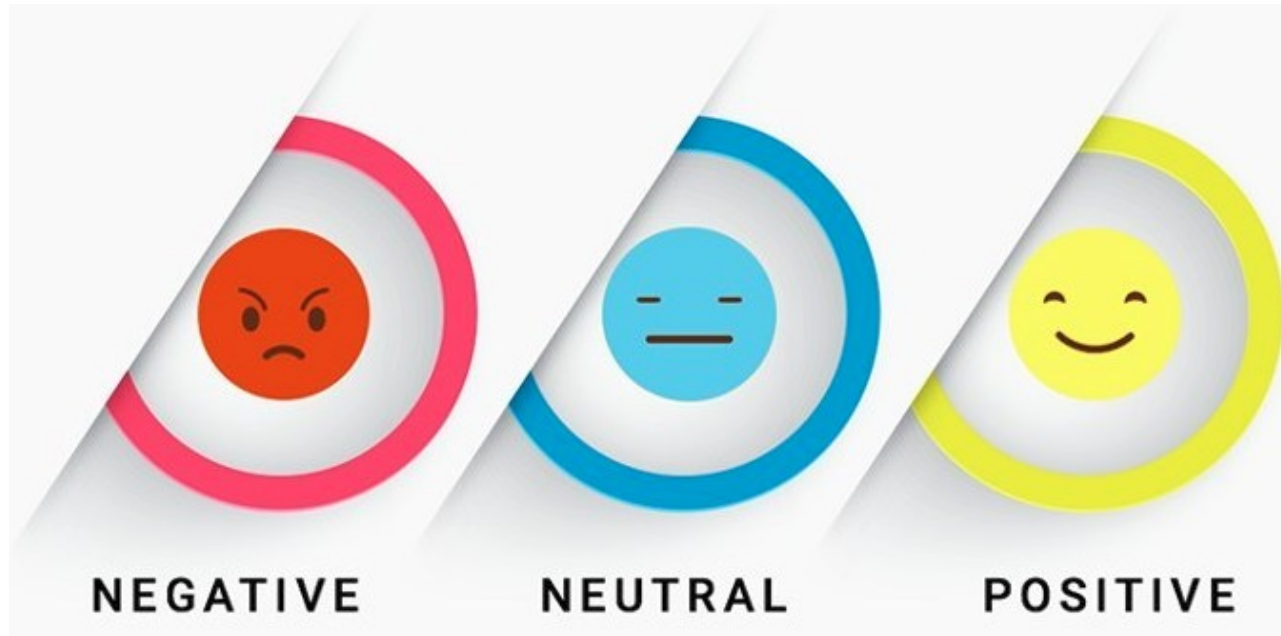


Taller

Análisis de sentimiento

¡Gracias redes sociales!



Minado de texto

Positivo o negativo?

Nostalgia, envidia, alegría, odio...?

Spam?

...

Subjetivo u objetivo?

...

Científico, literatura, mensaje personal...?

Inteligencia artificial

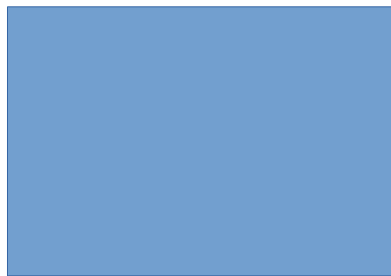
Aprendizaje automático



“haha buenísimo el rap de
rajoy contra pablo”

0	ingeniería
1	política
0	ciencia
0	humor

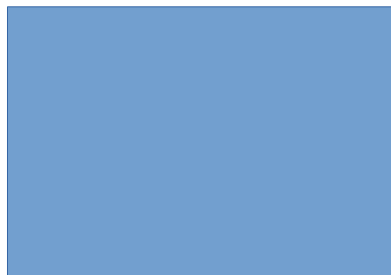
“haha buenísimo el
rap de rajoy contra
pablo”


$$\begin{bmatrix} 0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

ingeniería
política
ciencia
humor



“haha buenísimo el
rap de rajoy contra
pablo”



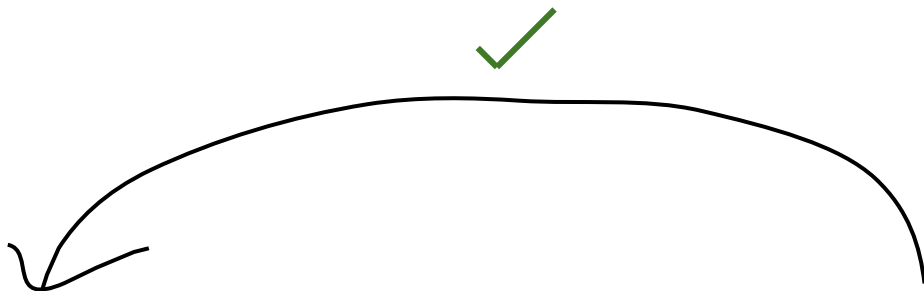
$$\begin{bmatrix} 0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

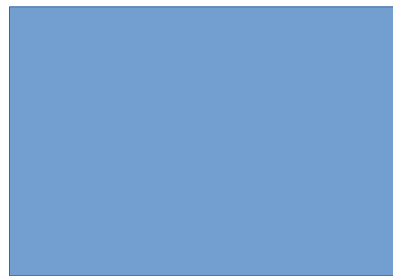
$$- \begin{bmatrix} 0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

