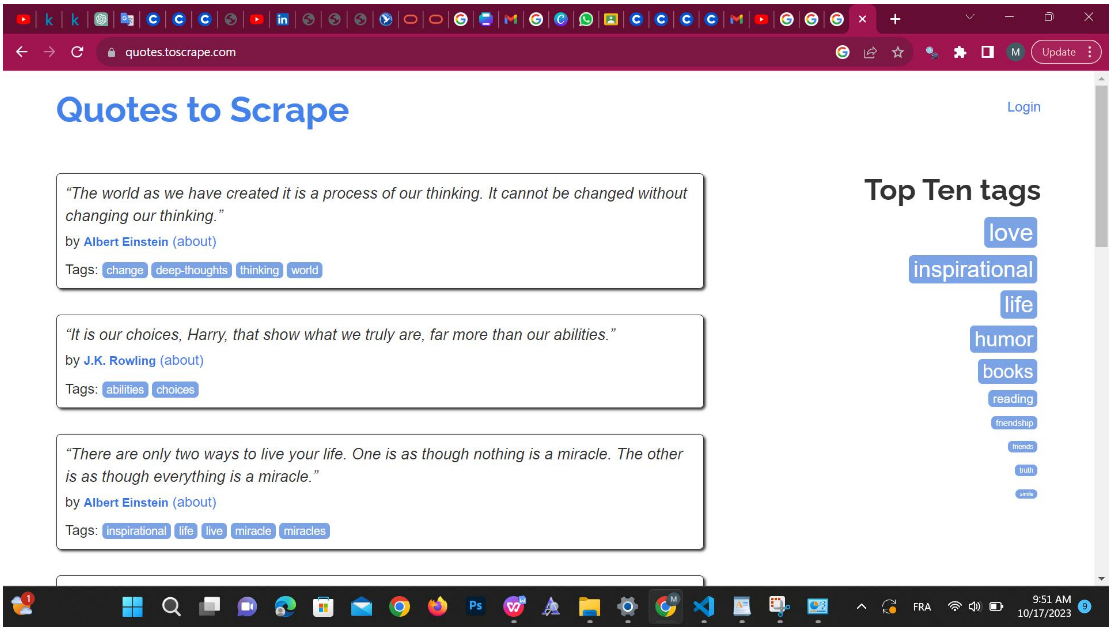
Atelier Web Scraping

On va travailler sur le site « https://quotes.toscrape.com/ » pour extraire des informations



Quotes to Scrape

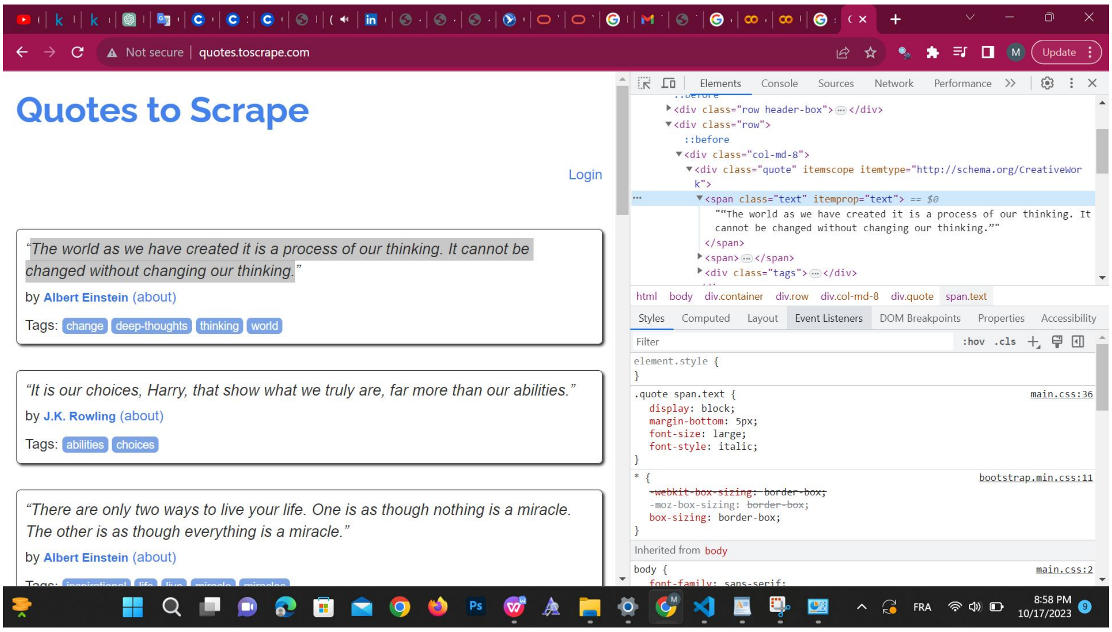
“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”
by Albert Einstein (about)
Tags: change deep-thoughts thinking world

“It is our choices, Harry, that show what we truly are, far more than our abilities.”
by J.K. Rowling (about)
Tags: abilities choices

“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”
by Albert Einstein (about)
Tags: inspirational life live miracle miracles

Top Ten tags
love
inspirational
life
humor
books
reading
friendship
friends
truth
simile

La notion de classe en web scraping est très important pour extraire des informations significatives



Quotes to Scrape

“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”
by Albert Einstein (about)
Tags: change deep-thoughts thinking world

“It is our choices, Harry, that show what we truly are, far more than our abilities.”
by J.K. Rowling (about)
Tags: abilities choices

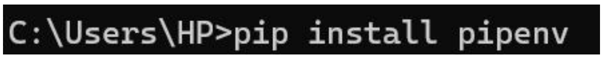
“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”
by Albert Einstein (about)

```
  ▶<div class="row header-box">⋯</div>
  ▼<div class="row">
     ::before
    ▼<div class="col-md-8">
      ▼<div class="quote" itemscope itemtype="http://schema.org/CreativeWor
        k">
...     ▼<span class="text" itemprop="text"> == $0
           "“The world as we have created it is a process of our thinking. It
           cannot be changed without changing our thinking.”"
         </span>
        ▶<span>⋯</span>
        ▶<div class="tags">⋯</div>
```

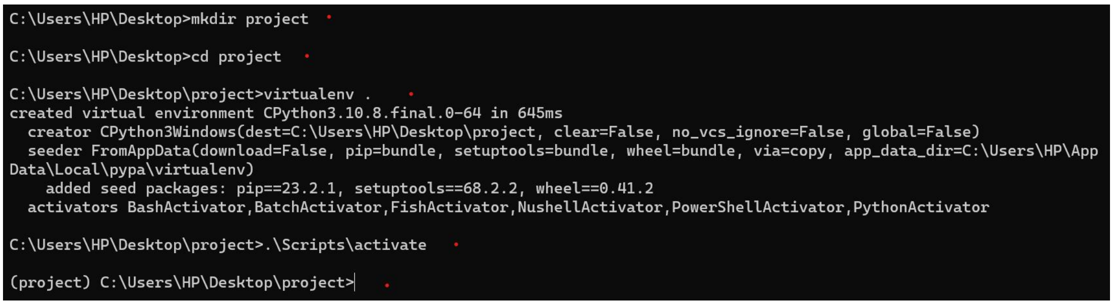
html body div.container div.row div.col-md-8 div.quote span.text

```
element.style {
}
.quote span.text {                                      main.css:36
   display: block;
   margin-bottom: 5px;
   font-size: large;
   font-style: italic;
}
* {                                              bootstrap.min.css:11
   ~~-webkit-box-sizing: border-box;~~
   -moz-box-sizing: ~~border-box~~;
   box-sizing: border-box;
}
Inherited from body
body {                                                   main.css:2
   font-family: sans-serif;
```

Tout d’abord on va commencer par l’installation et la configuration de l’environnement virtuelle

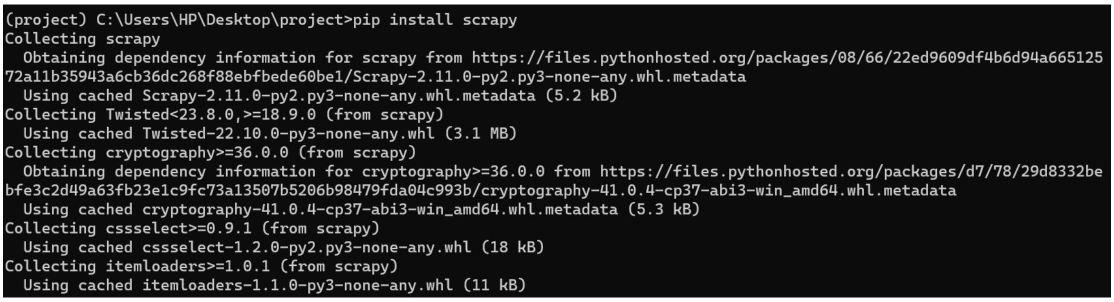


```
C:\Users\HP>pip install pipenv
```



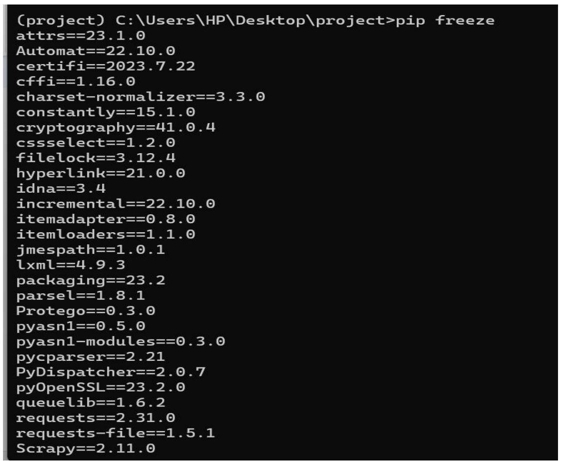
```
C:\Users\HP\Desktop>mkdir project

C:\Users\HP\Desktop>cd project

C:\Users\HP\Desktop\project>virtualenv .
created virtual environment CPython3.10.8.final.0-64 in 645ms
  creator CPython3Windows(dest=C:\Users\HP\Desktop\project, clear=False, no_vcs_ignore=False, global=False)
  seeder FromAppData(download=False, pip=bundle, setuptools=bundle, wheel=bundle, via=copy, app_data_dir=C:\Users\HP\App
Data\Local\pypa\virtualenv)
    added seed packages: pip==23.2.1, setuptools==68.2.2, wheel==0.41.2
  activators BashActivator,BatchActivator,FishActivator,NushellActivator,PowerShellActivator,PythonActivator

C:\Users\HP\Desktop\project>.\Scripts\activate

(project) C:\Users\HP\Desktop\project>
```

