FREAK DATA EXPERIENCE

Christian López - @christianlrcalo - info@christianlr.es

29-10-2022

SOBRE CHRISTIAN

- Ingeniero Informático - USC
- Master of Applied Data Science - Michigan University
- Desarrollador Drupal
- Devops - SysAdmin


FASE1 - ORIGEN DEL PROYECTO

Carrera terminada en 2016 (ATPSM)

Máster terminado en 2018

FASE1 - ORIGEN DEL PROYECTO

Carrera terminada en 2016 (ATPSM)

Máster terminado en 2018

Comienzo verano 2017

FASE1 - ORIGEN DEL PROYECTO

Carrera terminada en 2016 (ATPSM)

Máster terminado en 2018

Comienzo verano 2017

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

FASE1 - ORIGEN DEL PROYECTO

Carrera terminada en 2016 (ATPSM)

Máster terminado en 2018

Comienzo verano 2017

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

... y tenemos sobre un mes!


FIFA TV
BUENO, VAMOH A ELLO!

PRIMEROS PASOS

- Análisis - Diseño - Desarrollo - Pruebas => NOP!

PRIMEROS PASOS

- Análisis - Diseño - Desarrollo - Pruebas => NOP!
- Java - Php - Python? => Python - Tweepy

PRIMEROS PASOS

- Análisis - Diseño - Desarrollo - Pruebas => NOP!
- Java - Php - Python? => Python - Tweepy
- Facebook - Twitter - Instagram? => Twitter

PRIMEROS PASOS

- Análisis - Diseño - Desarrollo - Pruebas => NOP!
- Java - Php - Python? => Python - Tweepy
- Facebook - Twitter - Instagram? => Twitter
- SQL - NoSQL? => PostgreSQL

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

PRIMERA ENTREGA

Crawler de datos (cuenta gratuíta - límites)

Relación Tweet - Usuario

Búsqueda según hashtags y palabras clave

Optimización de almacenamiento (no repetidos, exclusión)

Principio de automatización - independencia

FEEDBACK?

FEEDBACK?

FEEDBACK?

Quiero más!

FEEDBACK?

Quiero más!

Y podemos?, y podemos?, y si?

FASE2 - MÁS DIMENSIONES

Geolocalización: tweets y usuarios

Datos macro y socioeconómicos

Datos meteorológicos

Datos de género

Configuración de búsqueda dinámica

Diferentes grupos de búsqueda

FASE2 - MÁS DIMENSIONES

Geolocalización: tweets y usuarios

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 => servicios/automática
- 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

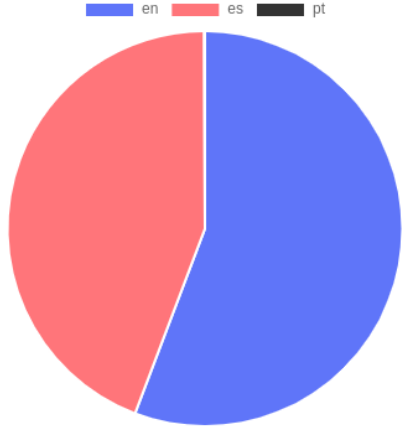
101894730	29038
TWEETS	AVERAGE LAST WEEK

