plain concepts

Politweets

VIRTUAL COFFEE

plain concepts

AI TECHNICAL LEAD

Apasionado de los datos, la Inteligencia Artifical 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.

@ikeinyyo

plain concepts

Politweets

1. The Idea
2. Politweets AI
   Natural Language Processing
   Information Retrieval
3. The solution

plain concepts

Politweets
The Idea

plain concepts

Politweets – The Idea

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

plain concepts

Politweets – The Idea

- Saber que piensan los politicos 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

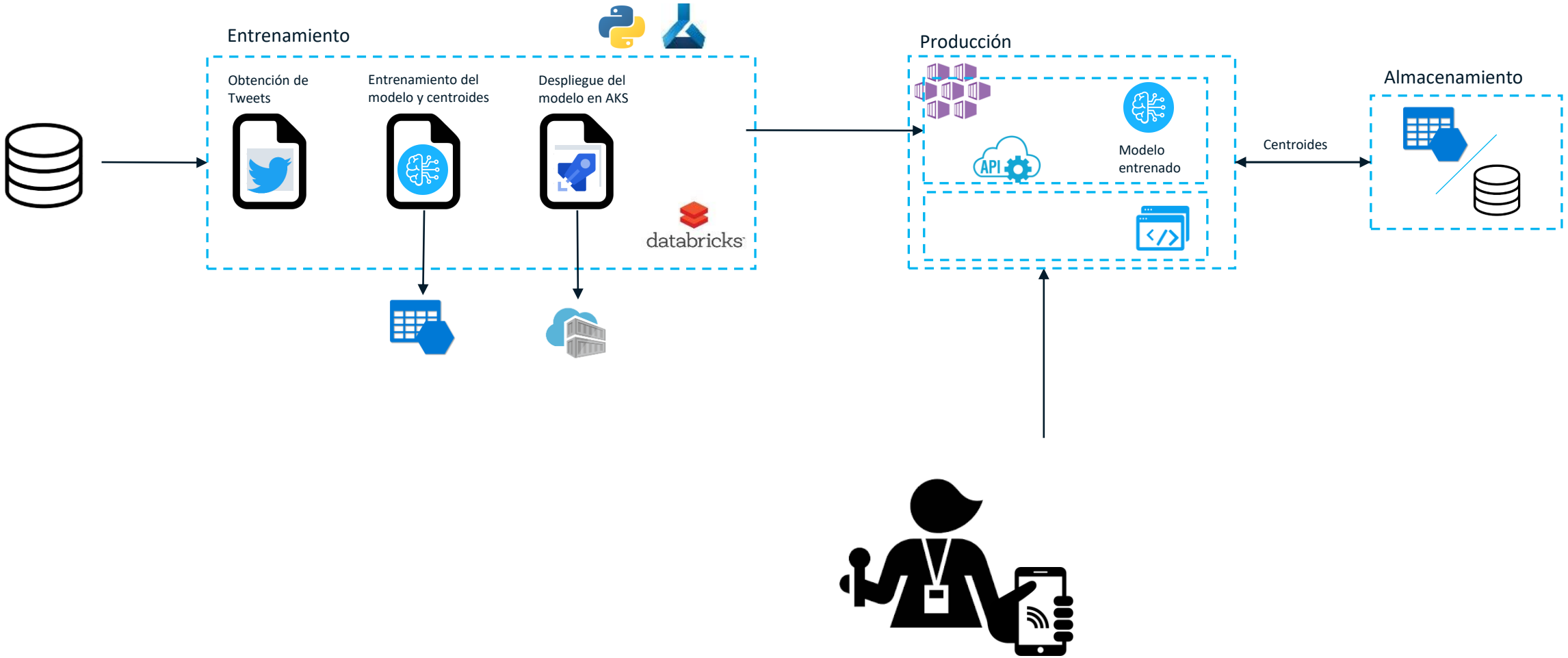
Query:

sanitario salud

TFIDF Semantic

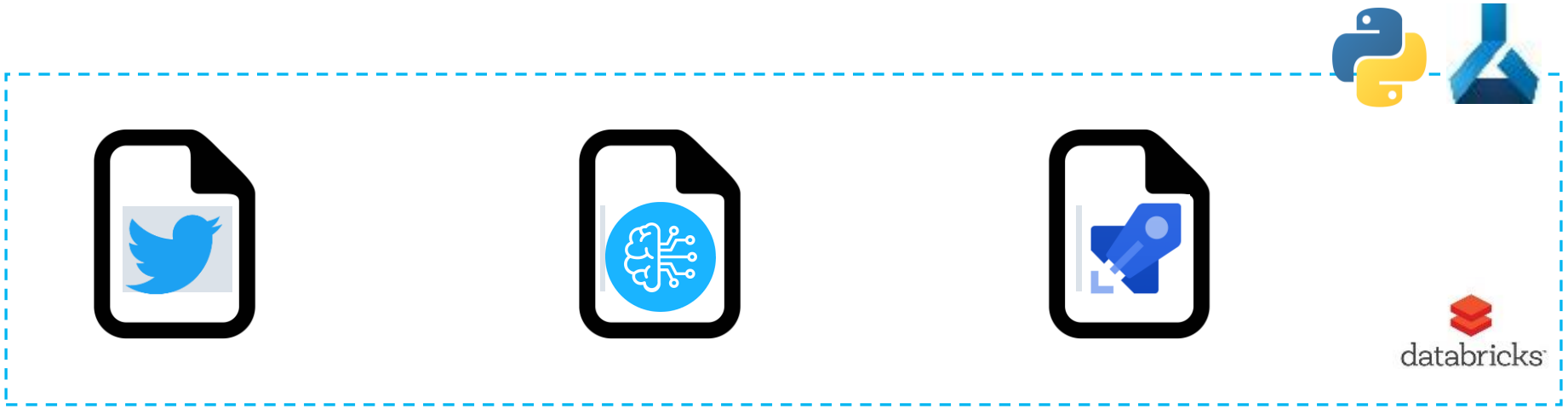
Submit

Politweets – The Idea



plain concepts

Politweets – The Idea



Extracción de datos

Entrenamiento del modelo

Despliegue del modelo

plain concepts

Politweets AI

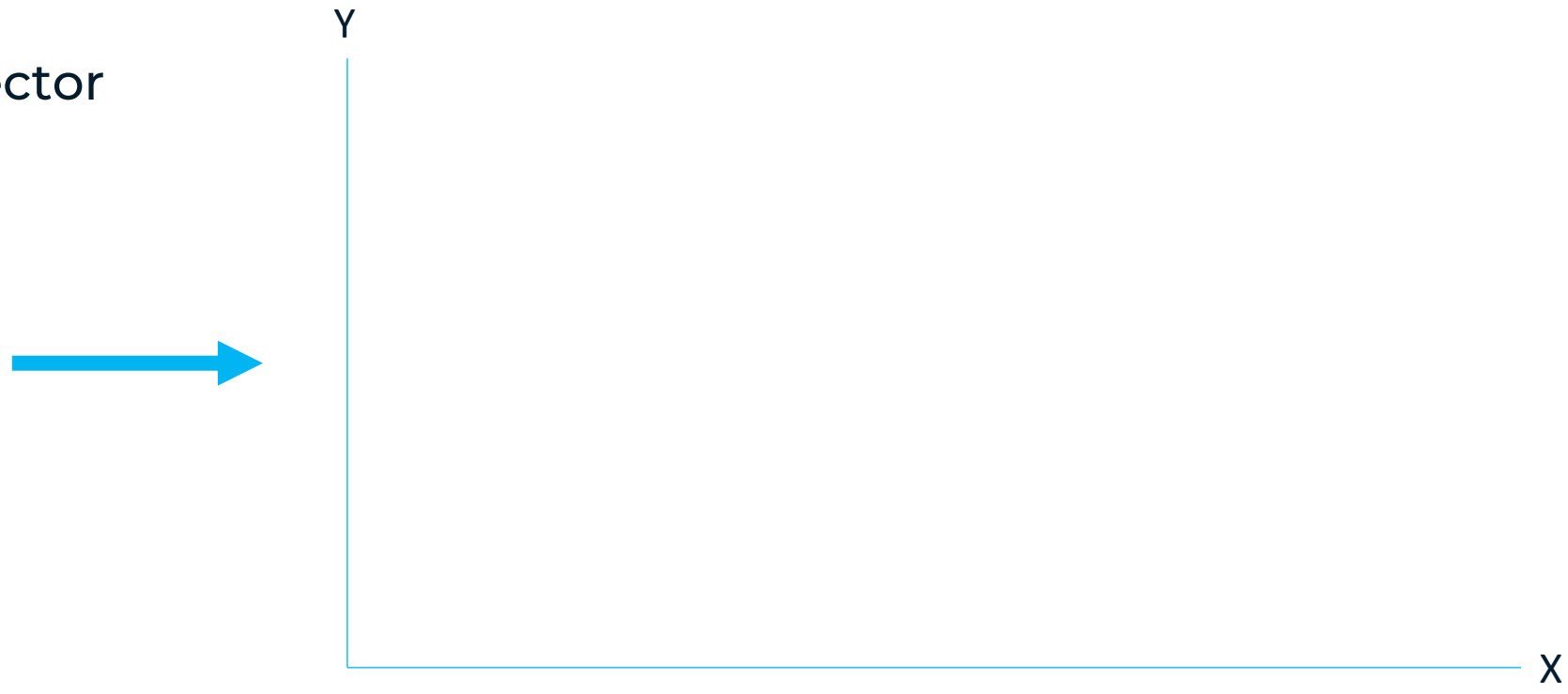
Natural
Language
Processing

plain concepts

NLP

Word to vector

blue
house
good
poor

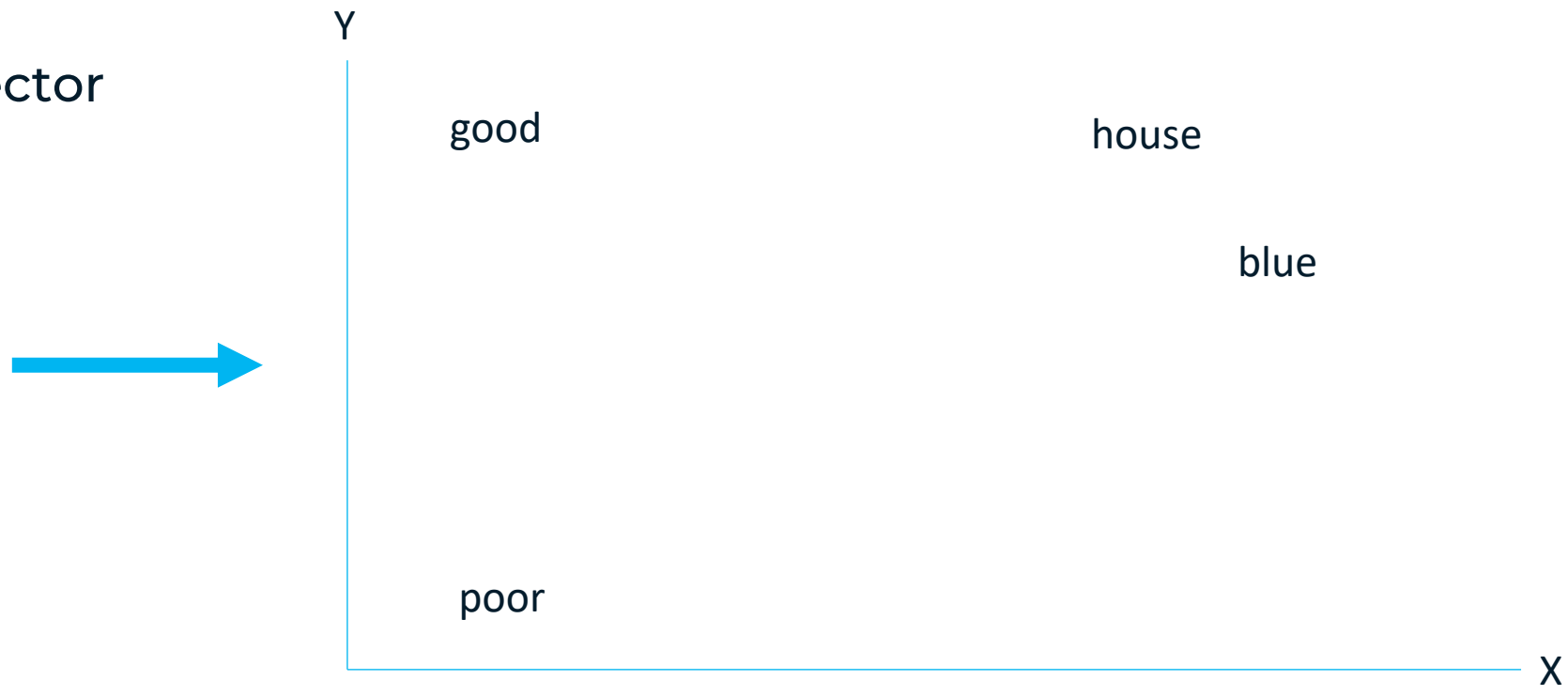


plain concepts

NLP

Word to vector

blue
house
good
poor



plain concepts

NLP

Word to vector

blue
house
good