=

$$\begin{bmatrix} -0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$



“haha buenísimo el rap de rajoy contra pablo”



$$\begin{bmatrix} 0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

=

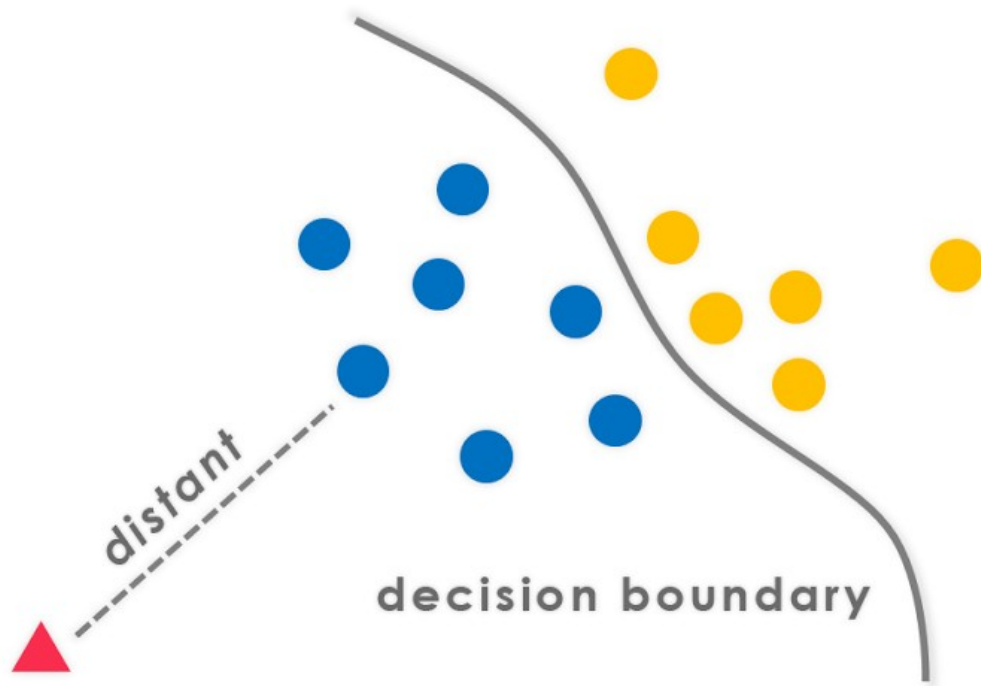
$$\begin{bmatrix} -0.8 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

“pos rajoy está más bueno q pedro xd”

