

Escuela Técnica superior de Ingeniería  
Universidad de Sevilla

# Curso de Python aplicado

Procesamiento de tweets  
Acceso a APIs públicas

Carlos Perales  
cperales@uloyola.es

5 de diciembre de 2018

## Introducción

- Presentación

- ¿Qué vamos a hacer?

- Herramientas y definiciones

## MongoDB

- Instalación

- pymongo

## Twitter

- API pública

- Escuchar tweets

## tweepy y pymongo

- Guardar tweets

- Recuperar tweets antiguos y análisis

## NLTK

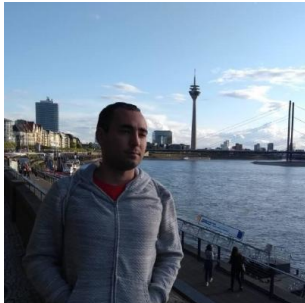
- Cómo usarlo

- NLTK, Twitter y MongoDB

# Introducción

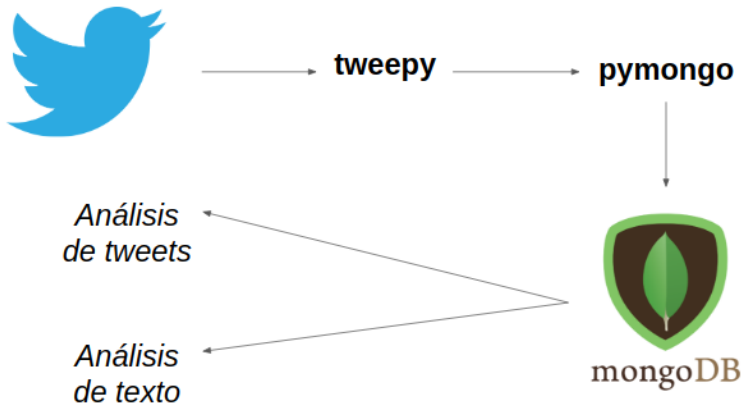
¿Quién soy?

2



Github: @cperales

- ▶ Graduado en Física (UCO)
- ▶ Máster en Ingeniería Matemática (UCM)
- ▶ Doctorando en Ciencia de los Datos (ULA)



# Twitter y MongoDB

¿Por qué estas herramientas?

4



- ▶ Red social de microblogging.
- ▶ 5 millones de usuarios activos.
- ▶ API pública y gratuita.



mongoDB

- ▶ Base de datos sencilla de montar.
- ▶ Fácil conexión con Python usando *pymongo*
- ▶ Teorema CAP.

# ¿Qué es una base de datos?

La organización Linux Information Project la define como

*A database is a set of data that has a regular structure and that is organized in such a way that a computer can easily find the desired information.*

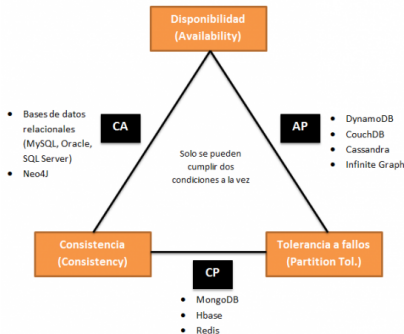
y los datos son

*Data is a collection of distinct pieces of information, particularly information that has been formatted (i.e., organized) in some specific way for use in analysis or making decisions.*

Ejemplos de bases de datos: MySQL, Cassandra, MongoDB

# Teorema CAP

Las bases de datos solo cumplen 2 de las 3 siguientes propiedades.



- **Consistency:** la consulta siempre da la misma información.
- **Availability:** todos los clientes puedan leer y escribir.
- **Partition Tolerance:** el sistema tiene que seguir funcionando aunque existan fallos parciales.

Mongo es una base de datos CP, pues garantiza consistencia y tolerancia a particiones. Para lograr la consistencia y replicar los datos a través de los nodos, sacrifican la disponibilidad.

