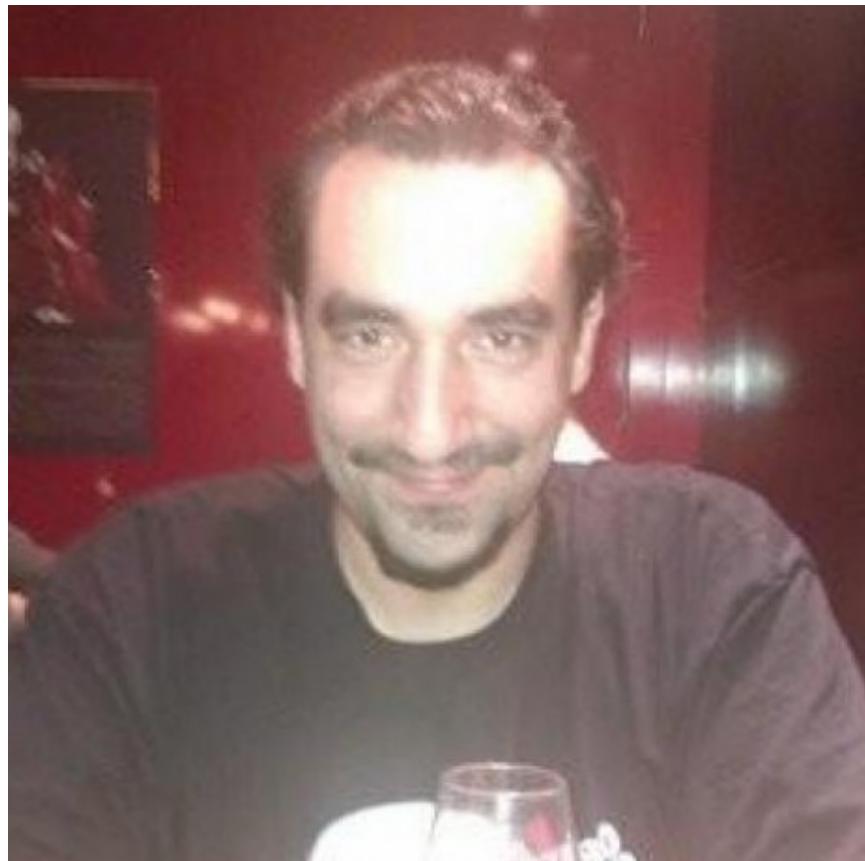


# Web Scrapando

LAS TECNOLOGÍAS MÁS USADAS EN LA WEB



# Hola!

**Soy David Vaquero  
Alias pepesan**

Desarrollador web desde el siglo pasado  
y formador del siglo XXI

Puedes encontrarme en <https://cursosdedesarrollo.com/>  
Podcasteando en [República web](#) y  
[Formadores en tiempos revueltos](#)  
<https://www.linkedin.com/in/davidvaquero/>

# OBJETIVOS

- ✓ Conocer de primera mano las tecnologías usadas en la web
- ✓ Aprender las técnicas de Web Scraping
- ✓ Compartir de datos generados
- ✓ Ayudar a generar mejores proyectos formativos

**De los dominios a la gráficas de resultados**

# Contenidos

- ✓ Introducción
- ✓ Obtención de Dominios
- ✓ Almacenamiento de información
- ✓ Análisis de Páginas Web
- ✓ Paralelizando que es gerundio
- ✓ Graficar resultados
- ✓ Resultados para Alexa
- ✓ Resultados Red.es
- ✓ Preguntas?
- ✓ Referencias

# 1

# INTRODUCCIÓN

**“Wordpress es usado en el 35% de la Web”**



# FUNDAMENTOS Web Scraping

**Elementos principales de la técnica y fases  
del proyecto**



“Para qué creerte  
los datos que te  
dan otros, si tú  
mismo puedes  
comprobarlos”