18042650	8839
USERS	AVERAGE LAST WEEK

22462455	201
GENDERS	AVERAGE LAST WEEK

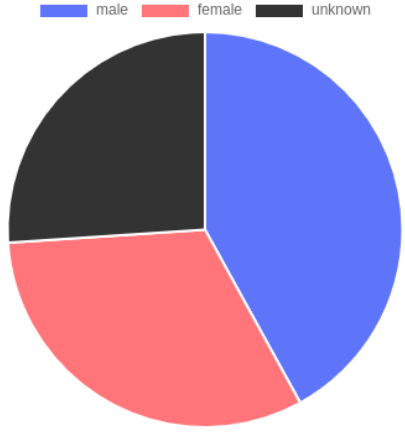
% of languages

% of top3 languages stored



% of genders

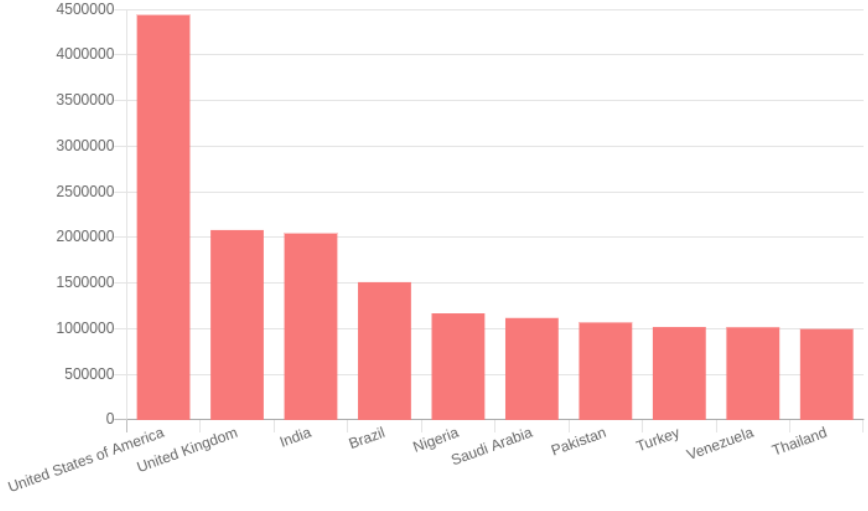
% of genders stored



Languages	User places	Keywords	Hashtags
0	12510135	18	27

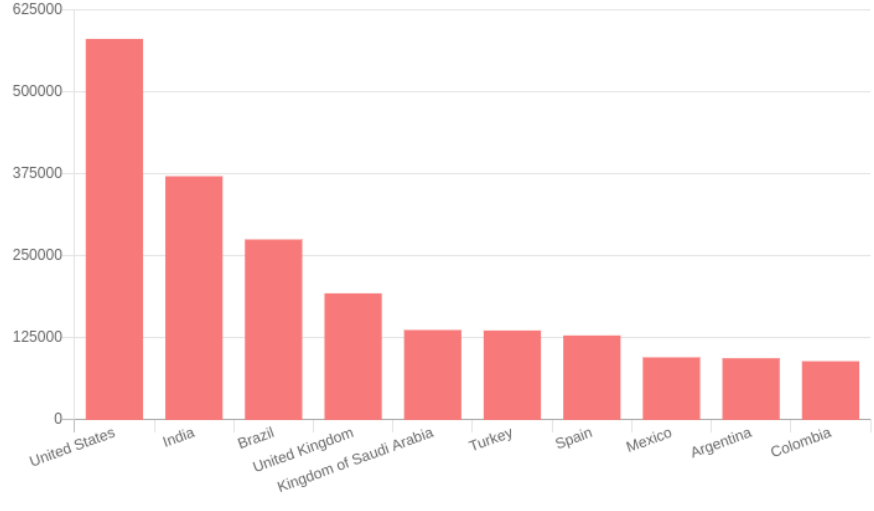
Top 10 user places

Top 10 of tweets by user location



Top 10 tweet places

Top 10 of tweets by tweet location



FASE3 - ML Y SENTIMENTANALYSIS

"Saber sentimiento/preocupación sobre un X en este dataset"

Si puede ser tanto en inglés como en español

Y además si puede ser dividido en regiones

Y claro, relacionando todas las dimesiones

FASE3 - ML Y SENTIMENTANALYSIS

"Saber sentimiento/preocupación sobre un X en este dataset"