# ¿Qué es una API?

El *Free On-Line Dictionary of Computing* la define como

*The API (application programming interface) is the interface (calling conventions) by which an application program accesses operating system and other services. An API is defined at source code level and provides a level of abstraction between the application and the kernel (or other privileged utilities) to ensure the portability of the code.*

Por otra parte, el diccionario *TechTerms* la define como

*An API is a set of commands, functions, protocols, and objects that programmers can use to create software or interact with an external system. It provides developers with standard commands for performing common operations so they do not have to write the code from scratch.*



# Uso de una API

Nosotros usaremos una web API para conectarnos con Twitter. Los resultados de las peticiones al servicio web Twitter serán en forma de *JavaScript Object Notation* o JSON.

```
1 {
2   "text": "RT @PostGradProblem: In preparation for the NFL lockout, I will be spending twice
3     as much time analyzing my fantasy baseball team during ...",
4   "truncated": true,
5   "in_reply_to_user_id": null,
6   "in_reply_to_status_id": null,
7   "favorited": false,
8   "source": "<a href='\"http://twitter.com/\"' rel='\"nofollow\"'>Twitter for iPhone</a>",
9   "in_reply_to_screen_name": null,
10  "in_reply_to_status_id_str": null,
11  "id_str": "54691802283900928",
12  "entities": {
13    "user_mentions": [
14      {
15        "indices": [
16          3,
17          19
18        ],
19        "screen_name": "PostGradProblem",
20        "id_str": "271572434",
21        "name": "PostGradProblems",
22        "id": "271572434"
23      }
24    ],
25    "urls": [ ],
26    "hashtags": [ ]
27  },
28  "contributors": null,
29  "retweeted": false,
30  "in_reply_to_user_id_str": null,
31  "place": null,
32  "retweet_count": 4,
33  "created_at": "Sun Apr 03 23:48:36 +0000 2011",
34  "retweeted_status": {
```

*Github* es un servicio en la nube que sirve como repositorio remote para *Git*. *Git* es un sistema de control de versiones open-source, que nos permite tener un control sobre nuestro código y las modificaciones que sobre este hacemos.

```
git clone https://github.com/cperales/  
curso_python_twitter_mongodb
```

The screenshot shows the GitHub interface for the repository 'cperales / Curso-Python-US'. At the top, there are navigation links for Pull requests, Issues, Marketplace, and Explore. Below the repository name, there are tabs for Code, Issues, Pull requests, Projects, Wiki, Insights, and Settings. The 'Code' tab is selected. The repository has 17 commits, 1 branch, 0 releases, and 1 contributor. A table lists the commit history, showing files changed and the time since the last commit.

Commit	Files	Time
cperales Renamed	images	Renamed 2 hours ago
	scripts	Comentarios en español 5 days ago
	.gignore	Write tweets 5 days ago
	beamercolorthemeFeather.sty	First commit 6 days ago
	beamerinnerthemeFeather.sty	First commit 6 days ago
	beamerouterthemeFeather.sty	Beamer theme modified 6 days ago
	beamerthemeFeather.sty	First commit 6 days ago
	curso_mongo_twitter.tex	Renamed 2 hours ago

Entrar en el *MongoDB Download Center*, seleccionar *Community Server* y nuestro sistema operativo

<https://www.mongodb.com/download-center/community>

## Ubuntu

```
sudo service mongod start
```

## Windows

```
C:\Program Files\MongoDB\Server\4.0  
\bin\mongo.exe
```

## MacOS

```
brew update  
brew install mongodb  
mongod
```

Existe un gestor virtual de MongoDB, muy fácil de usar

<https://robomongo.org/download>

Podemos instalar esta librería con *pip*. Es recomendable hacerlo en un entorno virtual.

```
virtualenv -p python3.6 env  
source env/bin/activate  
  
pip install pymongo
```

Tras haberse registrado en Twitter, accedemos a la plataforma de desarrolladores

`https://developer.twitter.com/`

Seguir el procedimiento para abrir una cuenta como investigador e indicar que el uso será únicamente para investigación. Crear una app, y guardar los siguientes campos en *config.ini*

```
[API]
consumer_key = ...
consumer_secret = ...
access_token = ...
access_token_secret = ...
```

Instalamos la librería *tweepy*.

```
pip install tweepy
```

Abrimos y probamos el código *listen\_tweets.py*. Este código

1. Establece una. conexión con tus credenciales con la API de Twitter
2. Te permite realizar una serie de acciones desde tu cuenta, como escribir mensajes, hacer retweets o escuchar tweets.
3. De momento imprimiremos por pantalla los tweets que recibimos.