- Installer par la suite scrapy par la commande : pip install scrapy



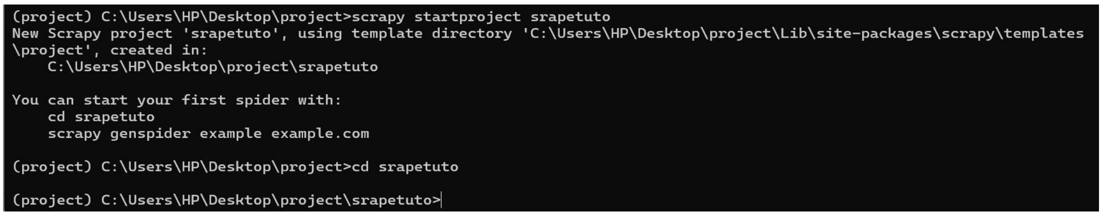
```
(project) C:\Users\HP\Desktop\project>pip install scrapy
Collecting scrapy
  Obtaining dependency information for scrapy from https://files.pythonhosted.org/packages/08/66/22ed9609df4b6d94a665125
72a11b35943a6cb36dc268f88ebfbede60be1/Scrapy-2.11.0-py2.py3-none-any.whl.metadata
  Using cached Scrapy-2.11.0-py2.py3-none-any.whl.metadata (5.2 kB)
Collecting Twisted<23.8.0,>=18.9.0 (from scrapy)
  Using cached Twisted-22.10.0-py3-none-any.whl (3.1 MB)
Collecting cryptography>=36.0.0 (from scrapy)
  Obtaining dependency information for cryptography>=36.0.0 from https://files.pythonhosted.org/packages/d7/78/29d8332be
bfe3c2d49a63fb23e1c9fc73a13507b5206b98479fda04c993b/cryptography-41.0.4-cp37-abi3-win_amd64.whl.metadata
  Using cached cryptography-41.0.4-cp37-abi3-win_amd64.whl.metadata (5.3 kB)
Collecting cssselect>=0.9.1 (from scrapy)
  Using cached cssselect-1.2.0-py2.py3-none-any.whl (18 kB)
Collecting itemloaders>=1.0.1 (from scrapy)
  Using cached itemloaders-1.1.0-py3-none-any.whl (11 kB)
```

La commande «pip freeze » est utilisé pour afficher la liste des bibliothèque installé après la commande précédente :



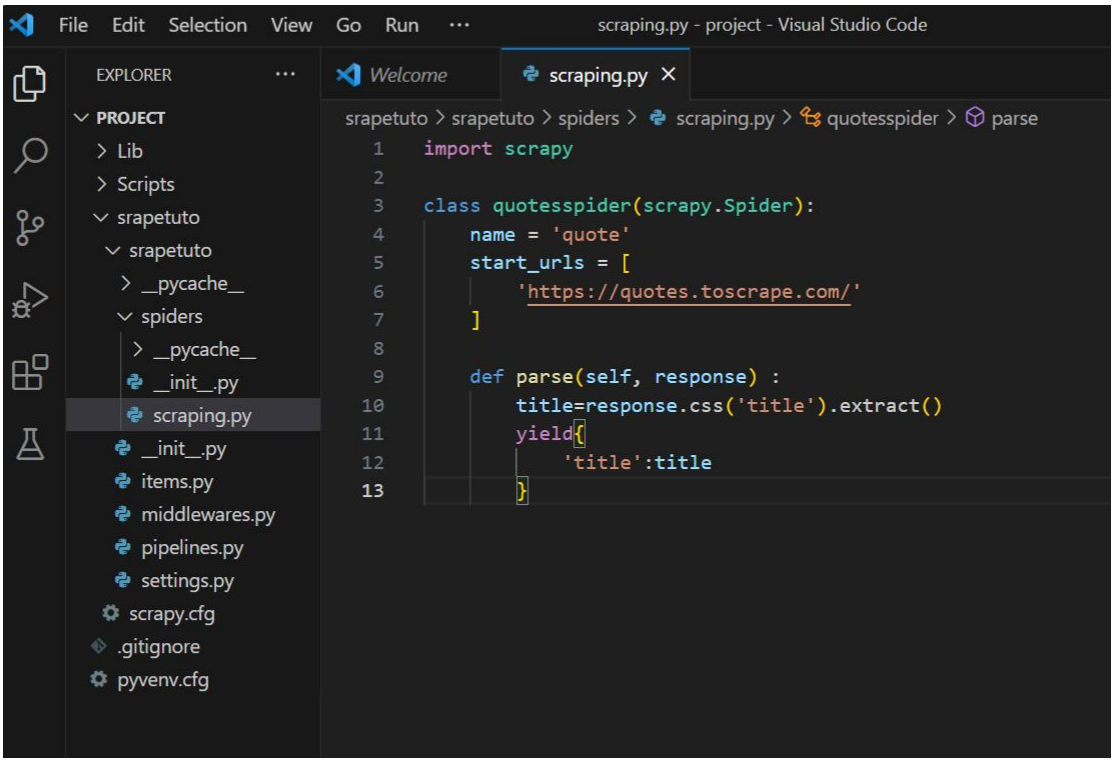
```
(project) C:\Users\HP\Desktop\project>pip freeze
attrs==23.1.0
Automat==22.10.0
certifi==2023.7.22
cffi==1.16.0
charset-normalizer==3.3.0
constantly==15.1.0
cryptography==41.0.4
cssselect==1.2.0
filelock==3.12.4
hyperlink==21.0.0
idna==3.4
incremental==22.10.0
itemadapter==0.8.0
itemloaders==1.1.0
jmespath==1.0.1
lxml==4.9.3
packaging==23.2
parsel==1.8.1
Protego==0.3.0
pyasn1==0.5.0
pyasn1-modules==0.3.0
pycparser==2.21
PyDispatcher==2.0.7
pyOpenSSL==23.2.0
queuelib==1.6.2
requests==2.31.0
requests-file==1.5.1
Scrapy==2.11.0
```

On va utilisé la commande suivante pour générer le dossier de chaque projet Scrapy qui contient les composants principales pour un projet web scraping (Scrapy)



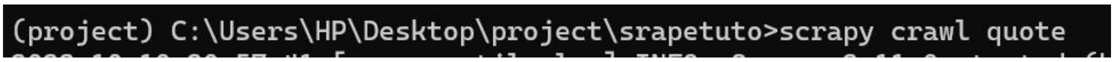
```
(project) C:\Users\HP\Desktop\project>scrapy startproject srapetuto
New Scrapy project 'srapetuto', using template directory 'C:\Users\HP\Desktop\project\Lib\site-packages\scrapy\templates
\project', created in:
    C:\Users\HP\Desktop\project\srapetuto

You can start your first spider with:
    cd srapetuto
    scrapy genspider example example.com

(project) C:\Users\HP\Desktop\project>cd srapetuto

(project) C:\Users\HP\Desktop\project\srapetuto>
```

On va essayez ici un petit code pour illustrer le fonctionnement de scrapy :

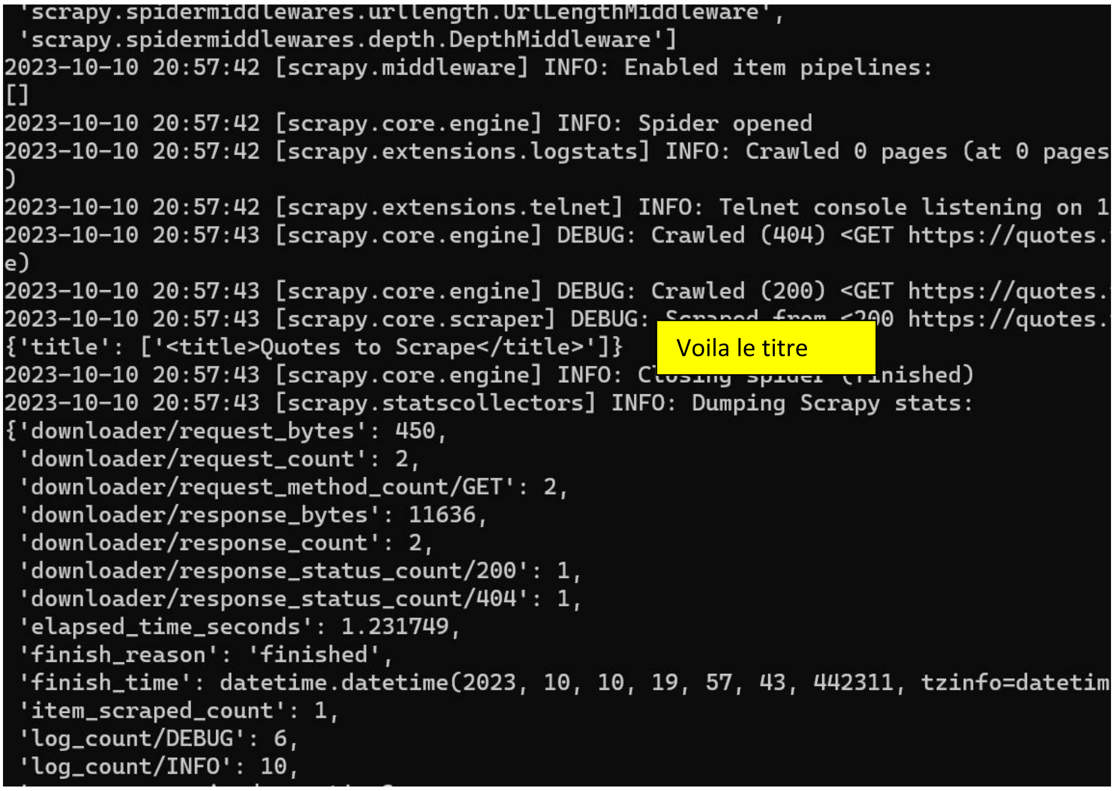


```python
import scrapy

class quotesspider(scrapy.Spider):
    name = 'quote'
    start_urls = [
        'https://quotes.toscrape.com/'
    ]

    def parse(self, response) :
        title=response.css('title').extract()
        yield{
            'title':title
        }
```

