



UNIVERSIDAD  
DE GRANADA

MÁSTER EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

# ANÁLISIS DE TENDENCIAS EN BIG DATA

AUTOR: JOSÉ ÁNGEL DÍAZ GARCÍA  
DEFENSA: 19/09/2019

# Contenidos

1. Motivación
2. Marco Teórico
3. Estado del arte
4. Objetivos
5. Planificación del proyecto
6. Metodología
  - A. Pre-procesado de datos
  - B. Minería de datos
7. Conclusiones y líneas futuras

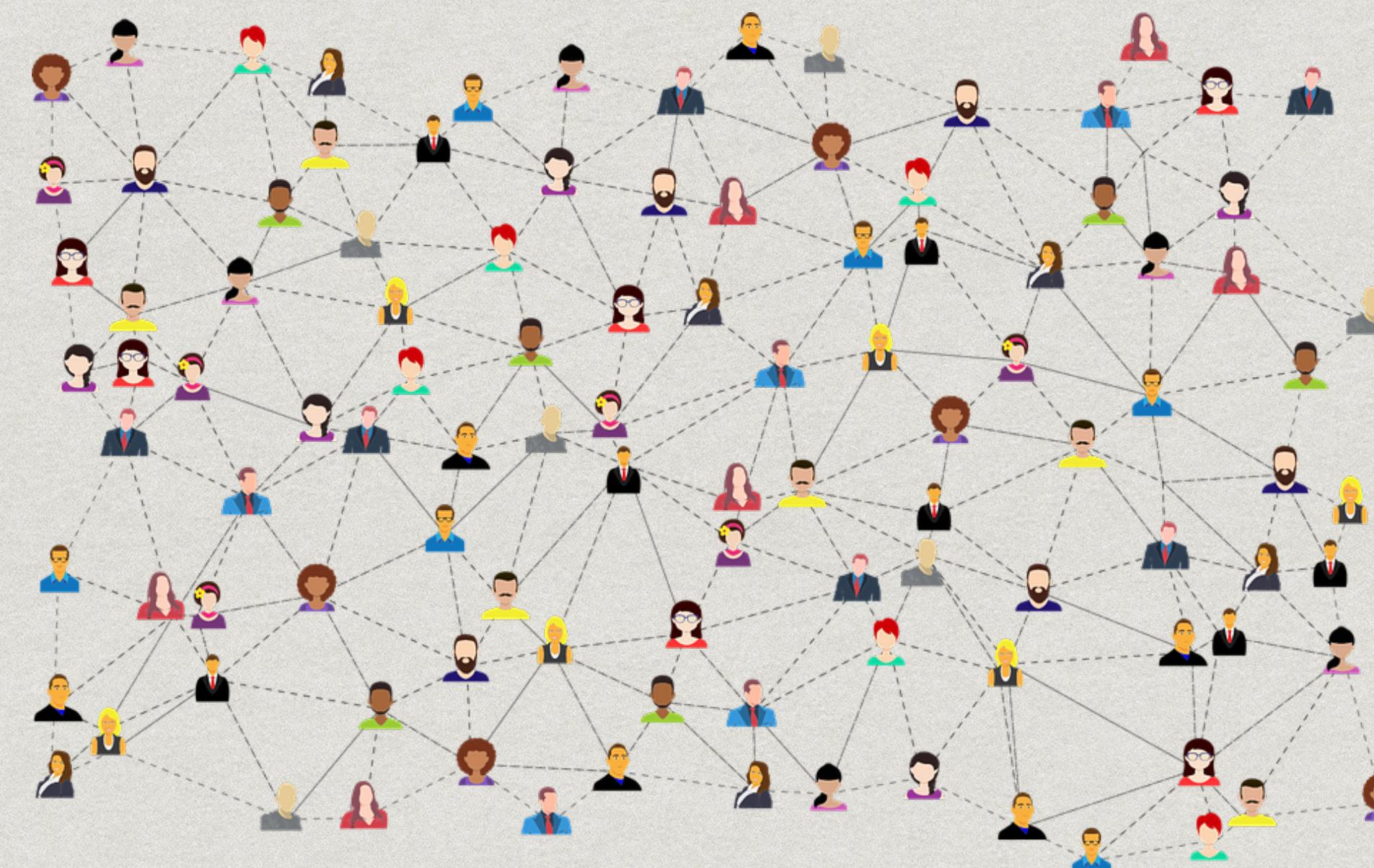
# Motivación

- \* Mundo actual muy influenciado por las redes sociales.
- \* Minería de datos y concretamente la minería de medios sociales muy influyente en los paradigmas actuales.
- \* Diversos casos de uso relevantes de la minería de medios sociales como el caso de la influencia de Rusia en las elecciones Americanas.
- \* Las redes sociales generan tal volumen de datos que cada vez es más necesario el paradigma del **Big Data** para obtener valor de las mismas.



