

Politweets

VIRTUAL COFFEE



AI TECHNICAL LEAD

Apasionado de los datos, la Inteligencia Artificial y las redes neuronales. No habría sobrevivido al Titanic. Coffee please!

@ericmcmc1



AI SOFTWARE DEV

Orfebre del software. Yo tampoco hubiera sobrevivido. Además, no puedo comprarme una casa en California.

@ikeinyo

Politweets

1. The Idea

2. Politweets AI

Natural Language Processing
Information Retrieval

3. The solution

Politweets

The Idea

Politweets – The Idea

- De la desinformación a la sobreinformación
AI for good
- Proyecto real end2end
AI for show

Politweets – The Idea

- Saber que piensan los políticos de un tema a través de sus tweets.
- Poder comparar fácilmente distintas opiniones
- Perfecto para:
 - Aprender los fundamentos de NLP
 - Aprender algunas técnicas de IR
 - Plantear un Proyecto “vivo”

politweets

Query:

sanitario salud

TFIDF Semantic

Submit

politweets

Query:

sanitario salud



Pablo Iglesias 
@PabloIglesias

La temporalidad y la precariedad laboral han alcanzado niveles alarmantes en los últimos años. La situación de esta enfermera del Servicio Gallego de Salud es un ejemplo vergonzoso, más aún cuando se da en la sanidad pública 🙄

pic.twitter.com/jfBZwkroGv

1,587 5:20 PM - Nov 29, 2018

1,406 people are talking about this



Pedro Sánchez 
@sanchezcastejon

Hemos recuperado la #sanidad universal y atendido las necesidades más urgentes. Ahora retomaremos el impulso definitivo a la ley de #eutanasia e iniciativas contra el cáncer infantil-juvenil, atención bucodental y salud mental. Escuchamos las demandas del sector sanitario.

pic.twitter.com/FZ9qDtDu22

1,528 6:53 PM - Aug 28, 2019

1,416 people are talking about this



Albert Rivera 
@Albert_Rivera

Aprobaremos la tarjeta sanitaria única para toda España en los primeros 100 días de Gobierno. Hay que garantizar el derecho de los ciudadanos a ser atendidos sin trabas ni burocracia en cualquier centro de salud u hospital del país. Necesitamos más igualdad y menos fragmentación. pic.twitter.com/GxGdCRrQm5

3,858 10:46 AM - Mar 22, 2019

1,805 people are talking about this



Pablo Casado Blanco 
@pablocasado_

Proponemos reforzar a nivel estatal la competencia sanitaria para que no haya 17 sistemas de salud distintos. Tarjeta sanitaria única, cartera básica de servicios y compras centralizadas. No puede haber desigualdades, ni ineficiencias y hay que acabar con el turismo sanitario.

pic.twitter.com/mB0GWam6Z

308 3:08 PM - Nov 27, 2018

254 people are talking about this



Santiago Abascal 
@Santi_ABASCAL

El discurso del odio, de la estigmatización y de los cordones sanitarios contra VOX tienen consecuencias: odio y violencia. Todos, desde Iglesias con su "alerta antifascista" hasta Casado con la cantinela de la "ultraderecha", están contribuyendo al pésimo clima moral y político. twitter.com/COPE/status/11...

