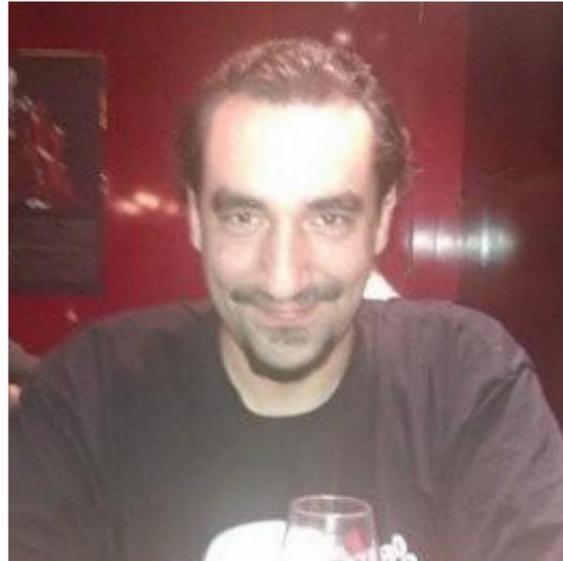


Web Scraping

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

“

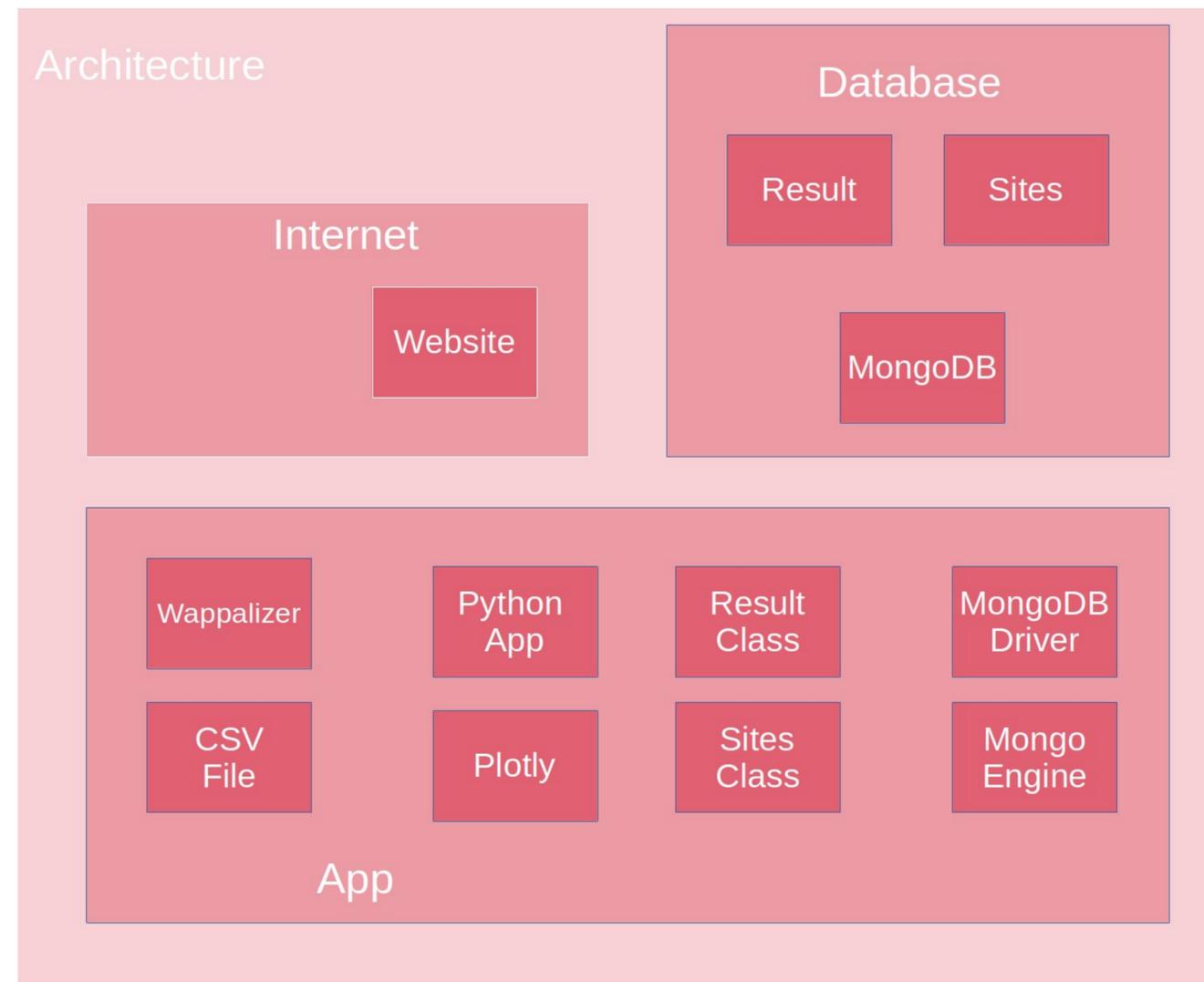
“Para qué creerte los datos que te dan otros, si tú mismo puedes comprobarlos”



