

---

# Programing

---

## FINAL PROJECT

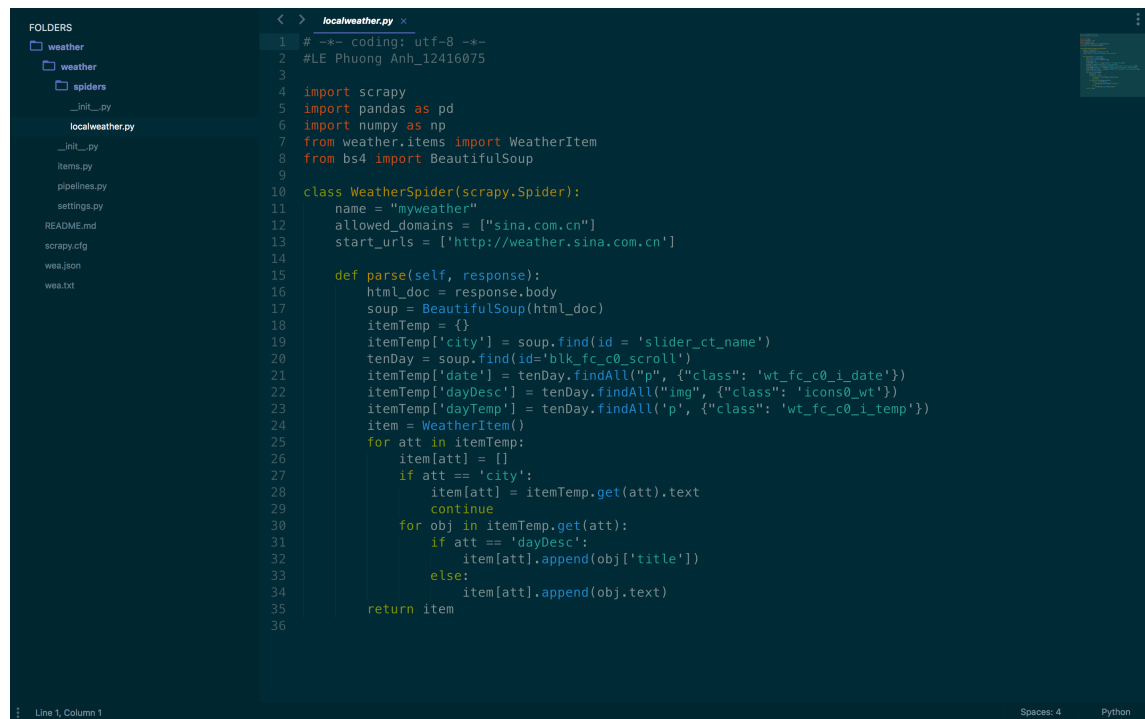
LE PHUONG ANH  
STUDENT ID: 12416075  
RITSUMEIKAN ASIA PACIFIC UNIVERSITY  
LPANH228@GMAIL.COM  
FEBRUARY 5, 2018

# 1 Intro

Web scraping using Python 3 and the BeautifulSoup library.  
A simple weather crawler using scrapy and BeautifulSoup.

# 2 Main

Project git repository at:  
<https://github.com/phuole16/weatherApp>



```
1 # -*- coding: utf-8 -*-
2 #LE Phuong Anh_12416075
3
4 import scrapy
5 import pandas as pd
6 import numpy as np
7 from weather.items import WeatherItem
8 from bs4 import BeautifulSoup
9
10 class WeatherSpider(scrapy.Spider):
11     name = "myweather"
12     allowed_domains = ["sina.com.cn"]
13     start_urls = ['http://weather.sina.com.cn']
14
15     def parse(self, response):
16         html_doc = response.body
17         soup = BeautifulSoup(html_doc)
18         itemTemp = {}
19         itemTemp['city'] = soup.find(id='slider_ct_name')
20         tenDay = soup.find(id='blk_fc_c0_scroll')
21         itemTemp['date'] = tenDay.findAll("p", {"class": 'wt_fc_c0_i_date'})
22         itemTemp['dayDesc'] = tenDay.findAll("img", {"class": 'icons0_wt'})
23         itemTemp['dayTemp'] = tenDay.findAll('p', {"class": 'wt_fc_c0_i_temp'})
24         item = WeatherItem()
25         for att in itemTemp:
26             item[att] = []
27             if att == 'city':
28                 item[att] = itemTemp.get(att).text
29                 continue
30             for obj in itemTemp.get(att):
31                 if att == 'dayDesc':
32                     item[att].append(obj['title'])
33                 else:
34                     item[att].append(obj.text)
35         return item
36
```

Figure 1: Main program.



```
1 |LE Phuong Anh_12416075
2
3
4 import scrapy
5
6
7 class WeatherItem(scrapy.Item):
8     # define the fields for your item here like:
9     # name = scrapy.Field()
10    city = scrapy.Field()
11    date = scrapy.Field()
12    dayDesc = scrapy.Field()
13    dayTemp = scrapy.Field()
14    pass
15
```

Line 1, Column 1: Detect Indentation: Setting indentation to 4 spaces

Tab Size: 4 Python

Figure 3: Weather Item

```
1 |
2 |
3 | BOT_NAME = 'weather'
4 |
5 | SPIDER_MODULES = ['weather.spiders']
6 | NEWSPIIDER_MODULE = 'weather.spiders'
7 | ITEM_PIPELINES = {'weather.pipelines.WeatherPipeline':1}
8 | # Crawl responsibly by identifying yourself (and your website) on the user-agent
9 | #USER_AGENT = 'weather (+http://www.yourdomain.com)'
10 | USER_AGENT = 'User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)'
11 |
12 | DEFAULT_REQUEST_HEADERS = {
13 |     'Referer': 'http://www.weibo.com'
14 | }
15 |
16 | ITEM_PIPELINES = {
17 |     'weather.pipelines.WeatherPipeline': 1
18 | }
19 |
20 | DOWNLOAD_DELAY = 0.5
21 |
```

Line 1, Column 1

Tab Size: 4

Python

Figure 4: Setting

```
< > Labwork.tex × wea.txt ×
1 city:漳
2
3 data:1-29      day:多云(29°C )      night:晴( 17°C)
4
5 data:1-30      day:多云(28°C )      night:晴( 18°C)
6
7 data:1-31      day:晴(29°C )       night:晴( 19°C)
8
9 data:2-1       day:晴(27°C )       night:晴( 20°C)
10
11 data:2-2      day:晴(27°C )       night:多云( 21°C)
12
13 data:2-3      day:多云(29°C )      night:晴( 19°C)
14
15 data:2-4      day:局部多云(29°C )  night:局部多云( 19°C)
16
17 data:2-5      day:少云(29°C )       night:局部多云( 20°C)
18
19 data:2-6      day:局部多云(28°C )  night:局部多云( 20°C)
20
21 data:2-7      day:晴(27°C )       night:晴( 20°C)
22
```

Figure 5: Output