Introducción a la ciencia de datos

Propuesta de scraping V1

Autor: Erick Humberto Cordova Gavilanes

**Objetivo:**

Extraer noticias de la web usando la librería scrapy de Python para la automatización de la descarga y el almacenamiento de datos referentes a las noticias, para así crear un dataset ordenado y organizado que nos permite analizarlo posteriormente.

**Restricciones:**

Las noticias deben ser referentes a la muerte de George Floyd y las consecuencias de este.

Las noticias deben ser extraídas por fuentes digitales confiables y veraces.

Las noticias deben pertenecer a fuentes pertenecientes a USA.

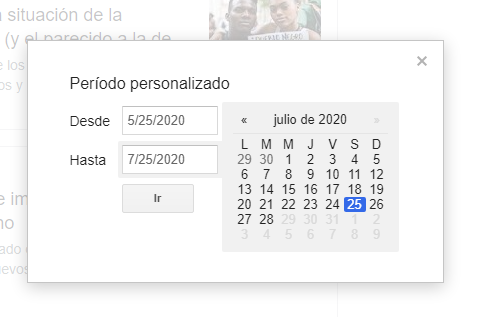
Las noticias deben ser extraídas de forma automatizada usando un framework como scrapy

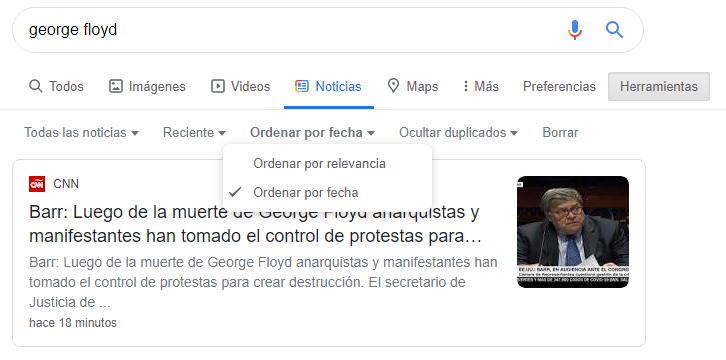
**Análisis de recursos**

Para obtener las noticias, usamos Google New el cual nos permite buscar noticias sobre un tema específico, filtrarla por un rango de fechas y ordenarlos por fechas.

Las noticias obtenidas en Google News nos permiten satisfacer las restricciones.



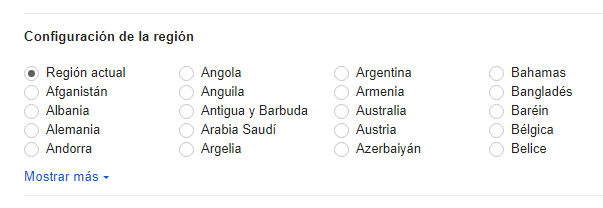




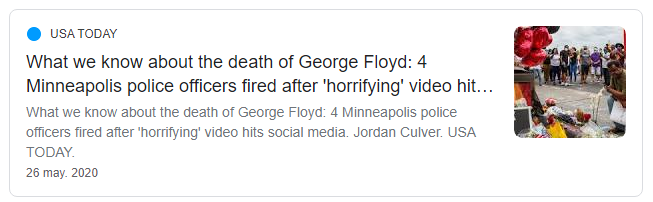
También permite configurar el número de noticias presentadas al buscar.



Se puede configurar la región de la cual se obtienen las noticias.



