#### Web Scrapping MERCADOLIBRE Manual de Construcción y Ejecución

Fiduciaria

Bancolombia

**PREREQUISITOS** 

Python 3.x + PIP Virtual ENV Code Editor

Solución en el Repositorio: https://github.com/ivanieves/env

# **Primer Paso**

Instalar entorno virtual (virtualenv vía librería PIP)

Un entorno virtual es una herramienta para mantener dependencias requeridas por diferentes proyectos en lugares separados. virtualenv crea un folder aislado el cual contiene todos los ejecutables necesarios para usar los paquetes que un proyecto de Python demanda.

para esto, nos dirigimos a la Terminal, y en ella navegamos a la carpeta donde usualmente guardamos los entornos virtuales. Con la orientación mkdir creamos el directorio y luego nos aseguramos de habilitar el entorno virtual.

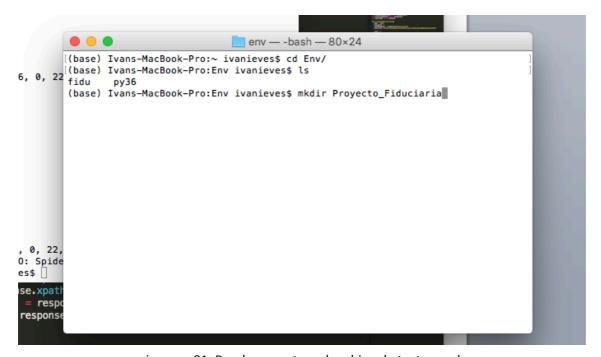


imagen 01. Puede encontrar el archivo de texto con la creación del entorno virtual (adjunto en: env/Soporte)

## Paso Dos

Instalar y Comenzar Scrapy Project

Scrapy es una plataforma colaborativa de código libre que corre en Python para extraer datos de páginas web. Dado que es un framework, Scrapy tiene una serie de herramientas poderosas para hacer el "scraping" o extraer información de webs de manera fácil y eficiente.

```
fidu - - bash - 80×24
                                                                     ⊗ (< >) Done
      Q~ install scrapy
xtrac
tions[-bash: cls: command not found
     (py36) (base) Ivans-MacBook-Pro:env ivanieves$ clear
     (py36) (base) Ivans-MacBook-Pro:env ivanieves$ ls
     pv36
     (py36) (base) Ivans-MacBook-Pro:env ivanieves$ pip install scrapy
     Requirement already satisfied: scrapy in ./py36/lib/python3.7/site-packages (1.7
     .3)
Requirement already satisfied: pyOpenSSL in ./py36/lib/python3.7/site-packages (
     from scrapy) (19.0.0)
     Requirement already satisfied: queuelib in ./py36/lib/python3.7/site-packages (f
     rom scrapy) (1.5.0)
Extra Requirement already satisfied: w3lib>=1.17.0 in ./py36/lib/python3.7/site-packag
Extra es (from scrapy) (1.21.0)
     Requirement already satisfied: six>=1.5.2 in ./py36/lib/python3.7/site-packages
     (from scrapy) (1.12.0)
     Requirement already satisfied: Twisted>=13.1.0; python_version != "3.4" in ./py3
     6/lib/python3.7/site-packages (from scrapy) (19.7.0)
    Requirement already satisfied: cssselect>=0.9 in ./py36/lib/python3.7/site-packa
Fide ges (from scrapy) (1.1.0)
     Requirement already satisfied: lxml; python_version != "3.4" in ./py36/lib/pytho
     n3.7/site-packages (from scrapy) (4.4.1)
     Requirement already satisfied: service-identity in ./py36/lib/python3.7/site-pac
    kages (from scrapy) (18.1.0)
numero_vendidos'] = response.xpath('normalize-space(//div[@class="item-condit
```

imagen 02.

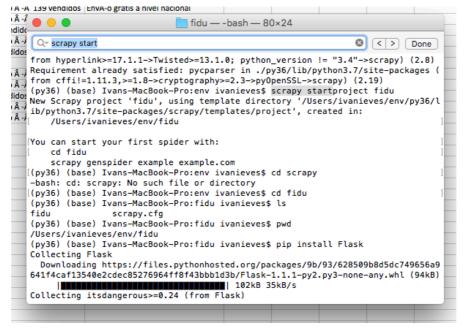


imagen 03.