Fases del Web Scraping



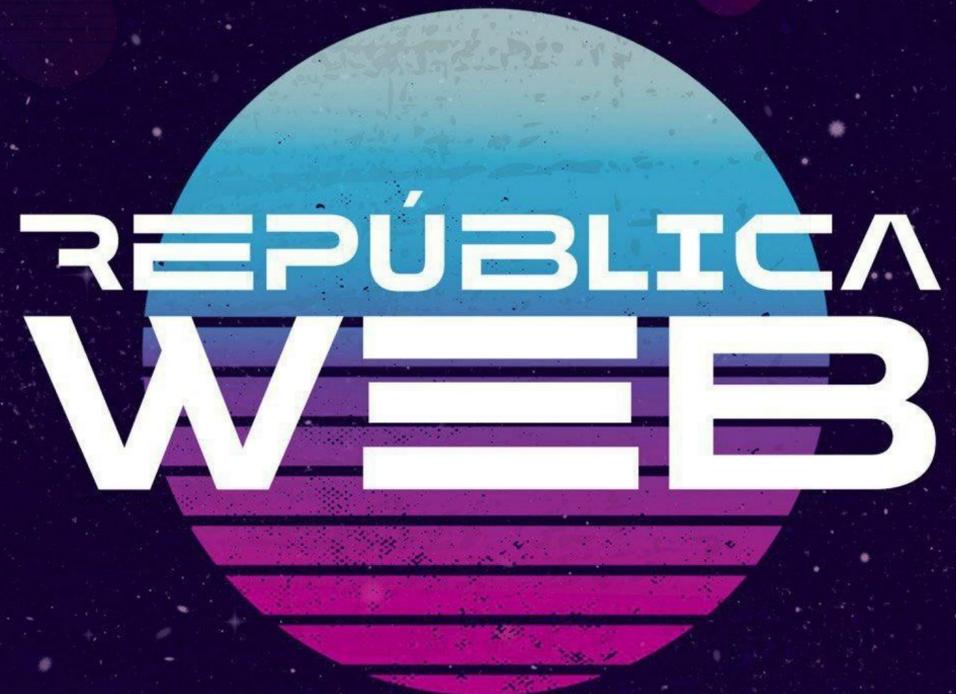
Arquitectura





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**

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 empleando principalmente dos recursos: Amazon Alexa Top Sites y Red.es RISP

The top 500 sites on the web

[Global](#) [By Country](#) [By Category](#)

Showing 50 of 500 results

Want access to the complete list?