**Web scrapeando las tecnologías más usadas de la Web**



# Fases del Web Scraping



Datos



Almacenamiento

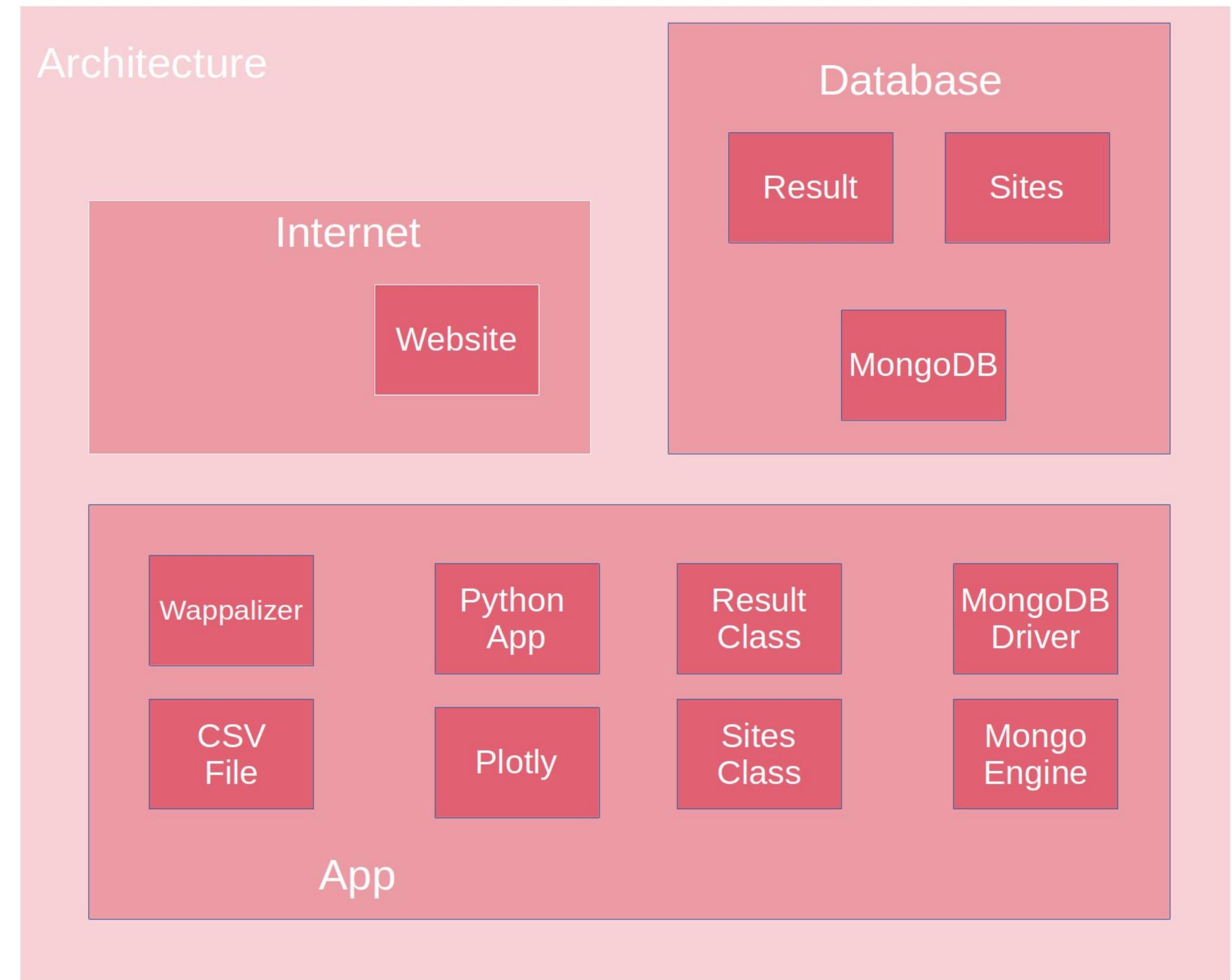


Análisis



Graficado

# Arquitectura





# Motivaciones

**Muchas veces ves los estudios sobre el uso de tecnologías y no sabes de donde vienen esos datos ni cómo se ha obtenido**