poor

plain concepts

NLP

Word to vector

vector_of(“king”) – vector_of(“men”) + vector_of(“woman”)

=

vector_of(“queen”)

plain concepts

NLP

Redes Neuronales Simples

Formas de leer o aprender

- C-BOW
- Skip-Gram

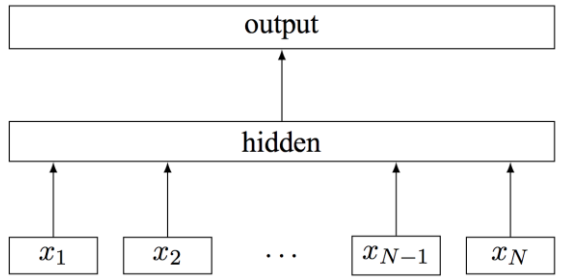
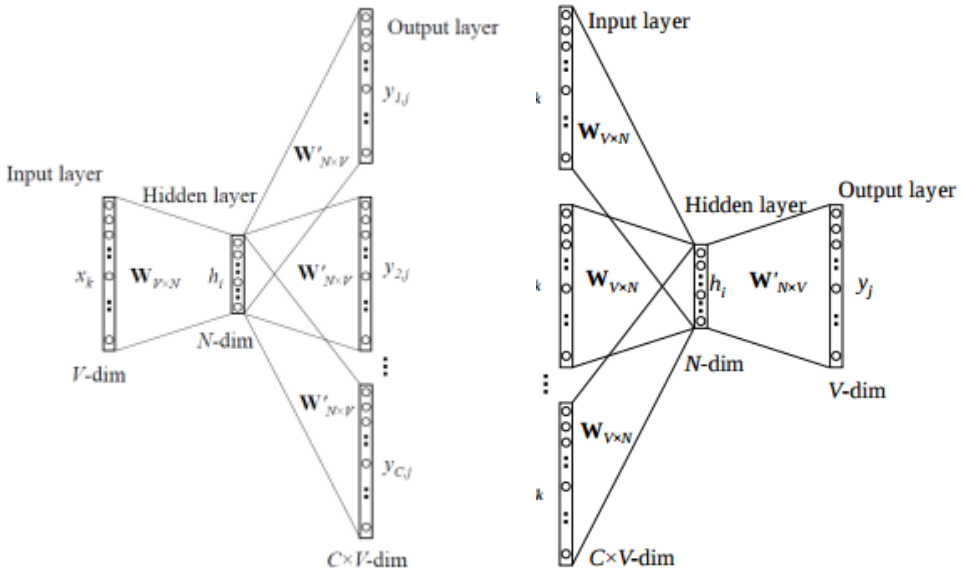


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.



fastText

PYTORCH

GENSIM
topic modelling for humans

TensorFlow

plain concepts

Word to vector

```
from gensim.models import Word2Vec

model = Word2Vec(mytexts)
```