[START YOUR FREE TRIAL](#)

	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

Alexa Top Sites

1M

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>

Castellano | Català | Euskera | Galego | English

dominios.es @administración electrónica

GOBIERNO DE ESPAÑA | MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL | SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

ONTSI | RedIRIS | AGENDA DIGITAL | Red.es

Buscar

INICIO | BUSCA Y REGISTRA TU DOMINIO | AGENTES REGISTRADORES | TODO LO QUE NECESITAS SABER | GESTIONA TU DOMINIO | ACTUALIDAD Y NOTICIAS | SOBRE NOSOTROS

Estás en: Inicio » Actualidad y noticias »

Ya está disponible el servicio de acceso al listado de los dominios ".es"

- Actualidad
- Notas de prensa
- Videos

red.es

30/06/2014

En el marco de la Ley 37/2007, de 16 de noviembre, sobre reutilización de información del sector público, ya se encuentra disponible el servicio de acceso al listado de dominios ".es". Este servicio posibilita realizar una descarga de un fichero de texto con el listado de los nombres de dominios activos ".es". Además, para aquellos casos en los que el titular sea persona jurídica, se incluyen el nombre del titular y su NIF. La actualización de este fichero se hará con carácter mensual. Pueden encontrarse más detalles en el siguiente enlace: [/es/todo-lo-que-necesitas-saber/normativa/risp](https://www.dominios.es/actualidad-y-noticias/comunicados/ya-est%C3%A1-disponible-el-servicio-de-acceso-al-listado-de-los)

Para realizar una solicitud, hay que acceder a la dirección <https://sede.red.gob.es> en el apartado específico de reutilización de información ("RISP"). Dicha solicitud debe ir firmada electrónicamente, mediante certificado reconocido.

dominios.es ANIVERSARIO

@administración electrónica | LACTUD | centr | www.sepe.es Trabajamos para ti

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

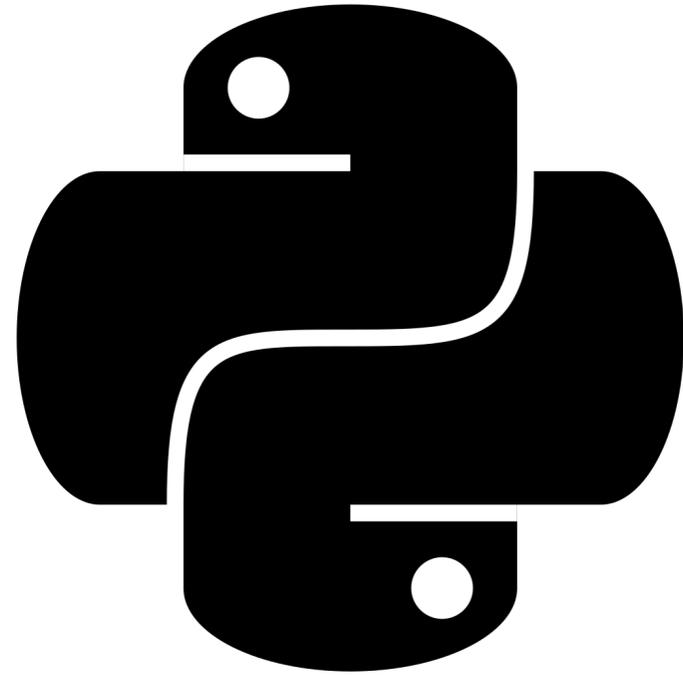
Almacenamiento de información

¿Tablas? Ja! ¿Quién las necesita?



