

实验 1：金融数据获取实验报告

Python环境安装

课程前已完成python3.7的安装与环境配置

IDE环境配置

课程前已完成pycharm安装

遇到的问题

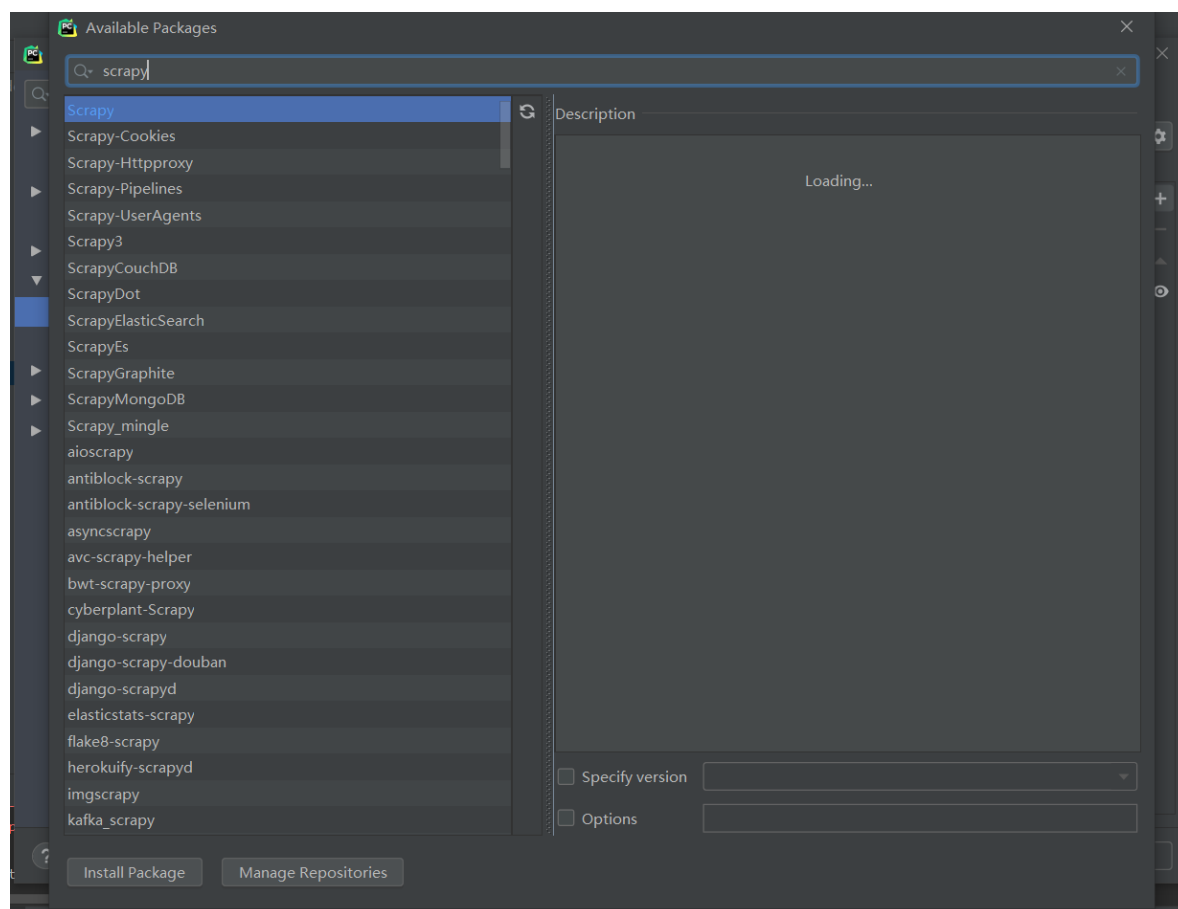
pycharm于去年暑假小学期（课程综合实践 I）安装，到现在正好一年，学生一年免费已到期

