

**MÁSTER UNIVERSITARIO EN**

**INGENIERÍA INFORMÁTICA**

**RECUPERACIÓN DE LA INFORMACIÓN**

**Y WEB SEMÁNTICA**

**PROYECTO DE WEB CRAWLING**

**Adriana Hernández Regueiro**

**Sofía Alejandra López Fernández**

**Estela Surís Graña**

Contenido

[Introducción 3](#_Toc24715024)

[Desarrollo del proyecto 3](#_Toc24715025)

[Definición del dominio 3](#_Toc24715026)

[Obtención de los resultados 3](#_Toc24715027)

[Postprocesado 4](#_Toc24715028)

[Tecnologías utilizadas 4](#_Toc24715029)

[Conclusión 4](#_Toc24715030)

# Introducción

Como parte de la asignatura Recuperación de la Información se ha realizado el proyecto correspondiente a la parte de recuperación de información. En el siguiente documento se expondrán los pasos por los que se han pasado para la elaboración del crawler web y la interfaz para su visualización.

# Desarrollo del proyecto

## Definición del dominio

Durante la fase de definición del dominio que se iba a crawlear se sopesó utilizar webs de noticias o crear un buscador de partituras. Por la imposibilidad de realizar el crawling de las páginas con información de partituras, se decidió buscar un dominio de noticias que permitiera obtener la información de la página a través de un crawler.

Finalmente, se optó por la subdivisión “Cinco Días” de la versión digital del periódico El País, este dominio recoge noticias importantes sobre economía, mercados, pymes y emprendedores, empresas, finanzas y tecnología.

## Obtención de los resultados

REGEX-URLFILTER.TXT

# accept URLS ended in \*number\*.html in cincodias.elpais.com

-^https://cincodias.elpais.com/agr/\w\*

-^https://cincodias.elpais.com/autor/\w\*

-^https://cincodias.elpais.com/especiales/\w\*

-^https://cincodias.elpais.com/herramientas/

-^https://cincodias.elpais.com/mercados/\w\*

-^https://cincodias.elpais.com/seccion/\w\*

-^https://cincodias.elpais.com/smartlife/\w\*

-^https://cincodias.elpais.com/smartlife

-^https://cincodias.elpais.com/estaticos/\w\*

-^https://cincodias.elpais.com/tag/\w\*

-^https://cincodias.elpais.com/rss/\w\*

-^https://cincodias.elpais.com/territorio\_pyme

-^https://cincodias.elpais.com/territorio\_pyme/\w\*

+^https://cincodias.elpais.com/

+^https://cincodias.elpais.com/cincodias/2019/[0-9]{2}/[0-9]{2}/\w\*/[0-9]\*\_[0-9]\*.html

# Postprocesado

# Visualización de resultados

# Tecnologías utilizadas

# Conclusión