

PRÁCTICA FINAL  
BOOTCAMP  
BIG DATA, AI & ML  
**EDICIÓN V**  
**EQUIPO**  
FERDALEAMOR

# TERTULIA MATUTINA

---

EXTRACCIÓN  
AUTOMÁTICA DE  
**SENTIDO**  
EN CONTENIDOS DE  
ACTUALIDAD

## EQUIPO



## FER DALE AMOR



ROBERTO MARTÍNEZ CAMACHO

MARCOS GUTIÉRREZ HERNÁNDEZ

ALBERTO MÉLIDA

ALEJANDRO LÓPEZ SÁNCHEZ

DAVID SÁNCHEZ MAYORAL

FERNANDO JARILLA VALIENTE



## BUSCANDO OPORTUNIDADES

‘Clipping’ sin derechos para medios digitales en las principales instituciones del Estado

Por Merca2.es - 20/04/2018

El cuasi monopolio de Kantar Media con el “clipping”: la Compañía se embolsa millones de euros gracias a elaborar resúmenes de prensa para organismos públicos

Clipping. Competencia multa a la AEDE con 225.000 euros por monopolio

## ¿EN QUÉ SECTORES PODRÍAMOS APLICAR BIG DATA, ML & AI DE FORMA DISRUPTIVA?

Las burbujas informativas: qué son y cómo afectan en nuestras vidas

### No vives en una burbuja informativa

Un nuevo estudio debilita la teoría de las cámaras de eco y muestra que los españoles consumen medios de todas las ideologías

### La posverdad de la burbuja informativa



¿Cómo se van a adaptar las empresas de clipping a la generalización de los muros de pago en la prensa digital?

Twitter y la cámara de eco

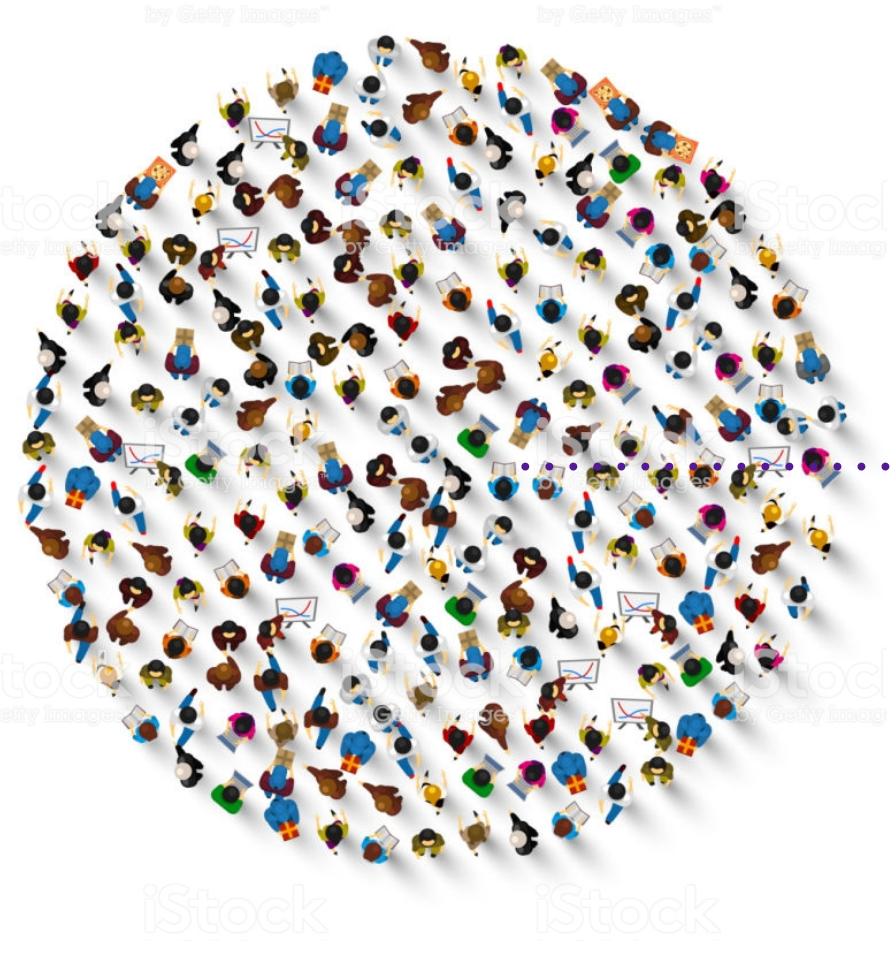
POLÍTICA



***El análisis de noticias en prensa digital y redes sociales nos permitirá poner en práctica de forma transversal buena parte de los conocimientos adquiridos.***



*Aunque a priori podría parecer un campo muy trillado, creemos que ofrece oportunidades muy interesantes tanto desde un punto de vista de la empresa como de cara al consumidor final de noticias.*