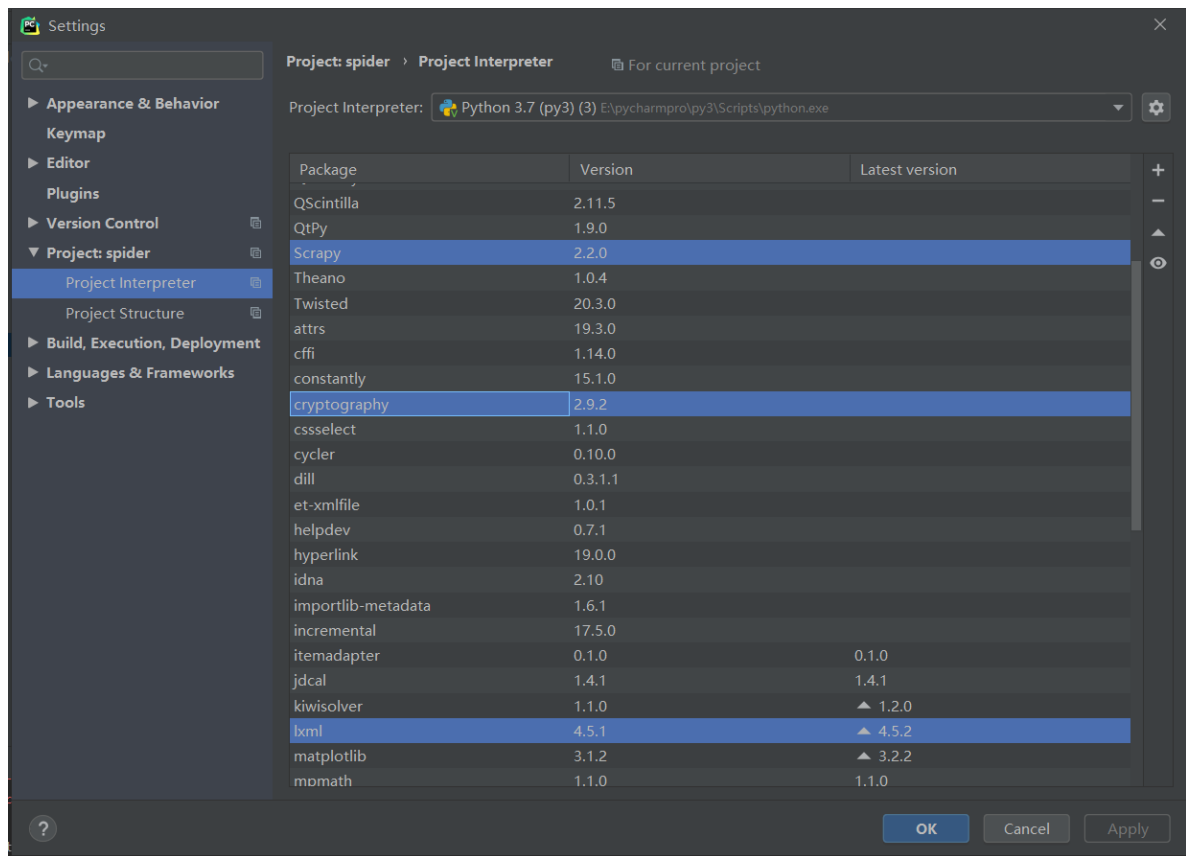
解决方法

更新学生证明失败，Jetbrains认为zju.edu.cn邮箱无效；搜索得知最近教育邮箱更新学生证明都遇到问题。尝试通过github学生包，一小时后未收到邮件，怀疑未通过机器审核，需要等待人工审核。由于时间关系搜索了序列码解决问题。