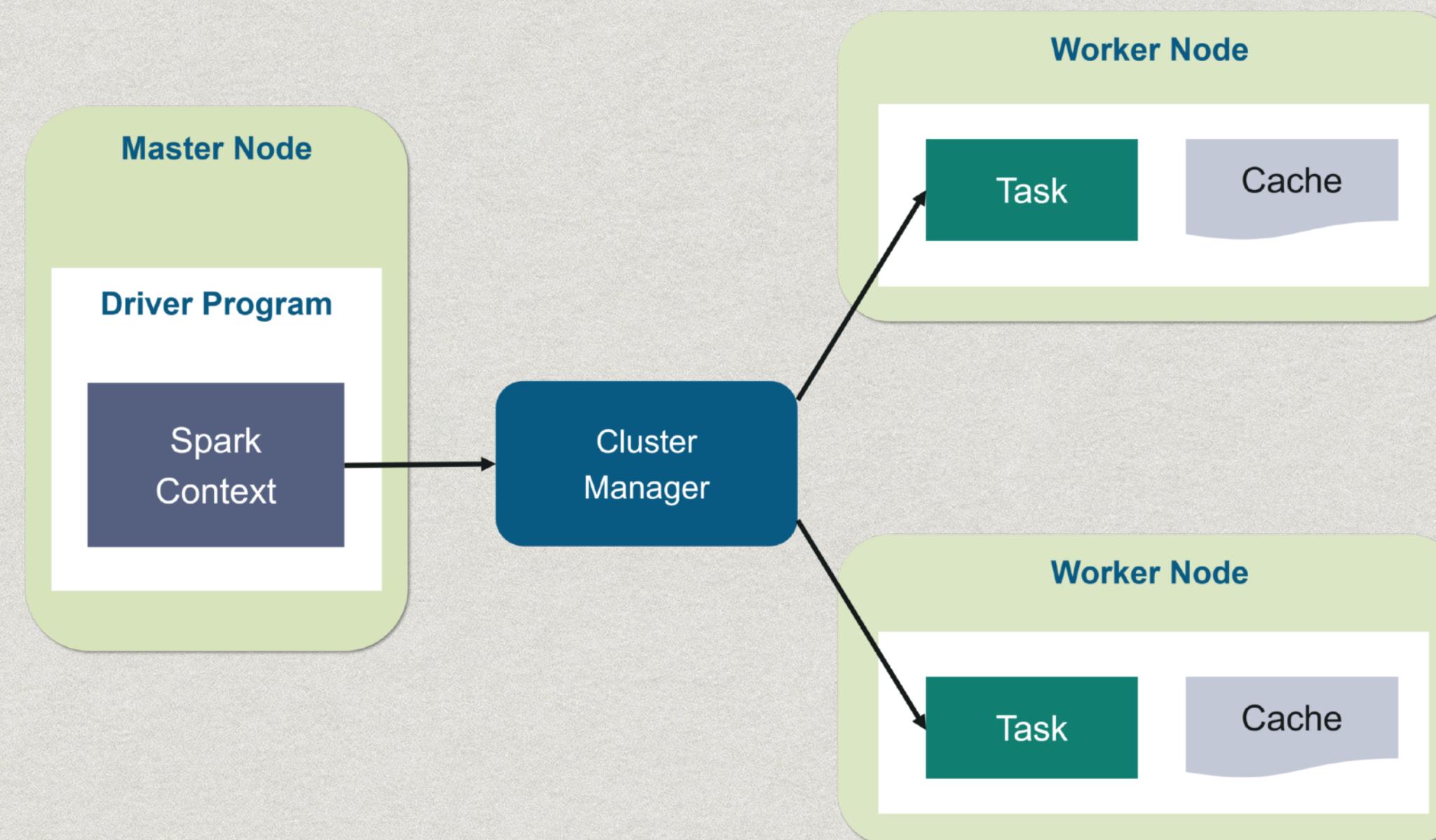
# Marco de trabajo

- \* Minería de medios sociales: Estudios de influencia, difusión de información, análisis de comunidades, minería de opiniones muy ligada al análisis de sentimientos y al ***text mining***.
- \* Dentro de la minería de opiniones y extracción de información de redes sociales hay muchos trabajos que usan aprendizaje supervisado, como clasificación, nosotros nos centraremos en el uso de **reglas de asociación** dentro de este ámbito.



# Marco de trabajo

- \* Una regla de asociación, en una base de datos transaccional, se representa como X->Y donde Y es un consecuente e X un antecedente definiendo una relación de co-ocurrencia entre ambos ítems. Para medir el ajuste, tenemos medidas como la **confianza**, el **soporte**, el **lift**, el **factor certeza**...
- \* Hay diversos algoritmos entre los que destacan **Apriori** o **FP-Growth** (usa una estructura interna **FP-TREE**, muy escalable aplicable a Big Data).
- \* En Big Data, tenemos distintos paradigmas, el más extendido actualmente es **Spark**.



# Estado del arte

- \* Hay bastantes estudios que usan las reglas de asociación sobre texto pero solo algunos sobre tuits.
- \* Los artículos que aplican reglas de asociación sobre tuits lo hacen sobre conjuntos de datos muy pequeños.
- \* Nuestra aportación al estado del arte del análisis de tendencias con reglas de asociación es por tanto un modelo capaz de trabajar con una cantidad de datos, que hasta donde nosotros sabemos, no ha sido aplicada anteriormente en trabajos de este tipo.

# Estado del arte

N tweets	Purpose	AS	PM
8275	Detección de patrones de cyberbullying.	No	Si
14000	Detección de patrones de cyberbullying.	No	Si