Pour éxécuter le code suivant on va utiliser la commande suivante : scrapy crawl quote



```
(project) C:\Users\HP\Desktop\project\srapetuto>scrapy crawl quote
```

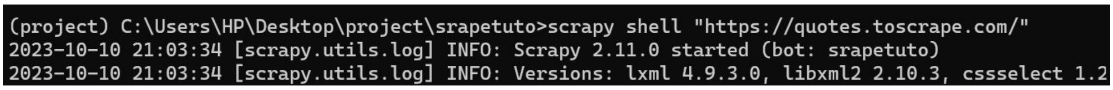


```
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2023-10-10 20:57:42 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2023-10-10 20:57:42 [scrapy.core.engine] INFO: Spider opened
2023-10-10 20:57:42 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages
)
2023-10-10 20:57:42 [scrapy.extensions.telnet] INFO: Telnet console listening on 1
2023-10-10 20:57:43 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://quotes.
e)
2023-10-10 20:57:43 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.
2023-10-10 20:57:43 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.
{'title': ['<title>Quotes to Scrape</title>']}
2023-10-10 20:57:43 [scrapy.core.engine] INFO: Closing spider (finished)
2023-10-10 20:57:43 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 450,
 'downloader/request_count': 2,
 'downloader/request_method_count/GET': 2,
 'downloader/response_bytes': 11636,
 'downloader/response_count': 2,
 'downloader/response_status_count/200': 1,
 'downloader/response_status_count/404': 1,
 'elapsed_time_seconds': 1.231749,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2023, 10, 10, 19, 57, 43, 442311, tzinfo=datetim
 'item_scraped_count': 1,
 'log_count/DEBUG': 6,
 'log_count/INFO': 10,
```

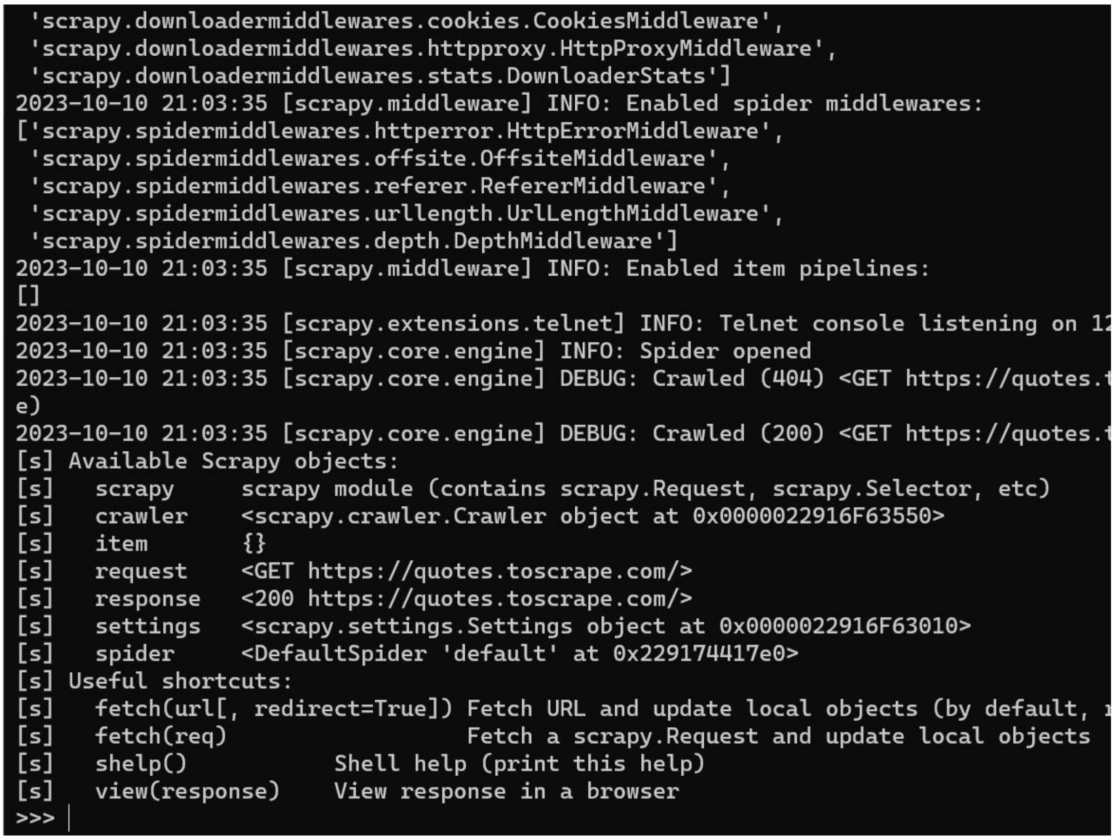
Voila le titre

CSS

La commande "scrapy shell http://www.quotetoscrape.com " ouvre une coquille interactive pour interagir avec le site web "http://www.quotetoscrape.com" en utilisant le framework de scraping Scrapy.

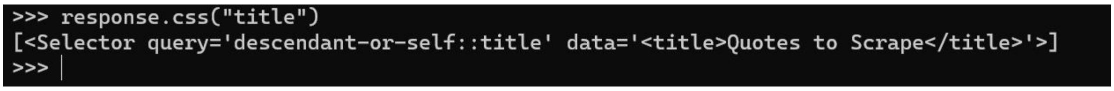


```
(project) C:\Users\HP\Desktop\project\srapetuto>scrapy shell "https://quotes.toscrape.com/"
2023-10-10 21:03:34 [scrapy.utils.log] INFO: Scrapy 2.11.0 started (bot: srapetuto)
2023-10-10 21:03:34 [scrapy.utils.log] INFO: Versions: lxml 4.9.3.0, libxml2 2.10.3, cssselect 1.2
```

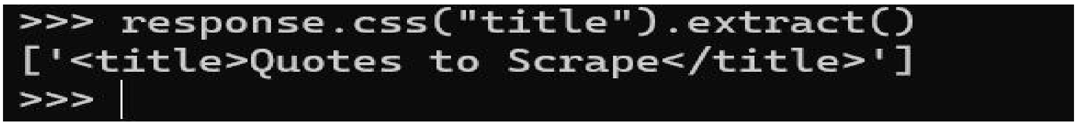


```
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2023-10-10 21:03:35 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2023-10-10 21:03:35 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2023-10-10 21:03:35 [scrapy.extensions.telnet] INFO: Telnet console listening on 12
2023-10-10 21:03:35 [scrapy.core.engine] INFO: Spider opened
2023-10-10 21:03:35 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://quotes.t
e)
2023-10-10 21:03:35 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.t
[s] Available Scrapy objects:
[s]   scrapy     scrapy module (contains scrapy.Request, scrapy.Selector, etc)
[s]   crawler    <scrapy.crawler.Crawler object at 0x0000022916F63550>
[s]   item       {}
[s]   request    <GET https://quotes.toscrape.com/>
[s]   response   <200 https://quotes.toscrape.com/>
[s]   settings   <scrapy.settings.Settings object at 0x0000022916F63010>
[s]   spider     <DefaultSpider 'default' at 0x229174417e0>
[s] Useful shortcuts:
[s]   fetch(url[, redirect=True]) Fetch URL and update local objects (by default, r
[s]   fetch(req)                  Fetch a scrapy.Request and update local objects
[s]   shelp()           Shell help (print this help)
[s]   view(response)    View response in a browser
>>>
```

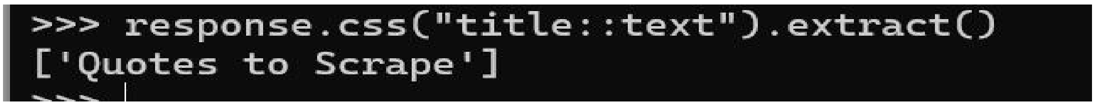
Ici on veut récupérer le titre de site web on utilisant un syntaxe different pour personnaliser l'affichage du titre



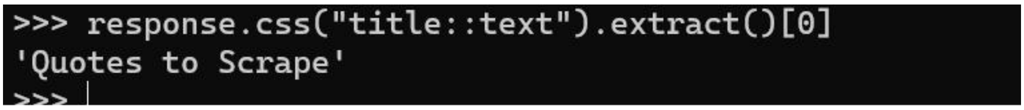
```
>>> response.css("title")
[<Selector query='descendant-or-self::title' data='<title>Quotes to Scrape</title>'>]
>>>
```