## PROPUESTA DE VALOR

**B2B**

**B2C**

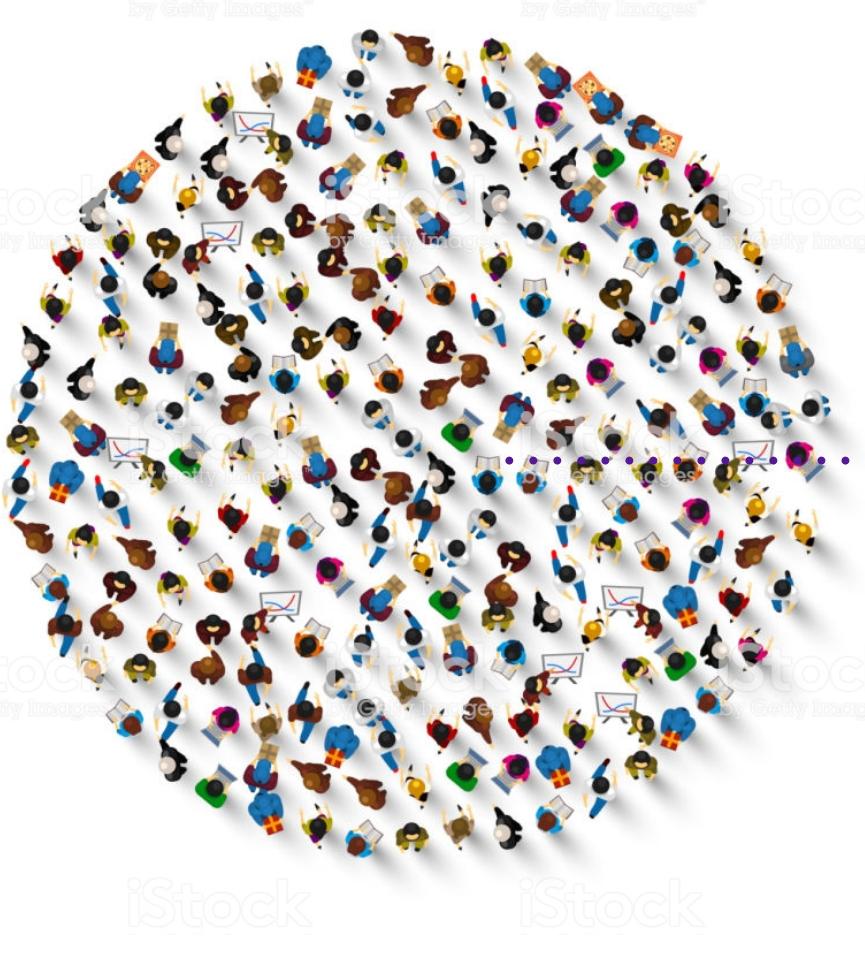
El servicio de press clipping hoy va mucho más allá de que escanear recortes de prensa.

Pero sigue siendo un mercado cerrado copado por un puñado de empresas que prestan un servicio muy caro a grandes empresas e instituciones.

El resto tira de Google News, Alertas de Google y periodistas precarios.

Si conseguimos crear **informes personalizados de seguimiento de la actualidad en medios y redes sociales, con extracción de topics y resúmenes** que faciliten el seguimiento de la reputación online, facilitaremos mucho el trabajo de los departamentos de comunicación.

Los expertos en comunicación y relaciones públicas podrán centrarse en generar valor añadido diseñando y ejecutando estrategias y campañas de PR y PA.



## PROPUESTA DE VALOR

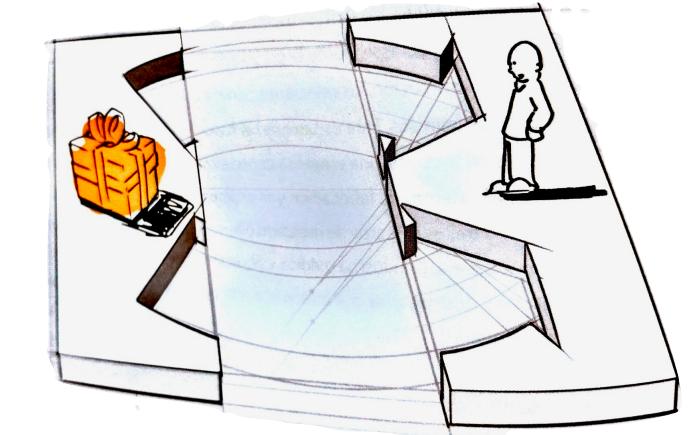
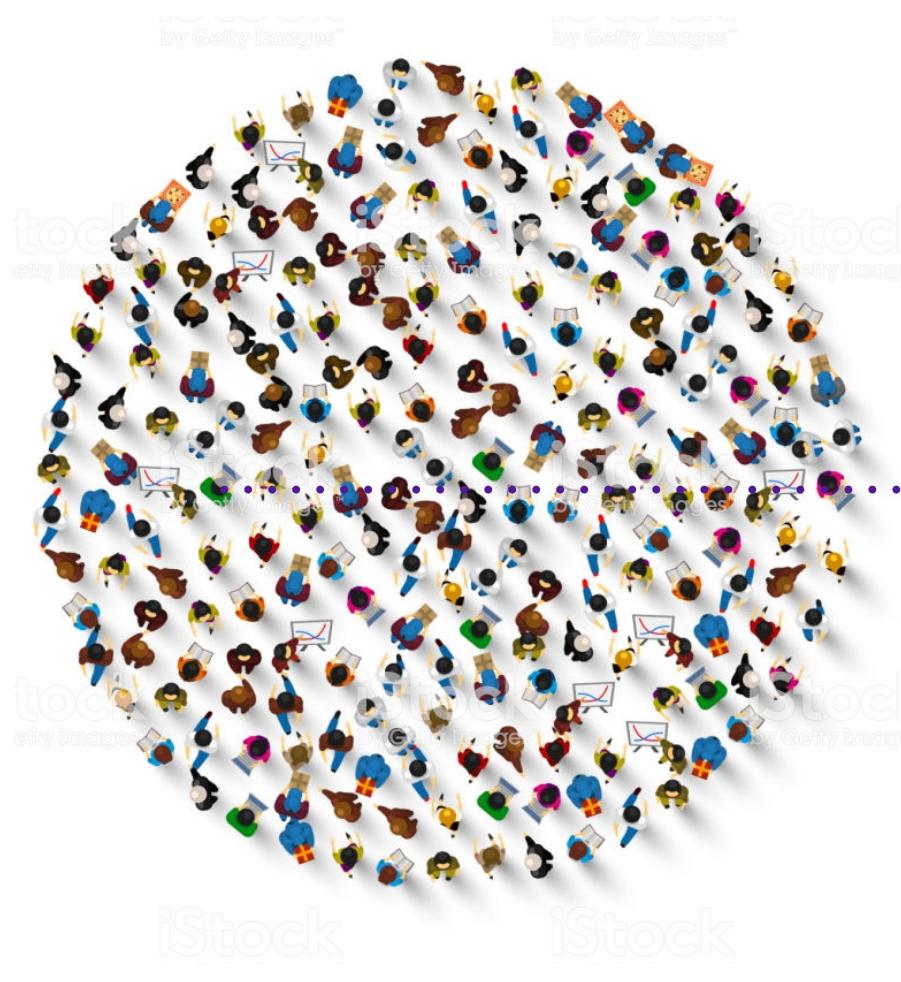
B2B

