EJECUTAR CÓDIGO DE SCRAPY SIN LA TERMINAL

(Jupyter Notebook, Google Colab, etc)

A lo largo del curso solamente ejecutaremos Scrapy desde la terminal para guardar los datos extraídos dentro de un archivo de manera automática. Sin embargo, yo también puedo correr Scrapy sin utilizar la terminal. Es decir, como normalmente corremos Scripts de Python **(Clic derecho + RUN en Pycharm)**. De la siguiente manera (la parte diferente e importante se encuentra resaltada en amarillo):

```
from scrapy.spiders import Spider
from scrapy.crawler import CrawlerProcess

class MiCrawler(Spider):
    name = "MiCrawler"
    start_urls = ["https://mi.url.semilla.com"]

    def parse(self, response):
        print ("TITULO", selector.xpath("//h1/text()").get())


if __name__ == "__main__": # Código que se va a ejecutar al dar clic en RUN
    process = CrawlerProcess()
    process.crawl(MiCrawler) # Nombre de la clase de mi Spider
    process.start()
```

Noten como, en esta modalidad al correr mi código, yo no tuve que definir una abstracción de datos. Al no estar descargando datos dentro de un archivo utilizando el comando: “scrapy runspider”, no necesito definirla. En este caso, solamente estoy imprimiendo el resultado de un XPATH por pantalla.

Aquí un ejemplo más complejo en donde yo envío las URLs semilla como parámetros de llamada a mi Spider. De esta manera puedo definir URLs semilla de manera dinámica desde un archivo, o desde parámetros por terminal.

```
from scrapy.spiders import Spider
from scrapy.crawler import CrawlerProcess

class MiCrawler(Spider):
    name = "MiCrawler"

    def __init__(self, args):
        Spider.__init__(self, self.name)
        self.start_urls = args[0] # En los argumentos viene la URL semilla

    def parse(self, response):
        print ("TITULO", response.xpath("//h1/text()").get())


if __name__ == "__main__": # Código que se va a ejecutar al dar clic en RUN
    TEST_URL = "https://mi.url.semilla.com"
    process = CrawlerProcess()
    process.crawl(MiCrawler, [[TEST_URL])
    process.start()
```

La documentación completa sobre esta funcionalidad se encuentra disponible aquí:

https://docs.scrapy.org/en/latest/topics/practices.html#run-scrapy-from-a-script