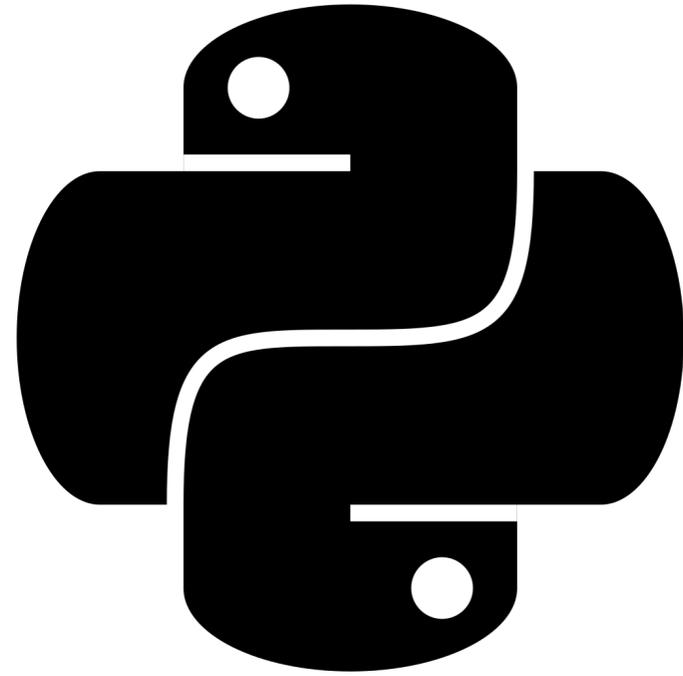
MongoDB

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



Python

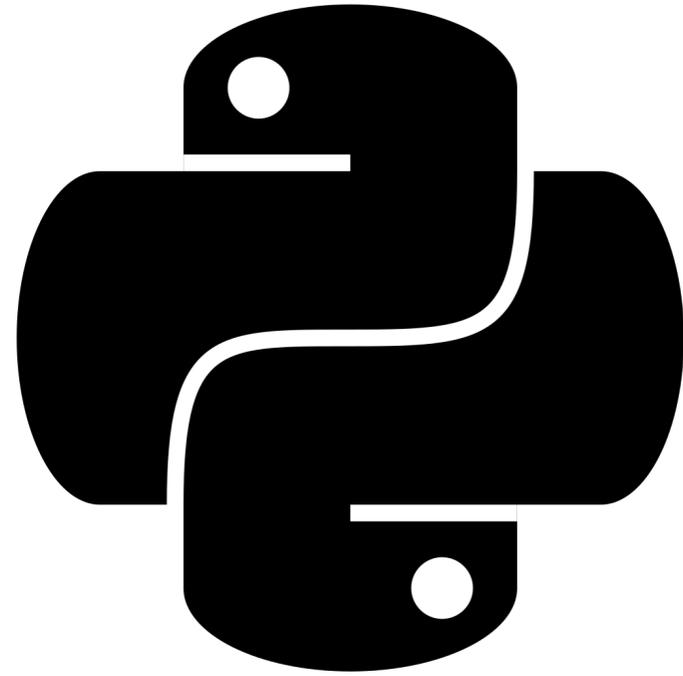
Python es el lenguaje principal del tratamiento masivo de datos, permite todos los niveles de procesado de información y graficado de datos



MongoEngine

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

<http://mongoengine.org/>



CSV

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

4

Análisis de dominios

¿Y tú de quien eres?



Wappalizer

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



Wappalizer Python

**Pero para usarlo desde Python 3 deberemos
usar una biblioteca que se basa en
request.urllib3 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/AliaSI0/wappalyzer/blob/master/src/technologies.json>