35000	Co-ocurrencia de hashtags para sistema experto contra el tabaco	No	Si
68370	Patrones en el ámbito de los seguros.	No	Si
24026	Identificación de usuarios activos durante ataques terroristas.	No	Si
500	Resumen de tweets sobre Obama.	No	Si
20000	Recomendación de películas.	No	Si
224291-3837291	Detección de eventos.	No	Si
57000	Análisis de las elecciones de Australia.	Si	No
80563	Obtención de patrones para promover el ciclismo.	Si	Si
8.772	Patrones sobre las mejores universidades de India.	Si	Si
150000	Predicción de stock de productos.	Si	Si
3000	Resumir conversaciones de Twitter.	Si	Si
450000	Detección de tópicos.	No	Si
450000	Estudios de propagación de información.	No	Si
<b>1517477</b>	<b>Obtener patrones y sentimientos sobre las elecciones del 28A.</b>	<b>Si</b>	<b>Si</b>

# Objetivos

1. Estudio del estado del arte en el campo de la minería de opinión basada en Big Data en plataformas de microblogging.
2. Desarrollo y aplicación de una metodología de preprocesado de datos eficaz para conjuntos de textos provenientes de plataformas de microblogging.
3. Aplicación de técnicas de minería de datos descriptiva para obtener patrones interesantes en los datos.
4. Obtención y salvado de un corpus de datos de gran tamaño que permitan elaborar la experimentación y validación del sistema final.
5. Experimentación, análisis de resultados y comparación de los mismos con posibles eventos políticos y sociales.
6. Puesta en valor del sistema mediante técnicas de visualización dinámicas.

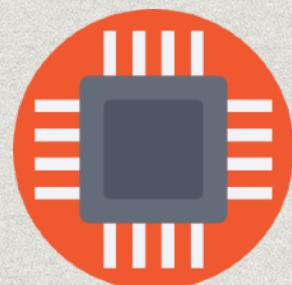
# Planificación del proyecto



**Timing:** Se dividió el proyecto en obtención de información, estudio estado del arte, obtención del data set, preprocesado, minería de datos y exploración de resultados, **outline** clásico de ciencia de datos.