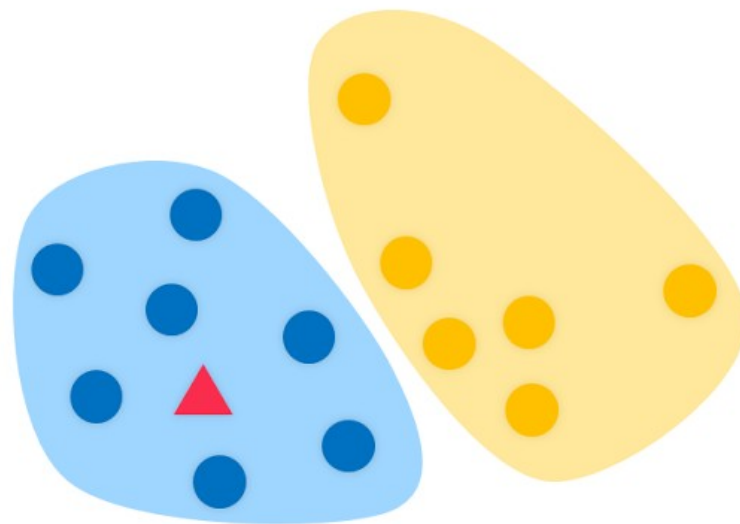
0.95 política



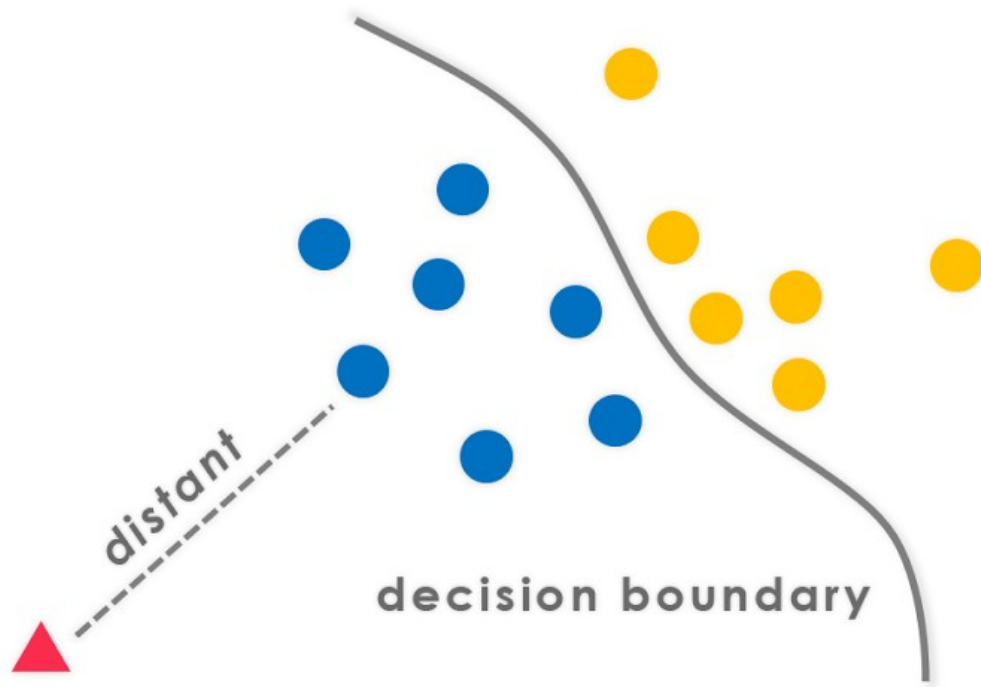
Discriminative



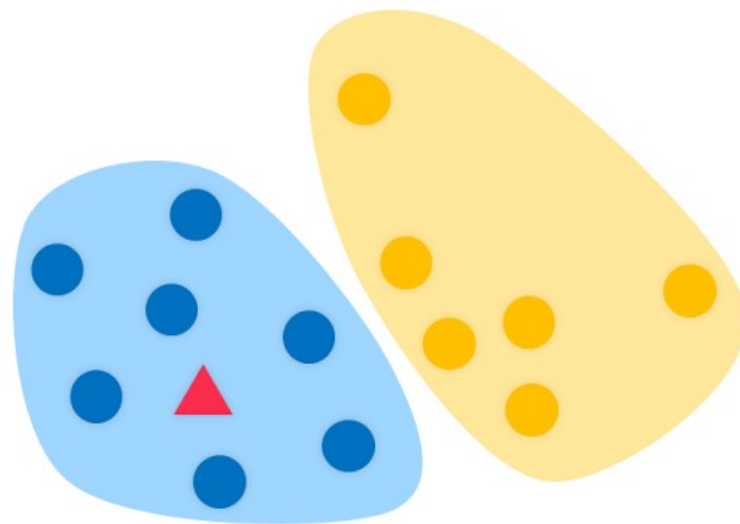
Generative



Discriminative



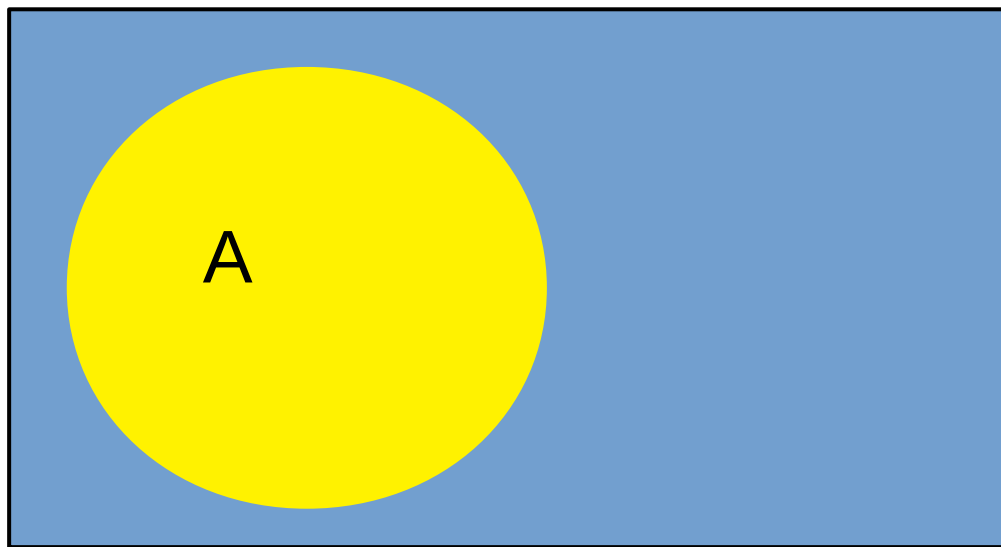
Generative



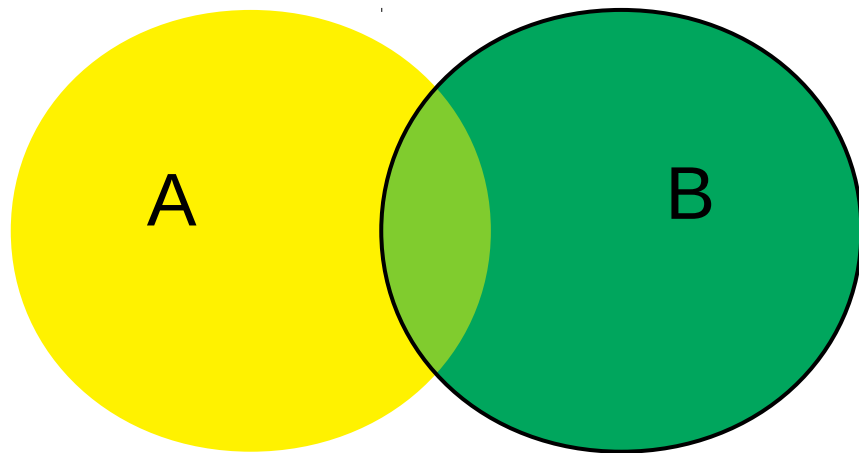
$$P(\text{"ciencias"} \mid \text{"proteina"}, \text{"lípidos"}) = ???$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



$$P(\text{ciencias} \mid \text{proteína, lípido}) = \frac{P(\text{proteína, lípido} \mid \text{ciencias}) \times P(\text{ciencias})}{P(\text{proteína, lípido})}$$

$$P(\text{ciencias} \mid \text{proteína, lípido}) = P(\text{proteína, lípido} \mid \text{ciencias}) \times P(\text{ciencias})$$