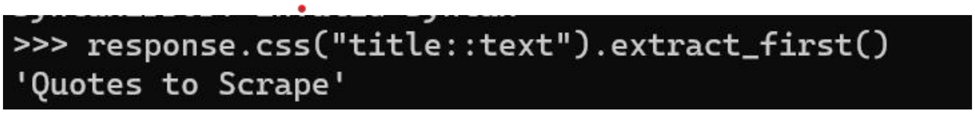
```
>>> response.css("title").extract()
['<title>Quotes to Scrape</title>']
>>>
```



```
>>> response.css("title::text").extract()
['Quotes to Scrape']
>>>
```

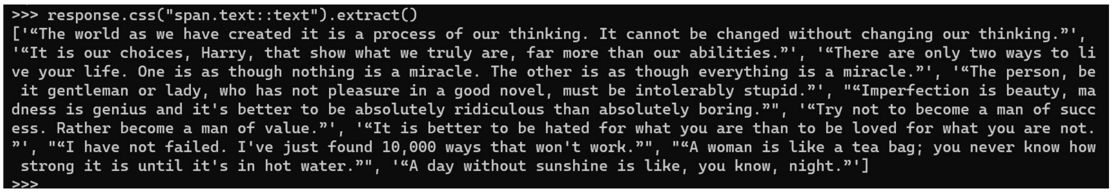


```
>>> response.css("title::text").extract()[0]
'Quotes to Scrape'
>>>
```



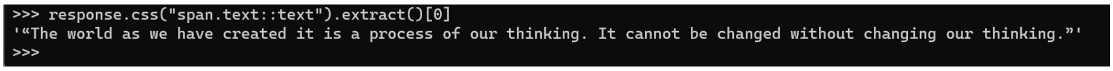
```
>>> response.css("title::text").extract_first()
'Quotes to Scrape'
```

Texte : on veut récupérer maintenant le texte du site web



```
>>> response.css("span.text::text").extract()
['“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”',
'“It is our choices, Harry, that show what we truly are, far more than our abilities.”', '“There are only two ways to li
ve your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”', '“The person, be
 it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”', "“Imperfection is beauty, ma
dness is genius and it's better to be absolutely ridiculous than absolutely boring.”", '“Try not to become a man of succ
ess. Rather become a man of value.”', '“It is better to be hated for what you are than to be loved for what you are not.
”', "“I have not failed. I've just found 10,000 ways that won't work.”", "“A woman is like a tea bag; you never know how
 strong it is until it's in hot water.”", '“A day without sunshine is like, you know, night.”']
>>>
```

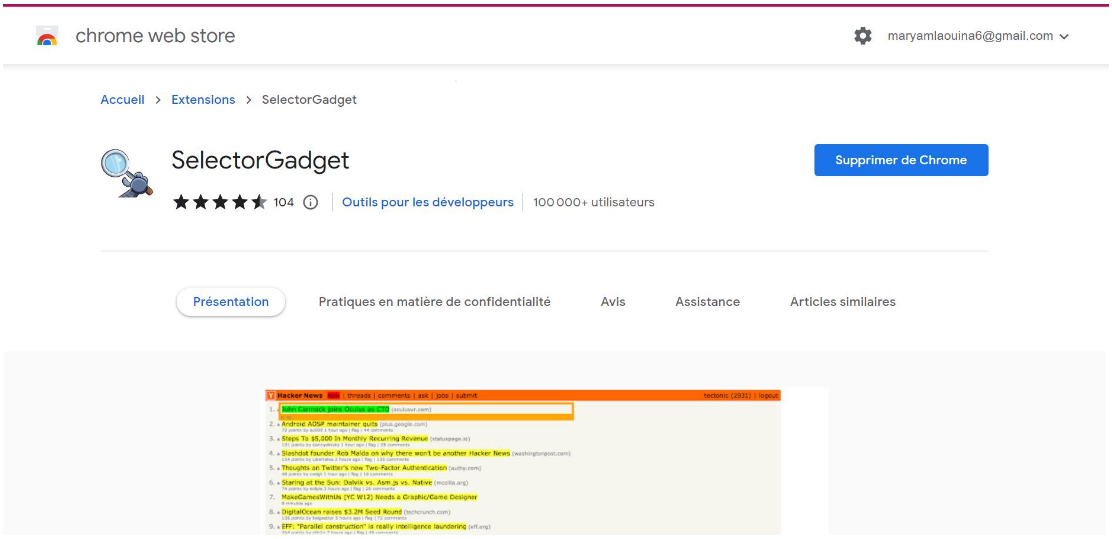
Ici on veut récupéré la première élément de la liste

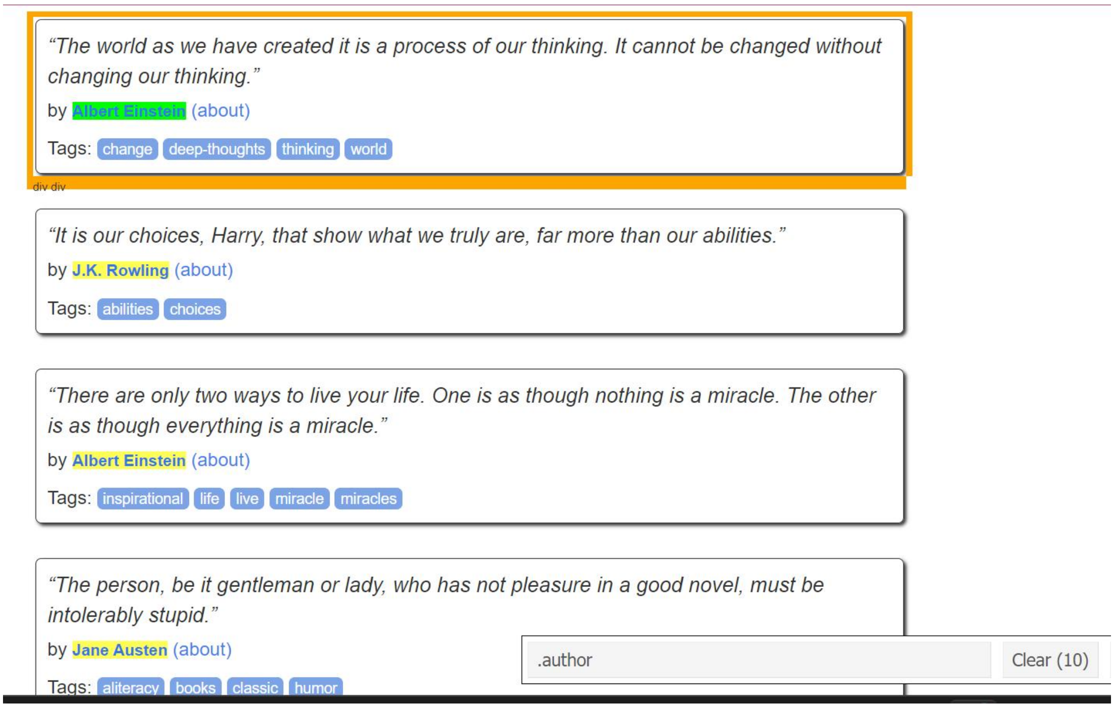


```
>>> response.css("span.text::text").extract()[0]
'“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”'
>>>
```

SelectorGadget est un outil de sélection d'éléments pour le web, sous forme d'une extension de navigateur, qui permet aux utilisateurs de point et de cliquer pour identifier et générer des sélecteurs CSS afin d'extraire des données spécifiques à partir de pages web ou de les cibler pour des actions de scraping ou de manipulation.

Voila ici on veut par exemple afficher la classe d’un élément sans faire traverser le code en entier pour trouver une classe





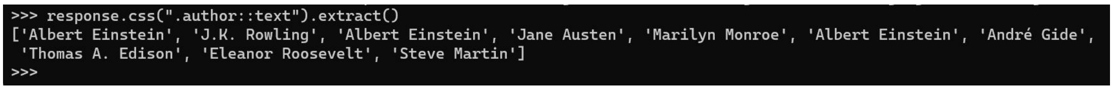
“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”
by Albert Einstein (about)
Tags: change deep-thoughts thinking world

“It is our choices, Harry, that show what we truly are, far more than our abilities.”
by J.K. Rowling (about)
Tags: abilities choices

“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”
by Albert Einstein (about)
Tags: inspirational life live miracle miracles

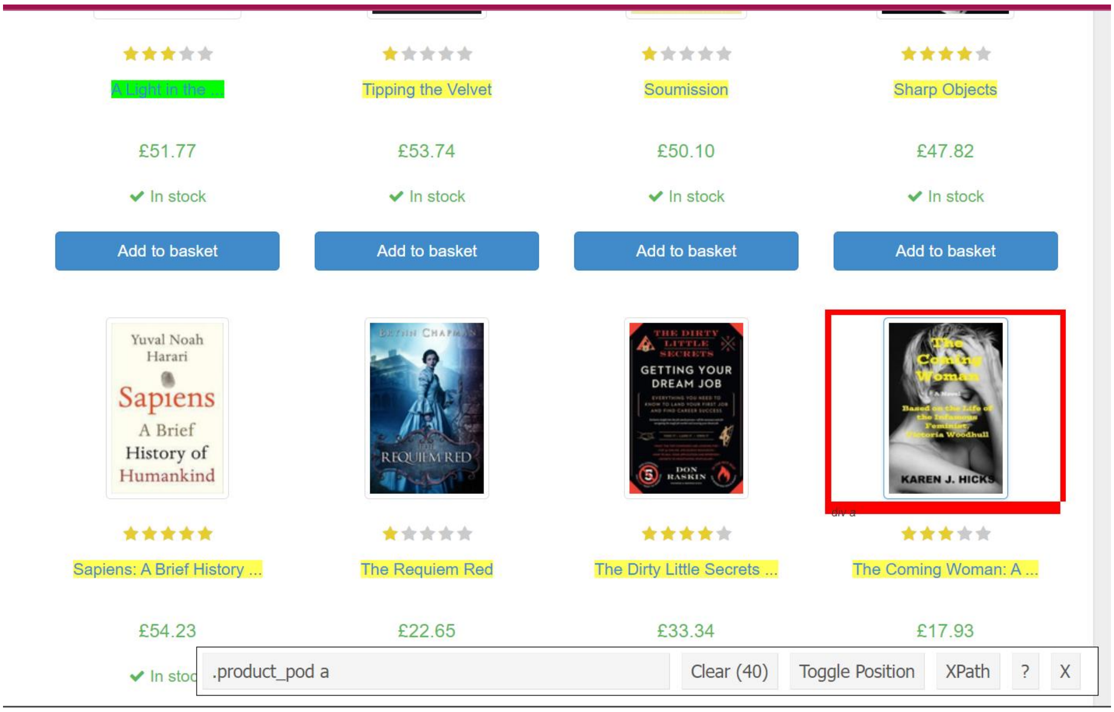
“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”
by Jane Austen (about)
Tags: aliteracy books classic humor