Scrapy框架安装

通过pycharm直接安装





可以看到安装scrapy的同时安装了需要的cryptography, lxml等其他包。

爬虫Demo编写

在terminal运行“scrapy startproject tutorial”语句，在spider目录下新建quotes_spider.py

```
import scrapy

class QuotesSpider(scrapy.Spider):
    name = "quotes"

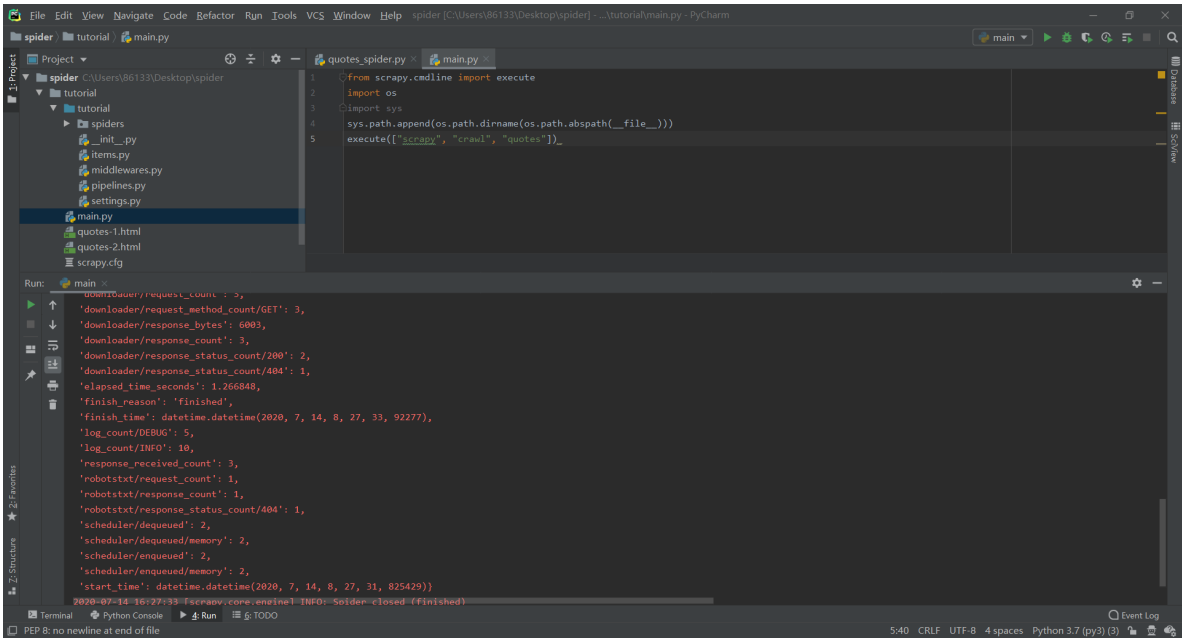
    def start_requests(self):
        urls = [
            'http://quotes.toscrape.com/page/1/',
            'http://quotes.toscrape.com/page/2/',
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = 'quotes-%s.html' % page
        with open(filename, 'wb') as f:
            f.write(response.body)
        self.log('Saved file %s' % filename)
```

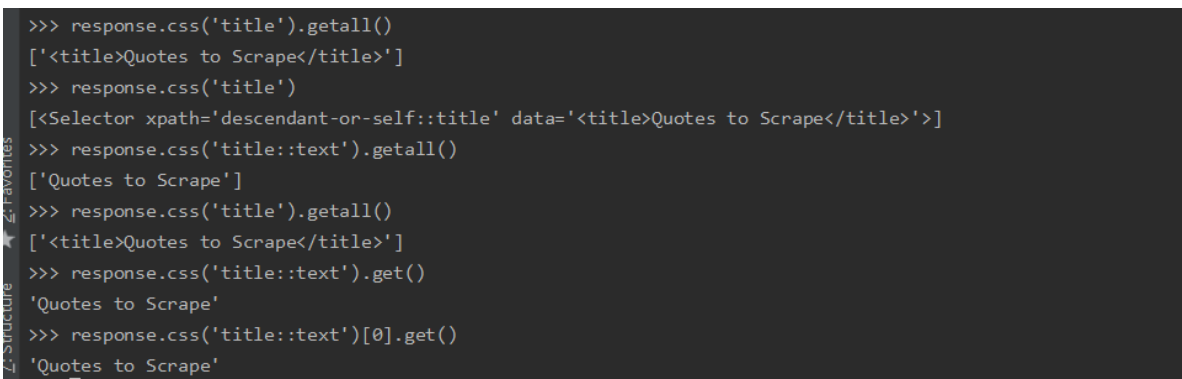
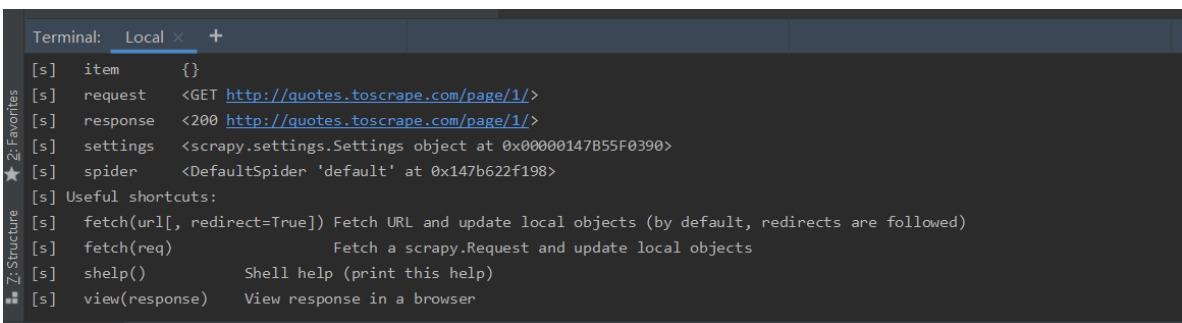
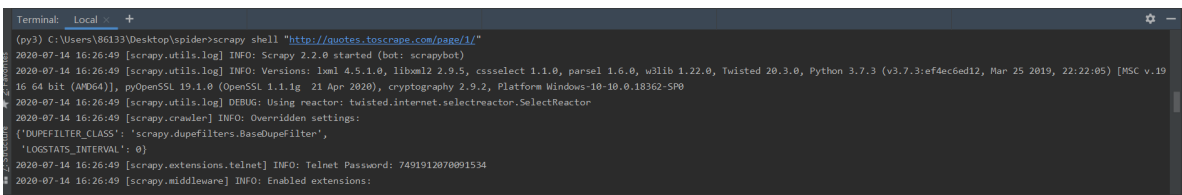
在tutorial目录（和cfg文件同目录）新建mian.py

```
from scrapy.cmdline import execute
import os
import sys
sys.path.append(os.path.dirname(os.path.abspath(__file__)))
execute(["scrapy", "crawl", "quotes"])
```