# República Web

**En uno de los programas de república web,  
se debatió el uso de WordPress en internet  
y salieron varios estudios, pero no  
quedaban claras las fuentes, así que  
tocaba investigar  
<https://republicaweb.es/>**

2

# Obtención de dominios

“Mierda, el nombre está cogido”



# Obtención de dominios

Una de las tareas principales es la de obtener los dominios registrados a analizar para ello se han empleado principalmente dos recursos:  
[Amazon Alexa Top Sites](#) y [Red.es RISP](#)

The screenshot shows the Alexa homepage with the title "The top 500 sites on the web". It includes navigation links for "SOLUTIONS", "TOOLS", "PRICING", "TAKE TOUR", "GUIDES", "BLOG", and "LOG IN". A prominent orange button says "START YOUR FREE TRIAL". Below the main title, there are filters for "Global", "By Country", and "By Category". A message encourages users to "Want access to the complete list?" and provides a "START YOUR FREE TRIAL" button. The main content area displays a table of the top 2 sites:

	Site	Daily Time on Si...	Daily Pageview...	% of Traffic Fro...	Total Sites Linki...
1	Google.com	14:57	16.46	0.40%	1,312,830
2	Youtube.com	16:05	8.65	15.10%	1,001,614



**El primer recurso ha sido el millón de páginas más visitadas según Alexa Top Sites**

**<http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>**

The screenshot shows the official website for .es domains. At the top, there's a banner with the Spanish flag, the text 'GOBIERNO DE ESPAÑA' and 'MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL', and the 'red.es' logo. Below the banner, the main navigation menu includes links for 'INICIO', 'BUSCA Y REGISTRA TU DOMINIO', 'AGENTES REGISTRADORES', 'TODO LO QUE NECESITAS SABER', 'GESTIONA TU DOMINIO', 'ACTUALIDAD Y NOTICIAS' (which is currently selected), and 'SOBRE NOSOTROS'. A search bar and social media links are also present. The main content area features a news item titled 'Ya está disponible el servicio de acceso al listado de los dominios ".es"' dated 30/06/2014. The news item discusses the RISP (Reutilización de Información Pública) service, which allows users to download a list of active '.es' domains. It provides instructions on how to request this service via the 'sede.red.gob.es' website. The footer of the page includes logos for LACLD, Centr, SEPE, and Confidencialidad.

# Red.es RISP

**Es necesario hacer una solicitud a Red.es  
indicando el uso que se dará a estos datos**

<https://www.dominios.es/dominios/es/actualidad-y-noticias/comunicados/ya-est%C3%A1-disponible-el-servicio-de-acceso-al-listado-de-los>