5

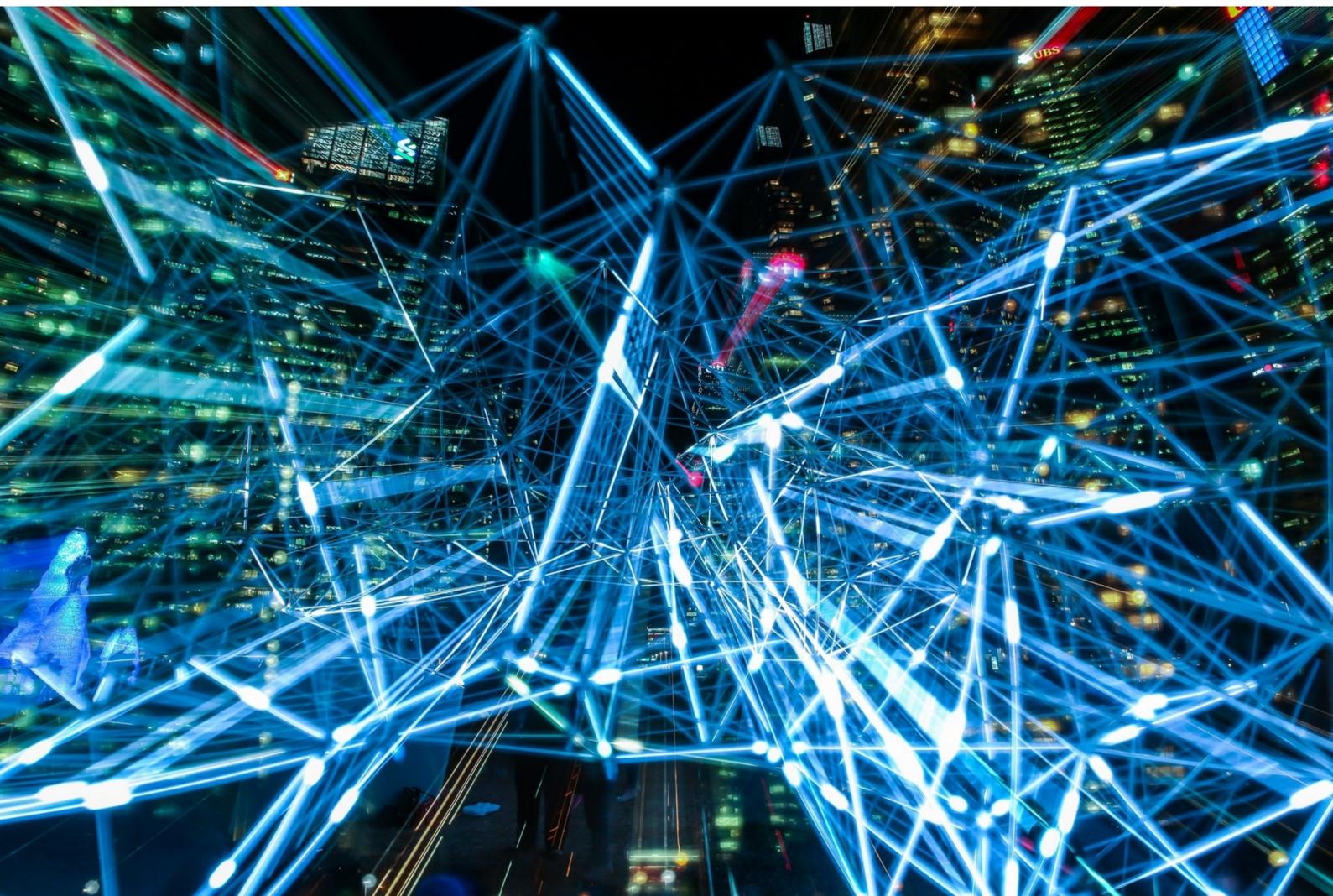
Paralelizando

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



Paralelizando 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 caído), si ya lo hemos analizado y el array de tecnologías



