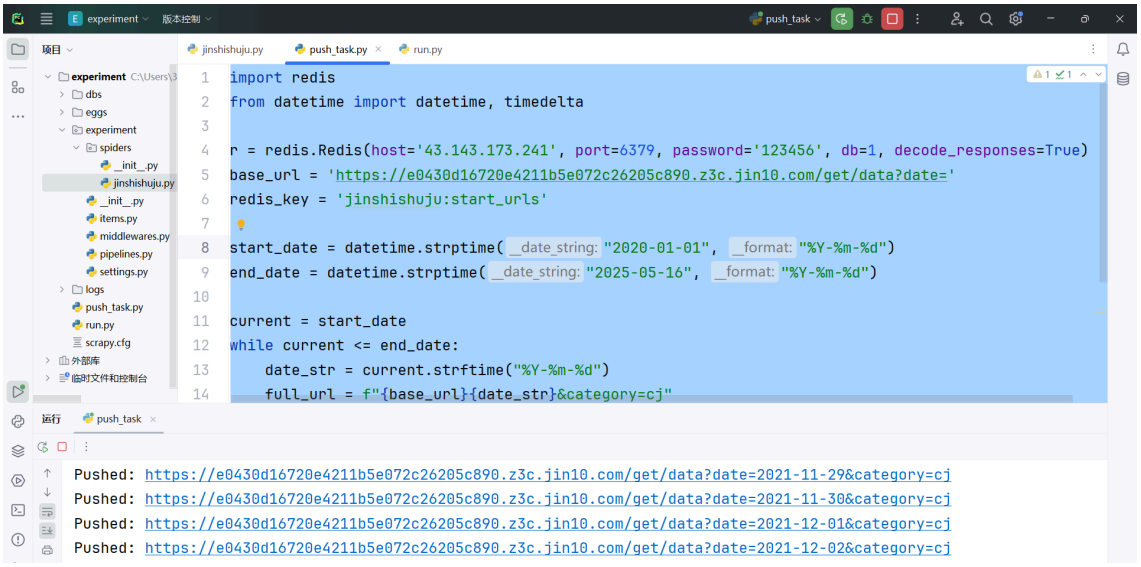
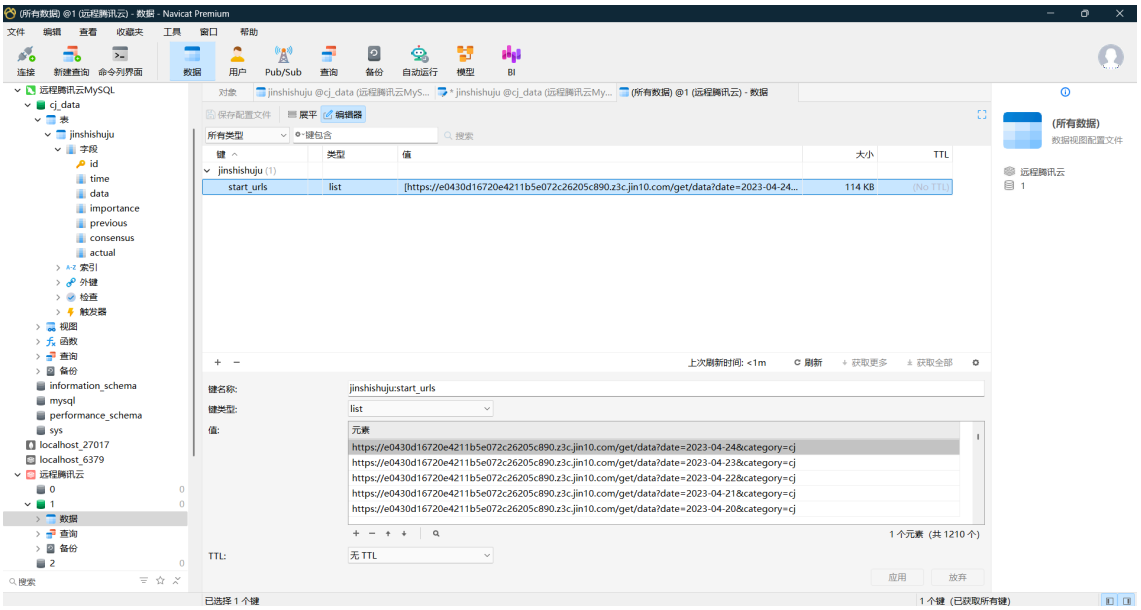
使用crawl模板的分布式爬虫