# 3

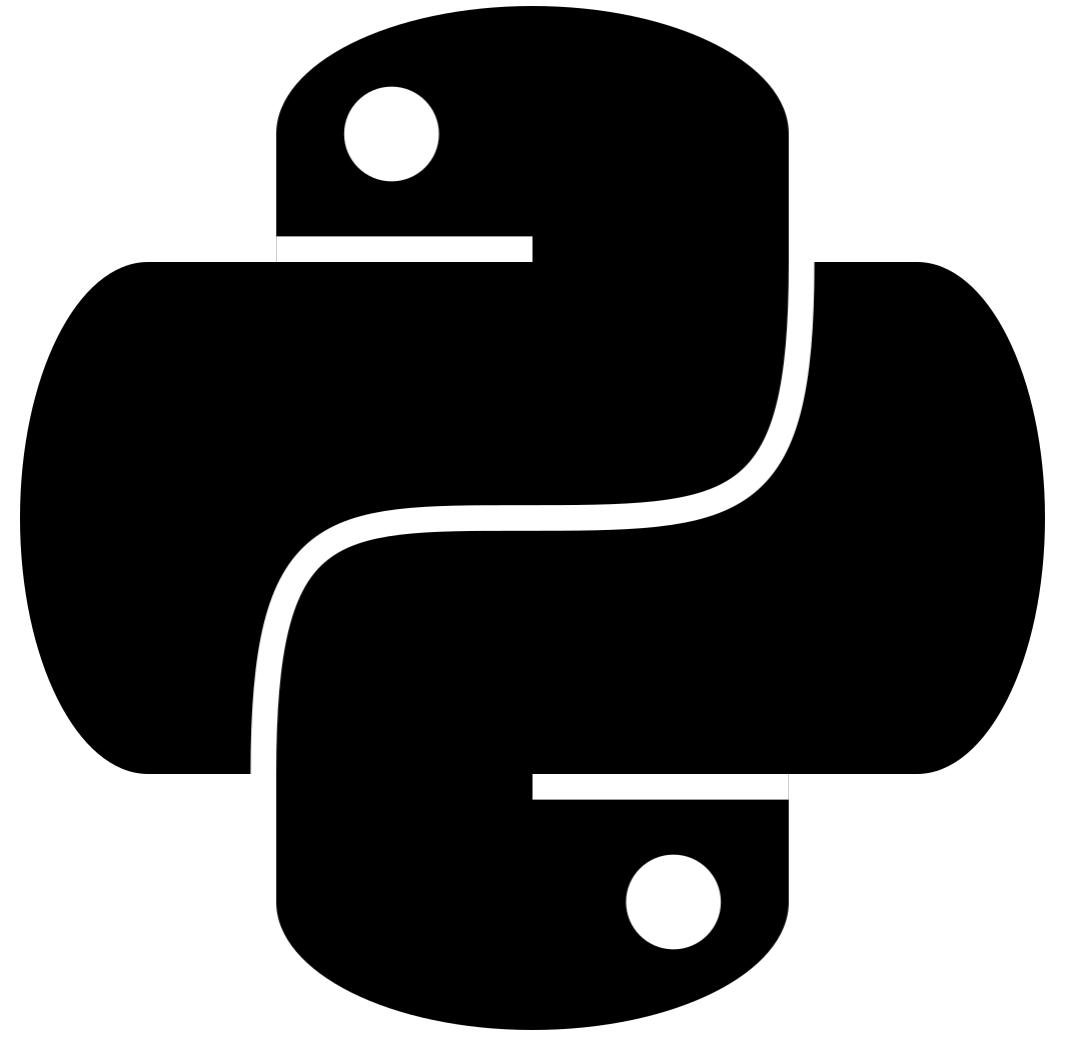
# Almacenamiento de información

¿Tablas? Ja! ¿Quién las necesita?