On va essayer d’afficher les auteurs on utilisant cette outil pour identifier la classe des auteurs .



```
>>> response.css(".author::text").extract()
['Albert Einstein', 'J.K. Rowling', 'Albert Einstein', 'Jane Austen', 'Marilyn Monroe', 'Albert Einstein', 'André Gide',
 'Thomas A. Edison', 'Eleanor Roosevelt', 'Steve Martin']
>>>
```

Ici après la commande : scrapy shell « https://books.toscrape.com/ »

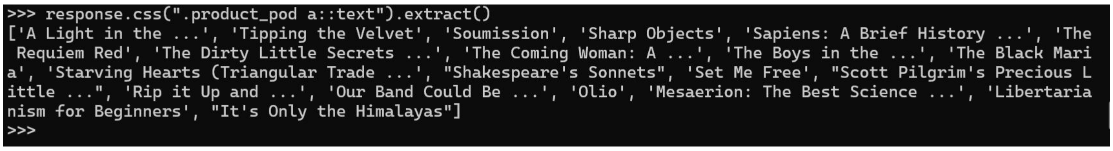


A Light in the ... — £51.77 — In stock
Tipping the Velvet — £53.74 — In stock
Soumission — £50.10 — In stock
Sharp Objects — £47.82 — In stock

Sapiens: A Brief History ... — £54.23
The Requiem Red — £22.65
The Dirty Little Secrets ... — £33.34
The Coming Woman: A ... — £17.93

On va extraire les titres des livres après l'utilisation de cette tool pour extraire la classe des titres des livres

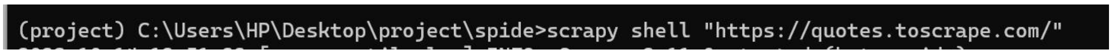
Voila le resultat :



```
>>> response.css(".product_pod a::text").extract()
['A Light in the ...', 'Tipping the Velvet', 'Soumission', 'Sharp Objects', 'Sapiens: A Brief History ...', 'The
 Requiem Red', 'The Dirty Little Secrets ...', 'The Coming Woman: A ...', 'The Boys in the ...', 'The Black Mari
a', 'Starving Hearts (Triangular Trade ...', "Shakespeare's Sonnets", 'Set Me Free', "Scott Pilgrim's Precious L
ittle ...", 'Rip it Up and ...', 'Our Band Could Be ...', 'Olio', 'Mesaerion: The Best Science ...', 'Libertaria
nism for Beginners', "It's Only the Himalayas"]
>>>
```

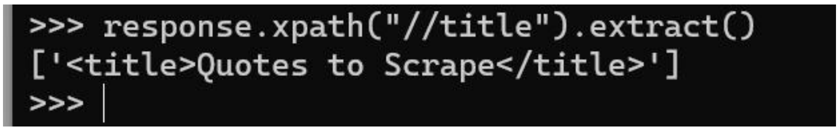
XPATH

On va ici de retravailler sur le site quote to scrape on utilisant cette fois le sélecteur XPATH

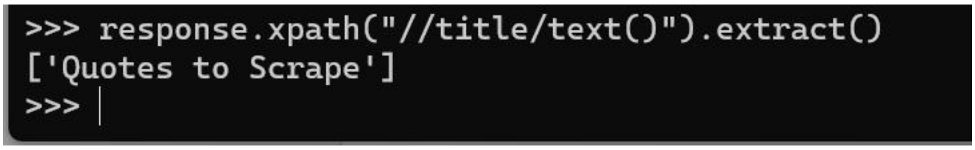


```
(project) C:\Users\HP\Desktop\project\spide>scrapy shell "https://quotes.toscrape.com/"
```

Extraire le titre de site on utilisant xpath :

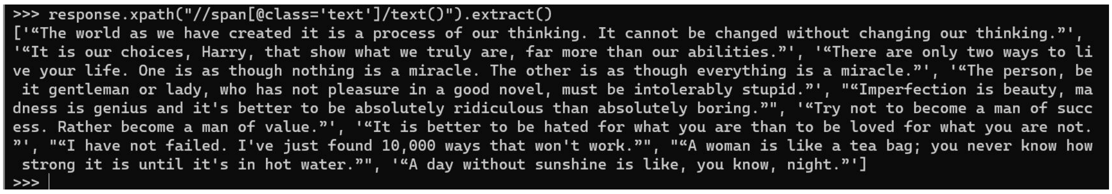


```
>>> response.xpath("//title").extract()
['<title>Quotes to Scrape</title>']
>>>
```



```
>>> response.xpath("//title/text()").extract()
['Quotes to Scrape']
>>>
```

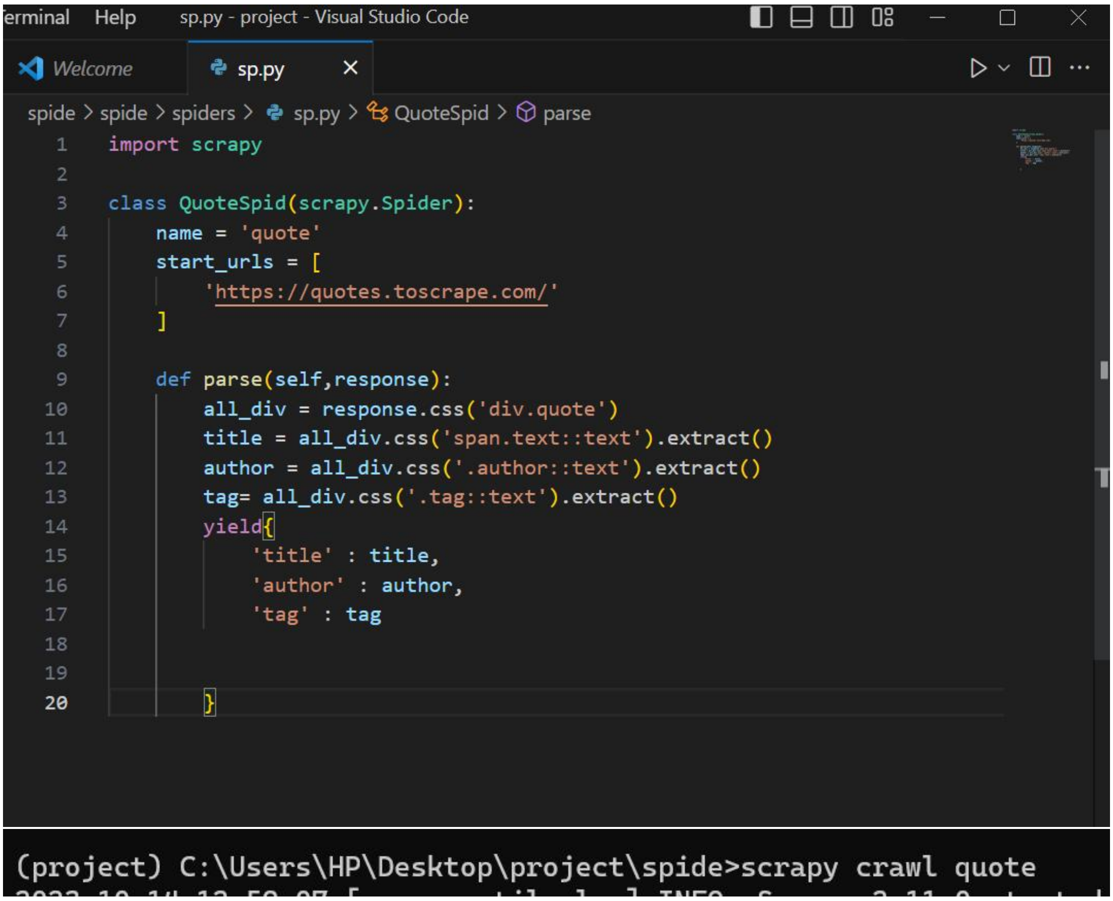
Extraction du texte :



```
>>> response.xpath("//span[@class='text']/text()").extract()
['“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”',
'“It is our choices, Harry, that show what we truly are, far more than our abilities.”', '“There are only two ways to li
ve your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”', '“The person, be
 it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”', "“Imperfection is beauty, ma
dness is genius and it's better to be absolutely ridiculous than absolutely boring.”", '“Try not to become a man of succ
ess. Rather become a man of value.”', '“It is better to be hated for what you are than to be loved for what you are not.
”', "“I have not failed. I've just found 10,000 ways that won't work.”", "“A woman is like a tea bag; you never know how
 strong it is until it's in hot water.”", '“A day without sunshine is like, you know, night.”']
>>>
```

Programme python :

