Contents

[**WEB SCRAPING USING PYTHON** 2](#_Toc110036157)

[**1.** **Build up spider** 2](#_Toc110036158)

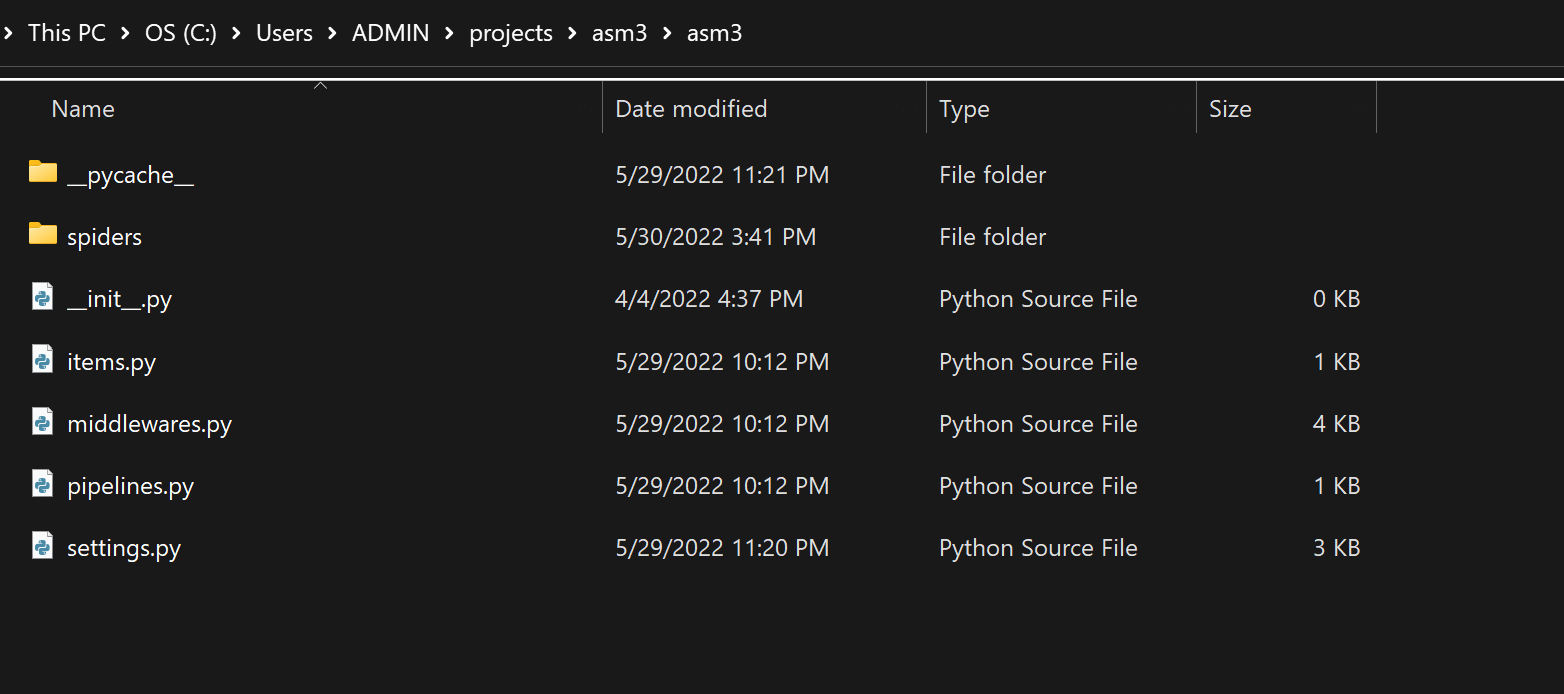
[**2.** **Use spider to handle the switch webpage to get data of previous days** 3](#_Toc110036159)

[**3.** **Get data of Date and Total cases in day** 3](#_Toc110036160)

[**4.** **Archive data on json format** 5](#_Toc110036161)

**WEB SCRAPING USING PYTHON**

1. **Build up spider**

* Run Terminal as ADMINISTRATION
* Create folder: “mkdir projects”
* Go to folder projects just created: “**cd projects**”, and create new project : **scrapy startproject asm3**
* Go to folder contain project asm3: “**cd asm3**”
* Create new spider: **scrapy genspider covid19 web.archive.org/web/20210907023426/https://ncov.moh.gov.vn/vi/web/guest/dong-thoi-gian**
* Open file covid19.py in folder spider to process next

The display of covid19.py

1. **Use spider to handle the switch webpage to get data of previous days**

next\_page = response.xpath("//ul[@class ='lfr-pagination-buttons pager']/li[2]/a/@href").get()

        if next\_page:

            yield scrapy.Request(url=next\_page, callback= self.parse)