plain concepts

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

plain concepts

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

plain concepts

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

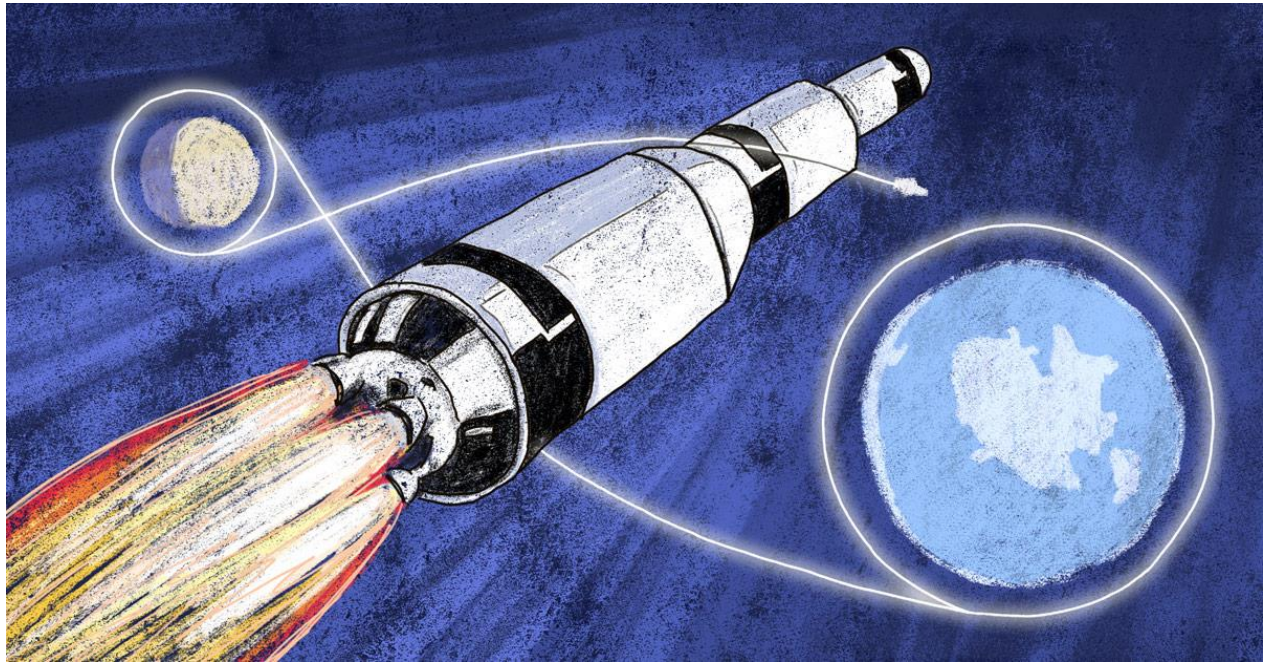
sánchez
si
españoles
ciudadanos
mundo
familias
gran
hacer
apoyo
ley

hacer
españa
si
país
familias
gente
políticas
derechos
política
ser

plain concepts

Hiperparámetros

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



plain concepts

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

plain concepts

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

Google News

WIKIPEDIA
The Free Encyclopedia

plain concepts

Palabras más próximas a Economía

económica
económico
comercio
política
globalización
desempleo
económicos
sostenible
industria
gobernabilidad

económica
económico
agricultura
ganadería
competitividad
comercio
política
recesión
económicos
pib

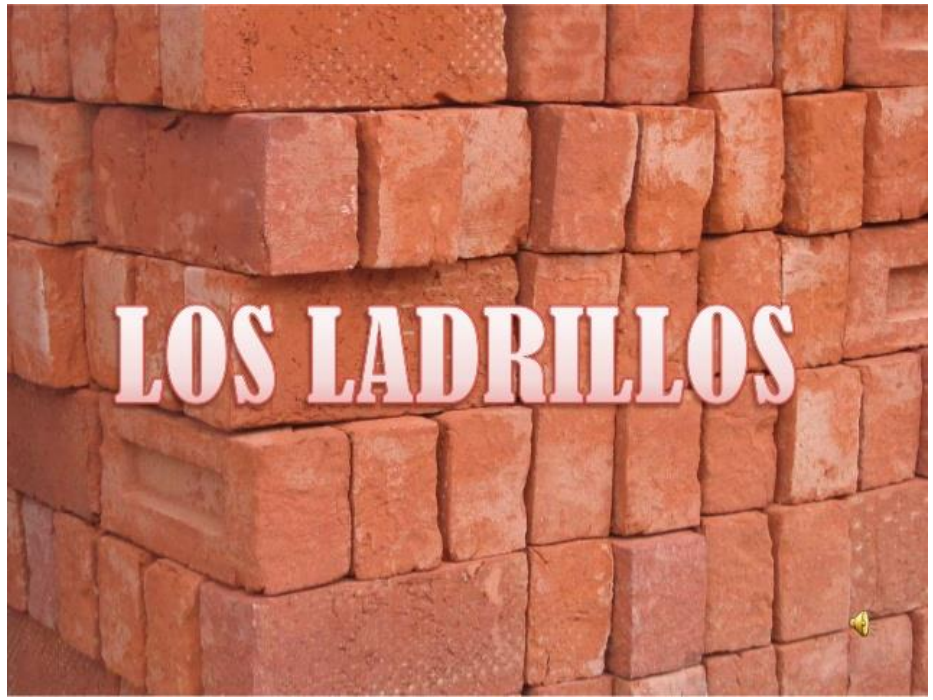
política
industria
políticas
impuestos
democracia
gobierno
socialismo
presupuestos
inmigración
agricultores

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?