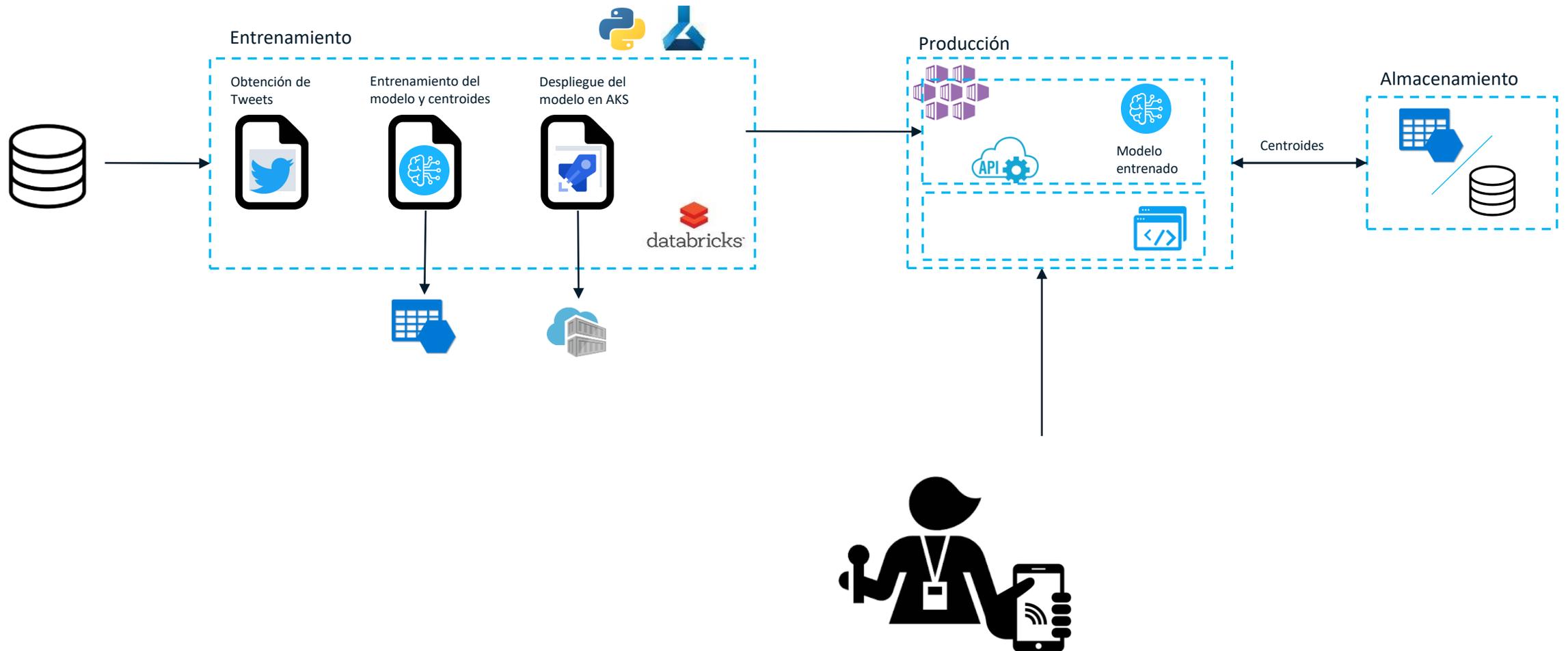
COPE  @COPE

Una alumna, a una profesora militante de Vox Málaga: "Facha, tú no entras". La docente, Inmaculada Enriquez, fue apoderada de Vox el 28-A y forma parte de la lista para las municipales en Rincón de la Victoria www.cope.es/snerz2