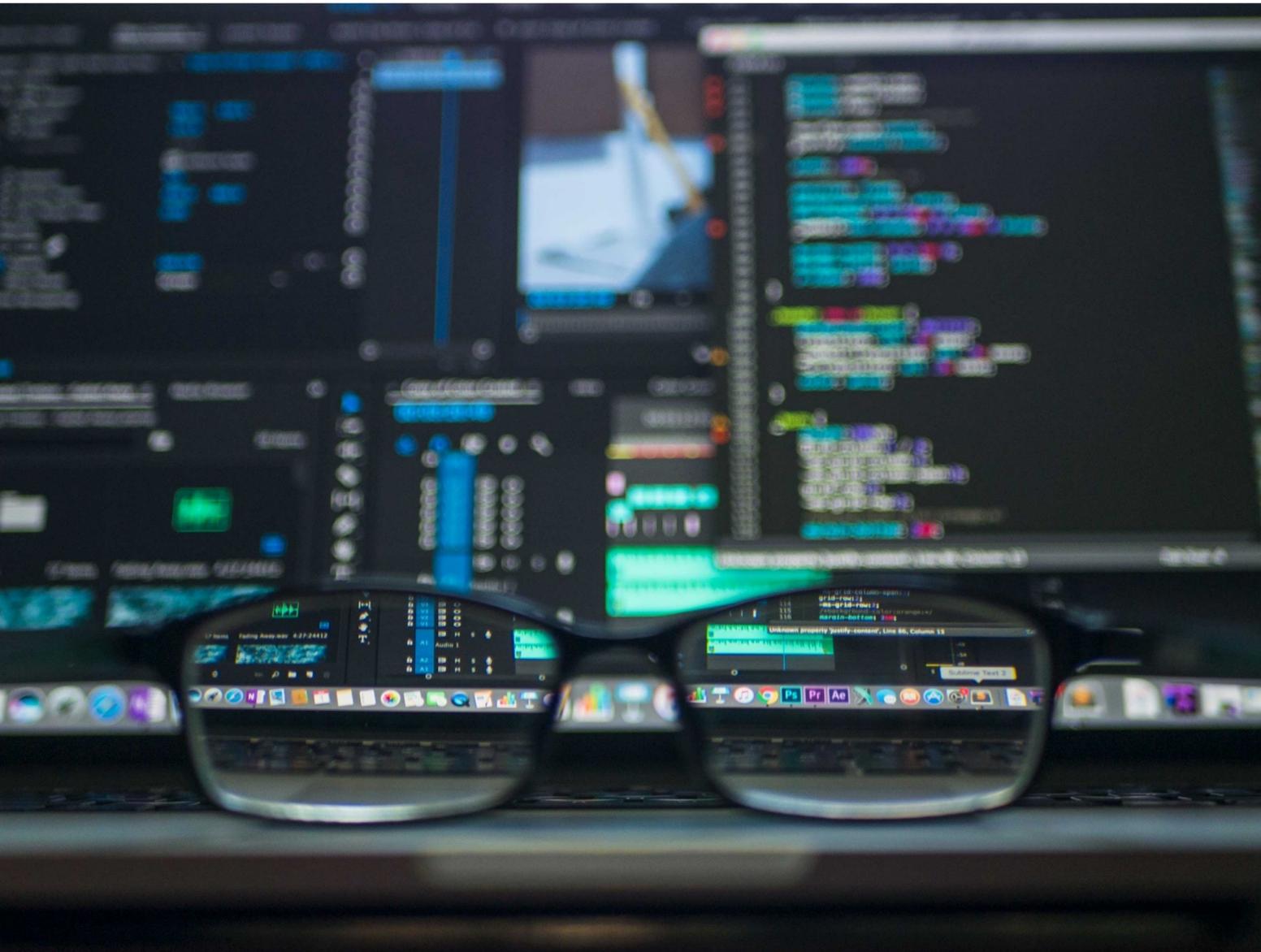
¿Cuanto tarda?

**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

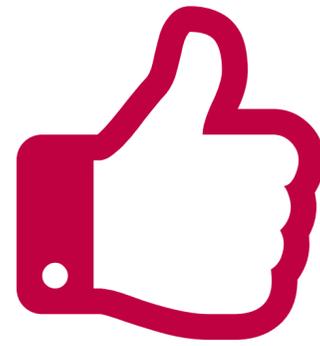
Resultados para Red.es

“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](mailto: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

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