**Personal:** El autor del proyecto y las directoras del mismo.

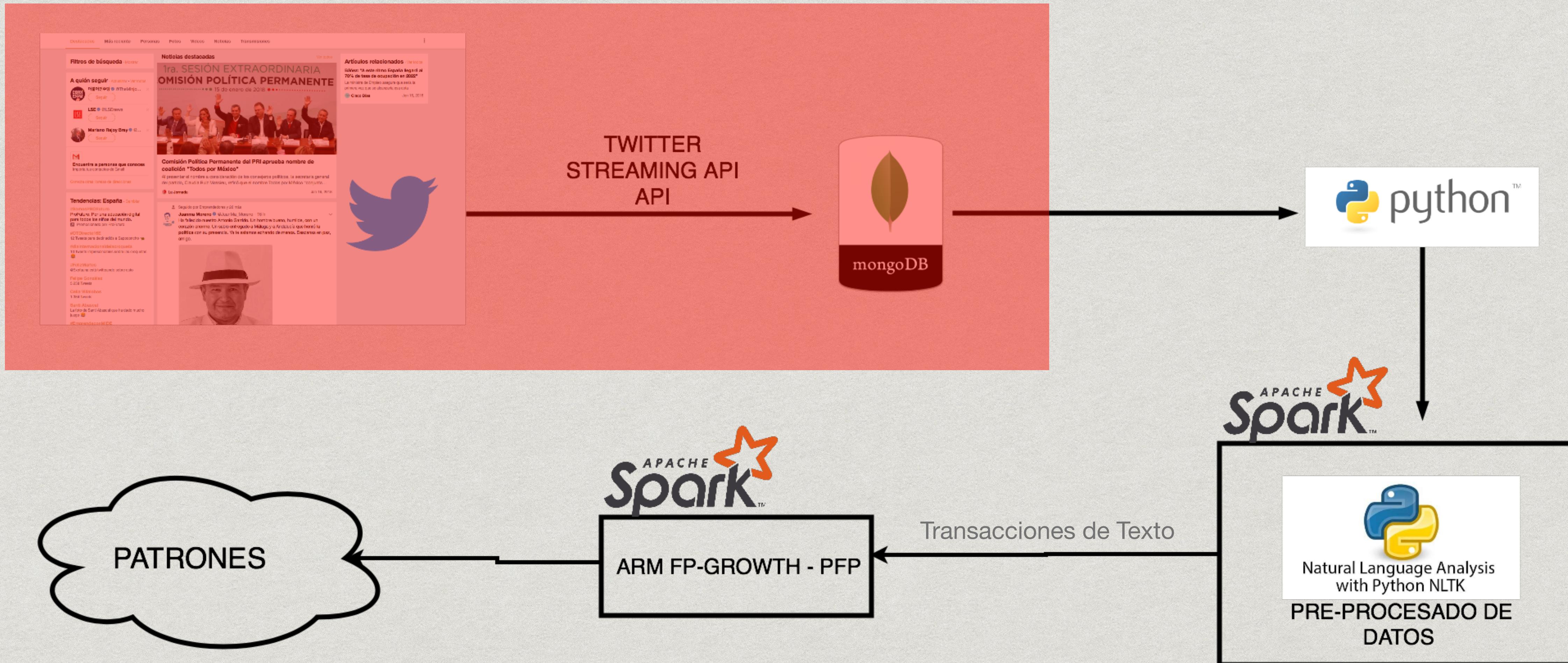


**Hardware:** Mac Pro 2,6 GHz Intel Core i5 8 GB 1600 MHz DDR3. Cluster con 4 nodos de Intel Xeon E5-2665 de 8 núcleos.



**Software:** TexShop, TextEdit, Rstudio, Spark, PySpark, PyCharm, nltk, Numbers, MongoDB.