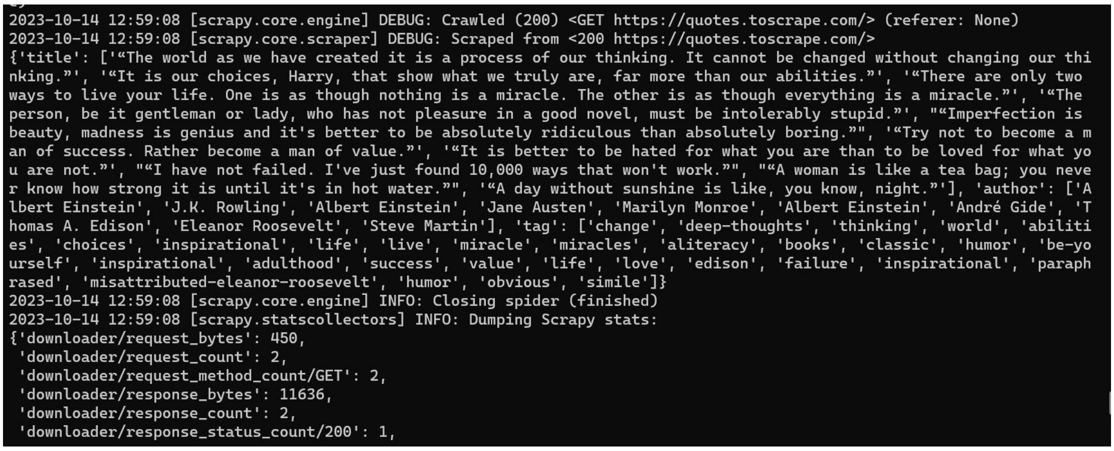
Script python (Spider) suivant nous a permet d'extraire les textes, les auteurs, les tags du site web quotes to scrape après l'exécution de la commande scrapy crawl quote


erminal Help sp.py - project - Visual Studio Code
Welcome
sp.py
spide > spide > spiders > sp.py > QuoteSpid > parse

```python
import scrapy

class QuoteSpid(scrapy.Spider):
    name = 'quote'
    start_urls = [
        'https://quotes.toscrape.com/'
    ]

    def parse(self,response):
        all_div = response.css('div.quote')
        title = all_div.css('span.text::text').extract()
        author = all_div.css('.author::text').extract()
        tag= all_div.css('.tag::text').extract()
        yield{
            'title' : title,
            'author' : author,
            'tag' : tag


        }
```

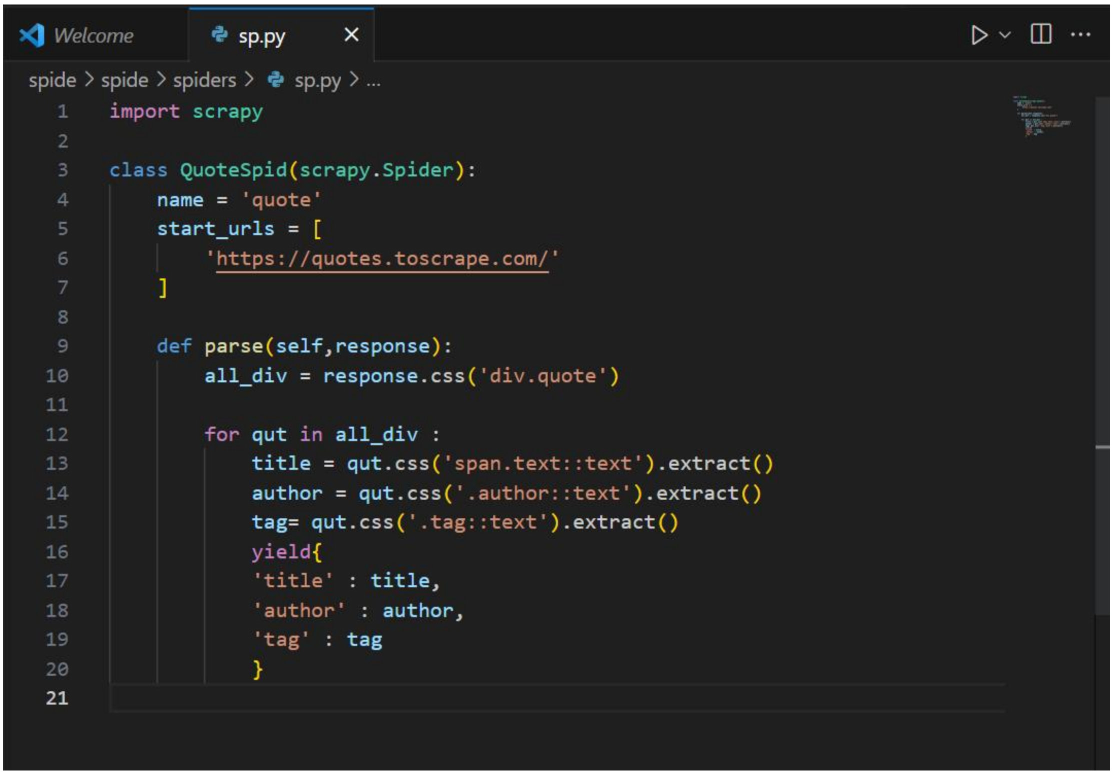
```
(project) C:\Users\HP\Desktop\project\spide>scrapy crawl quote
```

Voila le resultat obtenue



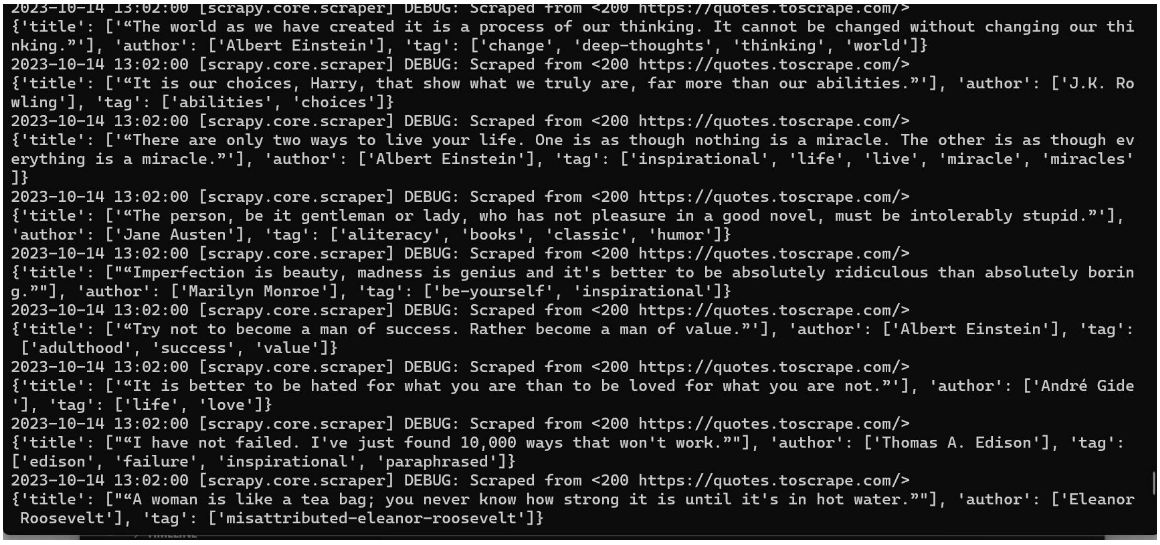
```
2023-10-14 12:59:08 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.toscrape.com/> (referer: None)
2023-10-14 12:59:08 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“The world as we have created it is a process of our thinking. It cannot be changed without changing our thi
nking.”', '“It is our choices, Harry, that show what we truly are, far more than our abilities.”', '“There are only two
ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”', '“The
person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”', "“Imperfection is
beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.”", '“Try not to become a m
an of success. Rather become a man of value.”', '“It is better to be hated for what you are than to be loved for what yo
u are not.”', "“I have not failed. I've just found 10,000 ways that won't work.”", "“A woman is like a tea bag; you neve
r know how strong it is until it's in hot water.”", '“A day without sunshine is like, you know, night.”'], 'author': ['A
lbert Einstein', 'J.K. Rowling', 'Albert Einstein', 'Jane Austen', 'Marilyn Monroe', 'Albert Einstein', 'André Gide', 'T
homas A. Edison', 'Eleanor Roosevelt', 'Steve Martin'], 'tag': ['change', 'deep-thoughts', 'thinking', 'world', 'abiliti
es', 'choices', 'inspirational', 'life', 'live', 'miracle', 'miracles', 'aliteracy', 'books', 'classic', 'humor', 'be-yo
urself', 'inspirational', 'adulthood', 'success', 'value', 'life', 'love', 'edison', 'failure', 'inspirational', 'paraph
rased', 'misattributed-eleanor-roosevelt', 'humor', 'obvious', 'simile']}
2023-10-14 12:59:08 [scrapy.core.engine] INFO: Closing spider (finished)
2023-10-14 12:59:08 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 450,
 'downloader/request_count': 2,
 'downloader/request_method_count/GET': 2,
 'downloader/response_bytes': 11636,
 'downloader/response_count': 2,
 'downloader/response_status_count/200': 1,
```

On va utiliser ici la boucle for pour structurer l’affichage du résultat, pour afficher l’auteur et les tags de chaque texte



```
import scrapy

class QuoteSpid(scrapy.Spider):
    name = 'quote'
    start_urls = [
        'https://quotes.toscrape.com/'
    ]

    def parse(self,response):
        all_div = response.css('div.quote')

        for qut in all_div :
            title = qut.css('span.text::text').extract()
            author = qut.css('.author::text').extract()
            tag= qut.css('.tag::text').extract()
            yield{
            'title' : title,
            'author' : author,
            'tag' : tag
            }