**Estructura de la noticia**

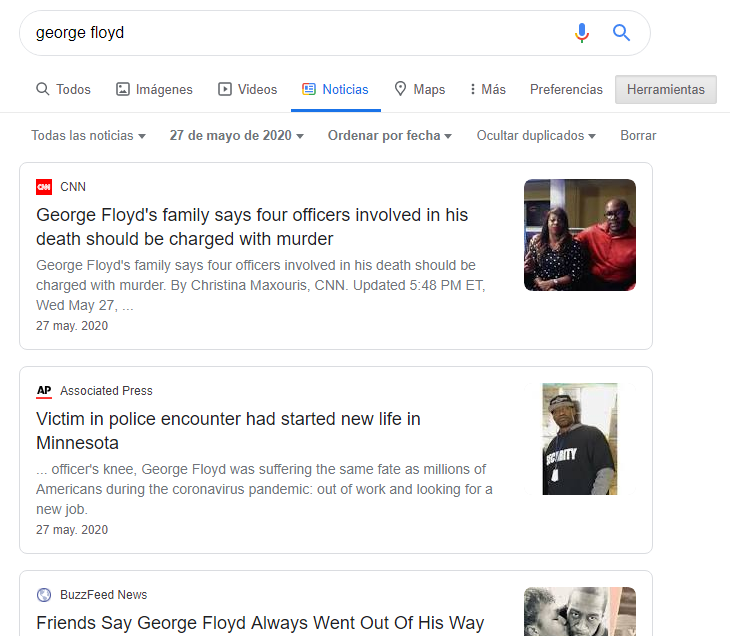


Vemos que la noticia se compone de:

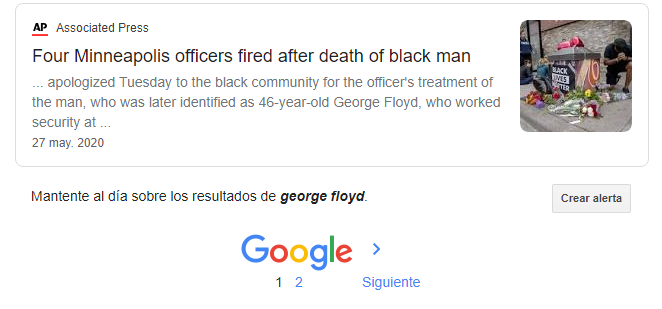
1. El nombre del medio digital que publica la noticia
2. El título de la noticia
3. El resumen de la noticia
4. La fecha en la que fue publicada la noticia

**Estructura de la página web**

La página se compone del buscador, las opciones de configuración y el contenedor de noticias.



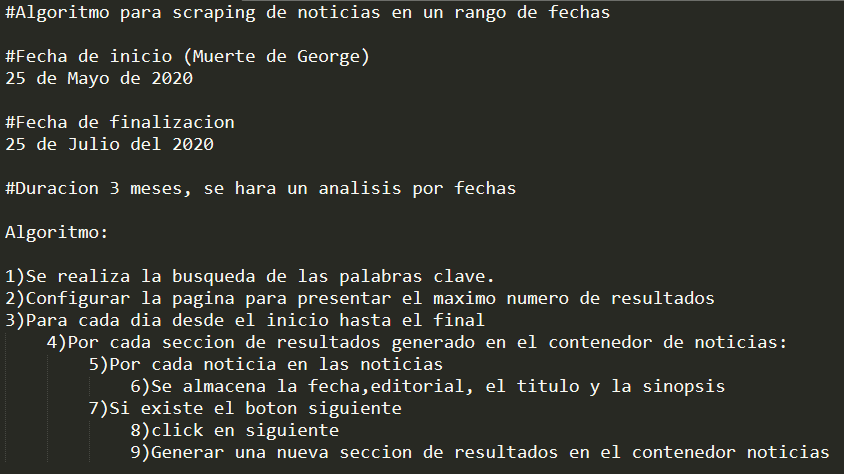
En caso de que los resultados de las noticias superen el número máximo, se presenta un ~~seleccionador de página,~~ lo que permite cargar los siguientes resultados dinámicamente dentro del mismo contenedor.



**Herramientas para utilizar para realizar scraping**

* Python 3
* Driver Selenium
* Librería Scrapy

Algoritmo para realizar scraping sobre la página de Google News.



Columnas del dataset como resultado

|  |  |  |  |
| --- | --- | --- | --- |
| fecha | fuente | titulo | resumen |

Aproximación de filas

60 días x 200 noticias en promedio al día = 12000 filas.