B2C

De cara al usuario final, con la llegada (esta vez sí) de las pasarelas de pago. Los usuarios agradecerán por un servicio que se encargue de leer toda la prensa y realice un resumen de lo más interesante.

Este resumen, a modo de tertuliano artificial, se encargaría de extraer y mostrar los diferentes puntos de vista de las noticias, ayudando a evitar las burbujas y cámaras de eco que fomentan los algoritmos de las redes sociales.

“Los algoritmo de Facebook o Twitter está optimizados para robarte tu tiempo. Con el Tertuliano Matutino, dinos cuánto tiempo quieres dedicar al día y nos aseguraremos de que estás al tanto de los temas que te interesan con información amena, fiable y plural.”

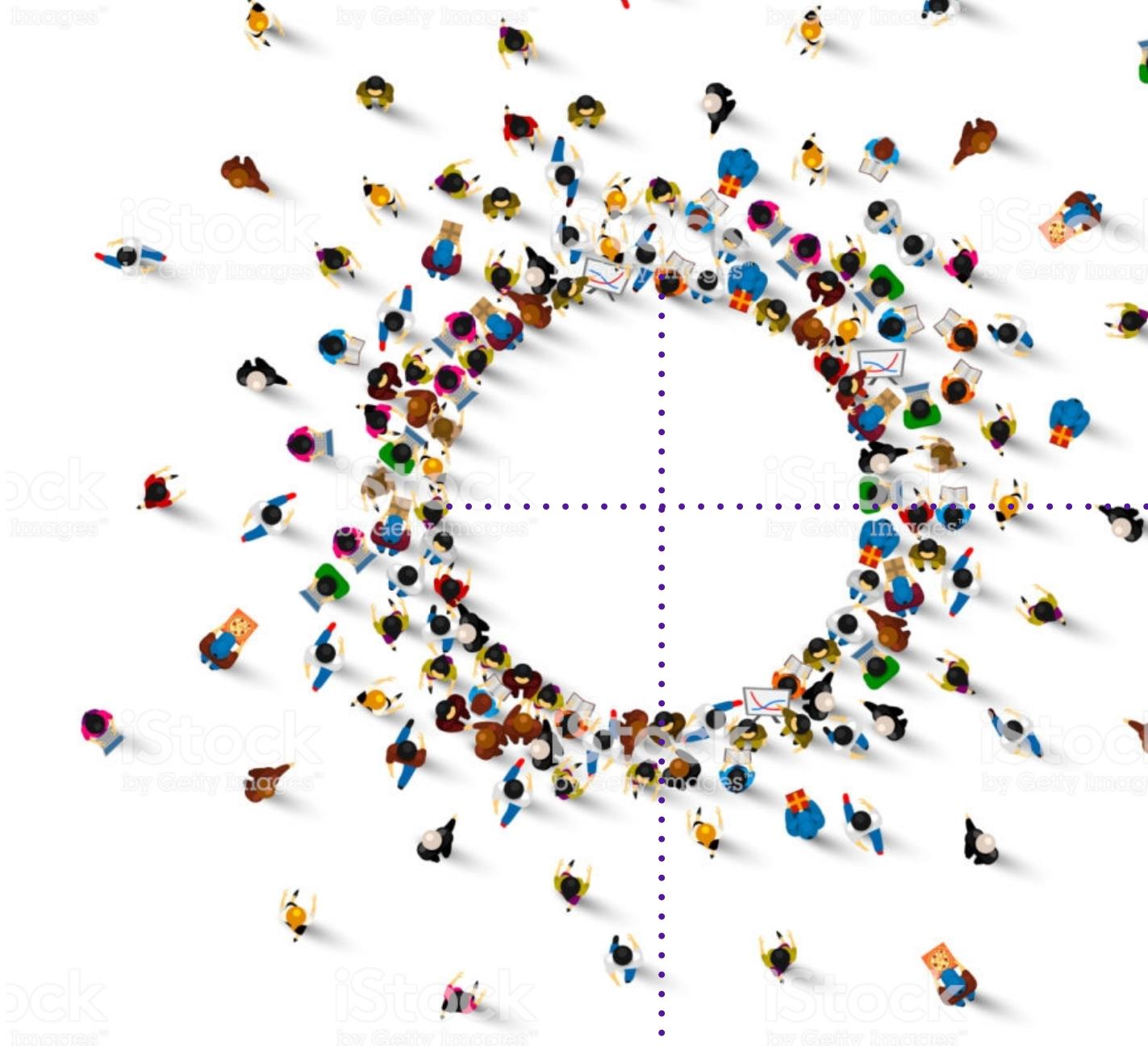


# ALCANCE DEL PROYECTO

**SCOPE**

En esta práctica vamos a:

- ✓ Recopilar las noticias a diario de una serie de periódicos
- ✓ De ahí sacar los topics más relevantes del día
- ✓ Cruzarlos con la repercusión que están teniendo en Twitter
- ✓ Obtener una presentación de los temas de actualidad del día



# FUERA DEL ALCANCE DE LA PRÁCTICA

**OUT OF SCOPE**

Para continuar con el proyecto tras la práctica necesitaríamos:

- Aumentar las fuentes de datos de prensa online, incluyendo medios de pago (€)
- Obtener más y mejor acceso a APIs de redes sociales (€)
- Escalar la arquitectura para gestionar un mayor volumen de datos
- Entrenar con más noticias y datos hasta desarrollar modelos con diferentes “ideologías” basadas en el sesgo de las distintas fuentes de información.