Prior

$$P(\text{ciencias} \mid \text{proteína, lípido}) = P(\text{proteína, lípido} \mid \text{ciencias}) \times P(\text{ciencias})$$



Posterior



Prior

$P(\text{proteína, lípido, ciencia}) =$

$$P(\text{proteína, lípido, ciencia}) = P(\text{proteína} \mid \text{lípido, ciencia}) \times P(\text{lípido, ciencia})$$

$$\begin{aligned} P(\text{proteína, lípido, ciencia}) &= P(\text{proteína} \mid \text{lípido, ciencia}) \times P(\text{lípido, ciencia}) \\ &= P(\text{proteína} \mid \text{lípido, ciencia}) \times P(\text{lípido} \mid \text{ciencia}) \times P(\text{ciencia}) \end{aligned}$$

Solución: **Naive Bayes**

$$P(A, B) = P(A|B) \times P(B) \dots = P(A) \times P(B)$$

Solución: **Naive Bayes**

$$P(A, B) = P(A | B) \times P(B) \dots = P(A) \times P(B)$$

$$P(\text{proteína, lípido, ciencia}) = P(\text{proteína} | \text{ciencia}) \times P(\text{lípido} | \text{ciencia}) \times P(\text{ciencia})$$

Solución: **Naive Bayes**

$$P(A, B) = P(A | B) \times P(B) \dots = P(A) \times P(B)$$

$$P(\text{proteína, lípido, ciencia}) = P(\text{proteína} | \text{ciencia}) \times P(\text{lípido} | \text{ciencia}) \times P(\text{ciencia})$$

Manos a la obra

Pasos

1. Obtener un data set para inferir algo
- 2. Obtener un texto a ser inferido**
3. Limpiar los textos...
4. Mejorar textos
5. Vectorizar textos
4. Entrenar un modelo (en nuestro caso, clasificador Naïve Bayes)
5. Evaluar la eficacia del modelo
6. ¡Predecir!

Dataset



TASS: Workshop on

Dataset Download

General Corpus (2012)

- [Train set](#) (tagged with entities, 5-level global and aspect-based sentiment and topics)
- [Train set](#) (tagged with entities, 3-level global and aspect-based sentiment and topics)

Politics corpus (2013)

Social-TV Corpus (2014)

- [Train set](#) (tagged with aspects and 3-level aspect-based sentiment)

STOMPOL Corpus (2015)

- [Train set](#) (tagged with aspects and 3-level aspect-based sentiment)

InterTASS corpus (tass 2018, task1)

Spanish dataset

- [Train set](#) (tagged with 3-level global sentiment)
- [Development set](#) (tagged with 3-level global sentiment)
- [Test](#)

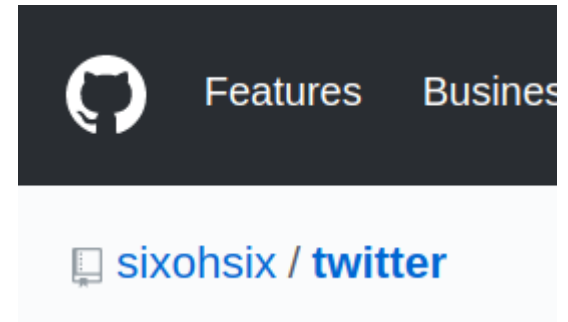
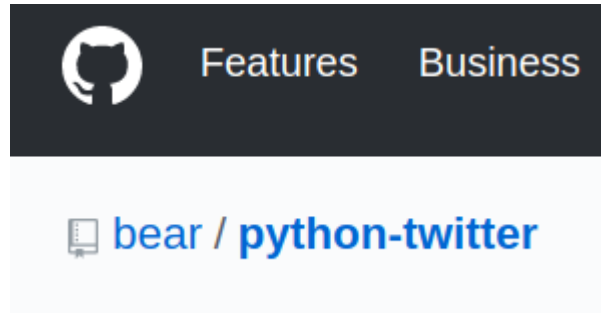
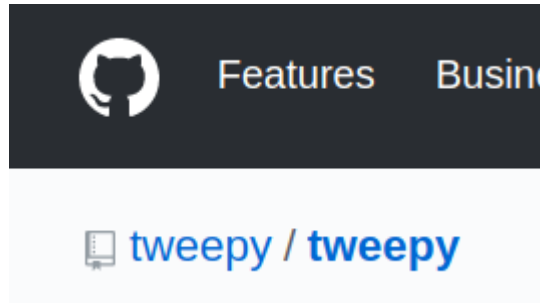
```
<tweet>
  <tweetid>771092111429083136</tweetid>
  <user>esskuu94</user>
  <content>Al final han sido 3h Bueno, mañana tengo fiesta así que.. No me quejo </content>
  <date>2016-08-31 21:07:40</date>
  <lang>es</lang>
  <sentiment>
    <polarity><value>P</value></polarity>
  </sentiment>
</tweet>
<tweet>
  <tweetid>771092070572449796</tweetid>
  <user>__ariadna9</user>
  <content>@Jorge_Ruiz14 yo no tengo tiempo para esas cosas ahora mismo </content>
  <date>2016-08-31 21:07:30</date>
  <lang>es</lang>
  <sentiment>
    <polarity><value>N</value></polarity>
  </sentiment>
</tweet>
```

Pasos

1. Obtener un data set para inferir algo
- 2. Obtener un texto a ser inferido**
3. Limpiar los textos...
4. Mejorar para la clasificación
5. Vectorizar textos
4. Entrenar un modelo (en nuestro caso, clasificador Naïve Bayes)
5. Evaluar la eficacia del modelo
6. ¡Predecir!