# Paso Tres

Editor de Texto (Sublime Text3)

En este punto, se deja un poco de lado la Terminal. Nos dirigimos a la carpeta que con la creación del proyecto Scrapy se acaba de crear, y la cargamos en el editor de texto de código de preferencia. En este caso Sublime Text. Aquí se modificarán los .py

```
items.py - fidu
                                                                                                                        UNREGISTERED
FOLDERS
 __pycache__
    spiders
    ▶ ■ _pycache_
       /* __init__.py
                                  import scrapy
       /* spider.py
     /* __init__.py
                                   class FiduItem(scrapy.Item):
    /* items.py
     /* middlewares.py
     /* pipelines.py
                                       titulo = scrapy.Field()
precio = scrapy.Field()
numero_vendidos = scrapy.Field()
envio_gratis = scrapy.Field()
     /* settings.py
Line 1, Column 1
                                                                                                        Tab Size: 4
                                                                                                                            Python
```

Imagen 04. Ítems.py se crean los ítems según los campos que queramos extraer de la web

```
\delta spider.py — fidu
                                                                                                                                                                                              UNREGISTERED
                                                         spider.py
FOLDERS
 ▼ 📄 fidu
                                                                rt scrapy
scrapy.spiders import CrawlSpider, Ru
scrapy.linkextractors import LinkExtr
serv excentions import CloseSpider
                                                                  scrapy
                                                                                                      ort CrawlSpider, Rule
ors import LinkExtractor
   __pycache__
                                                               scrapy.exceptions import C fidu.items import FiduItem
    __pycache__
                                                      class FiduSpider(CrawlSpider):
   name = 'fidu'
          /* __init__.py
          /* spider.py
                                                               item_count= 0
                                                              allowed_domain = ['www.mercadolibre.com.co']
start_urls= ['https://listado.mercadolibre.com.co/televisores#D[A:te'
       /* __init__.py
       /* items.py
       /* middlewares.py
       /* pipelines.pv
                                                                      Rule(LinkExtractor(allow =(), restrict_xpaths = ('//span[@class=
Rule(LinkExtractor(allow =(), restrict_xpaths = ('//*[@class="it
callback = 'parse_item', follow = False)
       /* settings.pv
                                           16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
                                                              def parse_item(self, response):
    ml_item = FiduItem()
                                                                     #info de los productos
ml_item['titulo'] = response.xpath('normalize-space(//*[@id="shormalitem['precio'] = response.xpath('normalize-space(//*[@id="procml_item['numero_vendidos'] = response.xpath('normalize-space(//d:ml_item['envio_gratis'] = response.xpath('normalize-space(//*[@id="shormalize-space"])
                                                                            f.item_count
                                                                            self.item_count >20:
    raise CloseSpider('item_exceeded')
                                                                       yield ml_item
Line 33, Column 1
                                                                                                                                                                    Tab Size: 4
                                                                                                                                                                                                   Python
```

Imagen 05. Se crea el spider en spider.py en la carpeta spiders

```
🍯 pipelines.py — fidu
     FOLDERS
                                                                                                                                                   pon t rorget to add your pipetine to the ilem_mirelines setting
See: https://docs.scrapy.org/en/latest/topics/item-pipeline.html
     ▼ 🚞 fidu
                                                                                                                                                                          scrapy
            __pycache
                                                                                                                                                                                                                            t signals
                                                                                                                                                                  scrapy i
              spiders
                                                                                                                                                                                                                                                                            rt CsvItemExporter
                                                                                                                                                                  scrapy.exporters impor
                    __pycache_
                                                                                                                                                                      t csv
                                                                                                               11
12
                             /* __init__.py
                              /* spider.py
                                                                                                               13
14
15
                                                                                                                                          class FiduPipeline(object):
                                                                                                                                                            def __init__(self):
    self.files = {}
                      /* init .pv
                      /* items.pv
                                                                                                                                                              @classmethod
                       /* middlewares.py
                                                                                                                                                            def from_crawler(cls, crawler):
    pipeline = cls()
                      /* pipelines.py
                                                                                                                                                                                crawler.signals.connect(pipeline.spider_opened, signals.spider_opened, signals.connect(pipeline.spider_closed, signals.spider_closed, sig
                                                                                                              20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
                     /* settings.pv
                                                                                                                                                            def spider_opened(self, spider):
    file = open('%s_items.csv' % spider.name, 'w+b')
    self.files[spider] = file
    self.exporter = CsvItemExporter(file)
    self.exporter.fields_to_export = ['titulo', 'precio', 'numero_verself.exporter.start_exporting()
                                                                                                                                                            def spider_closed(self, spider):
    self.exporter.finish_exporting()
    file = self.files.pop(spider)
    file.close()
                                                                                                                                                            def process_item(self, item, spider
    self.exporter.export_item(item)
                                                                                                                                                                                            turn item
   Line 1, Column 1
```