## A MEDIO PLAZO

- Multimedia: Añadir fuentes de TV/Radio y podcasts a resúmenes de texto
- Presentar resúmenes en diferentes formatos: Texto, Audio, Dashboard, Alertas de texto corto
- Ampliar interfaces: Apps, chatbot, Skill de Alexa (Ej: que lea el resumen al desayunar o en el coche)
- Recoger feedback del usuario para desarrollar contenidos basados en su perfil
- Investigar implicaciones legales y licencias de reproducción con CEDRO y AEDE
- Monetizar en B2C y/o B2B

# PIPELINE DEL PROYECTO

---



## DEFINIR EL DATASET

El dataset lo formamos nosotros scrapeando a diario diferentes periódicos en varios idiomas.

En posteriores pasos será cruzado con información obtenida de Twitter.



# ARQUITECTURA

## TERTULIA MATUTINA

SSH →

Google Cloud Platform

- VM INSTANCE
- UBUNTU
- PYTHON

SCRIPTS  
DE FUNCIONAMIENTO

STORAGE

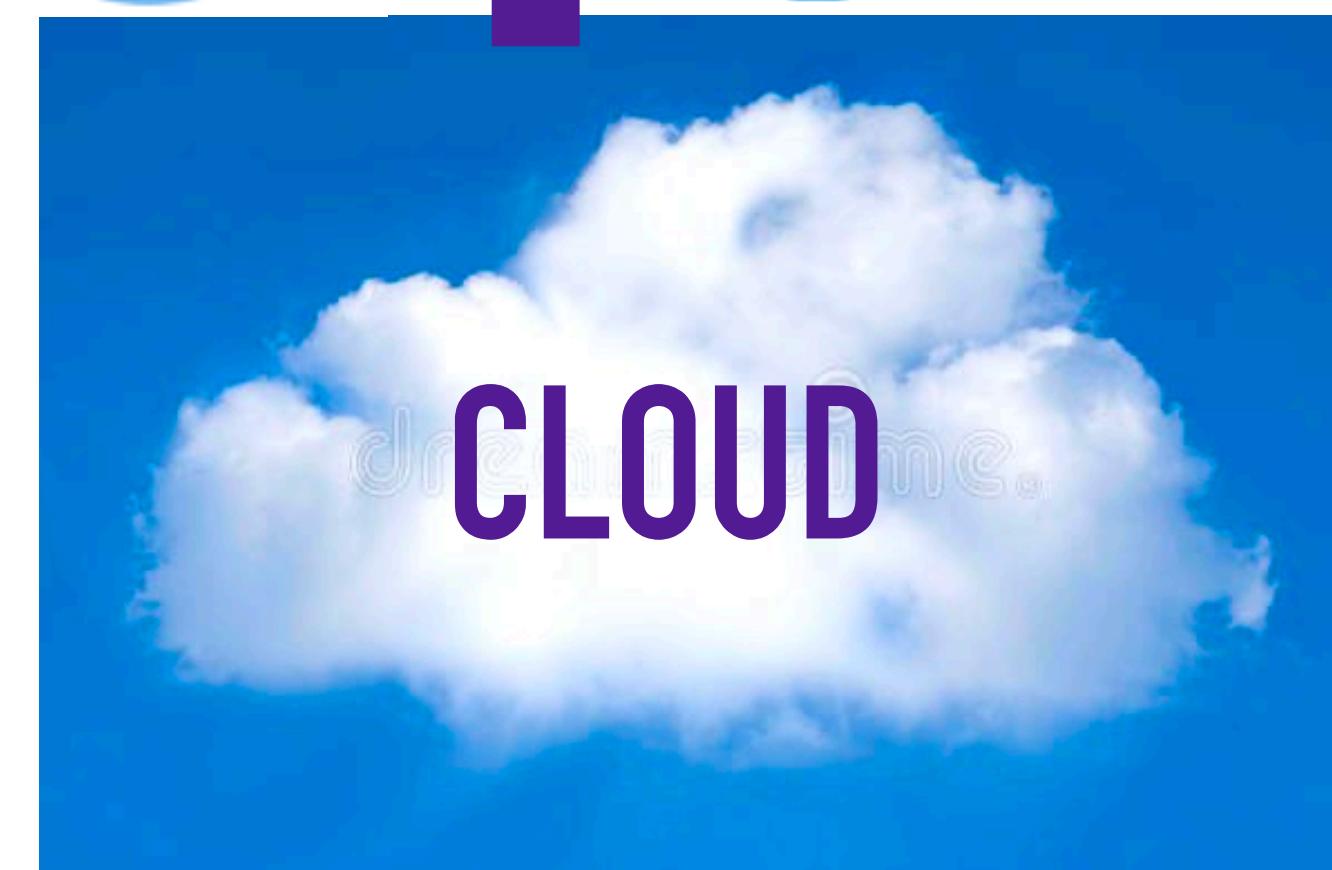
- SEGMENT

DATOS RESULTANTES

END USER



+a|bleau



FERNANDO



# SCRAPING

Elements    Console    Sources    Network    Performance

```
<section></section>
▼<section class="board board-a">
  ▼<div class="board-a4 " id="m70-69-71">
    ▶<div>...</div>
    ▼<div class="blq-3">
      ▼<ul>
        ▼<li>
          ▼<article class="media " id="m131-130-132">
            ▼<div>
              ▶<figure class="figure">...</figure>
              ▼<div class="media-content">
                ▼<header>
                  ▼<h1>
                    ▶<a id="m145-144-146" href="https://www.20minutos.es/noticia/4375110/0/madrid-duplica-sus-casos-...apenas-un-dia-1-728-nuevos-contagios-de-los-4-410-reportados-por-sanidad/">...</a> == $0
                  </h1>
                ...
              ...
            ...
          ...
        ...
      ...
    ...
  ...
</div>
```