Si puede ser tanto en inglés como en español

Y además si puede ser dividido en regiones

Y claro, relacionando todas las dimesiones

AHH! y tenemos un par de meses

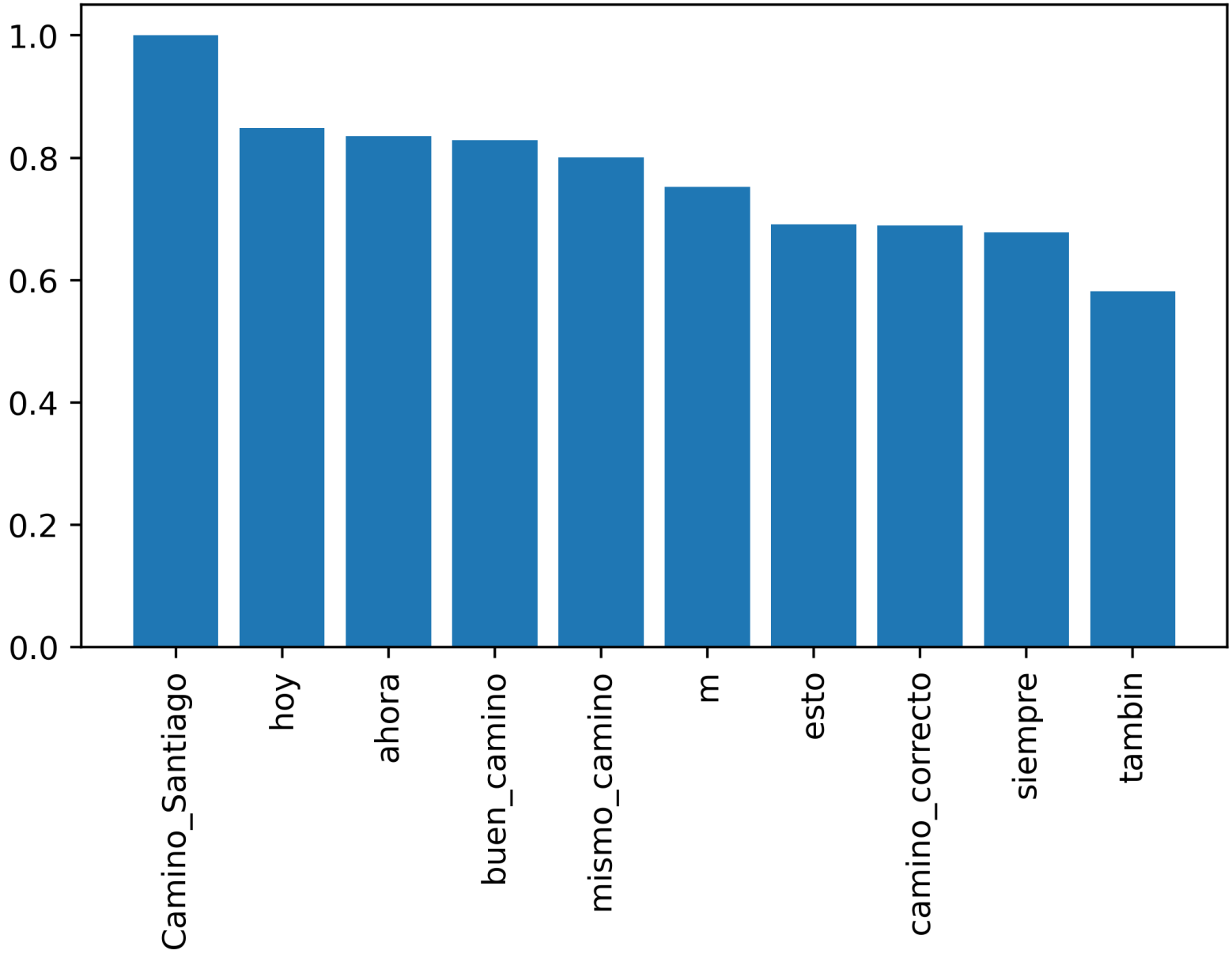

NOOP!, ESTA VEZ NOP!