```

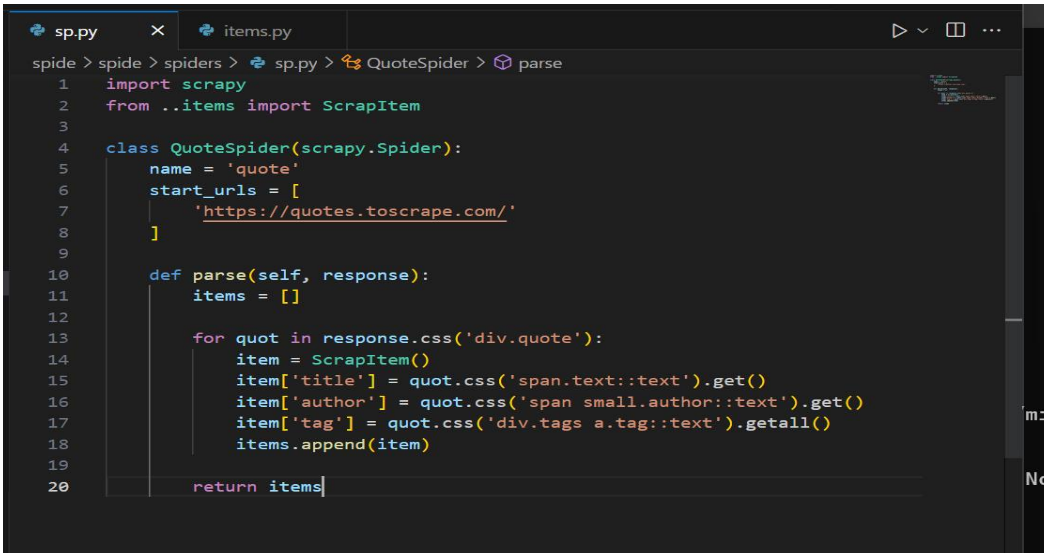
Après la commande : scrapy crawl quote , voila les resultats obtenues



```
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“The world as we have created it is a process of our thinking. It cannot be changed without changing our thi
nking.”'], 'author': ['Albert Einstein'], 'tag': ['change', 'deep-thoughts', 'thinking', 'world']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“It is our choices, Harry, that show what we truly are, far more than our abilities.”'], 'author': ['J.K. Ro
wling'], 'tag': ['abilities', 'choices']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though ev
erything is a miracle.”'], 'author': ['Albert Einstein'], 'tag': ['inspirational', 'life', 'live', 'miracle', 'miracles'
]}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”'],
'author': ['Jane Austen'], 'tag': ['aliteracy', 'books', 'classic', 'humor']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ["“Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely borin
g.”"], 'author': ['Marilyn Monroe'], 'tag': ['be-yourself', 'inspirational']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“Try not to become a man of success. Rather become a man of value.”'], 'author': ['Albert Einstein'], 'tag':
 ['adulthood', 'success', 'value']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ['“It is better to be hated for what you are than to be loved for what you are not.”'], 'author': ['André Gide
'], 'tag': ['life', 'love']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ["“I have not failed. I've just found 10,000 ways that won't work.”"], 'author': ['Thomas A. Edison'], 'tag':
['edison', 'failure', 'inspirational', 'paraphrased']}
2023-10-14 13:02:00 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'title': ["“A woman is like a tea bag; you never know how strong it is until it's in hot water.”"], 'author': ['Eleanor
 Roosevelt'], 'tag': ['misattributed-eleanor-roosevelt']}
```

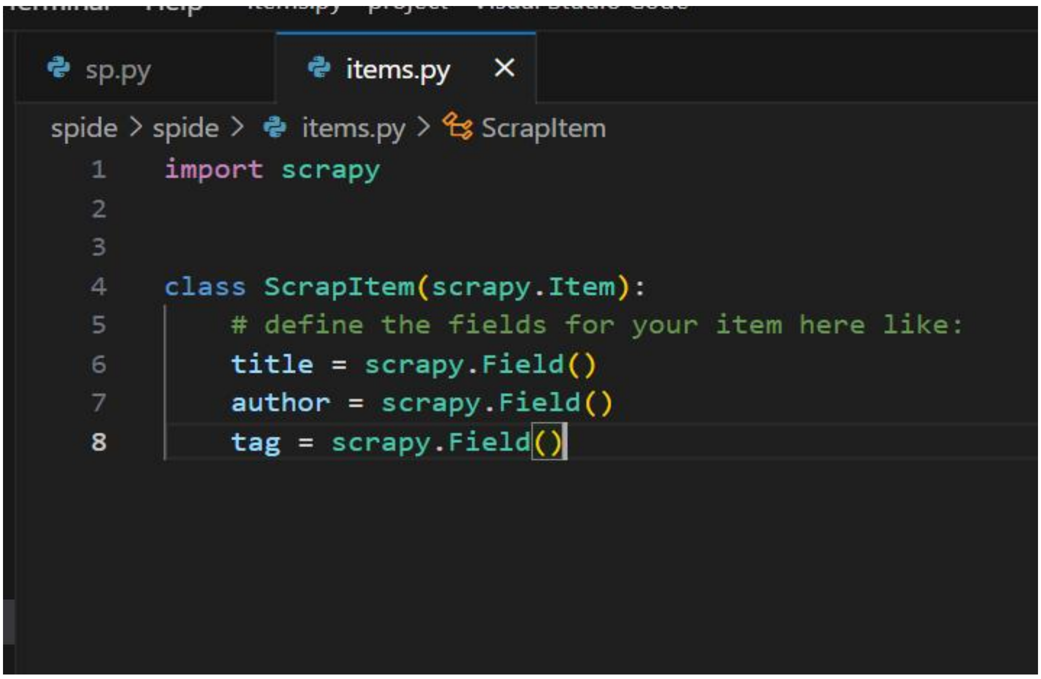
Utilisation des items

- Pour stocker de grandes quantités de données en utilisant Scrapy, on passe par des pipelines et des items Scrapy.
- Scrapy utilise des items pour structurer et stocker les données extraites.
- Donc, pour stocker de grandes quantités de données en utilisant Scrapy, vous n'avez pas besoin de passer par les items directement, mais plutôt de les utiliser pour structurer vos données, et ensuite configurer des pipelines Scrapy pour stocker ces données de la manière qui vous convient le mieux.



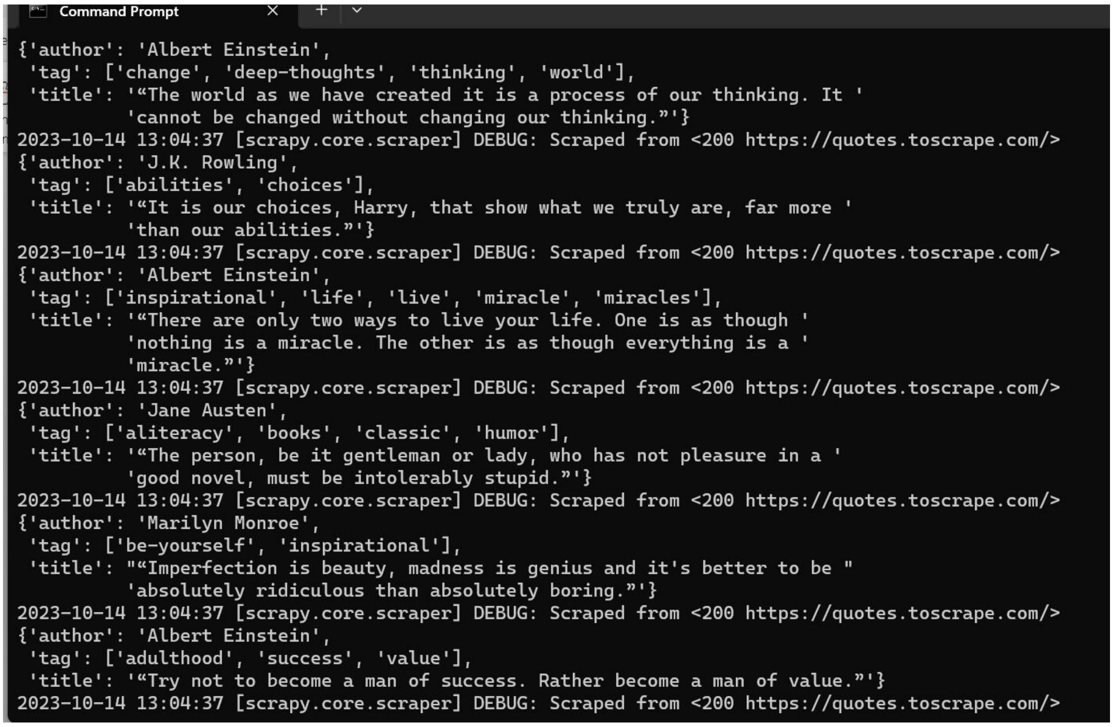
```python
import scrapy
from ..items import ScrapItem

class QuoteSpider(scrapy.Spider):
    name = 'quote'
    start_urls = [
        'https://quotes.toscrape.com/'
    ]

    def parse(self, response):
        items = []

        for quot in response.css('div.quote'):
            item = ScrapItem()
            item['title'] = quot.css('span.text::text').get()
            item['author'] = quot.css('span small.author::text').get()
            item['tag'] = quot.css('div.tags a.tag::text').getall()
            items.append(item)

        return items
```

Vous allez vers le fichier python «items.py» et specifier les champs que vous voulez extraire comme illustre l’exemple suivant



```python
import scrapy


class ScrapItem(scrapy.Item):
    # define the fields for your item here like:
    title = scrapy.Field()
    author = scrapy.Field()
    tag = scrapy.Field()
```

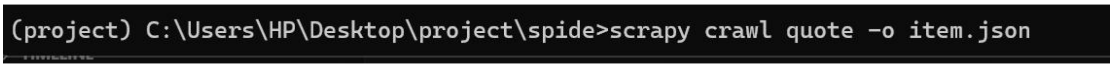
Pour exécuter le spider precèdent utiliser la commande suivante : Scrapy crawl quote