运行main, 可以看到运行结果和创建的两个文件



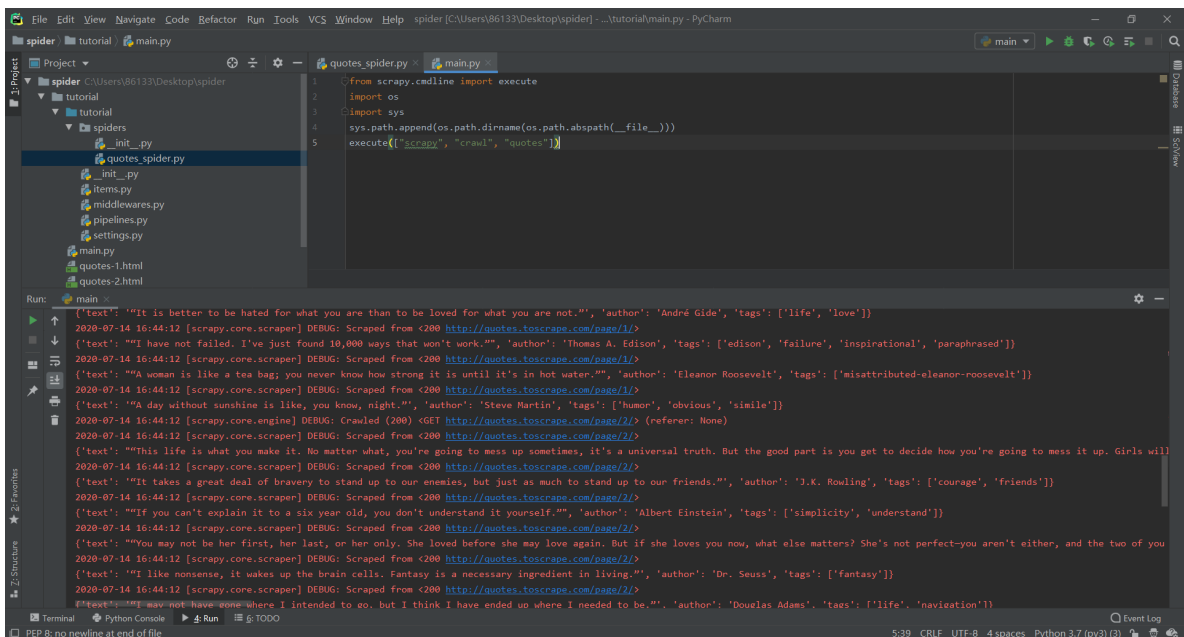
在terminal按照教程测试抓取结果



```
>>> response.css('title::text').re(r'Quotes.*')
['Quotes to Scrape']
>>> response.css('title::text').re(r'Q\w+')
['Quotes']
>>> response.css('title::text').re(r'(\w+) to (\w+)')
['Quotes', 'Scrape']
```

```
>>> quote = response.css("div.quote")[0]
>>> text = quote.css("span.text::text").get()
>>> text
""The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.""
>>> author = quote.css("small.author::text").get()
>>> author
'Albert Einstein'
>>> tags = quote.css("div.tags a.tag::text").getall()
>>> tags
['change', 'deep-thoughts', 'thinking', 'world']
```

```
>>> for quote in response.css("div.quote"):
...     text = quote.css("span.text::text").get()
...     File "<console>", line 2
...         text = quote.css("span.text::text").get()
...         ^
...     IndentationError: expected an indented block
>>> for quote in response.css("div.quote"):
...     text = quote.css("span.text::text").get()
...     author = quote.css("small.author::text").get()
...     tags = quote.css("div.tags a.tag::text").getall()
...     print(dict(text=text, author=author, tags=tags))
...
('text': '"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."', 'author': 'Albert Einstein', 'tags': ['change', 'deep-thoughts', 'thinking', 'world'])
('text': '"It is our choices, Harry, that show what we truly are, far more than our abilities."', 'author': 'J.K. Rowling', 'tags': ['abilities', 'choices'])
('text': '"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."', 'author': 'Albert Einstein', 'tags': ['inspirational', 'life', 'live', 'miracle', 'miracles'])
('text': '"The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid."', 'author': 'Jane Austen', 'tags': ['alterity', 'books', 'classic', 'humor'])
('text': '"Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring."', 'author': 'Marilyn Monroe', 'tags': ['be-yourself', 'inspirational'])
('text': '"Try not to become a man of success. Rather become a man of value."', 'author': 'Albert Einstein', 'tags': ['adulthood', 'success', 'value'])
('text': '"It is better to be hated for what you are than to be loved for what you are not."', 'author': 'André Gide', 'tags': ['life', 'love'])
('text': '"I have not failed. I've just found 10,000 ways that won't work."', 'author': 'Thomas A. Edison', 'tags': ['edison', 'failure', 'inspirational', 'paraphrased'])
('text': '"A woman is like a tea bag; you never know how strong it is until it's in hot water."', 'author': 'Eleanor Roosevelt', 'tags': ['misattributed-eleanor-roosevelt'])
('text': '"A day without sunshine is like, you know, night."', 'author': 'Steve Martin', 'tags': ['humor', 'obvious', 'simile'])
```