plain concepts

Politweets AI

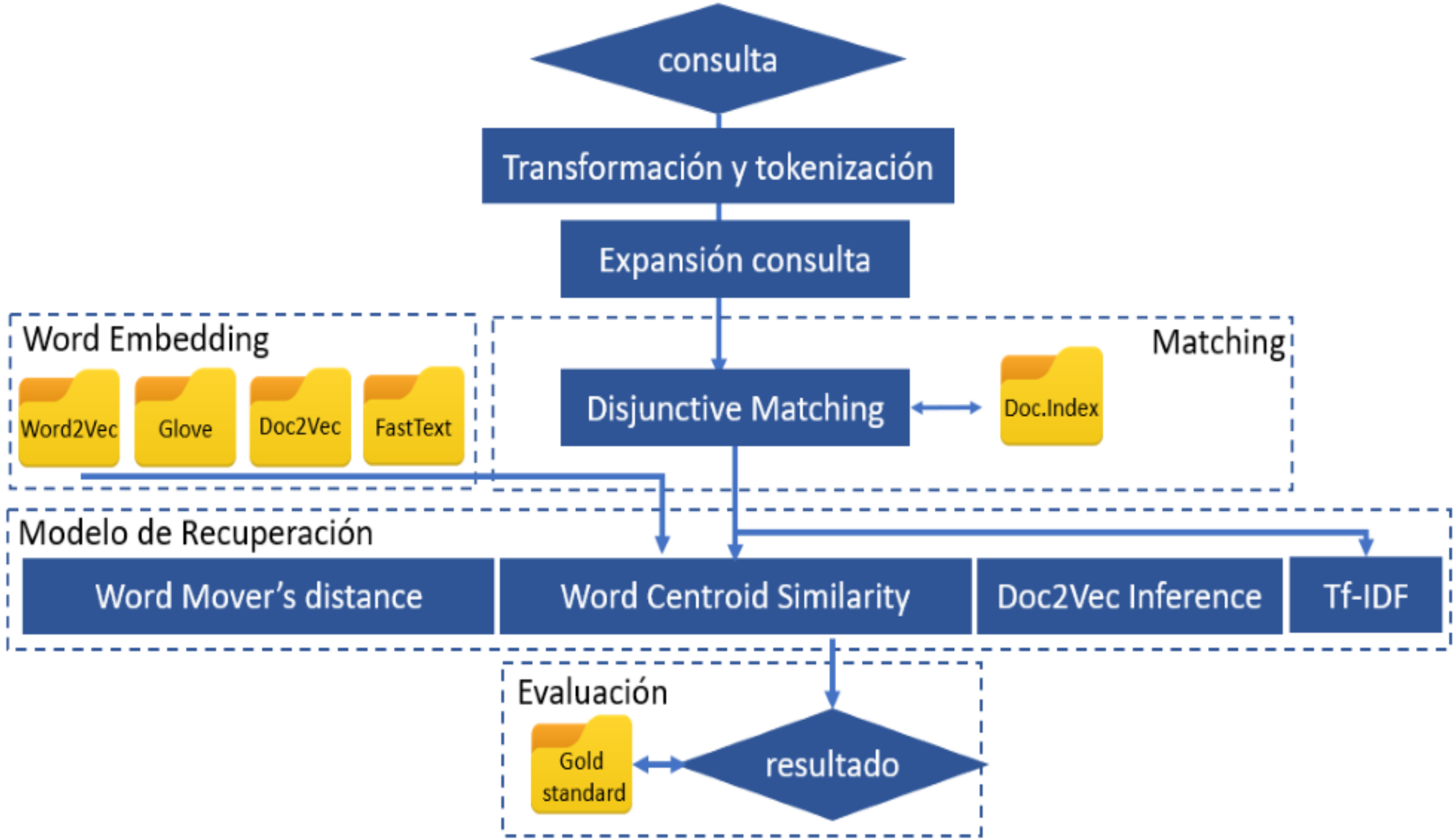
Information Retrieval

plain concepts

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?


consulta
Transformación y tokenización
Expansión consulta
Word Embedding
Word2Vec
Glove
Doc2Vec
FastText
Matching
Disjunctive Matching
Doc.Index
Modelo de Recuperación
Word Mover's distance
Word Centroid Similarity
Doc2Vec Inference
Tf-IDF
Evaluación
Gold standard
resultado

plain concepts

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_{ij} = 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