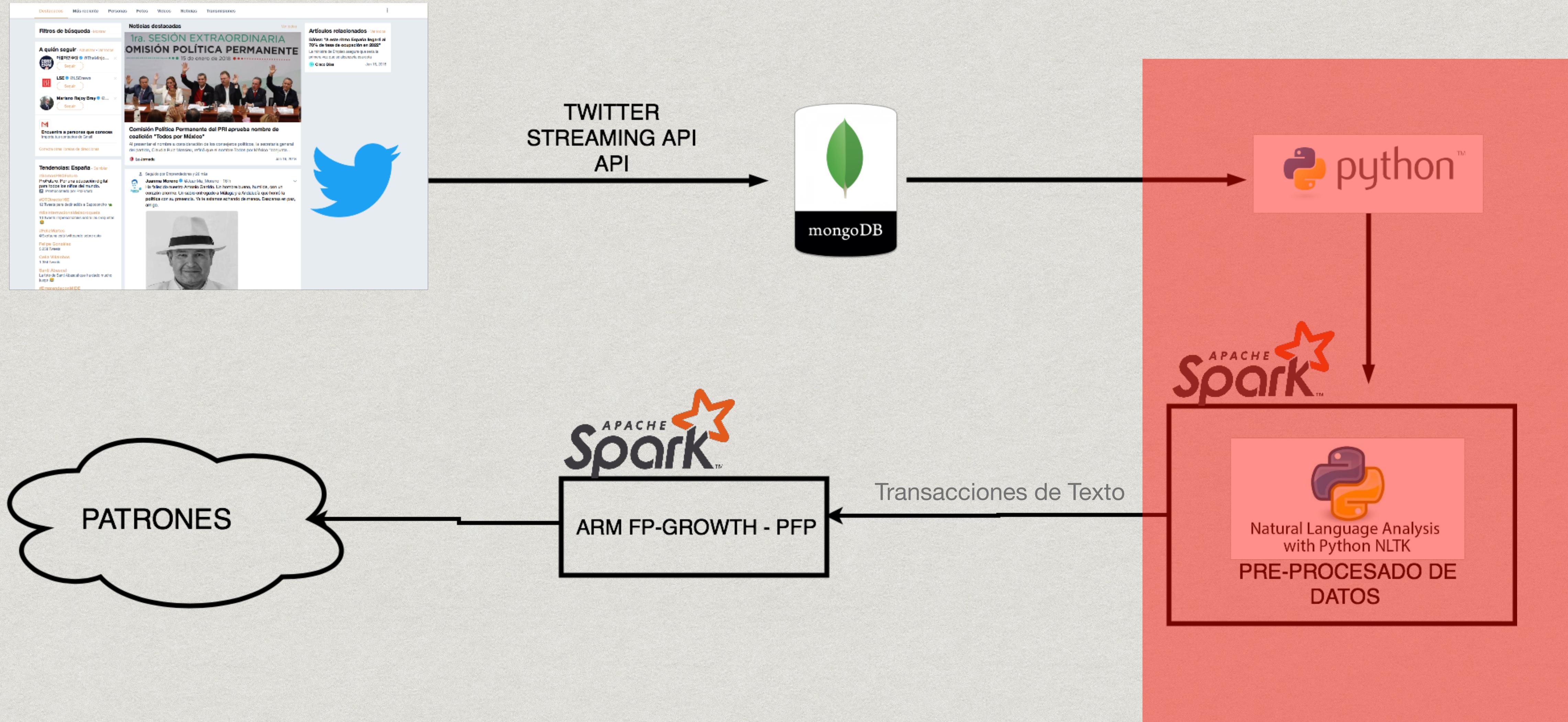
# Metodología



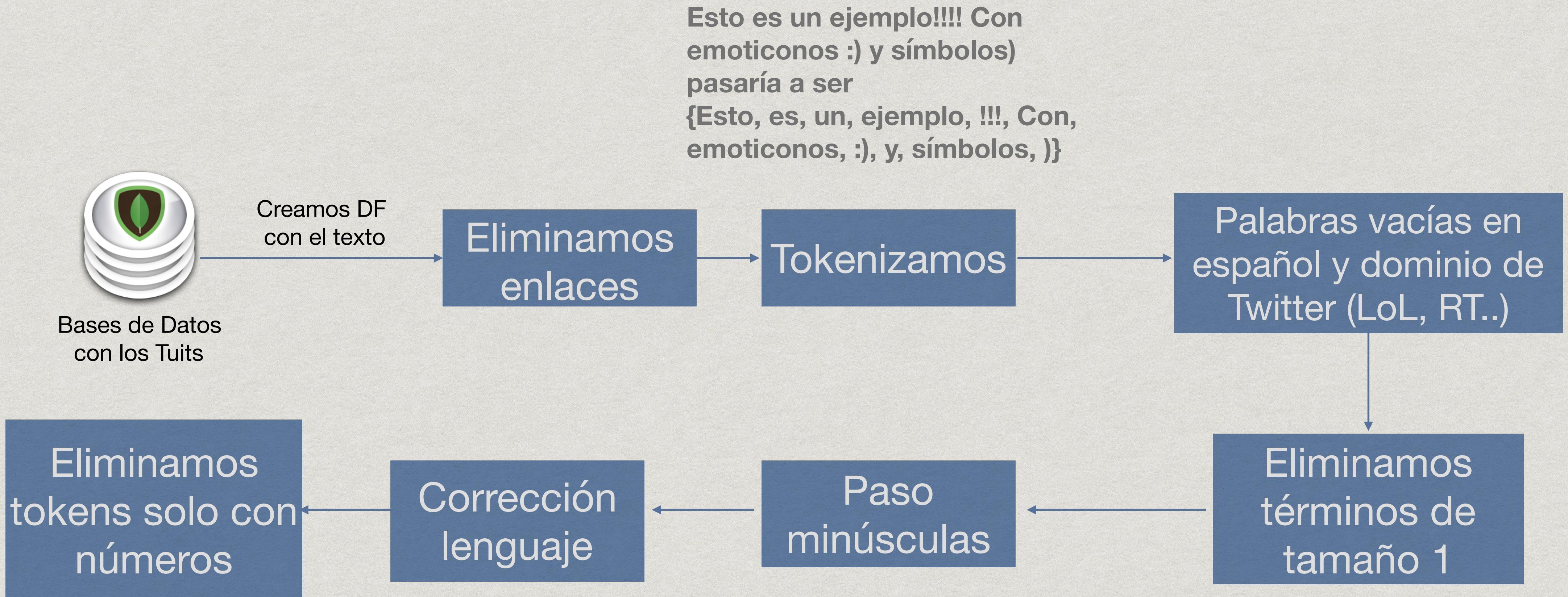
# Dataset

- \* El dataset proviene de la red social Twitter.
- \* Está formado por **1517477** tuits de como máximo 280 caracteres.
- \* Para la obtención, se creo una aplicación de Twitter que estuvo guardando en tuits en una base de datos noSQL de tipo MongoDB durante el mes de Abril de 2019.
- \* Se hizo uso de la API de **streaming** de Twitter.
- \* El filtrado viene dado por hashtags **#28A, #28Abril, #Elecciones2019**