- push_task.py（向远程主机推送任务）

```python
import redis
from datetime import datetime, timedelta

r = redis.Redis(host='43.143.173.241', port=6379, password='123456', db=1, decode_responses=True)
base_url = 'https://e0430d16720e4211b5e072c26205c890.z3c.jin10.com/get/data?date='
redis_key = 'jinshishuju:start_urls'

start_date = datetime.strptime("2020-01-01", "%Y-%m-%d")
end_date = datetime.strptime("2025-05-16", "%Y-%m-%d")

current = start_date
while current <= end_date:
    date_str = current.strftime("%Y-%m-%d")
    full_url = f"{base_url}{date_str}&category=cj"
    r.lpush(redis_key, full_url)
    print("Pushed:", full_url)
    current += timedelta(days=1)
```

- 运行结果



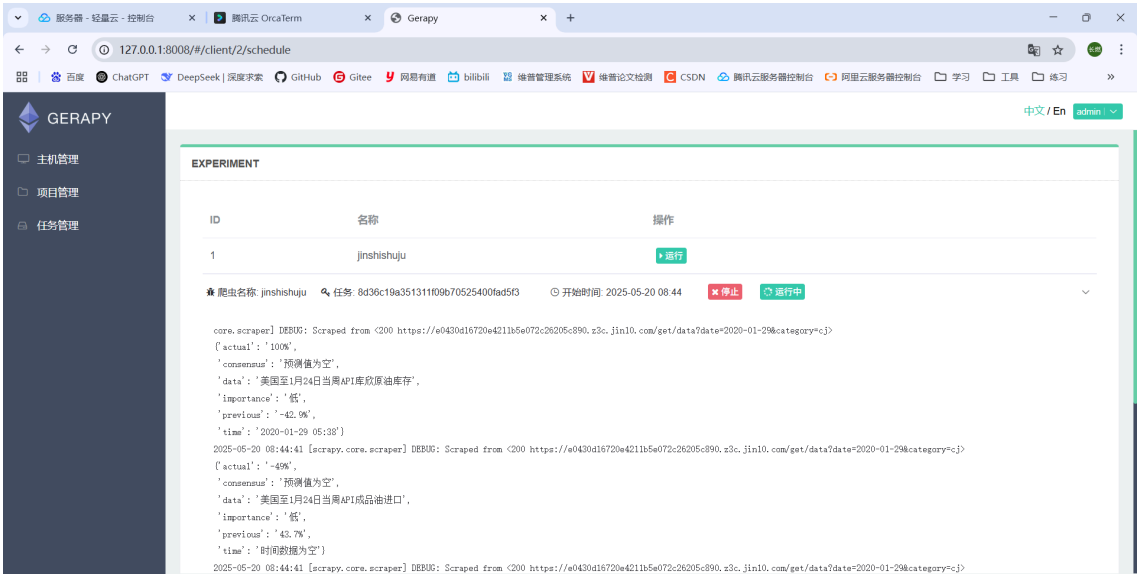


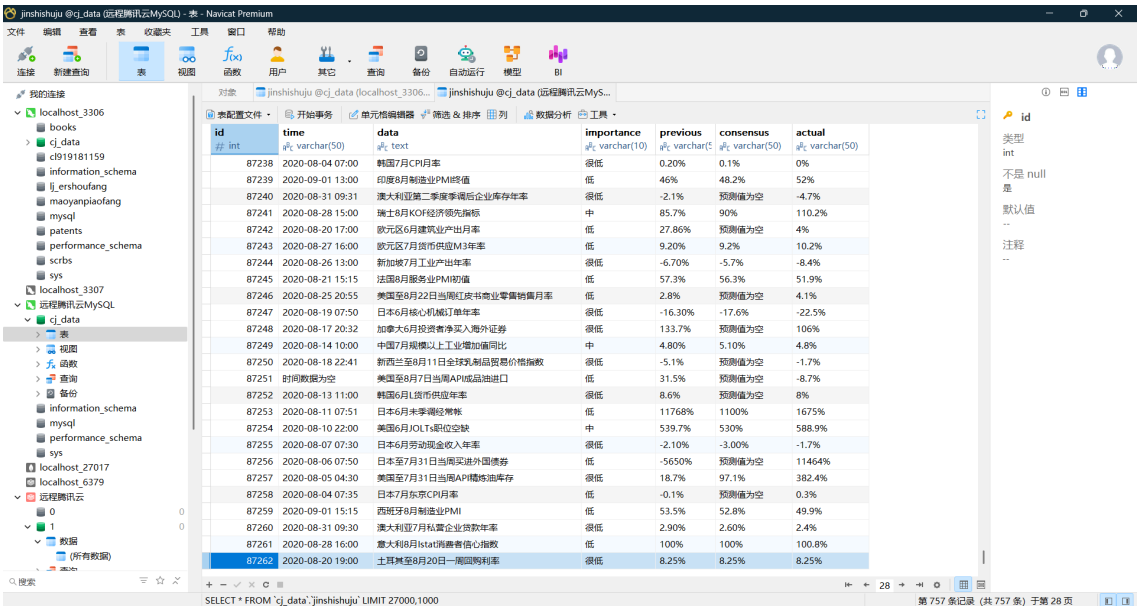
- jinshishuju.py

```
import json
import scrapy
from scrapy_redis.spiders import RedisSpider
from experiment.items import JinshishujuItem


class JinshishujuSpider(RedisSpider):
    name = "jinshishuju"
    redis_key = "jinshishuju:start_urls"

    custom_settings = {
        'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)...',
        'ROBOTSTXT_OBEY': False,
        'ITEM_PIPELINES': {
            'experiment.pipelines.MySQLPipeline': 300,
        },
        'SCHEDULER': 'scrapy_redis.scheduler.Scheduler',
        'DUPEFILTER_CLASS': 'scrapy_redis.dupefilter.RFPDupeFilter',
        'SCHEDULER_PERSIST': True,
        'REDIS_HOST': 'localhost',
        'REDIS_PORT': 6379,
        'REDIS_PARAMS': {
            'password': '123456',
            'db': 1,
            'decode_responses': False
        },
        'SCHEDULER_QUEUE_CLASS': 'scrapy_redis.queue.SpiderQueue',
        'DOWNLOAD_TIMEOUT': 15,  # 设置请求超时，避免长时间挂起
        'RETRY_TIMES': 3,  # 重试次数
    }

    headers = {
        "accept": "application/json, text/plain, */*",
        "origin": "https://rili.jin10.com",
        "referer": "https://rili.jin10.com/",
        "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36",