5,301 2:37 PM - May 2, 2019

3,216 people are talking about this

Politweets – The Idea



Politweets – The Idea



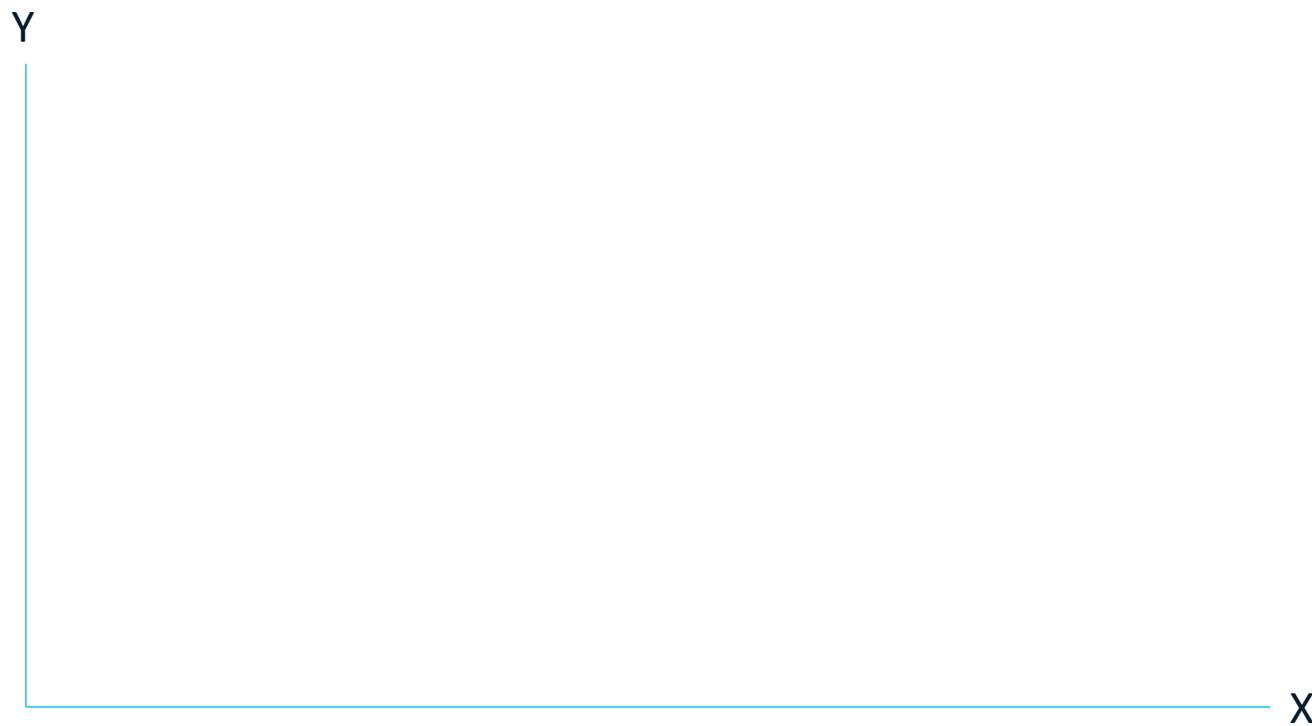
Politweets AI

Natural
Language
Processing

NLP

Word to vector

blue
house
good
poor



NLP

Word to vector

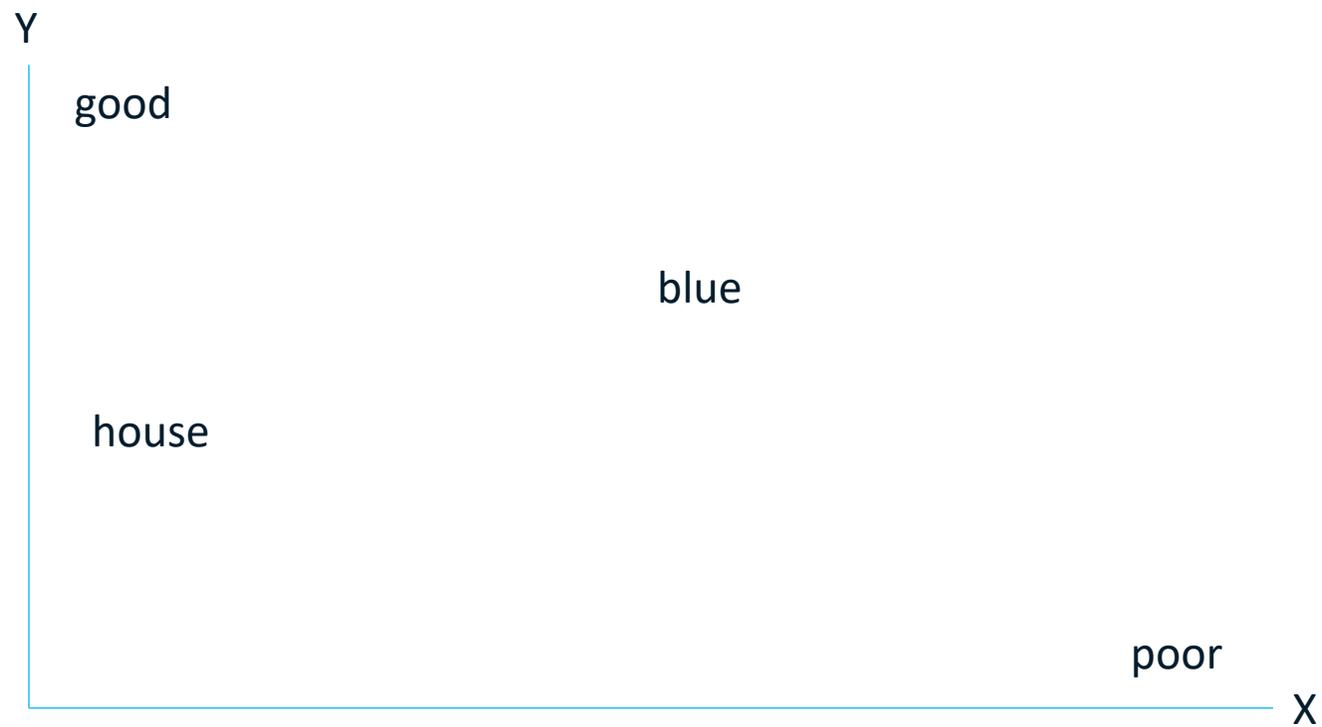
blue
house
good
poor



NLP

Word to vector

blue
house
good
poor



NLP

Word to vector

$\text{vector_of}(\text{"king"}) - \text{vector_of}(\text{"men"}) + \text{vector_of}(\text{"woman"})$

=

$\text{vector_of}(\text{"queen"})$

NLP

Redes Neuronales Simples
Formas de leer o aprender

- C-BOW
- Skip-Gram

fastText

PYTORCH

GENSIM
topic modelling for humans

TensorFlow

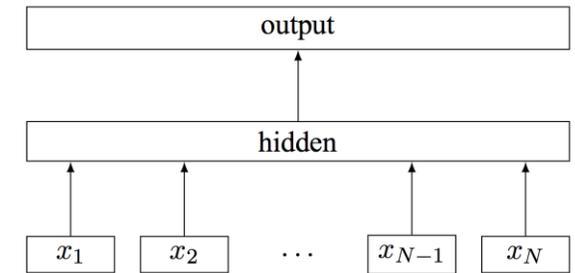
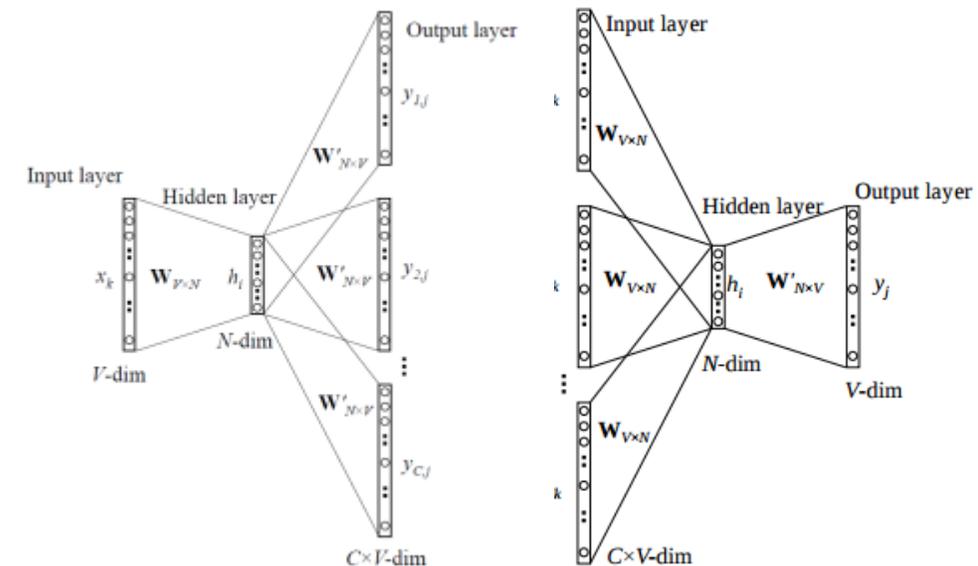