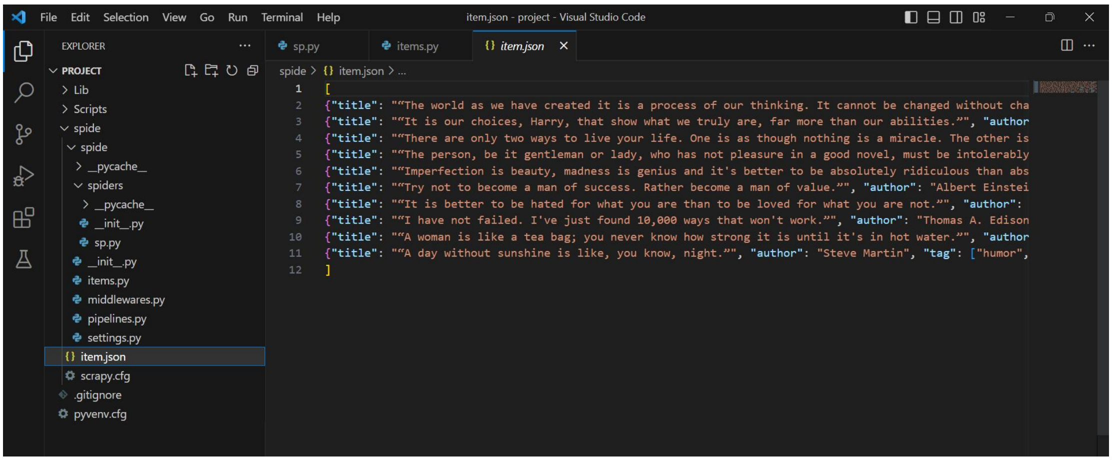
```
{'author': 'Albert Einstein',
 'tag': ['change', 'deep-thoughts', 'thinking', 'world'],
 'title': '“The world as we have created it is a process of our thinking. It '
          'cannot be changed without changing our thinking.”'}
2023-10-14 13:04:37 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'author': 'J.K. Rowling',
 'tag': ['abilities', 'choices'],
 'title': '“It is our choices, Harry, that show what we truly are, far more '
          'than our abilities.”'}
2023-10-14 13:04:37 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'author': 'Albert Einstein',
 'tag': ['inspirational', 'life', 'live', 'miracle', 'miracles'],
 'title': '“There are only two ways to live your life. One is as though '
          'nothing is a miracle. The other is as though everything is a '
          'miracle.”'}
2023-10-14 13:04:37 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'author': 'Jane Austen',
 'tag': ['aliteracy', 'books', 'classic', 'humor'],
 'title': '“The person, be it gentleman or lady, who has not pleasure in a '
          'good novel, must be intolerably stupid.”'}
2023-10-14 13:04:37 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'author': 'Marilyn Monroe',
 'tag': ['be-yourself', 'inspirational'],
 'title': "“Imperfection is beauty, madness is genius and it's better to be "
          'absolutely ridiculous than absolutely boring.”'}
2023-10-14 13:04:37 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
{'author': 'Albert Einstein',
 'tag': ['adulthood', 'success', 'value'],
 'title': '“Try not to become a man of success. Rather become a man of value.”'}
2023-10-14 13:04:37 [scrapy.core.scraper] DEBUG: Scraped from <200 https://quotes.toscrape.com/>
```

Stockage des données :

Par exemple, pour stocker les données dans un fichier JSON, vous pouvez utiliser la commande suivante :



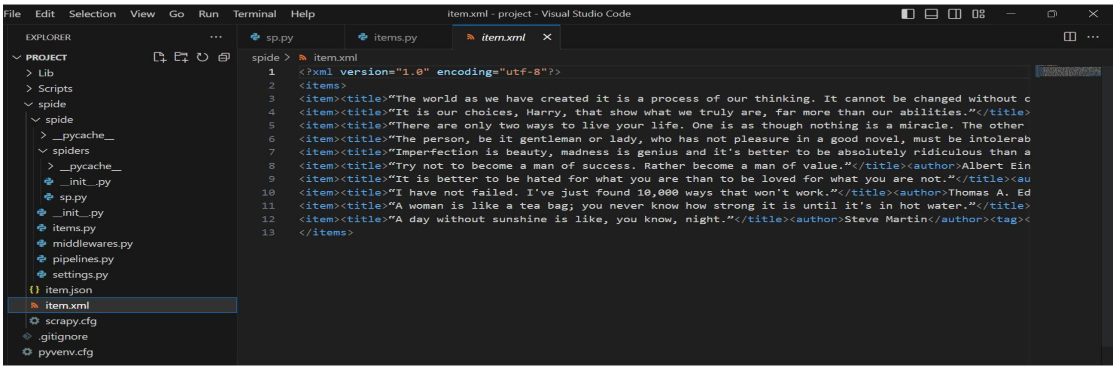
```
(project) C:\Users\HP\Desktop\project\spide>scrapy crawl quote -o item.json
```



```
[
{"title": "“The world as we have created it is a process of our thinking. It cannot be changed without cha
{"title": "“It is our choices, Harry, that show what we truly are, far more than our abilities.”", "author
{"title": "“There are only two ways to live your life. One is as though nothing is a miracle. The other is
{"title": "“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably
{"title": "“Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than abs
{"title": "“Try not to become a man of success. Rather become a man of value.”", "author": "Albert Einstei
{"title": "“It is better to be hated for what you are than to be loved for what you are not.”", "author":
{"title": "“I have not failed. I've just found 10,000 ways that won't work.”", "author": "Thomas A. Edison
{"title": "“A woman is like a tea bag; you never know how strong it is until it's in hot water.”", "author
{"title": "“A day without sunshine is like, you know, night.”", "author": "Steve Martin", "tag": ["humor",
]
```

Vous pouvez stocker les données dans un fichier csv on utilisant la commande suivante et le fichier apparaître a gauche dans le dossier crée

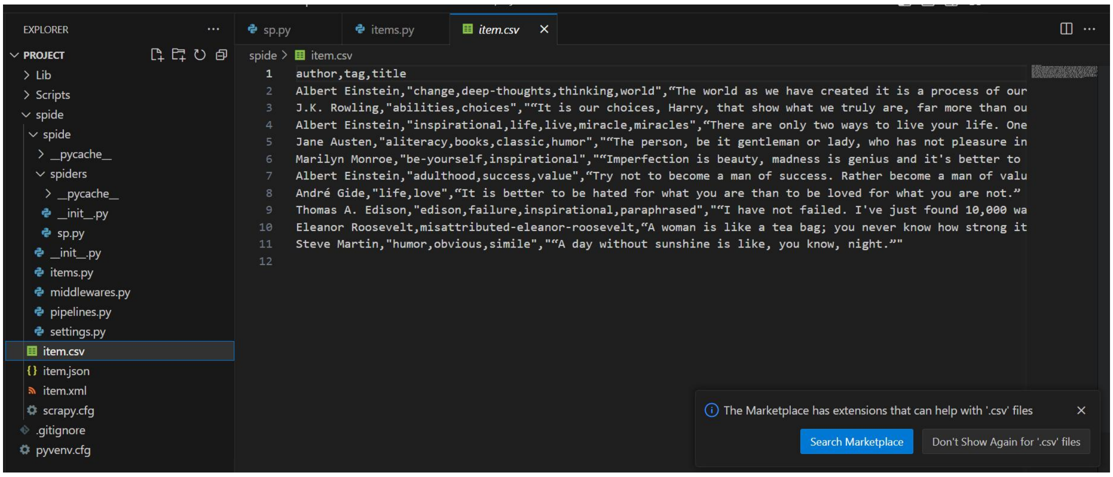
Scrapy crawl quote -o item.csv



```
<?xml version="1.0" encoding="utf-8"?>
<items>
<item><title>“The world as we have created it is a process of our thinking. It cannot be changed without c
<item><title>“It is our choices, Harry, that show what we truly are, far more than our abilities.”</title>
<item><title>“There are only two ways to live your life. One is as though nothing is a miracle. The other
<item><title>“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerab
<item><title>“Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than a
<item><title>“Try not to become a man of success. Rather become a man of value.”</title><author>Albert Ein
<item><title>“It is better to be hated for what you are than to be loved for what you are not.”</title><au
<item><title>“I have not failed. I've just found 10,000 ways that won't work.”</title><author>Thomas A. Ed
<item><title>“A woman is like a tea bag; you never know how strong it is until it's in hot water.”</title>
<item><title>“A day without sunshine is like, you know, night.”</title><author>Steve Martin</author><tag><
</items>
```

Vous pouvez aussi stocker les données dans un fichier xml on utilisant la commande suivante :

Scrapy crawl quote -o item.xml



```
author,tag,title
Albert Einstein,"change,deep-thoughts,thinking,world",“The world as we have created it is a process of our
J.K. Rowling,"abilities,choices","“It is our choices, Harry, that show what we truly are, far more than ou
Albert Einstein,"inspirational,life,live,miracle,miracles",“There are only two ways to live your life. One
Jane Austen,"aliteracy,books,classic,humor","“The person, be it gentleman or lady, who has not pleasure in
Marilyn Monroe,"be-yourself,inspirational","“Imperfection is beauty, madness is genius and it's better to
Albert Einstein,"adulthood,success,value",“Try not to become a man of success. Rather become a man of valu
André Gide,"life,love",“It is better to be hated for what you are than to be loved for what you are not.”
Thomas A. Edison,"edison,failure,inspirational,paraphrased","“I have not failed. I've just found 10,000 wa
Eleanor Roosevelt,misattributed-eleanor-roosevelt,“A woman is like a tea bag; you never know how strong it
Steve Martin,"humor,obvious,simile","“A day without sunshine is like, you know, night.”"
```

The Marketplace has extensions that can help with '.csv' files