plain concepts

TFIDF

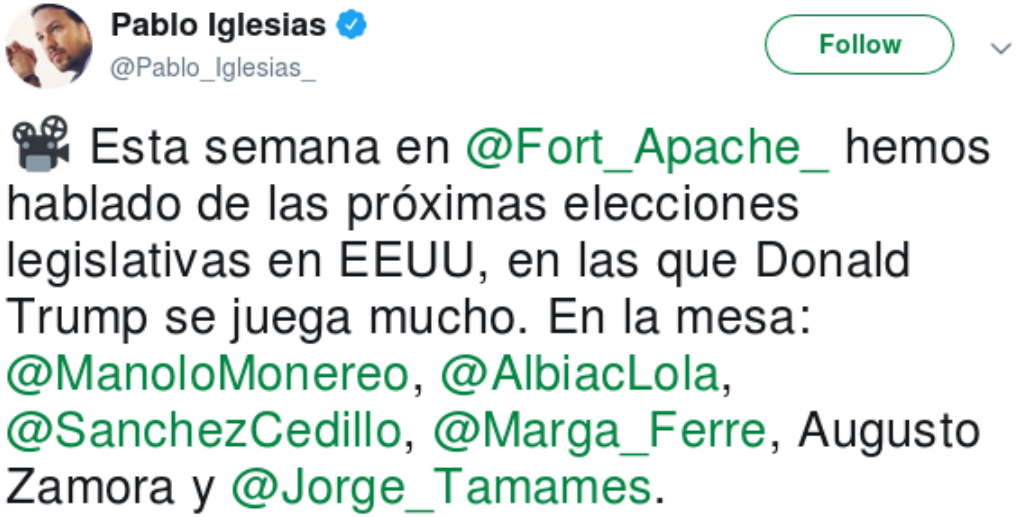
In [21]:
```
sentence = "The oldest human fossil is the skull discovered in the Cave of Aroeira in Almonda."

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

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

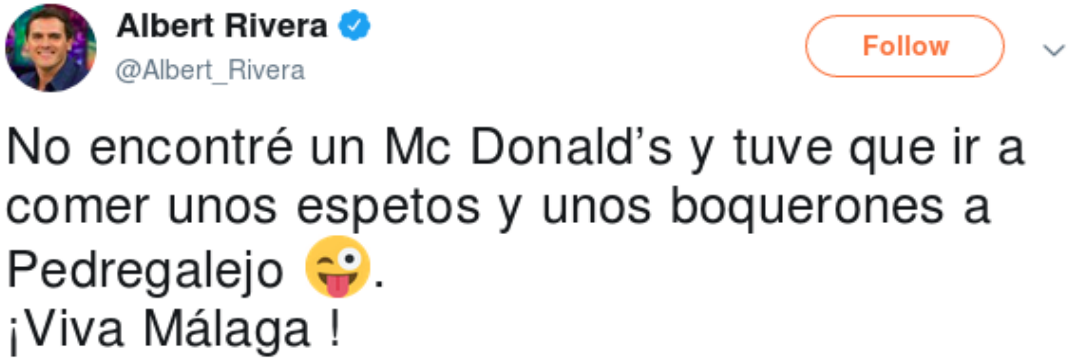
Out[21]:

	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

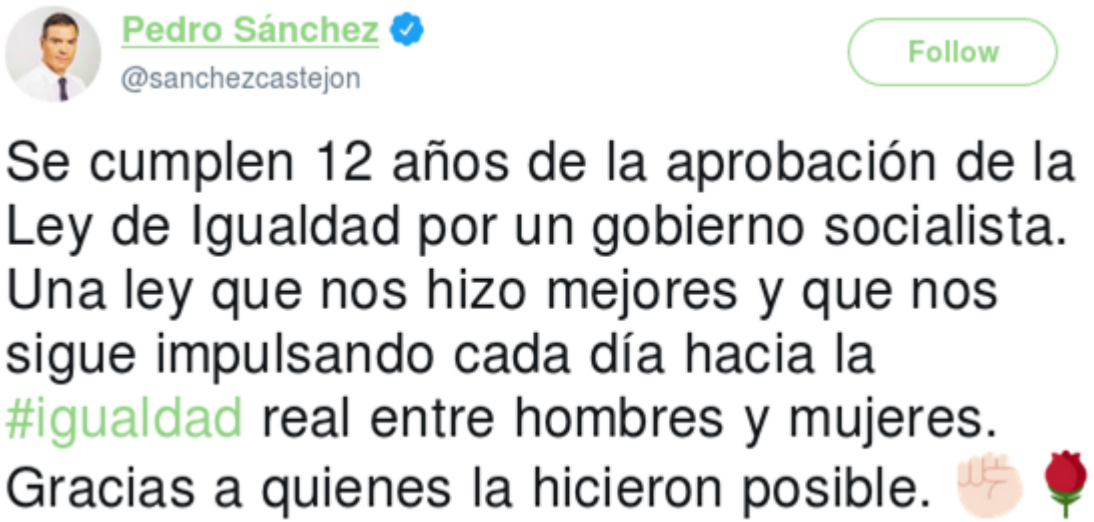
- Consulta: “**Donald Trump elecciones EEUU**”

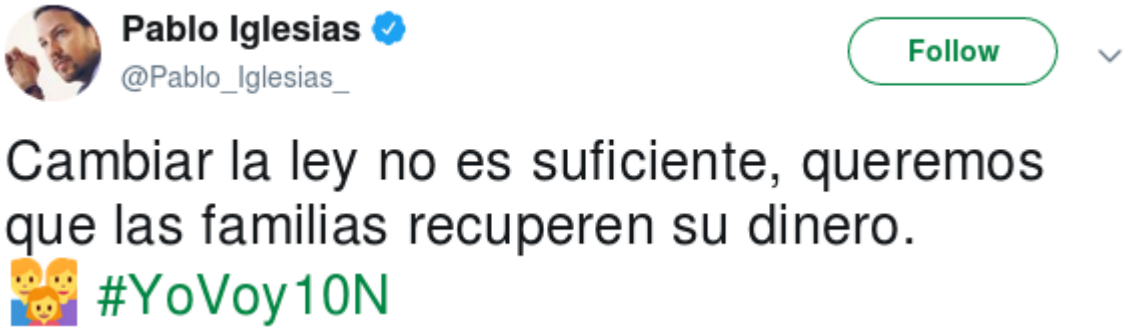




TFIDF

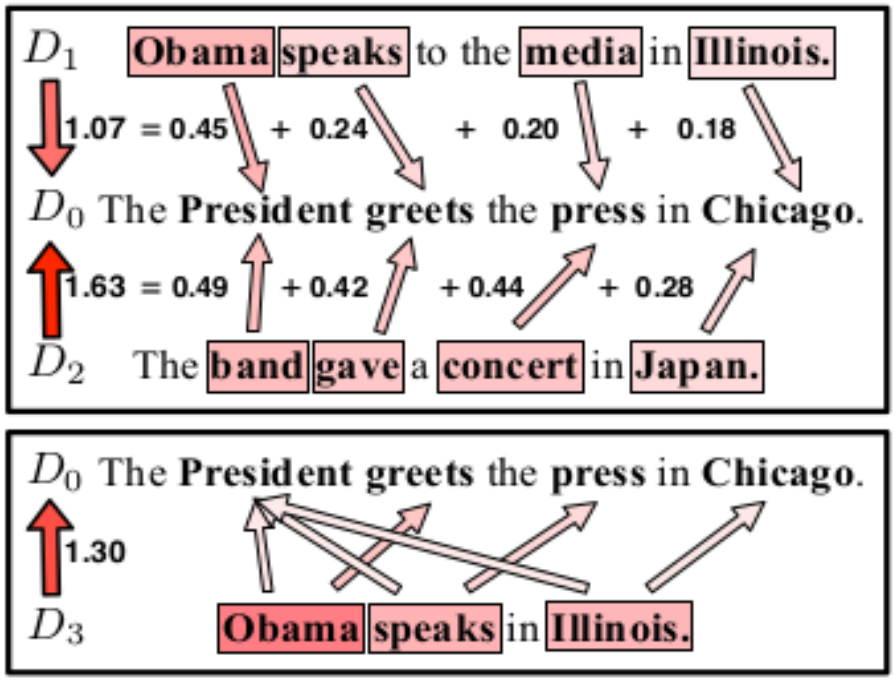