更改quotes_spider.py内容

```
import scrapy

class QuotesSpider(scrapy.Spider):
    name = "quotes"
    start_urls = [
        'http://quotes.toscrape.com/page/1/',
        'http://quotes.toscrape.com/page/2/',
    ]
```

```
def parse(self, response):
    for quote in response.css('div.quote'):
        yield {
            'text': quote.css('span.text::text').get(),
            'author': quote.css('small.author::text').get(),
            'tags': quote.css('div.tags a.tag::text').getall(),
        }
```

更改main.py内容:

```
from scrapy.cmdline import execute
import os
import sys
sys.path.append(os.path.dirname(os.path.abspath(__file__)))
execute(["scrapy", "crawl", "quotes", "-o", "quotes.json"])
```

可以看到运行结果和新建立的json文件

```
File Edit View Navigate Code Refactor Run Tools VCS Window Help spider [C:\Users\86133\Desktop\spider] - ...tutorial\main.py - PyCharm
Project: spider tutorial main.py
spiders
├── __init__.py
├── quotes_spider.py
├── __init__.py
├── items.py
├── middlewares.py
├── pipelines.py
├── settings.py
├── main.py
├── quotes.json
├── quotes-1.html
├── quotes-2.html
└── scrapy.cfg
External Libraries
Run: main
{
  'elapsed_time_seconds': 1.250425,
  'finish_reason': 'finished',
  'finish_time': datetime.datetime(2020, 7, 14, 8, 46, 29, 449180),
  'item_scraped_count': 20,
  'log_count/DEBUG': 23,
  'log_count/INFO': 11,
  'response_received_count': 3,
  'robotstxt/request_count': 1,
  'robotstxt/response_count': 1,
  'robotstxt/response_status_count/404': 1,
  'scheduler/dequeued': 2,
  'scheduler/dequeued/memory': 2,
  'scheduler/enqueued': 2,
  'scheduler/enqueued/memory': 2,
  'start_time': datetime.datetime(2020, 7, 14, 8, 46, 28, 198755)}
2020-07-14 16:46:29 [scrapy.core.engine] INFO: Spider closed (finished)
Process finished with exit code 0
```

```
quotes_spider.py main.py quotes.json
1 [{"text": "\u201cThe world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.\u201d", "author": "Albert Einstein", "tags": ["thinking", "world"]}]
2 [{"text": "\u201cIt is our choices, Harry, that show what we truly are, far more than our abilities.\u201d", "author": "J.K. Rowling", "tags": ["choices", "abilities"]}]
3 [{"text": "\u201cThere 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.\u201d", "author": "C.G. Jung", "tags": ["miracle", "life"]}]
4 [{"text": "\u201cThe person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.\u201d", "author": "Jane Austen", "tags": ["novel", "stupid"]}]
5 [{"text": "\u201cImperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.\u201d", "author": "Mark Twain", "tags": ["imperfection", "beauty", "madness", "genius", "ridiculous", "boring"]}]
6 [{"text": "\u201cTry not to become a man of success. Rather become a man of value.\u201d", "author": "Albert Einstein", "tags": ["success", "value"]}]
7 [{"text": "\u201cIt is better to be hated for what you are than to be loved for what you are not.\u201d", "author": "Andr\u00e9 Gide", "tags": ["hated", "loved"]}]
8 [{"text": "\u201cI have not failed. I've just found 10,000 ways that won't work.\u201d", "author": "Thomas A. Edison", "tags": ["failure", "ways"]}]