Imagen 06. Pipeline para exportar a csv

## Paso Cuarto

Correr scrapy desde la Terminal

```
A 139 vendidos | Enva-o gratis a nivei nacional
                                fidu — -bash — 80×24
                                                                  Q~ scrapy crawl
 2019-09-25 19:19:21 [scrapy.core.engine] INFO: Spider closed (item_exceeded)
 (py36) (base) Ivans-MacBook-Pro:fidu ivanieves$ scrapy crawl fidu -t csv
 2019-09-25 19:22:41 [scrapy.utils.log] INFO: Scrapy 1.7.3 started (bot: fidu)
 2019-09-25 19:22:41 [scrapy.utils.log] INFO: Versions: lxml 4.4.1.0, libxml2 2.9
 .9, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21.0, Twisted 19.7.0, Python 3.7.3 (d
 efault, Mar 27 2019, 16:54:48) - [Clang 4.0.1 (tags/RELEASE_401/final)], pyOpenS
 SL 19.0.0 (OpenSSL 1.1.1c 28 May 2019), cryptography 2.7, Platform Darwin-15.6.
 0-x86 64-i386-64bit
 2019-09-25 19:22:41 [scrapy.crawler] INFO: Overridden settings: {'BOT_NAME': 'fi
 du', 'NEWSPIDER_MODULE': 'fidu.spiders', 'ROBOTSTXT_OBEY': True, 'SPIDER_MODULES
': ['fidu.spiders']}
 2019-09-25 19:22:41 [scrapy.extensions.telnet] INFO: Telnet Password: 8d7aa03cc4
 79cc53
 2019-09-25 19:22:42 [scrapy.middleware] INFO: Enabled extensions:
 ['scrapy.extensions.corestats.CoreStats'
  'scrapy.extensions.telnet.TelnetConsole'
  'scrapy.extensions.memusage.MemoryUsage',
  'scrapy.extensions.logstats.LogStats']
 2019-09-25 19:22:43 [scrapy.middleware] INFO: Enabled downloader middlewares:
 ['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
   'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
  'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
  'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
  'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
```

Imagen 07.

Volvemos a la Terminal esta vez para correr el scrapy modificado, que dados los cambios realizados en los archivos .py va a mercadolibre y trae la información requerida. En este caso el número de registros a investigar y el artículo buscado no tienen dinámica front End. En vez de esto, están declarados de manera muy sencilla en el archivo spider.py que creamos en la carpeta spiders de nuestro entorno.

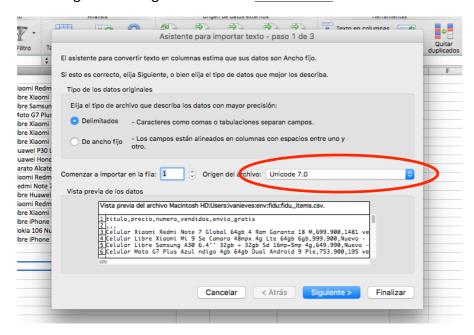
# Por Último

Visualización de Nuestra Consulta Mercadolibre

Siguiendo los pasos anteriormente descritos, se va a generar un archivo con nombre fidu\_items.csv en la carpeta fidu, a su vez dentro de la carpeta fidu alojada en nuestro repositorio de entornos virtuales env. Este archivo lo podemos cargar a Excel mediante la opción: importación de datos desde un archivo de texto.



Nota: Los datos van delimitados por comas. Para una mejor visualización de estos, se debe elegir como origen de archivo: Unicode 7.0



#### **Conclusiones Adicionales**

La realización de este proyecto se abordó mediante la instalación de módulos en Python. Para el Web Scrapping, se usó la plataforma colaborativa libre: scrapy de la biblioteca de cogido PIP. En el repositorio encontrará en: env/fidu/fidu, los archivos .py modificados y en: env/fidu/fidu/spiders, el spider.py que llama los datos desde mercadolibre.com

Adicionalmente encontrará este documento junto a otros archivos .txt adjuntos con los códigos de la Terminal de los pasos Uno, Dos y Cuatro en: env/Soporte