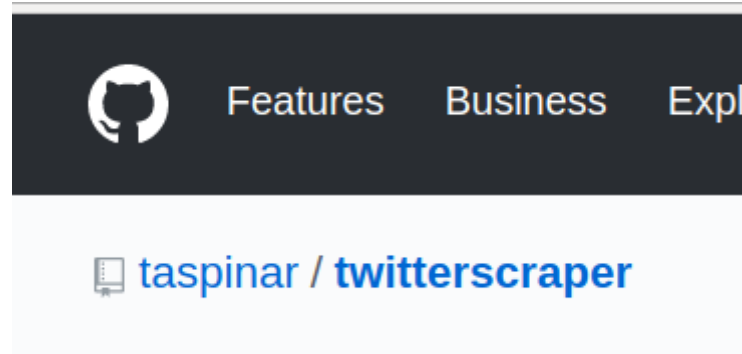
Sacar tweeeets

¿API de twitter?



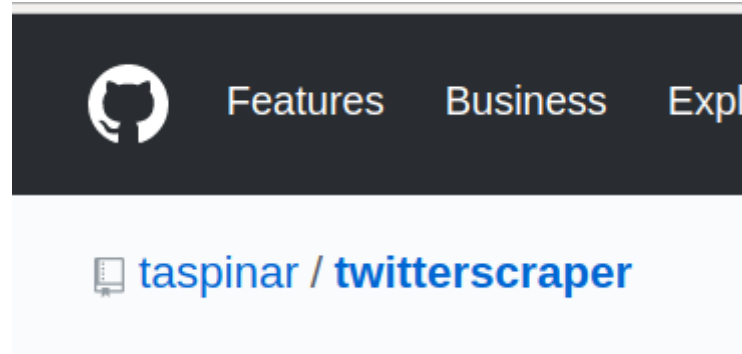
Sacar tweeeets

Meh, mejor un scrapper



Sacar tweeeets

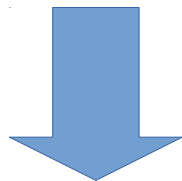
Meh, mejor un scrapper



```
bytes@bytes-GL752VM:~/Desktop/Simbiosis/Text Sentiment/scrap$ twitterscraper @LaCasaInvisible -l 1000 -o tweets.csv --csv
INFO: queries: ['@LaCasaInvisible since:2006-03-21 until:2006-11-01', '@LaCasaInvisible since:2006-11-01 until:2007-06-15', '@LaCasaInvisible since:2007-06-15 until:2008-01-26', '@LaCasaInvisible since:2008-01-26 until:2008-09-08', '@LaCasaInvisible since:2008-09-08 until:2009-04-21', '@LaCasaInvisible since:2009-04-21 until:2009-12-03', '@LaCasaInvisible since:2009-12-03 until:2010-07-16', '@LaCasaInvisible since:2010-07-16 until:2011-02-27', '@LaCasaInvisible since:2011-02-27 until:2011-10-10', '@LaCasaInvisible since:2011-10-10 until:2012-05-23', '@LaCasaInvisible since:2012-05-23 until:2013-01-03', '@LaCasaInvisible since:2013-01-03 until:2013-08-17', '@LaCasaInvisible since:2013-08-17 until:2014-03-30', '@LaCasaInvisible since:2014-03-30 until:2014-11-11', '@LaCasaInvisible since:2014-11-11 until:2015-06-24', '@LaCasaInvisible since:2015-06-24 until:2016-02-05', '@LaCasaInvisible since:2016-02-05 until:2016-09-17', '@LaCasaInvisible since:2016-09-17 until:2017-05-01', '@LaCasaInvisible since:2017-05-01 until:2017-12-12', '@LaCasaInvisible since:2017-12-12 until:2018-07-26']
INFO: Querying @LaCasaInvisible since:2006-03-21 until:2006-11-01
INFO: Querying @LaCasaInvisible since:2006-11-01 until:2007-06-15
INFO: Querying @LaCasaInvisible since:2007-06-15 until:2008-01-26
INFO: Querying @LaCasaInvisible since:2008-01-26 until:2008-09-08
INFO: Querying @LaCasaInvisible since:2008-09-08 until:2009-04-21
INFO: Querying @LaCasaInvisible since:2009-04-21 until:2009-12-03
INFO: Querying @LaCasaInvisible since:2009-12-03 until:2010-07-16
INFO: Querying @LaCasaInvisible since:2010-07-16 until:2011-02-27
INFO: Querying @LaCasaInvisible since:2011-02-27 until:2011-10-10
INFO: Querying @LaCasaInvisible since:2011-10-10 until:2012-05-23
INFO: Querying @LaCasaInvisible since:2012-05-23 until:2013-01-03
INFO: Querying @LaCasaInvisible since:2013-01-03 until:2013-08-17
INFO: Querying @LaCasaInvisible since:2013-08-17 until:2014-03-30
INFO: Querying @LaCasaInvisible since:2014-03-30 until:2014-11-11
INFO: Querying @LaCasaInvisible since:2014-11-11 until:2015-06-24
INFO: Querying @LaCasaInvisible since:2015-06-24 until:2016-02-05
INFO: Querying @LaCasaInvisible since:2016-02-05 until:2016-09-17
INFO: Querying @LaCasaInvisible since:2016-09-17 until:2017-05-01
INFO: Querying @LaCasaInvisible since:2017-05-01 until:2017-12-12
INFO: Querying @LaCasaInvisible since:2017-12-12 until:2018-07-26
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2009-04-21%20until%3A2009-12-03.
INFO: Got 0 tweets (0 new).
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2017-05-01%20until%3A2017-12-12.
INFO: Got 0 tweets (0 new).
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2006-11-01%20until%3A2007-06-15.
INFO: Got 0 tweets (0 new).
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2008-01-26%20until%3A2008-09-08.
INFO: Got 0 tweets (0 new).
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2006-03-21%20until%3A2006-11-01.
INFO: Got 0 tweets (0 new).
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2008-09-08%20until%3A2009-04-21.
INFO: Got 0 tweets (0 new).
INFO: Got 0 tweets for @LaCasaInvisible%20since%3A2007-06-15%20until%3A2008-01-26.
INFO: Got 0 tweets (0 new).
INFO: Got 60 tweets for @LaCasaInvisible%20since%3A2010-07-16%20until%3A2011-02-27.
INFO: Got 60 tweets (60 new).
INFO: Got 60 tweets for @LaCasaInvisible%20since%3A2014-11-11%20until%3A2015-06-24.
INFO: Got 120 tweets (60 new).
INFO: Got 60 tweets for @LaCasaInvisible%20since%3A2014-03-30%20until%3A2014-11-11.
INFO: Got 180 tweets (60 new).
INFO: Got 60 tweets for @LaCasaInvisible%20since%3A2016-09-17%20until%3A2017-05-01.
INFO: Got 240 tweets (60 new).
INFO: Got 60 tweets for @LaCasaInvisible%20since%3A2017-12-12%20until%3A2018-07-26.
INFO: Got 300 tweets (60 new).
INFO: Got 60 tweets for @LaCasaInvisible%20since%3A2009-12-03%20until%3A2010-07-16.
INFO: Got 360 tweets (60 new).
```