1. **Get data of Date and Total cases in day**

* Check the allowed\_domains have to follow format www.
* Call the time value is **time** and number of new case is **newCases**
* Both attributes are sub of ("//div[@class = 'timeline-detail']") 🡪 we take main selector and run loop to get data

time = row.xpath(".//div/h3/text()").get()

newCases = row.xpath(".//div/p[2]/text()").get()

* **Call name of each city correspond to new case**
  + **city\_cases** is a list with whole content of new case follow cities
  + Use function **no\_accent\_vietnamese** to remove string mark
  + Thu nhỏ Reduce string element to get final data

city\_cases = re.findall(regex2,city\_cases)

* + Only select which case same format, then other case with be NULL
  + Extract city name as **city**, and according number of case as **cases** using loop and add to dictionary **res\_dst**

if city\_cases != []:

                city =re.findall(regex3, city\_cases[0])

                cases = re.findall(regex4, city\_cases[0])

                for i in range(len(city)):

                    res\_dst[city[i]] = cases[i]

* + COMPLETE:

class Covid19Spider(scrapy.Spider):

    name = 'covid19'

    allowed\_domains = ['www.web.archive.org/']

    start\_urls = ['https://web.archive.org/web/20210907023426/https://ncov.moh.gov.vn/vi/web/guest/dong-thoi-gian']

    def parse(self, response):

        regex1 = r'\S ([0-9]+.[0-9]+) \S' # Sử dụng regex

        regex2 = r'(TP\s\*.\*\s[\)])'

        regex3 = r'([a-zA-Z]\*\s.\*?)\s[\(]'

        regex4 = r'\d+[0-9.]\*'

        for row in response.xpath("//div[@class = 'timeline-detail']"):

            res\_dst = dict()

            time = row.xpath(".//div/h3/text()").get()

            newCases = row.xpath(".//div/p[2]/text()").get()

            newCases = re.findall(regex1, newCases)

# city\_cases đang ở dạng list

            city\_cases = row.xpath(".//div[@class ='timeline-content']/p[3]/text()").get()

            city\_cases = no\_accent\_vietnamese(city\_cases)

city\_cases = re.findall(regex2,city\_cases)

            if city\_cases != []:

                city =re.findall(regex3, city\_cases[0])

                cases = re.findall(regex4, city\_cases[0])

                for i in range(len(city)):