```

```
        "x-app-id": "sKKYe29sFuJaeOCJ",
        "x-version": "2.0"
    }

    def start_requests(self):
        while True:
            url = self.server.lpop(self.redis_key)  # self.server 是 Redis 连
接实例
            if not url:
                break
            if isinstance(url, bytes):
                url = url.decode('utf-8')
            self.logger.info(f"开始请求URL: {url}")
            print('-----------------------------------')
            yield scrapy.Request(url=url, headers=self.headers,
callback=self.parse)

    def parse(self, response, **kwargs):
        self.logger.info(f"响应状态码: {response.status}，URL: {response.url}")
        if response.status != 200:
            self.logger.warning(f"非200响应，跳过: {response.status}")
            return
        try:
            data = json.loads(response.text)
        except json.JSONDecodeError as e:
            self.logger.error(f"JSON解析失败: {e}，内容片段:
{response.text[:200]}")
            return
        data_dict = data.get('data', [])
        if not data_dict:
            self.logger.warning(f"接口未返回数据，URL: {response.url}")
            return

        for i in data_dict:
            item = JinshishujuItem()
            item['time'] = i.get('actual_time') or '时间数据为空'

            country = i.get('country') or ''
            time_period = i.get('time_period') or ''
            indicator_name = i.get('indicator_name') or ''
            item['data'] = country + time_period + indicator_name or '数据为
空'

            star_map = {1: '很低', 2: '低', 3: '中', 4: '高', 5: '很高'}
            item['importance'] = star_map.get(i.get('star'), '重要性为空')

            item['previous'] = (i.get('previous') + '%') if i.get('previous')
else '前值为空'
            item['consensus'] = (i.get('consensus') + '%') if
i.get('consensus') else '预测值为空'
            item['actual'] = (i.get('actual') + '%') if i.get('actual') else
'公布值为空'

            yield item
```

