.66, 1.66, 1.64, 1.62, 1.54]
['cáncer', 'matar', 'matando', 'murió', 'muerte']
```

In [7]:
```
# EXAMPLE
# 1. BASIC EMOTION (INCLUDING STOPWORDS)
sentmient, matrix = emotion('Hola muy bien que tal todo bien?', labMT_dic, shift=True, happsList=labMT_matrix)
print(sentmient)
```

```
7.196
```

In [8]:
```
# 2. Get matrix of stop words
stopWords_matrix = stopper(matrix, labMT_matrix, labMTwordList, stopVal=1.0)
#print(stopWords_matrix)
```

In [9]:
```
# 3. EXECUTE NEW EMOTION WITH STOPWORDS MATIRX
sentmient_no_stopwords = emotionV(stopWords_matrix, labMT_matrix)
print(sentmient_no_stopwords)
```

```
7.196000000000001
```


maana
nico camino
claro
aunque
peor
pasado
amor
otros
punto
tema
junto
quiere
sigue
lugar
est
dice
mucho
seguir camino
hijo
luego
qu
vez
poco
sera
nadie
llegar
paso
est camino
cmo
pueblo
ahora
pue
debe
mejor
mal camino
amigo
gran
casa
nada
adem
gracia
hombre
decir
estn
bueno
vida
toda veces
hizo
mismo
camino
mano
m
problema
siempre
despu
otra
camino vida
son
esto
camino seguir
mundo
parece
seguro
aqu
persona
camino ms
pa
familia
mitad camino
buen camino
nuevo
grande
idea
historia
sigue camino
Camino Santiago
camino camino
poltica
nunca
gente
hecho
final
alguien
tambin
ah
falta
partido
hoy
gobierno
tena
camino correcto
camino casa
equipo
meno
mejor camino
importante
calle
creo
slo
momento

¿CONCLUSIONES?

- Hay demasiado humo y sobredimesión en internet
- La planificación es imposible
- El rendimiento siempre queda en 2º plano
- Lo que importa es la visualización
- Nunca vas a tener tiempo a hacerlo cómo te quieres (limpieza, formateo, etc...)
- SIEMPRE FALTAN DATOS

¡GRACIÑAS!

Dudas, ideas, mejoras, etc..

info@christianlr.es

@christianlrcalo