Figure 1: Model architecture of `fastText` for a sentence with N ngram features x_1, \dots, x_N . The features are embedded and averaged to form the hidden variable.



Word to vector

```
from gensim.models import Word2Vec  
model = Word2Vec(mytexts)
```



Palabras más próximas a Economía



en
o
con
para
se
y
que
a
como
una



en
de
y
para
se
la
o
como
a
con



en
que
de
y
la
para
los
con
las
se



en
para
de
o
se
con
a
que
el
como



en
se
a
y
de
el
que
para
la
con

Preprocessing

Facilitar la vida al modelo

Según la tarea aplicaremos un preprocesado u otro

Técnicas de preprocesado

- Tokenize
- Standarize
- Punkt
- Stop Words
- Lemmatization
- Stemming



spaCy

Preprocessing

Frase Original

'HEMOS RECIBIDO PETICIONES DE CLIENTES QUE NOS PIDEN PANTALONES ANCHOS CON ATADURAS EN TOBILLO.UN SALUDO .JAVIER .CABALLERO '

Estandarización y Eliminación signos de puntuación

'hemos recibido peticiones de clientes que nos piden pantalones anchos con ataduras en tobillo un saludo javier caballero'

Eliminación stop words

'recibido peticiones clientes piden pantalones anchos ataduras tobillo saludo javier caballero'

Stemming

'recib peticion client pid pantalon anchos atadur tobill salud javi caballer

Palabras más próximas a Economía



debemos
españa
compromiso
lucha
europa
país
gracias
política
mundo
hoy



sánchez
españa
españoles
gobierno
apoyo
pp
empleo
si
ser
debe



hacer
sánchez
españoles
apoyo
españaviva
porespaña
miedo
sábado
progre
fronteras



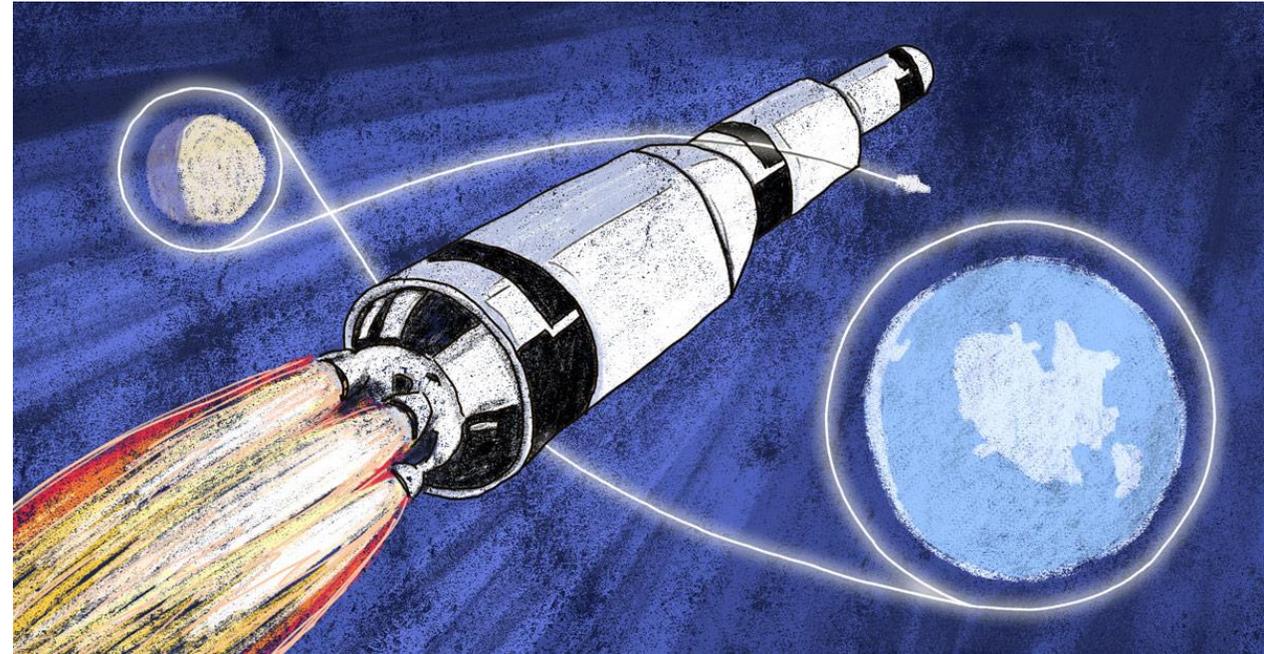
sánchez
si
españoles
ciudadanos
mundo
familias
gran
hacer
apoyo
ley