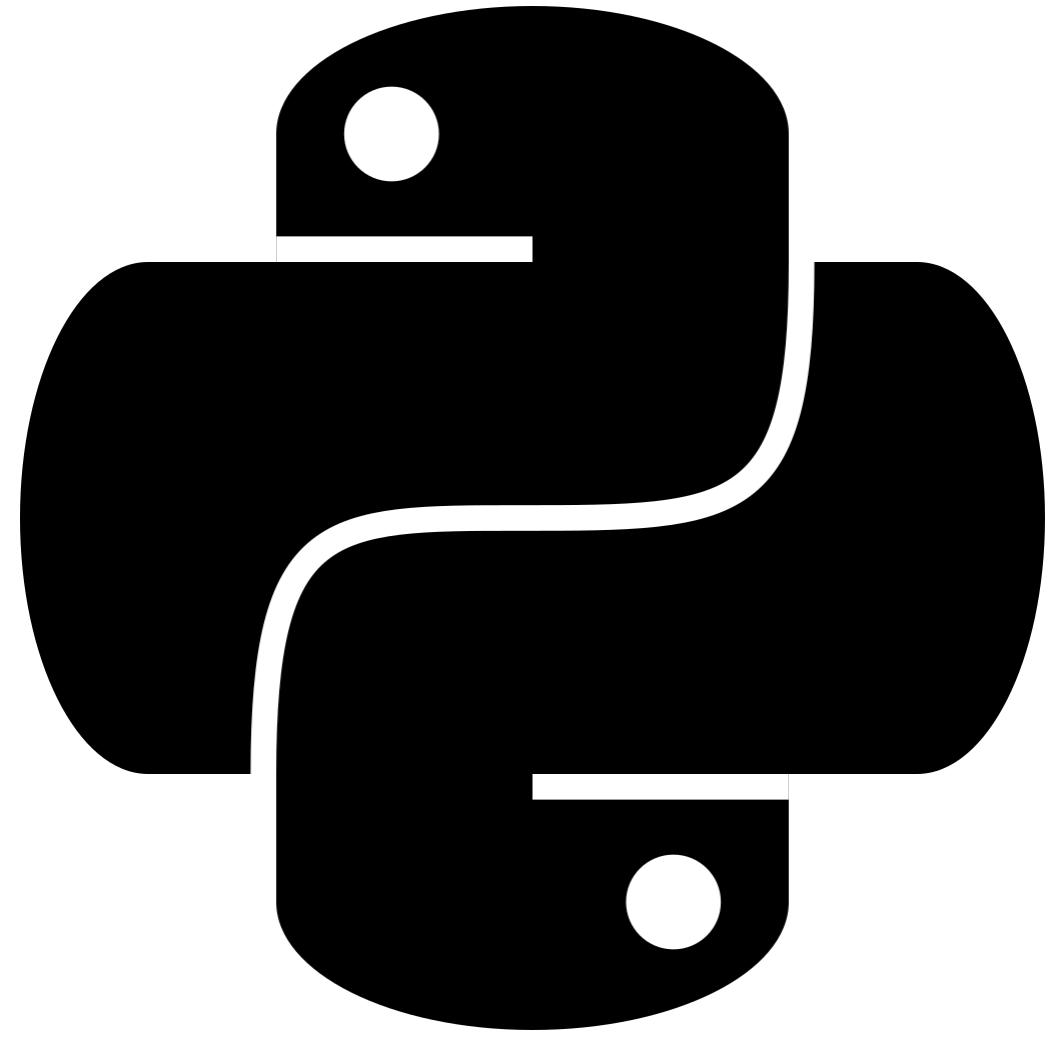
# MongoDB

**MongoDB sirve por la flexibilidad y rapidez a la hora de manejar millones de objetos**



# Python

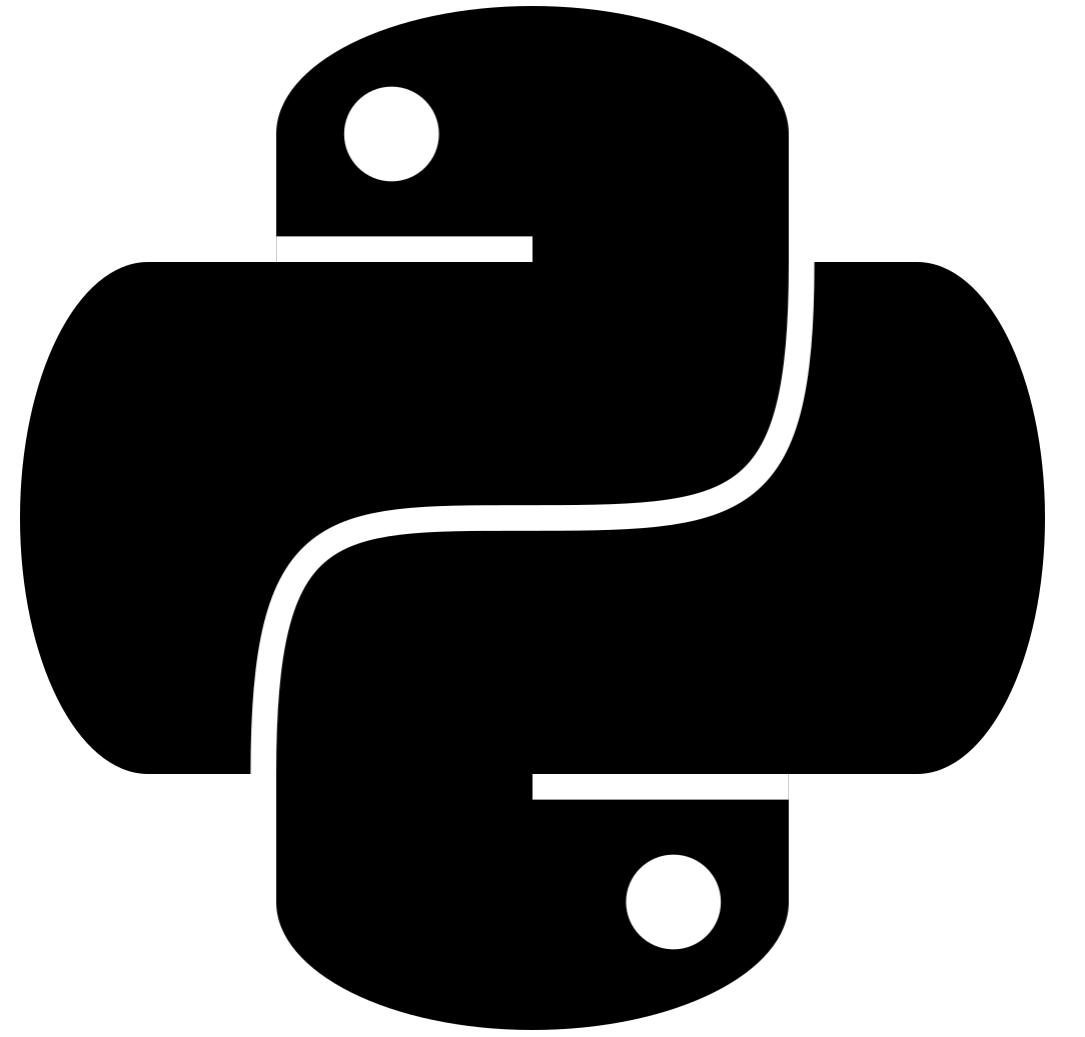
**Python es el lenguaje principal del tratamiento masivo de datos, permite todos los niveles de procesado de información y graficado de datos**



# MongoEngine

**Es una especie de motor ORM para MongoDB o  
cómo lo llaman ellos, Document-Object Mapper  
(DOM)**

**<http://mongoengine.org/>**



**Para manejar el fichero de dominios  
importaremos la biblioteca cvs**

4

# Análisis de dominios

¿Y tú de quien eres?

# Analizando las webs

**Con todos los dominios cargados deberemos visitarlos para analizarlos, se coge sólo la página principal, son millones de dominios y no podemos profundizar mucho de momento**



# Wappalyzer

**Es una extensión de Chrome y Firefox que analiza una web y permite extraer las tecnologías que utiliza**  
**<https://www.wappalyzer.com/>**



# Wappalyzer Python

Pero para usarlo desde Python 3 deberemos  
usar una biblioteca que se basa en  
request.urlib3 que hemos toqueteado  
<https://github.com/vincd/wappylyzer>



# Fichero de Expresiones

Para analizar las páginas se usa un fichero de definiciones de expresiones regulares que busca esos patrones, en este caso de la página principal del dominio <https://github.com/AliasIO/wappalyzer/blob/master/src/technologies.json>

5

# Paralelizando

**¿Qué Rei Ayanami es ésta? ¿La primera o la segunda?**

# Lanzando las peticiones

**Ya tenemos todo para lanzar las peticiones  
pero nos encontramos con el problema de lo  
que tarda en obtener resultados, hay que  
parallelizar**