## SCRAPING URL SCRAPING NOTICIAS

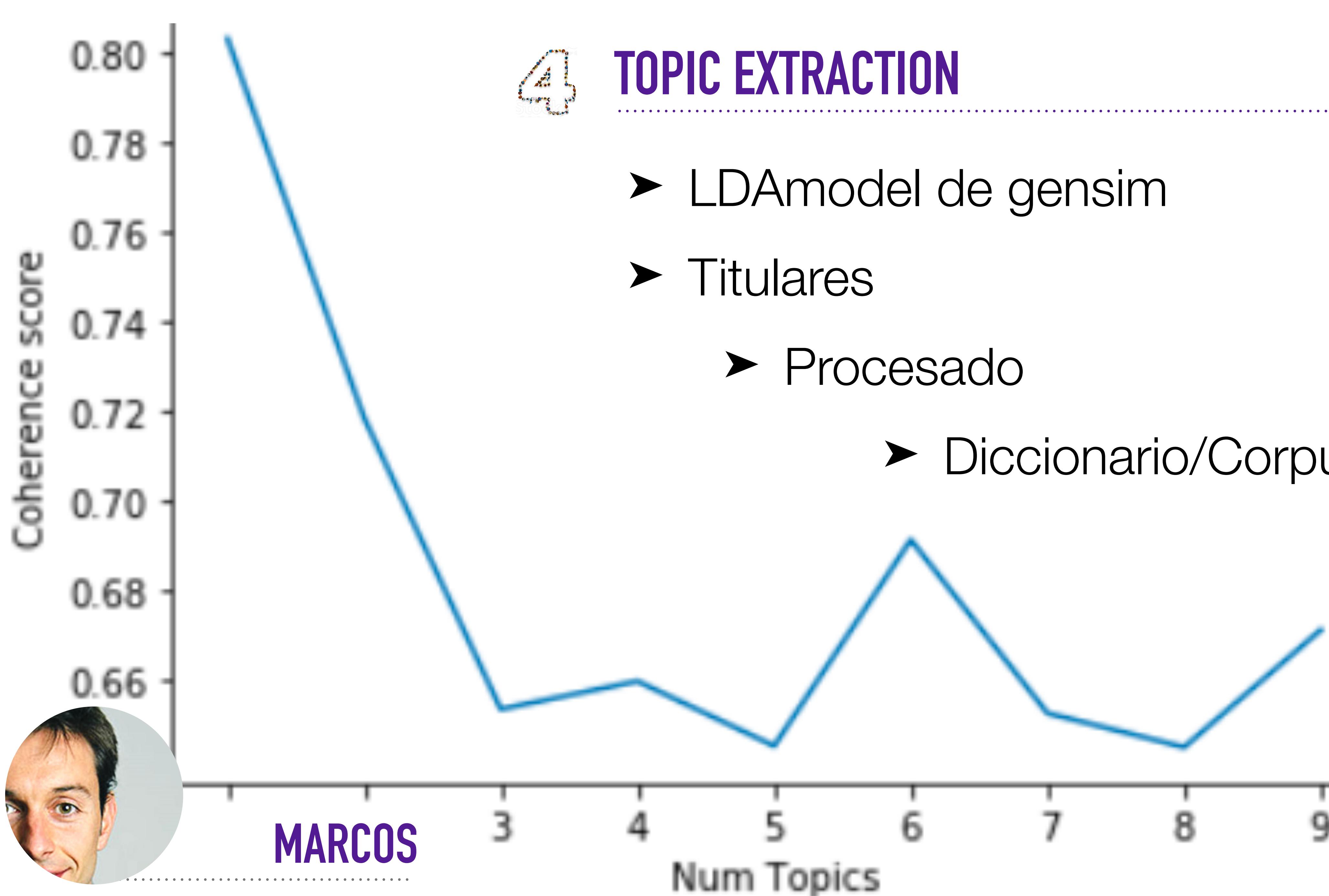
Titular / Texto completo



ALEX



MARCOS



4  
Stock

## TOPIC EXTRACTION

- LDAmodel de gensim
- Titulares
- Procesado
- Diccionario/Corpus



# TOPIC EXTRACTION

ANTES

	Topic #01	Topic #02	Topic #03
0	vario temer	millón euro	poder facilitar
1	facilitar fecho nacimiento	redar social	marcar vario
2	poder facilitar fecho	llevar cabo	fecho nacimiento
3	marcar vario temer	mejorar ofertar	poder ofrecerte contenido
4	breve preguntar	favor marcar vario	favor marcar
5	facilitar fecho	querer conocerte poder	poder ofrecerte
6	conocerte poder	poder facilitar fecho	vario temer
7	conocerte poder ofrecerte	uno breve preguntar	respondernos uno breve
8	uno breve	poder ofrecerte	respondernos uno
9	ofrecerte contenido	fecho nacimiento	conocerte poder
10	favor marcar vario	marcar vario	marcar vario temer