                    res\_dst[city[i]] = cases[i] # import vào dictionary

                yield {

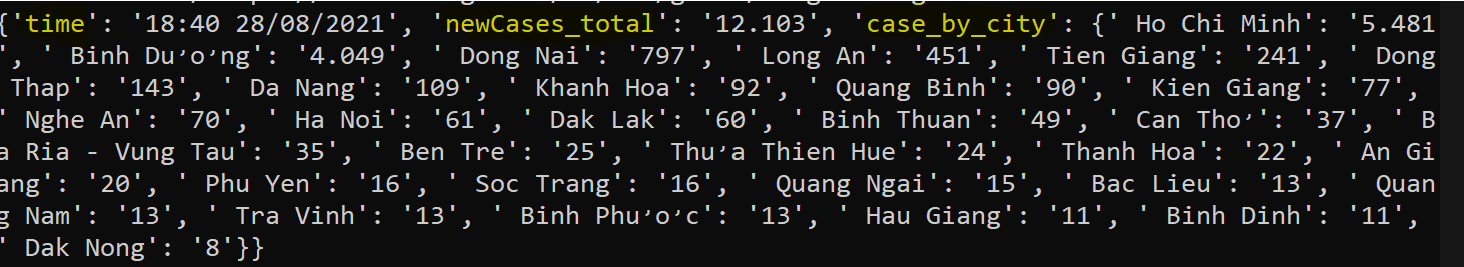
                    'time' : time,

                    'newCases\_total' : newCases[0],

                    'case\_by\_city' : res\_dst # xuất dictionary

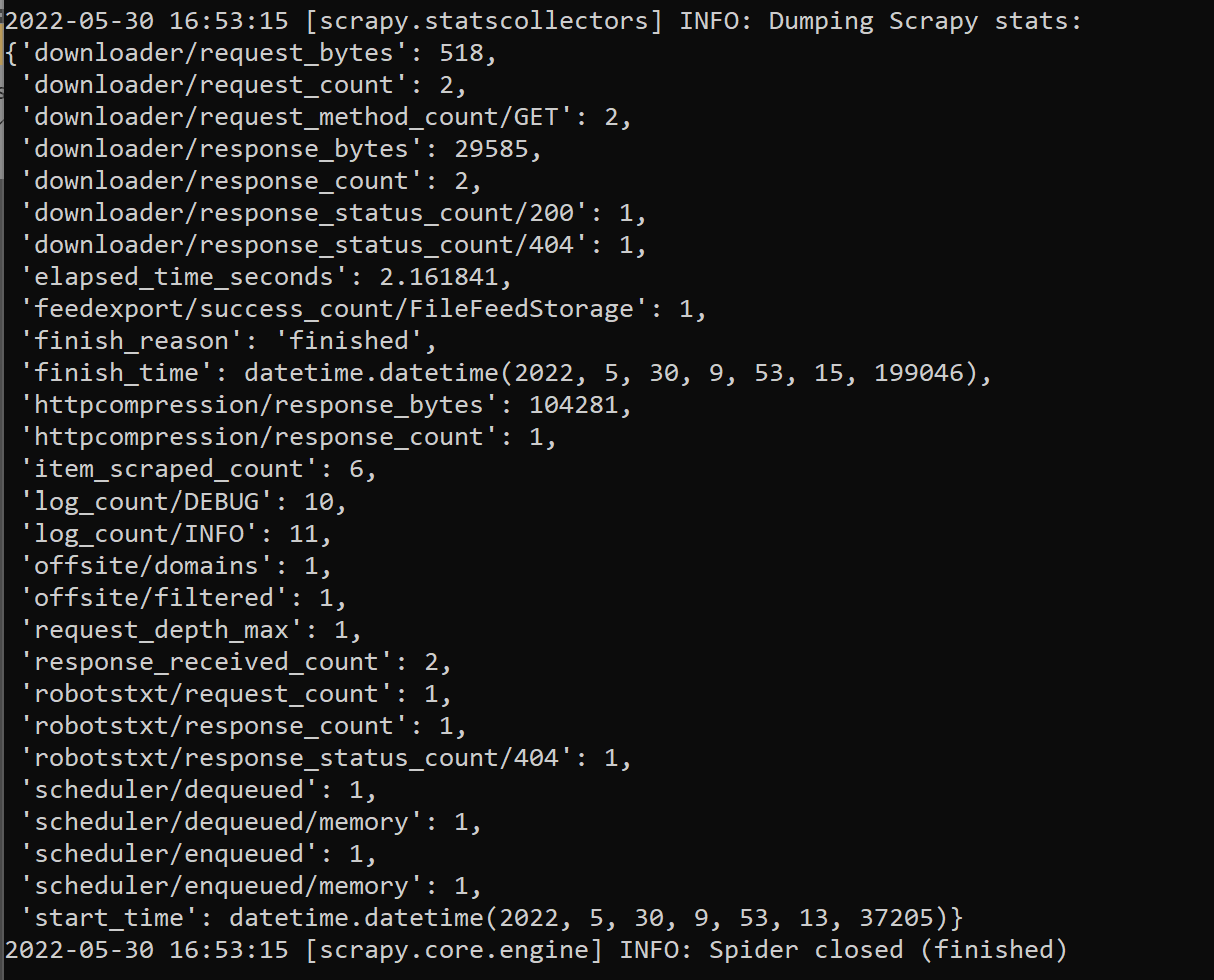
                }

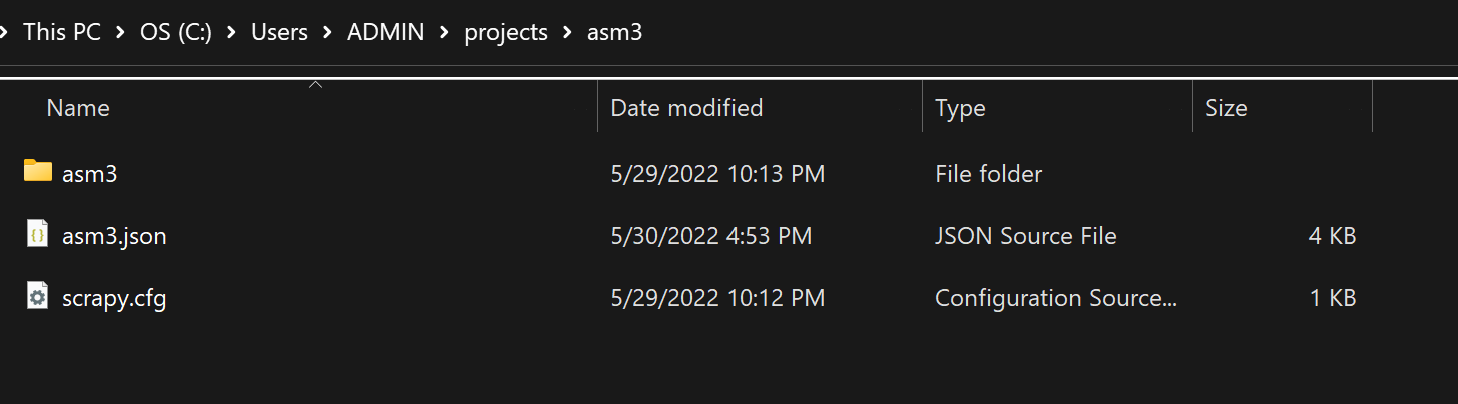
**RESULT**



1. **Archive data on json format**

Script: **“scrapy crawl covid19 -o asm3.json -t json”**

****

****