# Parallelizando python

Utilizamos el módulo Process de  
multiprocessing, conseguimos mejorar el  
rendimiento en un más de 10 veces  
<https://docs.python.org/3/library/multiprocessing.html>



# Almacenando los resultados

Para almacenar los resultados usamos una colección de mongoDB, donde indicamos campos para saber la fecha, el batch (pretendemos hacerlo muchas veces), el número de intentos (el posible que el sitio esté caido), si ya lo hemos analizado y el array de tecnologías



# ¿Cuanto tarda?

Después de los procesos de paralelización ejecutado desde un portátil un par de días, así que ármate de paciencia

# 6

# Graficado de resultados

“No necesitas otra cosas que matplotlib”



# Estructura de resultados

**Los resultados tiene una estructura donde se guarda para cada sitio analizado un listado de tecnologías y en el caso se alexa su posición en el top**



# Clasificando

Desde mongoengine podemos consultar usando toda la potencia de mongodb así que filtramos según tecnología, o grupos de tecnologías, en aquellos que superan el 0.1% de uso, así filtrando en porciones del top: 100, 1000, etc...



# Graficando

**La famosa matplotlib, nos ayuda a hacernos una idea de cómo está la cosa pero con plotly lo hacemos de una manera mucho más sencilla**  
<https://github.com/plotly/plotly.py>

7

# Resultados para Alexa

“El software libre ha ganado. ¿Y ahora qué?”

# Resultados para Alexa Top 1M

## Tecnologías más usadas

- Jquery 56%
- PHP 36%
- Google Font API 30%
- Nginx 29%
- Apache HTTPd 24%
- Cloudflare 22%
- Google Tag Manager 21%
- Mysql 21%
- Wordpress 20%
- FontAwesome 19%

8

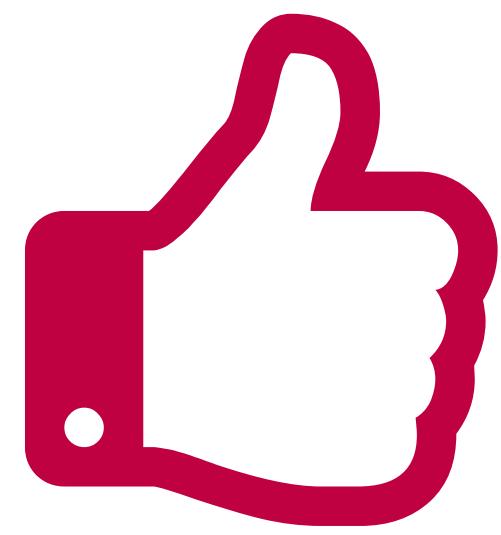
# Resultados para Red.es

**“Españita, ese país del que usted me habla”**

# Resultados para Red.es 2M

## Tecnologías más usadas

- Apache httpd: 43%
- Jquery 29%
- PHP 23%
- Google Font API 21%
- Mysql 17%
- Nginx 17%
- Wordpress 15%
- Jquery migrate 13%
- FontAwesome 10%
- Bootstrap 7%



# Gracias!

Preguntas?

Puedes encontrarme en:

Twitter @dvaquero

<https://github.com/pepesan>

pepesan(at)gmail.com

# Referencias

Donde buscar más información:

- Alexa Top Sites :<http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>
- Fotos usadas de Pexels: <https://www.pexels.com/>
- República Web: <https://republicaweb.es/>
- Foentire: <https://twitter.com/foentire?lang=es>
- Cursos de Desarrollo: <https://cursosdedesarrollo.com/>
- Repositorio de código:  
<https://github.com/pepesan/web-technology-stats>

# Licencia

El documento está liberado bajo la licencia CC-BY-SA-NC 4.0

[https://creativecommons.org/licenses/by-nc-sa/4.0/  
deed.es\\_ES](https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es_ES)

El autor es David Vaquero Santiago  
Con correo pepesan@gmail.com y cedido a  
<https://cursosdedesarrollo.com/>