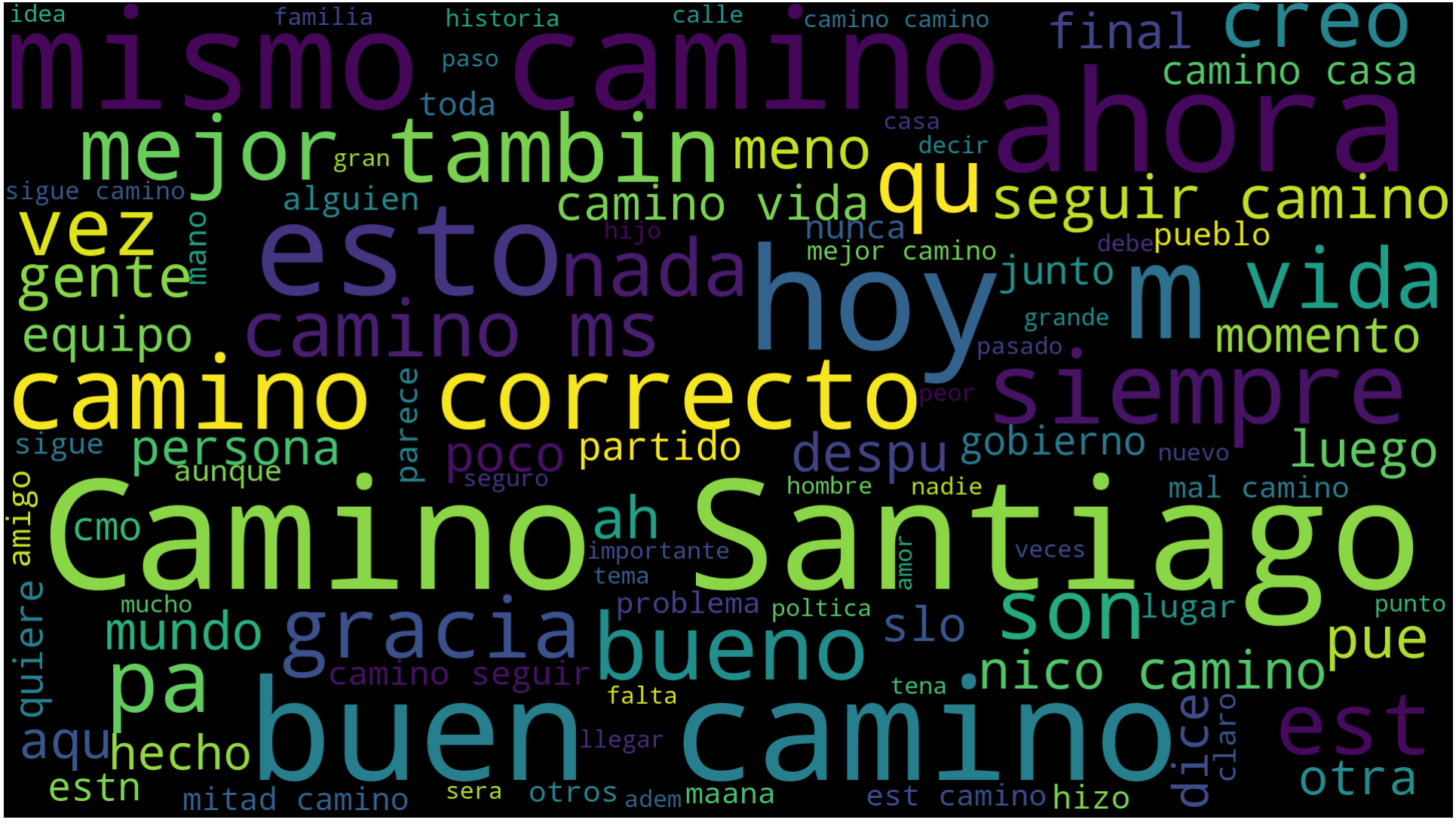
PROBLEMAS

- No hay nada formateado/ limpio
- Máquinas de crawler sin conexión entre ellas
- Necesidad de clasificación de datos => mejora de reporting
- Gran dispersidad de datos (culturas, idiomas, contextos, etc...)
- Diccionarios/corpus orientados a felicidad (+/-)

ENTREGA

- Reestructuración de las máquinas => una sola
- Reorganización de las ddbb y relaciones => una ddbb
- Restricción a principales países
- Restricción a 2 idiomas
- Reorientacion de preocupación hacia felicidad
- Apoyar el proceso mediante datos estadísticos


Frequency words of all dataset
1.0
0.8
0.6
0.4
0.2
0.0
Camino_Santiago
hoy
ahora
buen_camino
mismo_camino
m
esto
camino_correcto
siempre
tambin


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


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

BONUS - REQUISITOS EXTRA

BONUS - REQUISITOS EXTRA

Quiero un poco de todo en Twitter

BONUS - REQUISITOS EXTRA

Quiero un poco de todo en Twitter

ML con video

BONUS - REQUISITOS EXTRA

Quiero un poco de todo en Twitter

ML con video

Interactuación con mapas

BONUS - REQUISITOS EXTRA

Quiero un poco de todo en Twitter

ML con video

Interactuación con mapas

Encuestras embebidas dinámicas

¿CONCLUSIONES?

¿CONCLUSIONES?

- Hay demasiado humo (foros, documentación, ...)

¿CONCLUSIONES?

- Hay demasiado humo (foros, documentación, ...)
- La planificación es imposible

¿CONCLUSIONES?

- Hay demasiado humo (foros, documentación, ...)
- La planificación es imposible
- El rendimiento siempre queda en 2º plano

¿CONCLUSIONES?

- Hay demasiado humo (foros, documentación, ...)
- La planificación es imposible
- El rendimiento siempre queda en 2º plano
- Lo que importa es la visualización

¿CONCLUSIONES?

- Hay demasiado humo (foros, documentación, ...)
- La planificación es imposible
- El rendimiento siempre queda en 2º plano
- Lo que importa es la visualización
- SIEMPRE FALTAN DATOS

¡GRACIÑAS!

Dudas, ideas, mejoras, etc..

info@christianlr.es

@christianlrcalo