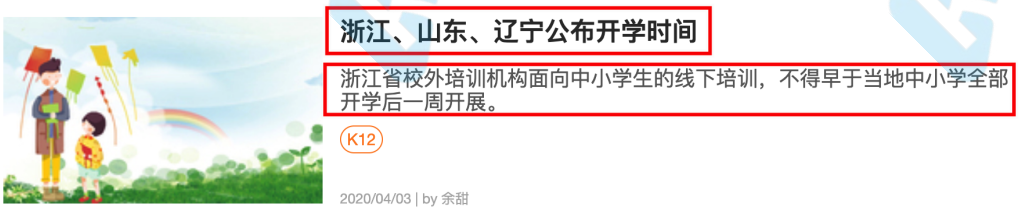
爬虫scrapy框架练习答案

题目要求

- 根据已知网址爬取数据,将本网址的第一页数据使用scrapy框架爬取到本地并写入excel表格中，具体爬取数据位置见图片，红色框部分是需要爬取的数据部分。

> http://www.duozhi.com/



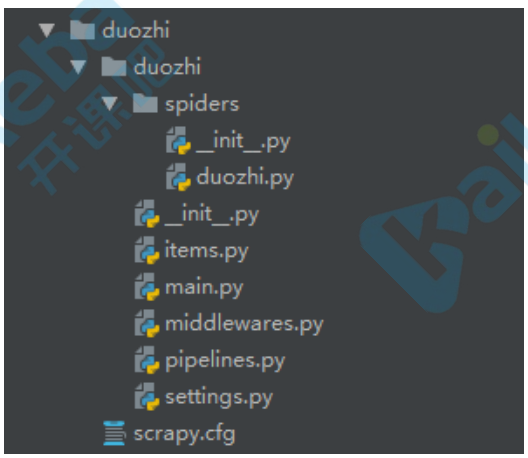
温馨提示：
1.用命令来创建项目
2.settings里面要修改配置

解题步骤：

1、创建项目

使用命令创建即可

```
scrapy
```



settings.py文件

```
# -*- coding: utf-8 -*-
# Scrapy settings for duozhi project
```

```
#
# For simplicity, this file contains only settings considered important or
# commonly used. You can find more settings consulting the documentation:
#
#     https://docs.scrapy.org/en/latest/topics/settings.html
#     https://docs.scrapy.org/en/latest/topics/downloader-middleware.html
#     https://docs.scrapy.org/en/latest/topics/spider-middleware.html
BOT_NAME = 'duozhi'
SPIDER_MODULES = ['duozhi.spiders']
NEWSPIDER_MODULE = 'duozhi.spiders'
# Crawl responsibly by identifying yourself (and your website) on the
user-agent
# USER_AGENT = 'duozhi (+http://www.yourdomain.com)'
# Obey robots.txt rules
ROBOTSTXT_OBEY = False
# Configure maximum concurrent requests performed by Scrapy (default: 16)
# CONCURRENT_REQUESTS = 32
# Configure a delay for requests for the same website (default: 0)
# See https://docs.scrapy.org/en/latest/topics/settings.html#download-
delay
# See also autothrottle settings and docs
# DOWNLOAD_DELAY = 3
# The download delay setting will honor only one of:
# CONCURRENT_REQUESTS_PER_DOMAIN = 16
# CONCURRENT_REQUESTS_PER_IP = 16
# Disable cookies (enabled by default)
# COOKIES_ENABLED = False
# Disable Telnet Console (enabled by default)
# TELNETCONSOLE_ENABLED = False
# Override the default request headers:
# DEFAULT_REQUEST_HEADERS = {
#   'Accept':
'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
#   'Accept-Language': 'en',
# }
# Enable or disable spider middlewares
# See https://docs.scrapy.org/en/latest/topics/spider-middleware.html
# SPIDER_MIDDLEWARES = {
#    'duozhi.middlewares.DuozhiSpiderMiddleware': 543,
# }
# Enable or disable downloader middlewares
# See https://docs.scrapy.org/en/latest/topics/downloader-middleware.html
# DOWNLOADER_MIDDLEWARES = {
#    'duozhi.middlewares.DuozhiDownloaderMiddleware': 543,
# }
# Enable or disable extensions
# See https://docs.scrapy.org/en/latest/topics/extensions.html
# EXTENSIONS = {
#    'scrapy.extensions.telnet.TelnetConsole': None,
```

```
# }
# Configure item pipelines
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
    'duozhi.pipelines.DuozhiPipeline': 300,
}
# Enable and configure the AutoThrottle extension (disabled by default)
# See https://docs.scrapy.org/en/latest/topics/autothrottle.html
# AUTOTHROTTLE_ENABLED = True
# The initial download delay
# AUTOTHROTTLE_START_DELAY = 5
# The maximum download delay to be set in case of high latencies
# AUTOTHROTTLE_MAX_DELAY = 60
# The average number of requests Scrapy should be sending in parallel to
# each remote server
# AUTOTHROTTLE_TARGET_CONCURRENCY = 1.0
# Enable showing throttling stats for every response received:
# AUTOTHROTTLE_DEBUG = False
# Enable and configure HTTP caching (disabled by default)
# See https://docs.scrapy.org/en/latest/topics/downloader-middleware.html#httpcache-middleware-settings
# HTTPCACHE_ENABLED = True
# HTTPCACHE_EXPIRATION_SECS = 0
# HTTPCACHE_DIR = 'httpcache'
# HTTPCACHE_IGNORE_HTTP_CODES = []
# HTTPCACHE_STORAGE = 'scrapy.extensions.httpcache.FilesystemCacheStorage'
```

duozhi.py文件

```
# -*- coding: utf-8 -*-
import scrapy
import bs4
from ..items import DuozhiItem
class DangdangBookSpider(scrapy.Spider):
    name = 'duozhi'
    allowed_domains = ['duozhi.com']
    start_urls = ['http://www.duozhi.com/']
    def parse(self, response):
        soup = bs4.BeautifulSoup(response.text, 'html.parser')
        elements = soup.find_all('div', class_="post-item")
        for element in elements:
            title = element.find('a',class_='post-title').text
            content = element.find('p',class_='post-desc').text
            item = DuozhiItem()
            item['title'] = title
            item['content'] = content
            yield item
```

items文件

```
# -*- coding: utf-8 -*-
```

```
# Define here the models for your scraped items
#
# See documentation in:
# https://docs.scrapy.org/en/latest/topics/items.html
import scrapy
class DuozhiItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    title = scrapy.Field()
    content = scrapy.Field()
```

piplines.py文件

```
# -*- coding: utf-8 -*-
# Define your item pipelines here
#
# Don't forget to add your pipeline to the ITEM_PIPELINES setting
# See: https://docs.scrapy.org/en/latest/topics/item-pipeline.html
import openpyxl
class DuozhiPipeline(object):
    def __init__(self):
        self.wb = openpyxl.Workbook()
        self.ws = self.wb.active
        self.ws.append(['标题', '内容'])
    def process_item(self, item, spider):
        line = [item['title'], item['content']]
        self.ws.append(line)
        return item
    def close_spider(self, spider):
        self.wb.save('duozhi.xlsx')
        self.wb.close()
```

main.py文件

```
from scrapy import cmdline
cmdline.execute(['scrapy', 'crawl', 'duozhi'])
```

其他文件里面不需要动