3. Limpiar textos

¡oah q guapo illo! Lo vamos a petar en la fiesta de #fiestarandom123,
organizada por @quiensea84. Visita nuestra página del evento
<http://algunapaginarandom.que/no/existe.html>



oah qué guapo illo vamos a petar fiesta organizada visita nuestra página del
evento

3. Limpiar textos

1. hashtags, links y demás cosas de redes sociales
2. símbolos como “!, #” o emoticonos (todo lo que no sea una palabra en realidad)
3. queremos todo en minúsculas
4. [podemos aplicar un corrector ortográfico]

Pasos

1. Obtener un data set para inferir algo
2. Obtener un texto a ser inferido
3. Limpiar los textos...
- 4. Mejorar para la clasificación**
5. Vectorizar textos
4. Entrenar un modelo (en nuestro caso, clasificador Naïve Bayes)
5. Evaluar la eficacia del modelo
6. ¡Predecir!

4. Mejorar para la clasificación

1. Podemos eliminar stopwords
2. Podemos aplicar un “lemmatizador” → saca raíz de las palabras
- [3. Extraer *features*: n-gramas]

Vectorizar

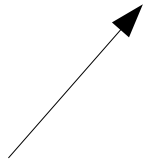
1. Obtener un data set para inferir algo
2. Obtener un texto a ser inferido
3. Limpiar los textos...
4. Mejorar para la clasificación