hacer
españa
si
país
familias
gente
políticas
derechos
política
ser

Hiperparámetros

- Vector size
- N-grams
- Window size
- Min count/ Max vocab/ Max count
- Parámetros de la red neuronal:
 - Epochs, Learning Rate, Decay...



Palabras más próximas a Economía



económico
enfriamiento
prestigiar
ecosistema
incompatible
modelo
vertebrador
impulsando
climateemergency
creciendo



desaceleración
industrial
laboral
crear
internacionales
gestionar
vuelve
reforma
déficit
banco



provocando
seguro
cristianas
gastos
polonia
difícil
quién
atentado
impuestos
debemos



podemizar
planea
sablazos
gasto
sablazo
déficit
conductores
podemización
cartera
consejería



pagan
impuestos
consumo
cuidados
económica
ricos
solidaria
justo
climática
permitirá

Modelos pre-entrenados

- Modelos pre-entrenados como base y reentrenar sobre mi vocabulario específico
- Permite al modelo entender perfectamente el idioma y especializarse en una tarea
- Tamaños estándar de embeddings: 100 y 300 sobre todo



Palabras más próximas a Economía



económica
económico
comercio
política
globalización
desempleo
económicos
sostenible
industria
governabilidad



económica
económico
agricultura
ganadería
competitividad
comercio
política
recesión
económicos
pib



política
industria
políticas
impuestos
democracia
gobierno
socialismo
presupuestos
inmigración
agricultores



económica
económico
comercio
política
productiva
cultura
turismo
políticas
educación
liberalismo



económica
económico
política
recesión
neoliberalismo
económicos
cultura
bienestar
políticas
educación

De vectores de palabras a aplicaciones del mundo real

- ¿Qué tenemos ahora?
- ¿Cuales son las aplicaciones de esto en el mundo real?