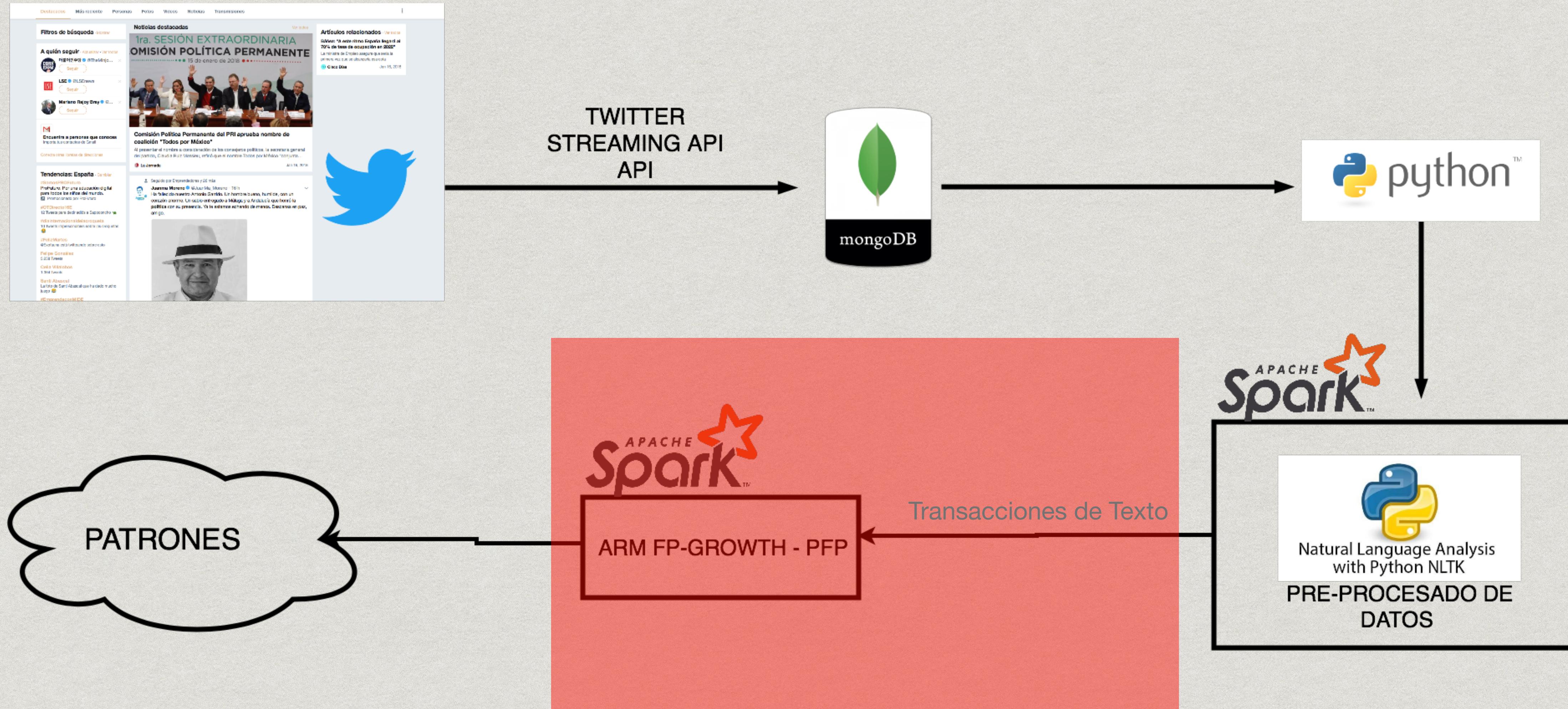
# Metodología



# Preprocesado



# Metodología



# Minería de datos: Transacciones textuales

- \* Para poder aplicar algoritmos de reglas de asociación sobre textos necesitamos transacciones textuales. La definición formal sería:
  - Cada tweet es una transacción.
  - Cada término es un ítem.
  - Su representación, para un ejemplo de 2 tuits (**transacciones**) y vocabulario de 6 palabras (**ítems**).