9 [{"text": "\u201cA woman is like a tea bag; you never know how strong it is until it's in hot water.\u201d", "author": "Eleanor Roosevelt", "tags": ["woman", "tea bag", "strong"]}]
10 [{"text": "\u201cA day without sunshine is like, you know, night.\u201d", "author": "Steve Martin", "tags": ["humor", "obvious", "simile"]}]
11 [{"text": "\u201cThis life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth. But the good part is you get to decide how you're going to mess up.\u201d", "author": "Woody Allen", "tags": ["life", "mess up", "universal truth"]}]
12 [{"text": "\u201cIt takes a great deal of bravery to stand up to our enemies, but just as much to stand up to our friends.\u201d", "author": "J.K. Rowling", "tags": ["bravery", "enemies", "friends"]}]
13 [{"text": "\u201cIf you can't explain it to a six year old, you don't understand it yourself.\u201d", "author": "Albert Einstein", "tags": ["simplicity", "understanding"]}]
14 [{"text": "\u201cYou may not be her first, her last, or her only. She loved before she may love again. But if she loves you now, what else matters?\u201d", "author": "Dr. Seuss", "tags": ["love", "matters"]}]
15 [{"text": "\u201cI like nonsense, it wakes up the brain cells. Fantasy is a necessary ingredient in living.\u201d", "author": "Dr. Seuss", "tags": ["nonsense", "fantasy", "living"]}]
16 [{"text": "\u201cI may not have gone where I intended to go, but I think I have ended up where I needed to be.\u201d", "author": "Douglas Adams", "tags": ["intention", "need"]}]
17 [{"text": "\u201cThe opposite of love is not hate, it's indifference. The opposite of art is not ugliness, it's indifference. The opposite of faith is not disbelief, it's indifference.\u201d", "author": "Friedrich Nietzsche", "tags": ["love", "hate", "indifference", "art", "faith", "disbelief"]}]
18 [{"text": "\u201cIt is not a lack of love, but a lack of friendship that makes unhappy marriages.\u201d", "author": "Friedrich Nietzsche", "tags": ["love", "friendship", "marriages"]}]
19 [{"text": "\u201cGood friends, good books, and a sleepy conscience: this is the ideal life.\u201d", "author": "Mark Twain", "tags": ["books", "conscience", "ideal life"]}]
20 [{"text": "\u201cLife is what happens to us while we are making other plans.\u201d", "author": "Allen Saunders", "tags": ["fate", "life", "plans"]}]
```

遇到的问题

运行main程序报错: no active project

Unknown command: crawl

问题分析

通过查找资料发现应该是根目录问题，尝试：

1.cmd运行

显示scrapy不是可以运行的指令

2.设置根目录

没有改变

3.在pycharm内为main配置config文件

报错变为找不到quotes文件

解决办法

考虑到查找的资料过多，尝试的方法可能互相影响造成了更大的混乱；在建立工程的过程中可能也发生了错误。考虑重装pycharm，重新建立工程，最终顺利解决了问题。

【不过最后思考了一下认为是第一次建工程可能没注意quotes_spider建立的目录位置。XD】

bonus：抓取网贷之家信息存入mysql数据库

按照上述方法获取网贷之家-数据网页（“<https://shuju.wdzj.com/>”）html信息，用chorm打开

The screenshot shows a web browser displaying a table of financial data. The table has columns for company names, various financial metrics, and a '追踪' (Tracking) column. The right side of the image shows the Chrome DevTools 'Elements' panel, which is open to the 'body' element. The 'Styles' pane is visible, showing the default 'display: block' style for the 'body' element. The 'Filter' input is empty, and the 'Show all' button is visible.