Politweets AI

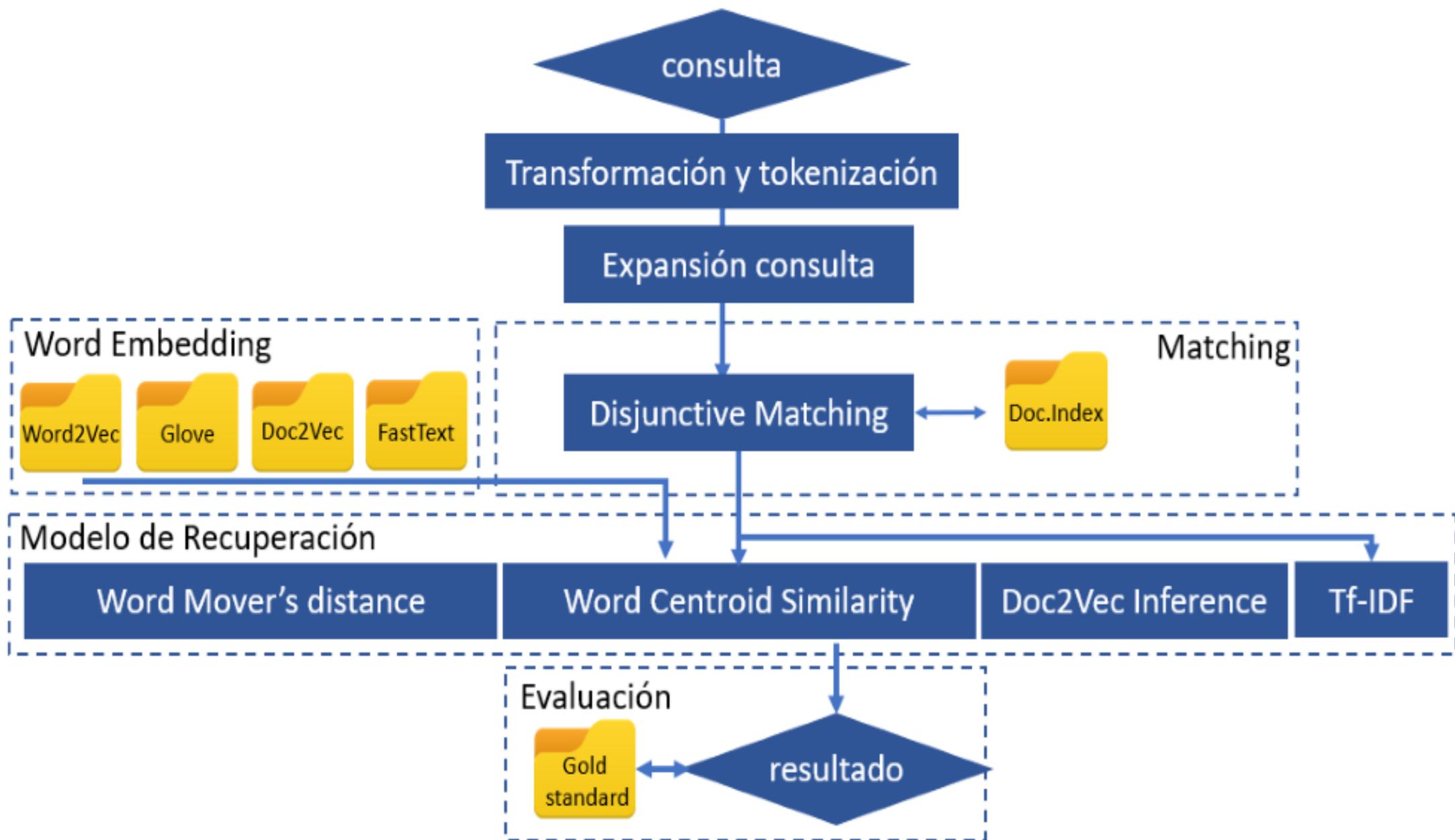
Information Retrieval

Information Retrieval

- Encontrar el documento más cercano a una query de entre una pila de documentos.



- Aprendizaje no supervisado
- Distintos algoritmos un objetivo: Calcular un vector que represente un documento
- ¿Como saber si está funcionando bien?



TFIDF

- Matriz de documentos x términos
- Basado en coincidencia exacta de términos
- Muy Bueno con vocabulario específico
- Sufre con consultas más complejas

TFIDF

For a term i in document j :

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

$tf_{i,j}$ = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents

	littl	hous	prairi	mari	lamb	silenc	twinkl	star
"Little House on the Prairie"	1	1	1	0	0	0	0	0
"Mary had a Little Lamb"	1	0	0	1	1	0	0	0
"The Silence of the Lambs"	0	0	0	0	1	1	0	0
"Twinkle Twinkle Little Star"	1	0	0	0	0	0	2	1

Document-Term Matrix

TFIDF

```
In [21]: sentence = "The oldest human fossil is the skull discovered in the Cave of Aroeira in Almonda."

TfidfVec = TfidfVectorizer()
tfidf = TfidfVec.fit_transform([sentence])

cols = TfidfVec.get_feature_names()
matrix = tfidf.todense()
pd.DataFrame(matrix, columns = cols, index=["Tf-Idf"])
```

Out[21]:

	almonda	aroeira	cave	discovered	fossil	human	in	is	of	oldest	skull	the
Tf-Idf	0.208514	0.208514	0.208514	0.208514	0.208514	0.208514	0.417029	0.208514	0.208514	0.208514	0.208514	0.625543

TFIDF

- Consulta: “Donald Trump elecciones EEUU”



Pablo Iglesias ✓
@Pablo_Iglesias_

Follow



🎬 Esta semana en [@Fort_Apache_](#) hemos hablado de las próximas elecciones legislativas en EEUU, en las que Donald Trump se juega mucho. En la mesa: [@ManoloMonereo](#), [@AlbiacLola](#), [@SanchezCedillo](#), [@Marga_Ferre](#), Augusto Zamora y [@Jorge_Tamames](#).



Albert Rivera ✓
@Albert_Rivera

Follow



No encontré un Mc Donald's y tuve que ir a comer unos espetos y unos boquerones a Pedregalejo 😜.
¡Viva Málaga !

TFIDF

- Consulta: “Ley de Reforestación”



Pedro Sánchez ✓

@sanchezcastejon

Follow

Se cumplen 12 años de la aprobación de la Ley de Igualdad por un gobierno socialista. Una ley que nos hizo mejores y que nos sigue impulsando cada día hacia la **#igualdad** real entre hombres y mujeres. Gracias a quienes la hicieron posible. 🤝 🌹