- Consulta: **“Ley de Reforestación”**





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



plain concepts

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
Doc1			
Doc2			
Doc3			

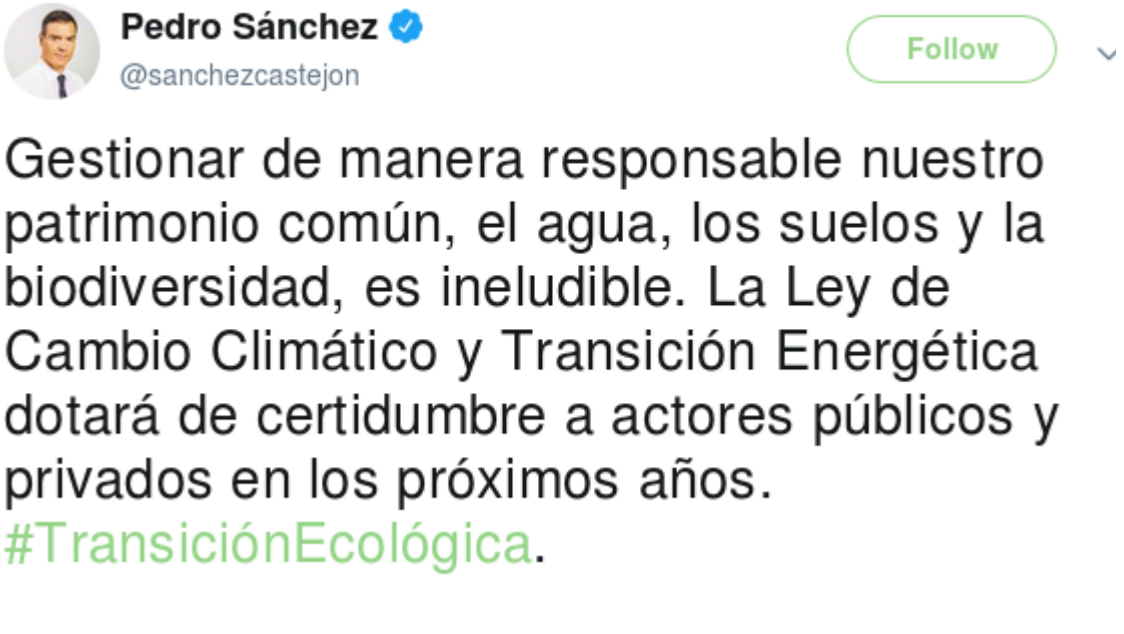
	dim1	dim2	dim3	dim4	dim5	dim6
W1						
W2						
W3						

	dim1	dim2	dim3	dim4	dim5	dim6
W1						
W2						
W3						

plain concepts

WCD

- Consulta: **“Ley de Reforestación”**



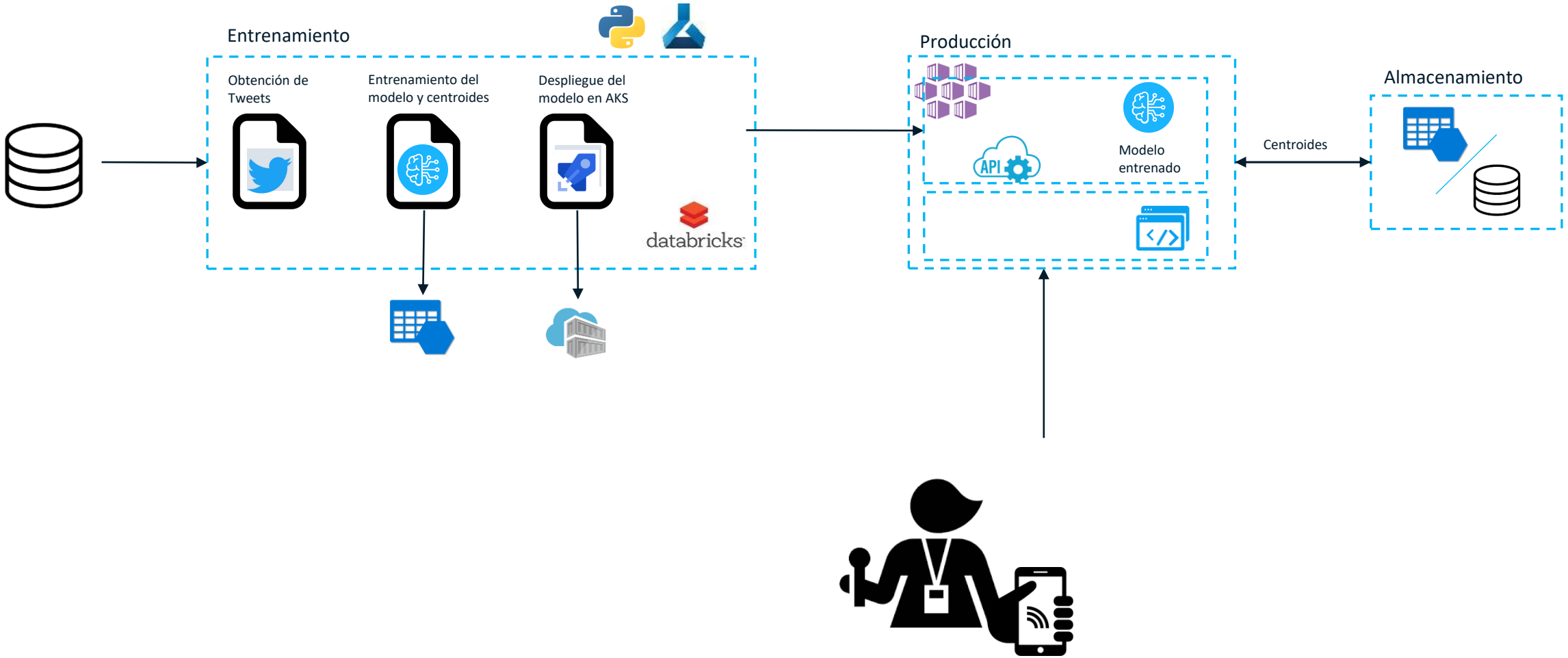
plain concepts

Politweets
The solution

Information Retrieval

plain concepts

Politweets – The Solution