公司	2017.35	6.65	3.31	140932.61	追踪
汇金	2017.35	6.65	3.31	140932.61	追踪
金股	1315.26	8.78	11.47	880735	追踪
凤凰	615.97	9.37	11.85	160575.81	追踪
智信	535.47	7.48	4.08	86234.2	追踪
融贝	474.66	8.62	11.79	23941.71	追踪
网	340.95	7.49	4.41	95758.06	追踪
白	279.38	11.77	19.52	229013.98	追踪
菜	157.86	9.58	6.54	38753.96	追踪
金融					
合众					
众					
普					
次					
出					
借					
退					
高					
返					
利					
380					
元					
2020-07-14					
开始					
的数					
据更					
新不					
完整					
连接					
错误					
参考					
博金					
和信					
爱投					
金融					

在右边界面选择要抓取的内容（公司名称以及其四个相关数据），右键选择复制选择器，得到css selector表达式，根据得到的表达式编写爬虫代码：

```
import scrapy

class QuotesSpider(scrapy.Spider):
    name = "quotes"
    start_urls = [
        'https://shuju.wdzj.com/',
    ]

    def parse(self, response):
```

```

for company in response.css('#platTable > tr'):
    yield {
        'name': company.css('td:nth-child(8) > div::attr(data-
platname)').get(),
        'money': company.css('td:nth-child(3) > div::text').get(),
        'ben': company.css('td:nth-child(4) > div::text').get(),
        'time': company.css('td:nth-child(5) > div::text').get(),
        'waiting': company.css('td:nth-child(6) > div::text').get(),
    }

```

试着讲读到的数据存入json文件，检查前半部分工作是否正确，从下图json文件截图可以看出，虽然由于utf-8显示不出部分中文字符，但是对照网页可以确认抓取信息正确。

```

1 [{"name": "91\u65fa\u8d22", "money": "7835.34", "ben": "12", "time": "1.33", "waiting": "143010.98"},
2 {"name": "\u7ffc\u9f99\u8d37", "money": "2906.38", "ben": "8.97", "time": "5.2", "waiting": "1028153.3"},
3 {"name": "\u6c47\u76c8\u91d1\u670d", "money": "2017.35", "ben": "6.65", "time": "3.31", "waiting": "140932.61"},
4 {"name": "\u51e4\u51f0\u667a\u4fe1", "money": "1315.26", "ben": "8.78", "time": "11.47", "waiting": "880735"},
5 {"name": "\u878d\u8d1d\u7f51", "money": "615.97", "ben": "9.37", "time": "11.85", "waiting": "160575.81"},
6 {"name": "\u767d\u83dc\u91d1\u878d", "money": "535.47", "ben": "7.48", "time": "4.08", "waiting": "86234.2"},
7 {"name": "\u5408\u4f17\u8d37", "money": "474.66", "ben": "8.62", "time": "11.79", "waiting": "23941.71"},
8 {"name": "\u535a\u91d1\u8d37", "money": "340.95", "ben": "7.49", "time": "4.41", "waiting": "95758.06"},
9 {"name": "\u548c\u4fe1\u8d37", "money": "279.38", "ben": "11.77", "time": "19.52", "waiting": "229013.98"},
10 {"name": "\u7231\u6295\u91d1\u878d", "money": "157.86", "ben": "9.58", "time": "6.54", "waiting": "38753.96"},
11 {"name": "\u5411\u4e0a\u91d1\u670d", "money": "8.47", "ben": "11.18", "time": "7.82", "waiting": "94094"},
12 {"name": "\u70b9\u878d", "money": "0", "ben": "0", "time": "0", "waiting": "197682.89"},
13 {"name": "\u5b89\u5fc3\u6295", "money": "0", "ben": "0", "time": "0", "waiting": "4648.36"},
14 {"name": "\u5fae\u8d37\u7f51", "money": "0", "ben": "0", "time": "0", "waiting": "610513.27"},
15 {"name": "\u96c6\u5229\u8d22\u5bcc", "money": "0", "ben": "0", "time": "0", "waiting": "5913.01"},
16 {"name": "\u7231\u94b1\u8fdb", "money": "0", "ben": "0", "time": "0", "waiting": "1152387.28"},
17 {"name": "\u878d\u8d44\u6613", "money": "0", "ben": "0", "time": "0", "waiting": "0"},
18 {"name": "\u94b1\u76c6\u7f51", "money": "0", "ben": "0", "time": "0", "waiting": "10254.85"},
19 {"name": "\u6d0b\u94b1\u7f50", "money": "0", "ben": "0", "time": "0", "waiting": "60989.27"},
20 ]

```

更改pipeline代码，连接数据库

```

import pymysql

class TutorialPipeline(object):
    def __init__(self):
        # connection database
        self.connect = pymysql.connect("localhost","root","jkwry4s45889","wdzj"
)

        # get cursor
        self.cursor = self.connect.cursor()
        print("连接数据库成功")

    def process_item(self, item, spider):
        # sql语句
        insert_sql = """
        insert into company(name, money, ben, time, waiting) VALUES
        (%s,%s,%s,%s,%s)
        """
        # 执行插入数据到数据库操作
        self.cursor.execute(insert_sql, (item['name'], item['money'],
item['ben'], item['time'],
item['waiting']))

        # 提交，不进行提交无法保存到数据库
        self.connect.commit()

    def close_spider(self, spider):