5. Vectorizar textos

4. Entrenar un modelo (en nuestro caso, clasificador Naïve Bayes)
5. Evaluar la eficacia del modelo
6. ¡Predecir!

Vectorizar

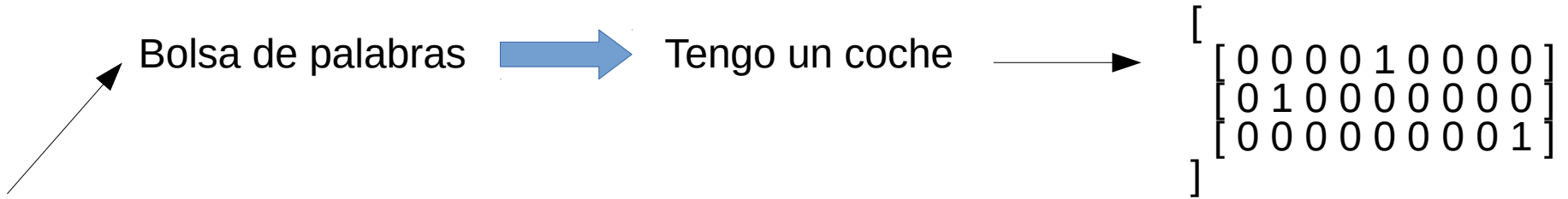
Bolsa de palabras



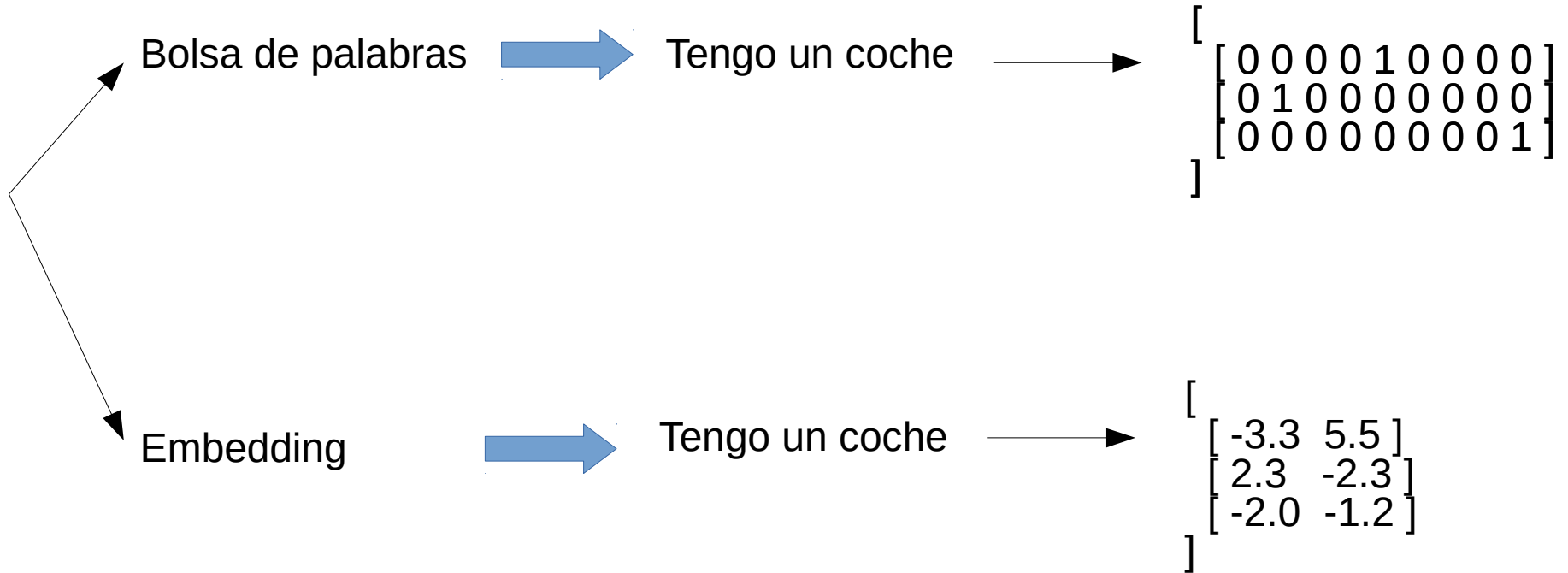
Bolsa de palabras

Armario	0	[1 0 0 0 0 0 0 0 0]
Almacen	1	[0 1 0 0 0 0 0 0 0]
Comida	2	[0 0 1 0 0 0 0 0 0]
Dedo	3	[0 0 0 1 0 0 0 0 0]
Elefante	4	[0 0 0 0 1 0 0 0 0]
Cosa	5	[0 0 0 0 0 1 0 0 0]
Moto	6	[0 0 0 0 0 0 1 0 0]
Coche	7	[0 0 0 0 0 0 0 1 0]
Árbol	8	[0 0 0 0 0 0 0 0 1]