Vamos a escribir tweets con *write\_tweet.py*.

Una vez visto *tweepy* y *pymongo* por separado, podemos ver cómo encadenar ambas librerías.

Abrimos *tweets\_to\_mongodb.py* y exploramos el código.

Tweepy es una librería que accede a la API oficial, y por ello, tiene algunas limitaciones. Solo podemos acceder a tweets publicados hace no más de 7 días.

```
python look_for_tweets.py
```

Abrir *analysis\_mongo.py*. Comentaremos qué análisis podemos hacer en MongoDB desde *pymongo*.

Como tareas, analizar las siguientes cosas de los tweets:

1. Contar cuántos tweets tenemos en la base de datos.
2. Determinar cuántos tweets son retweets y cuántos son originales.
3. Analizar los hashtags más utilizados.



# ¿Qué es NLTK?

Es una librería open source de procesamiento de lenguaje natural, que incluye:

- ▶ Análisis léxico (tokenizadores).
- ▶ Modelos de árboles y n-gramas a partir de textos.
- ▶ Machine learning.

```
pip install nltk
```

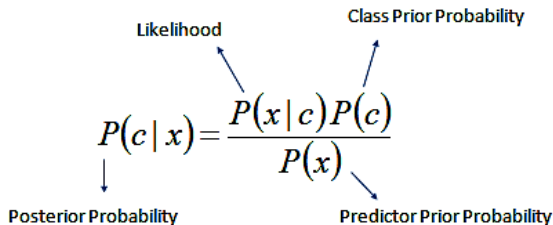
Tareas:

1. Lematizar (*stem*) el texto y contar las palabras más repetidas.

# Clasificador bayesiano ingenuo

17

Usaremos un clasificador bayesiano ingenuo



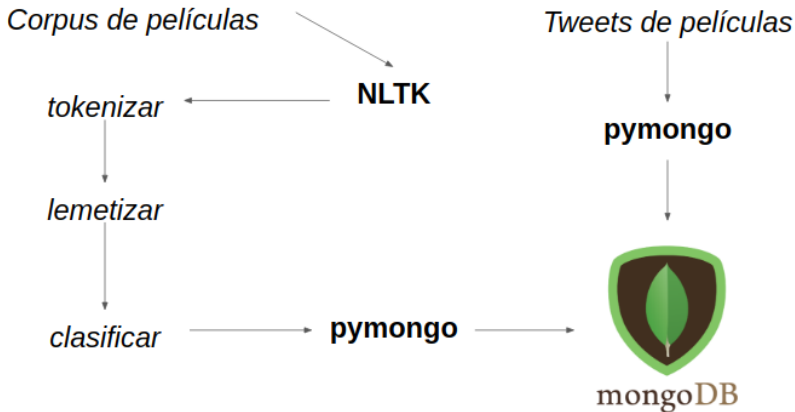
The diagram shows the Naive Bayes formula with arrows pointing from descriptive labels to the corresponding parts of the equation:

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

- Likelihood** points to  $P(x | c)$
- Class Prior Probability** points to  $P(c)$
- Posterior Probability** points to  $P(c | x)$
- Predictor Prior Probability** points to  $P(x)$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c)$$

Muy usado para análisis de texto pese a su simplicidad.



A stylized graphic of a wave or a series of flowing lines in shades of blue and white, curving around the text. The lines are smooth and have a slight gradient, giving it a sense of motion. There are some small white dots scattered along the curves, possibly representing water droplets or light reflections.

¡Gracias!