



P.11 - Borme


Pon en práctica todo lo aprendido sobre `requests`, `bs4`, `selenium` y `chromedriver` para descargar datos de empresas recién constituidas del Registro Mercantil.

Obtener datos públicos

- Vamos a crear un script que entre en la web del Registro Mercantil:

BOE.es - BORME de Noviembre de 2023


BORME de Noviembre de 2023

 https://www.boe.es/diario_borme/

- Monta un script al que se le puedan pasar en la ejecución la fecha del Boletín que se quiere descargar tal que así "python3 spyder.py YYYYMMDD" (donde YYYY es el año - la Y es de Year - en formato de 4 dígitos, MM es el mes - la M es de Month - en formato 2 dígitos y DD es el día - la D es de Day - en formato 2 dígitos. Por ejemplo, 20231127).
- En dicho script, parsea la fecha introducida y úsala para configurar una lógica que permita cargar el Boletín del día del calendario que se le ha pasado al script como argumentos al ejecutar. Por ejemplo, el Boletín de hoy:

BOE.es - Sumario del día 27/11/2023

Sumario del día 27/11/2023

 <https://www.boe.es/borme/dias/2023/11/27/>



¡Ojo! No todos los días se publica un boletín, protege tu código (con un try/except, por ejemplo) ante potenciales errores por ese lado.

- Desarrolla un código que permita descargar todos PDFs de “Actos inscritos” de la “Sección primera. Empresarios”. Hay uno por cada provincia pero no todos los días están presentes todas las provincias. Puedes hacerlo directamente o a través del Dropdown de la web:

Secciones



Todas

SECCIÓN PRIMERA. Empresarios

Actos inscritos

Otros actos publicados en el Registro Mercantil

SECCIÓN SEGUNDA. Anuncios y avisos legales

CONVOCATORIAS DE JUNTAS

FUSIONES Y ABSORCIONES DE EMPRESAS

AUMENTO DE CAPITAL

REDUCCIÓN DE CAPITAL

- Guarda todos los archivos en local organizados en carpetas por fecha. Puedes usar un código como el que os dejamos abajo para crear tu sistema de carpetas. Los

PDFs puedes guardarlos con el nombre que traigan por defecto.

```
from os import makedirs
from os.path import exists
if not exists(directory):
    makedirs(directory)
```



Si estás trabajando con `selenium` para descargar el archivo te bastará con hacer click en el link de cada uno. Si estás trabajando con `requests` tendrás que usar un código similar al que vimos en su día (solo que en este caso el contenido a imprimir es binario, de ahí el "wb"):

```
with open("tu/path/donde/quieras/guardar/el/archivo.pdf", "wb") as
output_file:
    output_file.write(response.content)
```

Extras

- Si has llegado hasta aquí sin problemas crea el código necesario para que el script de arriba pueda también recibir como input un archivo .txt con fechas y vaya línea a línea descargando los PDFs de cada uno de los días dados.
- Esta es una página del gobierno, recuerda que es muy importante poner `time.sleep()`.