Pablo Iglesias ✓

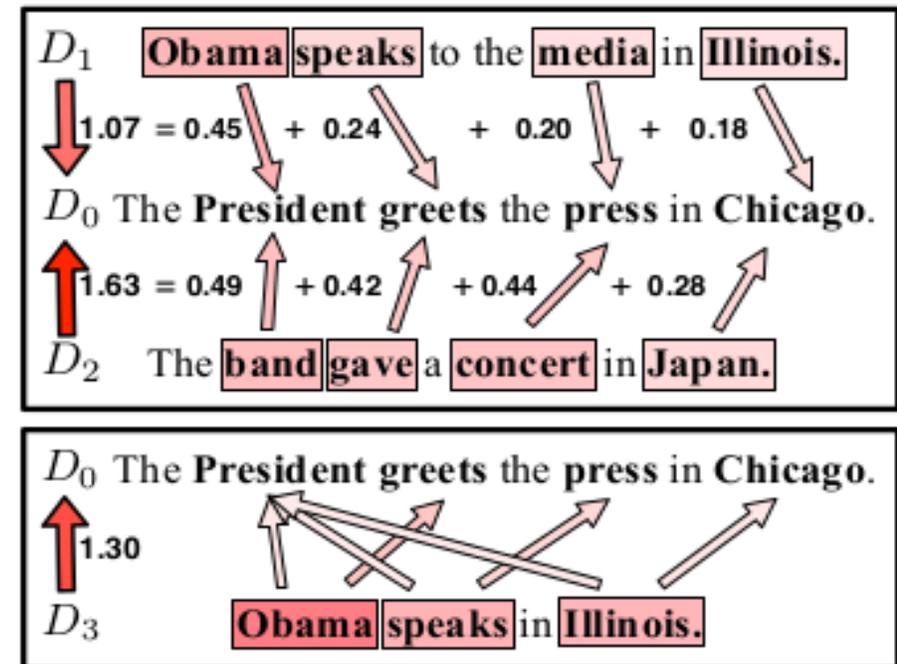
@Pablo_Iglesias_

Follow

Cambiar la ley no es suficiente, queremos que las familias recuperen su dinero. 🧑🧒🧒 **#YoVoy10N**

Semantic - WMD

- Parte de una matriz de Word Embeddings como la que hemos visto antes
- Busca la menor distancia para pasar de una oración a otra calculando las distancias entre sus palabras.
- Problema de transporte
- Matemáticamente este problema de optimización es bastante costoso, ya que además penaliza la diferencia de longitud entre documentos



Semantic - WCD

- Parte de una matriz de Word Embeddings y de un TF-IDF con el mismo vocabulario
- Simplificación de WMD
- Calculamos la media de los Word Embeddings de las palabras ponderados por su Tfidf
- Mucho menor coste computacional

	w1	w2	w3		dim1	dim2	dim3	dim4	dim5	dim6
Doc1				×	W1					
Doc2					W2					
Doc3					W3					

	dim1	dim2	dim3	dim4	dim5	dim6
W1						
W2						
W3						

WCD

- Consulta: “Ley de Reforestación”



Pedro Sánchez ✓

@sanchezcastejon

Follow

Gestionar de manera responsable nuestro patrimonio común, el agua, los suelos y la biodiversidad, es ineludible. La Ley de Cambio Climático y Transición Energética dotará de certidumbre a actores públicos y privados en los próximos años.

[#TransiciónEcológica.](#)



Pablo Iglesias ✓

@Pablo_Iglesias_

Follow

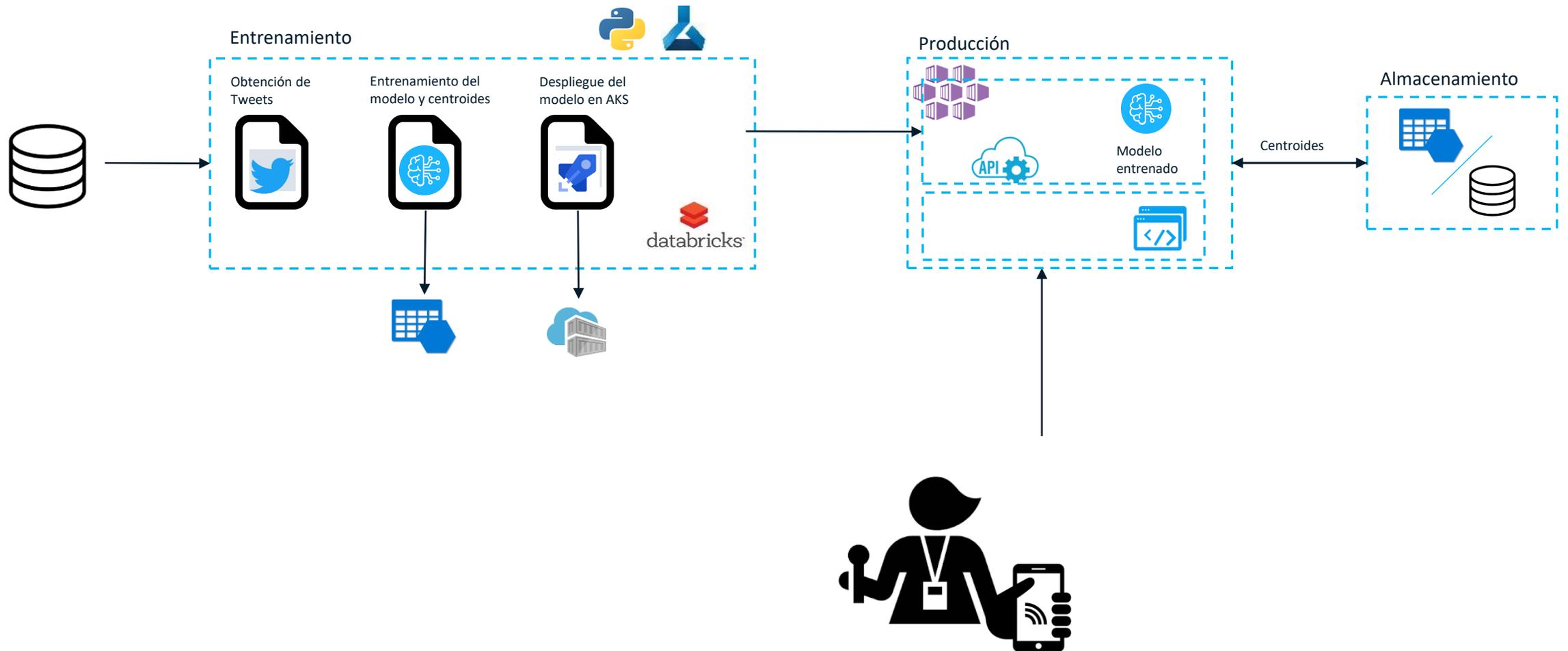
Para afrontar el cambio climático contamos con nuestro Plan Horizonte Verde para dar un impulso a las energías renovables y frenar las emisiones contaminantes, apoyado por una empresa pública de energía que garantice la transición ecológica. Se puede. farodevigo.es/sociedad/2019/ ...