MARCOS

DESPUÉS

	date	Topic #01
0	2020-09-07	sanidad notifica
1	2020-09-07	gobierno lleva
2	2020-09-07	sale coma
3	2020-09-07	djokovic descalificado
4	2020-09-07	navalni sale
5	2020-09-07	hacer frente okupacion
6	2020-09-07	dictara instrucion hacer
7	2020-09-07	dictara instrucion
8	2020-09-07	espana podria
9	2020-09-07	tres millones dosis
10	2020-09-07	navalni sale coma



# STREAMING



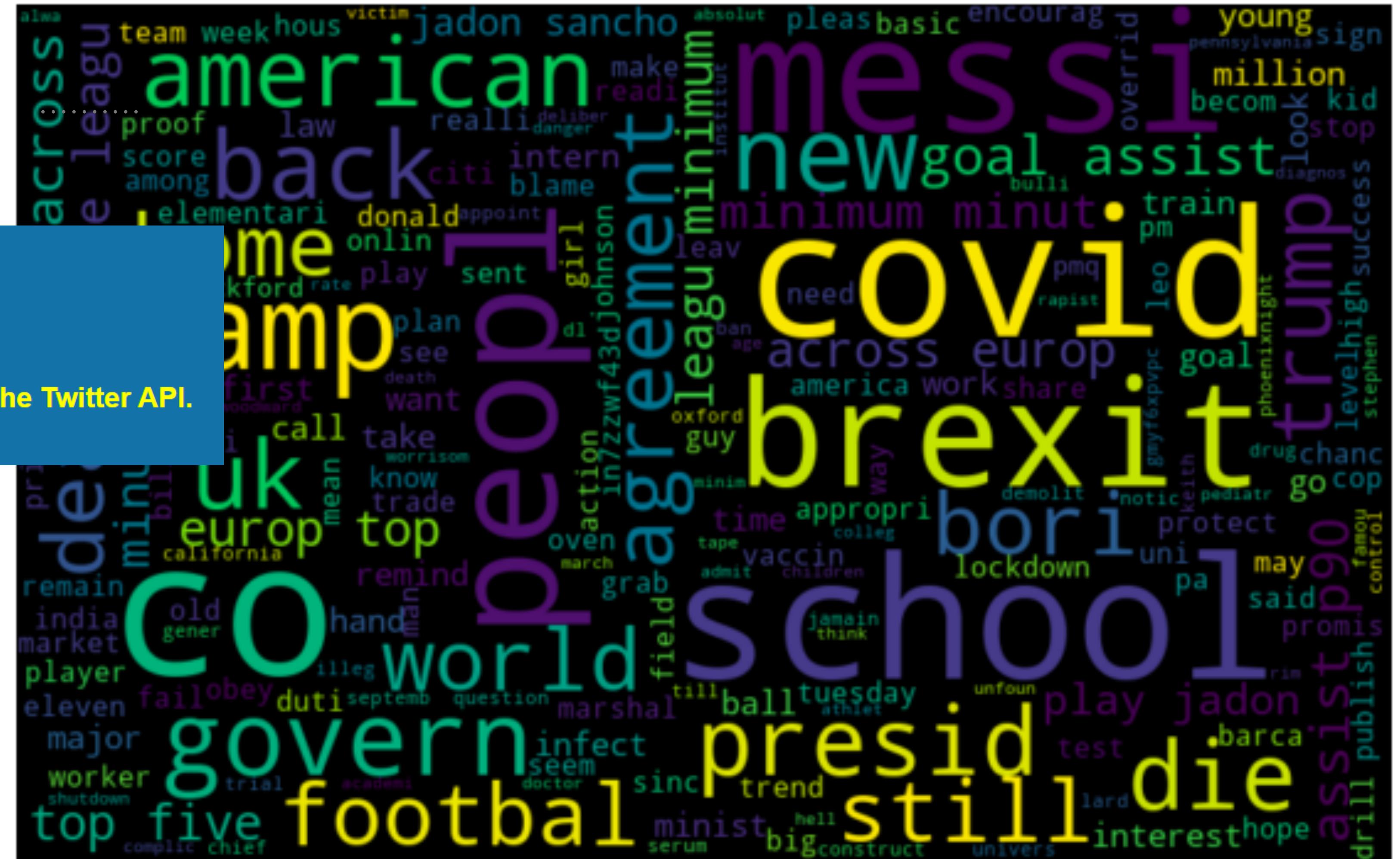
# An easy-to-use Python library for accessing the Twitter API