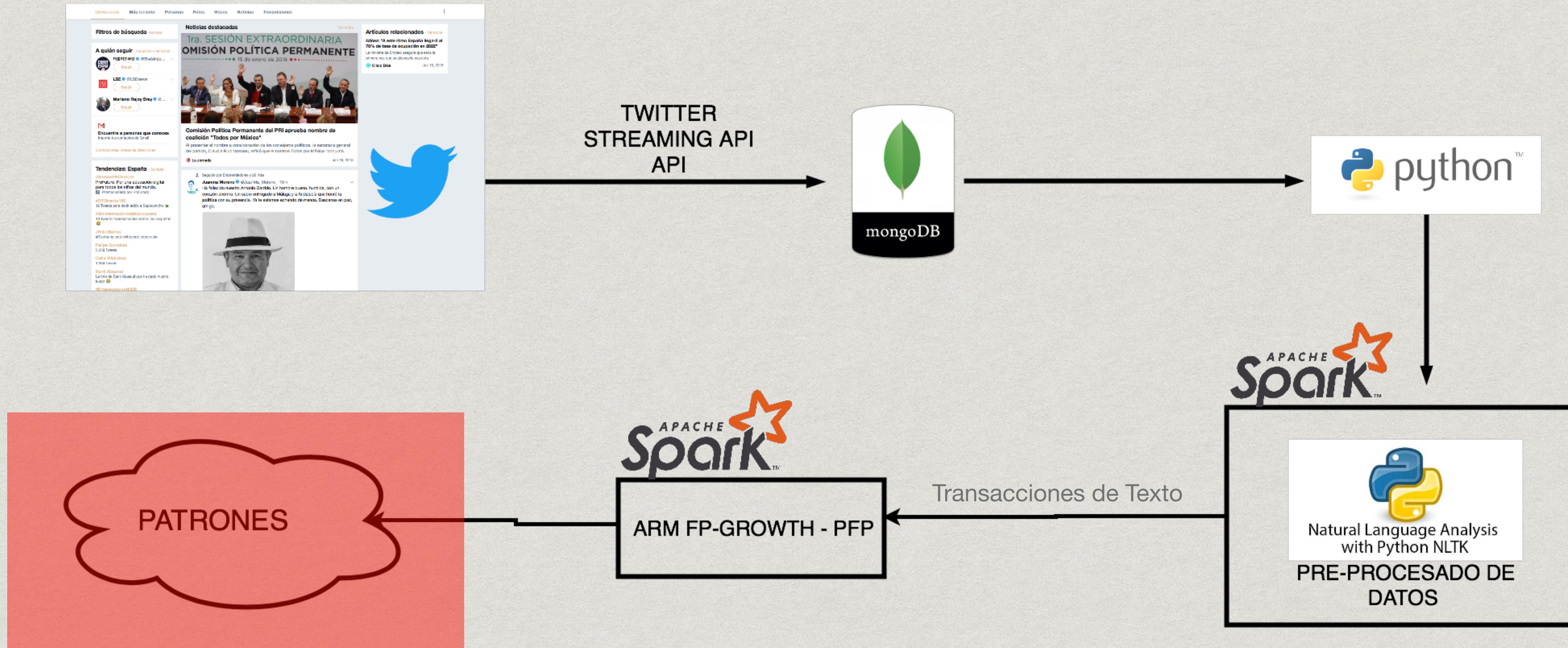
esto	es	un	ejemplo	muy	largo
1	1	1	1	0	0
1	1	1	1	1	1

# Minería de datos: Reglas de asociación



- \* Para las reglas de asociación se ha usado el algoritmo **FP-GROWTH** con mínimo valor de 0.8 para la confianza y 0.001 para el soporte.
- \* Con los patrones se han intentado dar respuesta a preguntas del ámbito de la política dado que los datos provienen del **28A**. ¿Qué preocupaba a la sociedad? ¿Qué términos se asocian con qué partidos?
- \* Se obtienen más de 2 millones de reglas.

# Metodología



# Patrones interesantes

## ¿Temas del discurso político?

Antecedente	Consecuente	Confianza
complicidad, socios, pnv, proetarras	sánchez	1.0
pnv, socios, sánchez	hostigando	1.0
junto, independencia	cataluña	0,993
miedo, nadie, 28a	vox_es	1.0

## ¿Términos de apoyo y en contra?

Antecedente	Consecuente	Confianza
mejor, candidato	pablo_casado	0.999
porespaña	vox	1.0
presagio, fantastico	santi_abascal	0,993
amenazado, agredido, simpatizantes	vox_es	1.0
blas, lezo, madrid, victoria	colon	0.999

# Patrones interesantes, visualización

buenas      socios      pnv  
hoy      el      no      san      separatismo  
hostigando  
**sánchez**  
complicidad      gente  
sebastián  
proetarras      los  
pedro

# Conclusiones y líneas futuras

- \* Se ha demostrado que las reglas de asociación pueden usarse en Big Data para obtener información de bases de datos de textos.
- \* Se ha constatado, la necesidad de paradigmas distribuidos cuando el volumen de datos es muy grande.
- \* Los datos provenientes de redes sociales pueden aportar mucho valor pero tienen mucho ruido.
- \* Sería interesante dar una aproximación del sistema en streaming.
- \* También sería interesante realizar una aproximación difusa a las transacciones textuales.
- \* En cuanto a la visualización queda mucho que hacer en reglas de asociación pues cuando tenemos muchas reglas los métodos habituales dejan de ser útiles.



UNIVERSIDAD  
DE GRANADA

**GRACIAS POR SU ATENCIÓN**

**ANÁLISIS DE TENDENCIAS EN BIG DATA**  
**JOSÉ ÁNGEL DÍAZ GARCÍA**