Politweets

The solution

Information Retrieval

Politweets – The Solution



Politweets – It´s alive!

- Mmmmmmm... vale no funcionaría



Politweets – It´s alive!

- A ver vamos a pensar ¿Qué espera mi cliente?

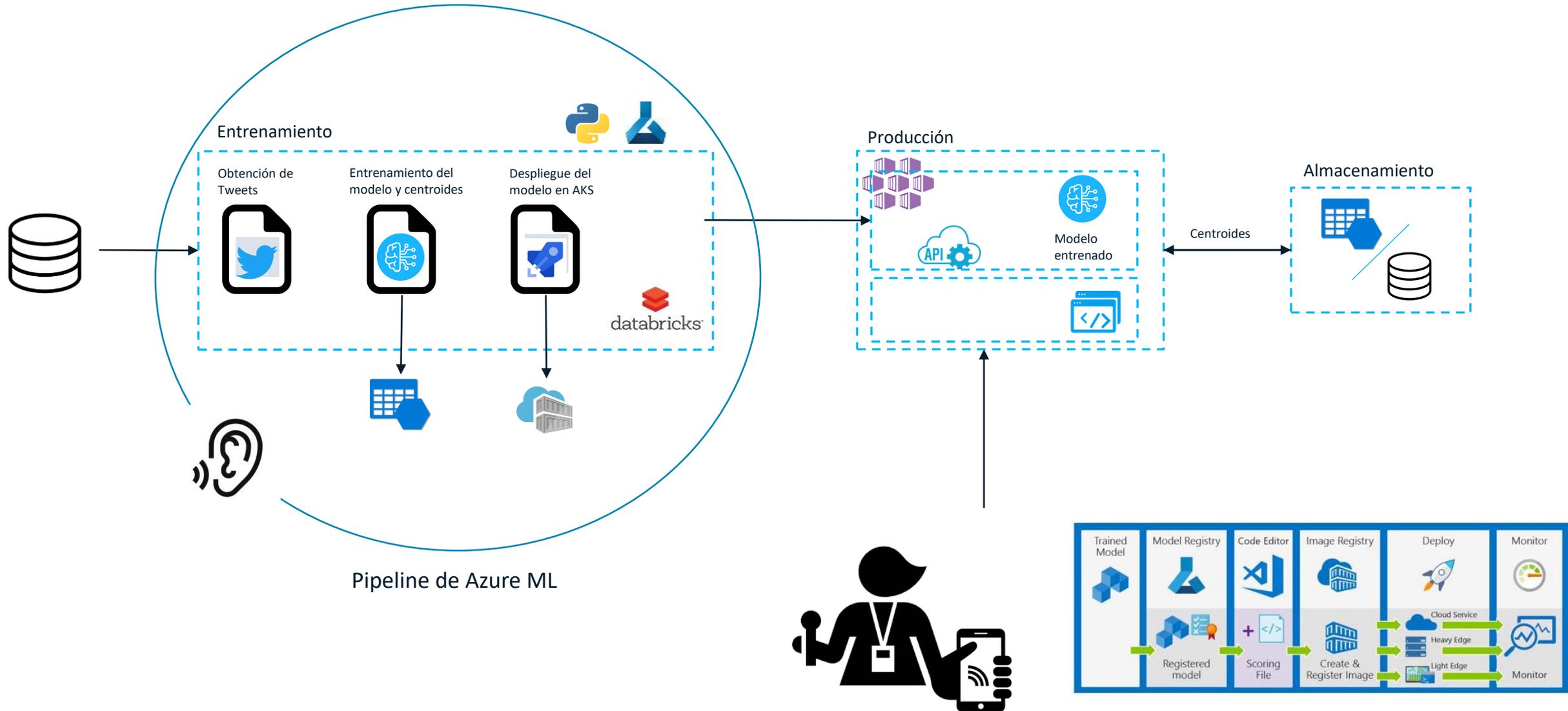


Politweets – It´s alive!

- Y si..... ¿Lo automatizo y creo un trigger en función de sus necesidades?



Politweets – The Solution



politweets



<https://github.com/madaidays/politweets-workshop>



Thank you

[@plainconcepts](#)

www.plainconcepts.com