 Fork 3,207  Star 7,080



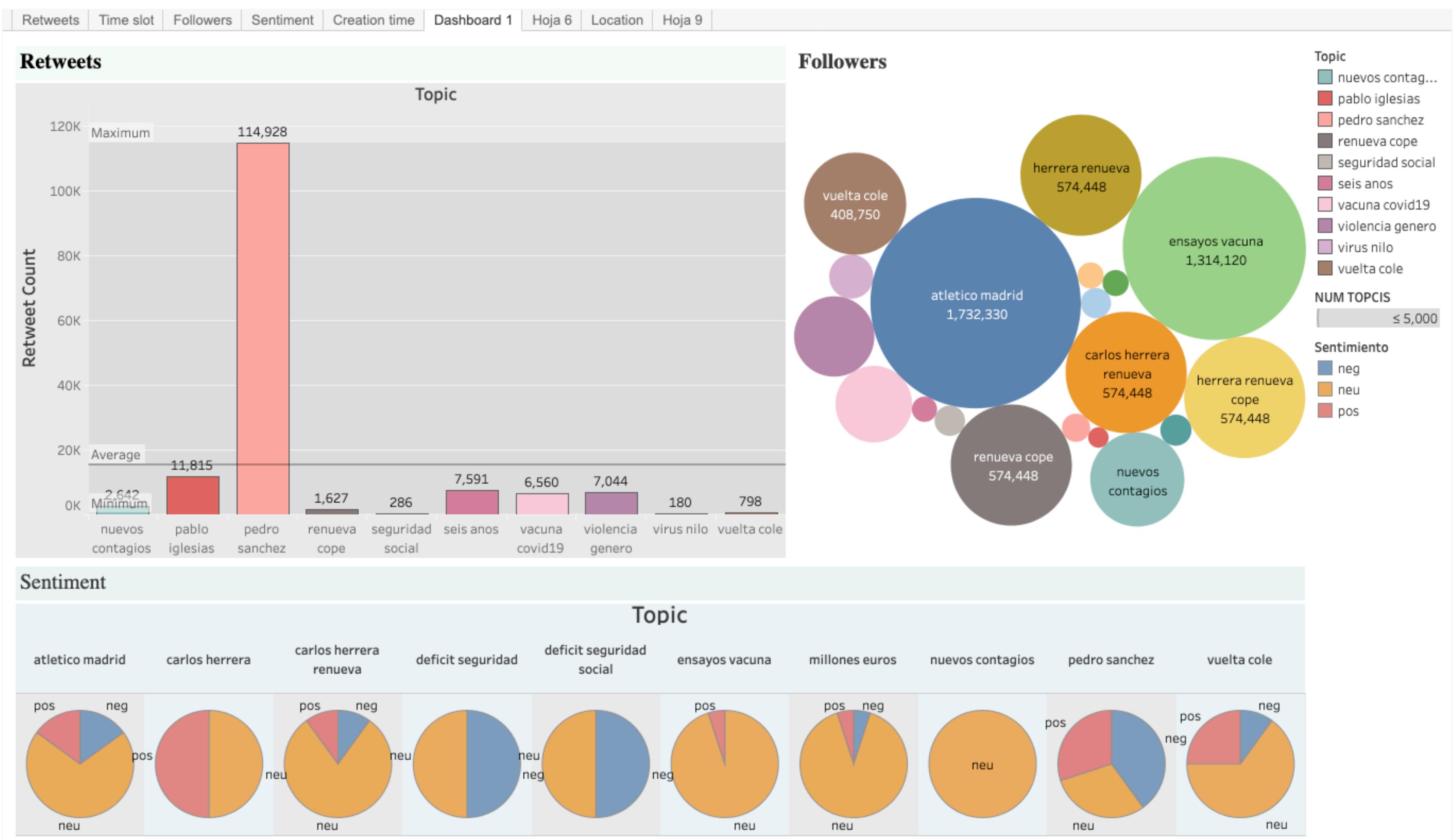
# ROBERTO

# PROCESADO DE TEXTO

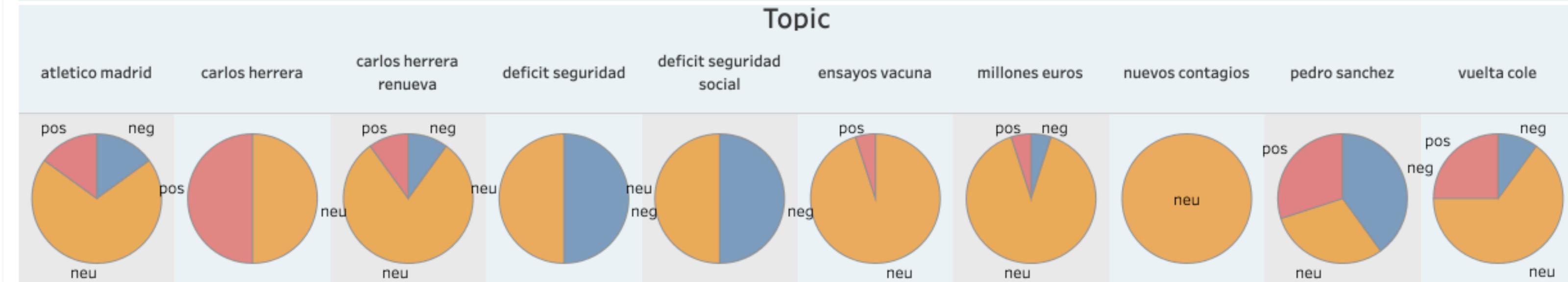




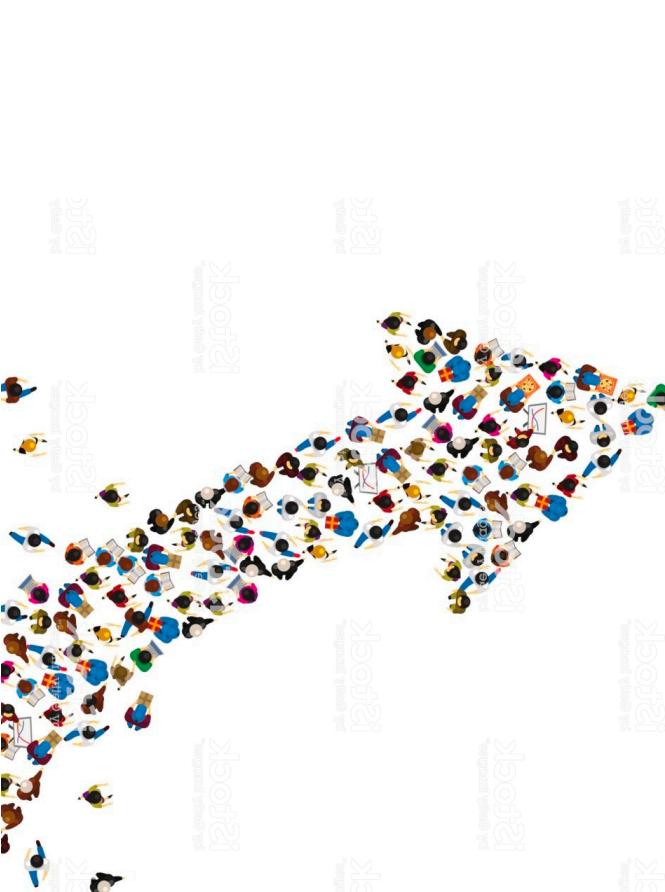
# VISUALIZACIÓN



### Sentiment



DAVID

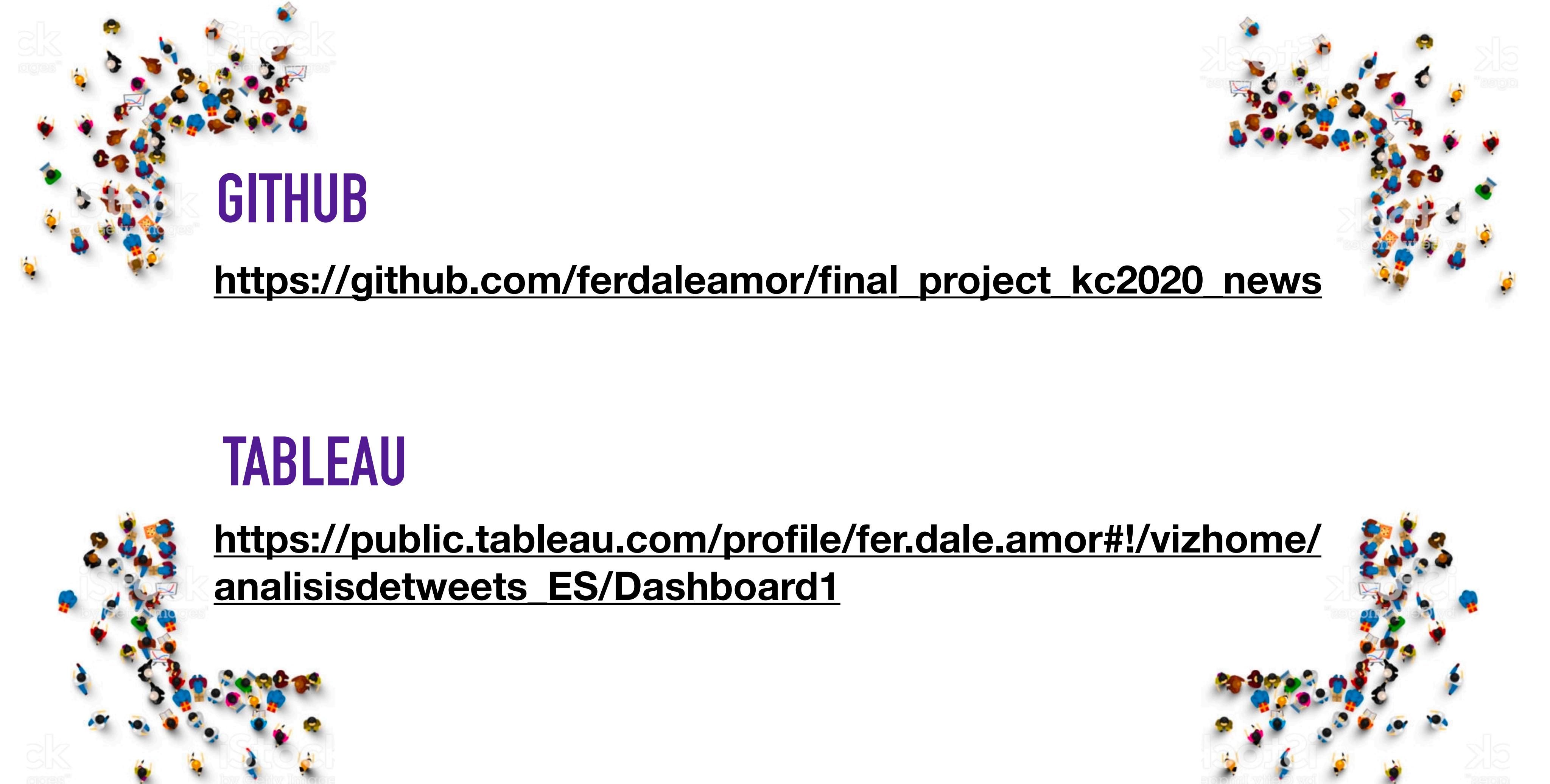


# AUTOEVALUACIÓN

1. Definir Data Set
  - Creamos nuestro propio dataset escarpando noticias en tres idiomas
  - Lo cruzamos con la API de Twitter
2. Arquitectura y validación de los datos
  - Arquitectura en Google Cloud
3. Análisis Exploratorio (iterativo)
  - Visualización en Tableau
  - Proceso constante de mejora, noticias nuevas día a día
4. Preprocesamiento
  - Procesamos el escarpado de las noticias para que nos sirva para el modelado de topics
  - De momento nos centramos en los titulares
5. Modelado
  - Usamos NLP con extracción de topics y análisis de sentimiento
6. Informe (Data Scientist/Desarrollador Big Data)
  - Visualización en Tableau

# ENTREGABLES

.....



GITHUB

<https://github.com/ferdaleamor/final project kc2020 news>

TABLEAU

[https://public.tableau.com/profile/fer.dale.amor#/vizhome/analisisdetweets\\_ES/Dashboard1](https://public.tableau.com/profile/fer.dale.amor#/vizhome/analisisdetweets_ES/Dashboard1)

*Gracias*