```

```
# 关闭游标和连接
self.cursor.close()
self.connect.close()
```

更改setting代码:

```
BOT_NAME = 'tutorial'

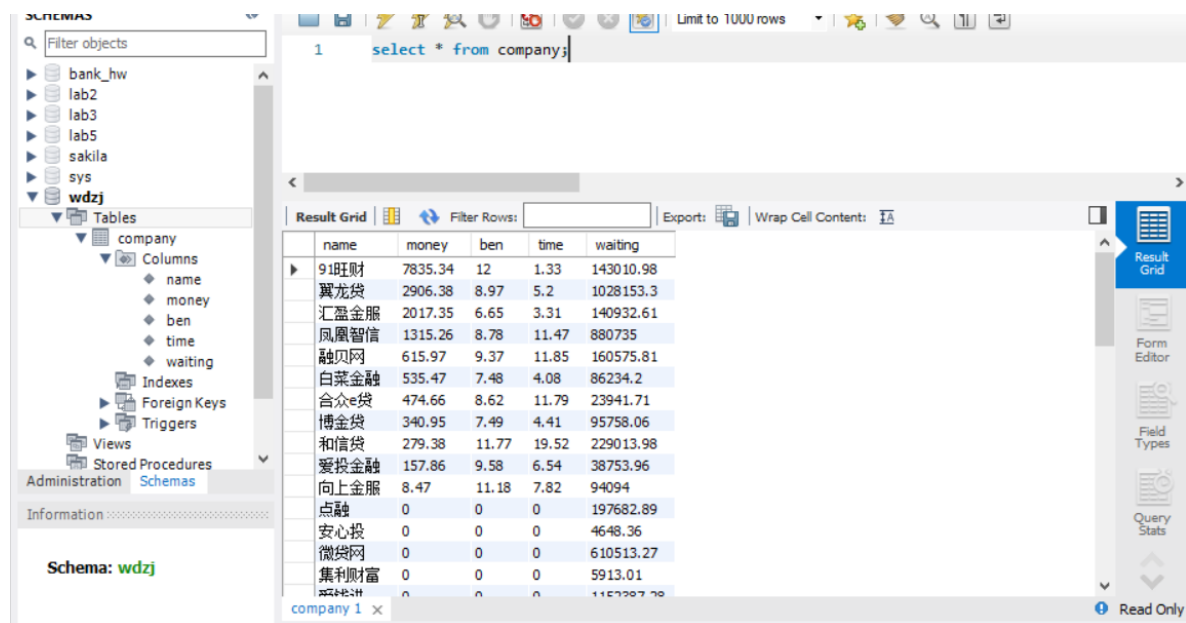
SPIDER_MODULES = ['tutorial.spiders']
NEWSPIDER_MODULE = 'tutorial.spiders'

ITEM_PIPELINES = {
    'pipelines.TutorialPipeline': 200,
}

ROBOTSTXT_OBEY = True
USER_AGENT = 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'
```

在数据库中建立新的数据库和表格, 运行爬虫程序。

在mysql中查询表格内容, 确认抓取成功



	name	money	ben	time	waiting
▶	91旺财	7835.34	12	1.33	143010.98
	翼龙贷	2906.38	8.97	5.2	1028153.3
	汇盈金服	2017.35	6.65	3.31	140932.61
	凤凰智信	1315.26	8.78	11.47	880735
	融贝网	615.97	9.37	11.85	160575.81
	白菜金融	535.47	7.48	4.08	86234.2
	合众e贷	474.66	8.62	11.79	23941.71
	博金贷	340.95	7.49	4.41	95758.06
	和信贷	279.38	11.77	19.52	229013.98
	爱投金融	157.86	9.58	6.54	38753.96
	向上金服	8.47	11.18	7.82	94094
	点融	0	0	0	197682.89
	安心投	0	0	0	4648.36
	微贷网	0	0	0	610513.27
	集利财富	0	0	0	5913.01
	众安贷	0	0	0	115287.78

遇到的问题

- 1.网贷之家网站设置了反爬虫, 无法直接爬取
- 2.mysql远程连接失败, 显示localhost没有权限连接

解决方法

- 1.在setting中增加“USER_AGENT = 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'”语句
- 2.多次restart mysql