plain concepts

Politweets – It´s alive!

- Mmmmmmmm... vale no funcionaría

plain concepts

Politweets – It´s alive!

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

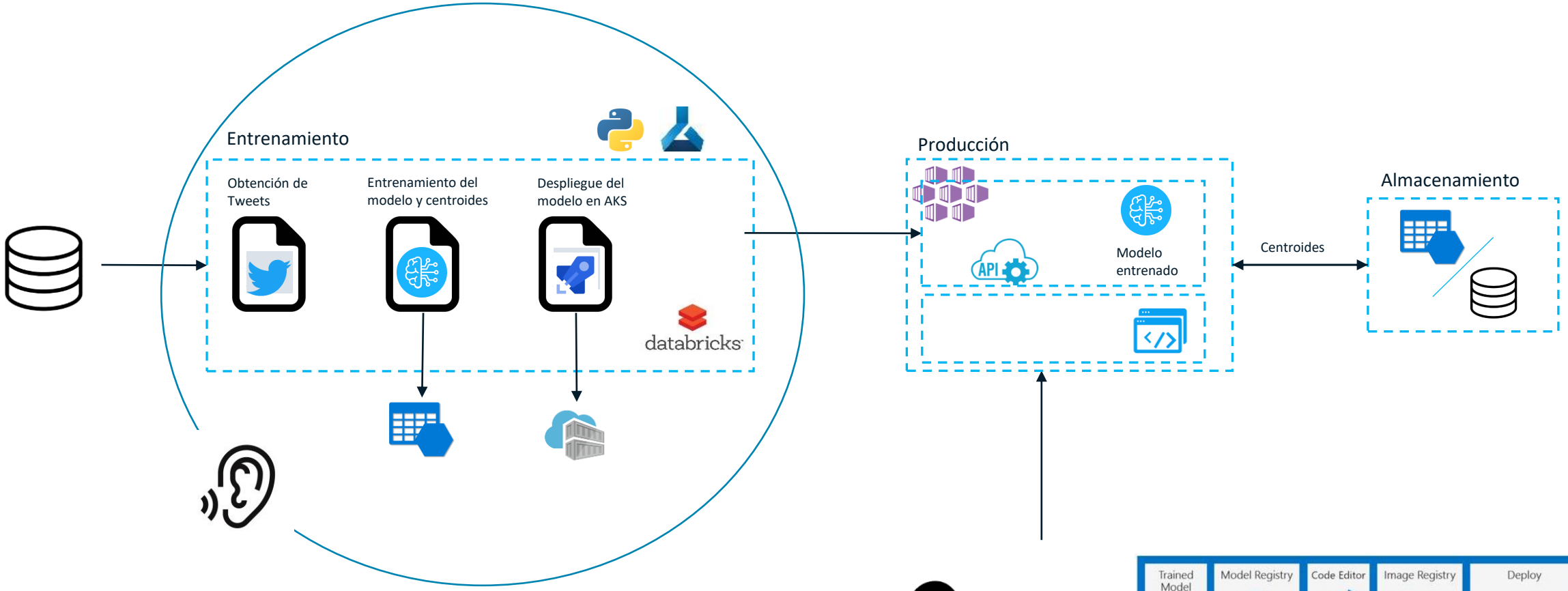
plain concepts

Politweets – It´s alive!

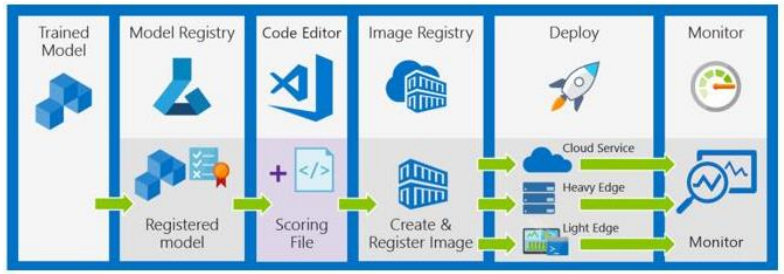
- Y si….. ¿Lo automatizo y creo un trigger en función de sus necesidades?

plain concepts

Politweets – The Solution



Pipeline de Azure ML



plain concepts

politweets

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

Thank you

@plainconcepts

www.plainconcepts.com