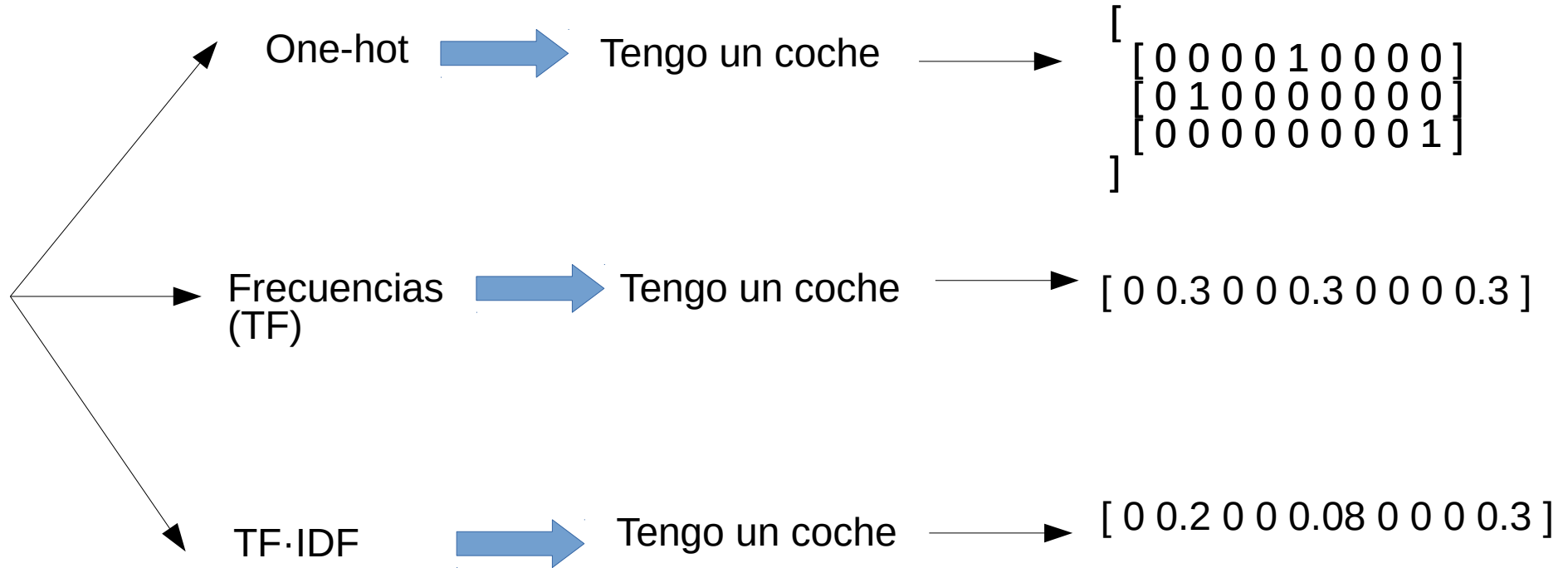
Vectorizar



Vectorizar



Vectorizar textos



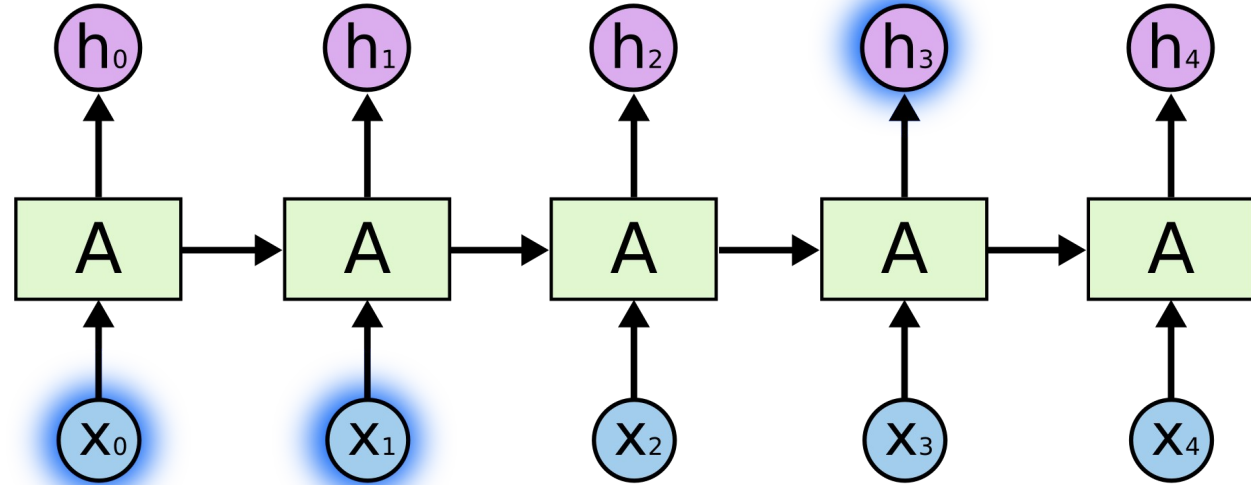
$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Vectorizar

1. Obtener un data set para inferir algo
2. Obtener un texto a ser inferido
3. Limpiar los textos...
4. Mejorar para la clasificación
5. Vectorizar textos
- 6. Entrenar un modelo** (en nuestro caso, clasificador Naïve Bayes)
7. Evaluar la eficacia del modelo
8. ¡Predecir!

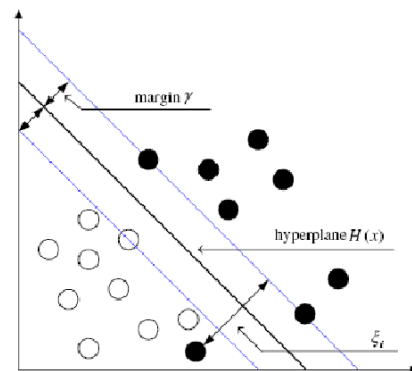
Modelos para clasificar

- Redes neuronales (una o más capas)



Modelos para clasificar

- Redes neuronales (una o más capas, RNN, CNN...)
- Clasificadores analógicos (SVM con o sin Kernel)



Modelos para clasificar

- Redes neuronales (una o más capas, RNN, CNN...)
- Clasificadores analógicos (SVM con o sin Kernel)
- Clasificador bayesiano → Naïve Bayes

Así de fácil

```
from sklearn.naive_bayes import MultinomialNB  
clf = MultinomialNB()  
clf.fit(X, y)
```


Vectorizar

1. Obtener un data set para inferir algo
2. Obtener un texto a ser inferido
3. Limpiar los textos...
4. Mejorar para la clasificación
5. Vectorizar textos
6. Entrenar un modelo (en nuestro caso, clasificador Naïve Bayes)
- 7. Evaluar la eficacia del modelo**
8. ¡Predecir!

Eficacia

¿Y predicho == Y conocido?

Vectorizar

1. Obtener un data set para inferir algo
2. Obtener un texto a ser inferido
3. Limpiar los textos...
4. Mejorar para la clasificación
5. Vectorizar textos
6. Entrenar un modelo (en nuestro caso, clasificador Naïve Bayes)
7. Evaluar la eficacia del modelo
- 8. ¡Predecir!**