- 运行结果


服务器 - 轻量云 - 控制台
腾讯云 OrcaTerm
Gerapy
127.0.0.1:8008/#/client/2/schedule
GERAPY
主机管理
项目管理
任务管理
中文 / En
admin
EXPERIMENT
ID
名称
操作
1
jinshishuju
运行
爬虫名称: jinshishuju
任务: 8d36c19a351311f09b70525400fad5f3
开始时间: 2025-05-20 08:44
停止
运行中
core.scraper] DEBUG: Scraped from <200 https://e0430d16720e4211b5e072c26205c890.z3c.jin10.com/get/data?date=2020-01-29&category=cj>
{'actual': '100%',
'consensus': '预测值为空',
'data': '美国至1月24日当周API库欣原油库存',
'importance': '低',
'previous': '-42.9%',
'time': '2020-01-29 05:38'}
2025-05-20 08:44:41 [scrapy.core.scraper] DEBUG: Scraped from <200 https://e0430d16720e4211b5e072c26205c890.z3c.jin10.com/get/data?date=2020-01-29&category=cj>
{'actual': '-49%',
'consensus': '预测值为空',
'data': '美国至1月24日当周API成品油进口',
'importance': '低',
'previous': '43.7%',
'time': '时间数据为空'}
2025-05-20 08:44:41 [scrapy.core.scraper] DEBUG: Scraped from <200 https://e0430d16720e4211b5e072c26205c890.z3c.jin10.com/get/data?date=2020-01-29&category=cj>


jinshishuju @cj_data (远程腾讯云MySQL) - 表 - Navicat Premium
文件 编辑 查看 表 收藏夹 工具 窗口 帮助
连接 新建查询 表 视图 函数 用户 其它 查询 备份 自动运行 模型 BI
我的连接
localhost_3306
books
cj_data
cl919181159
information_schema
lj_ershoufang
maoyanpiaofang
mysql
patents
performance_schema
scrbs
sys
localhost_3307
远程腾讯云MySQL
cj_data
表
视图
函数
查询
备份
information_schema
mysql
performance_schema
sys
localhost_27017
localhost_6379
远程腾讯云
0
1
数据
(所有数据)
对象
jinshishuju @cj_data (localhost_3306...
jinshishuju @cj_data (远程腾讯云MyS...
表配置文件
开始事务
单元格编辑器
筛选 & 排序
列
数据分析
工具
id int
time varchar(50)
data text
importance varchar(10)
previous varchar(5
consensus varchar(50)
actual varchar(50)
87238 2020-08-04 07:00 韩国7月CPI月率 很低 0.20% 0.1% 0%
87239 2020-09-01 13:00 印度8月制造业PMI终值 低 46% 48.2% 52%
87240 2020-08-31 09:31 澳大利亚第二季度季调后企业库存年率 很低 -2.1% 预测值为空 -4.7%
87241 2020-08-28 15:00 瑞士8月KOF经济领先指标 中 85.7% 90% 110.2%
87242 2020-08-20 17:00 欧元区6月建筑业产出月率 低 27.86% 预测值为空 4%
87243 2020-08-27 16:00 欧元区7月货币供应M3年率 低 9.20% 9.2% 10.2%
87244 2020-08-26 13:00 新加坡7月工业产出年率 很低 -6.70% -5.7% -8.4%
87245 2020-08-21 15:15 法国8月服务业PMI初值 低 57.3% 56.3% 51.9%
87246 2020-08-25 20:55 美国至8月22日当周红皮书商业零售销售月率 低 2.8% 预测值为空 4.1%
87247 2020-08-19 07:50 日本6月核心机械订单年率 很低 -16.30% -17.6% -22.5%
87248 2020-08-17 20:32 加拿大6月投资者净买入海外证券 很低 133.7% 预测值为空 106%
87249 2020-08-14 10:00 中国7月规模以上工业增加值同比 中 4.80% 5.10% 4.8%
87250 2020-08-18 22:41 新西兰至8月11日全球乳制品贸易价格指数 很低 -5.1% 预测值为空 -1.7%
87251 时间数据为空 美国至8月7日当周API成品油进口 低 31.5% 预测值为空 -8.7%
87252 2020-08-13 11:00 韩国6月L货币供应年率 很低 8.6% 预测值为空 8%
87253 2020-08-11 07:51 日本6月未季调经常帐 低 11768% 1100% 1675%
87254 2020-08-10 22:00 美国6月JOLTs职位空缺 中 539.7% 530% 588.9%
87255 2020-08-07 07:30 日本6月劳动现金收入年率 很低 -2.10% -3.00% -1.7%
87256 2020-08-06 07:50 日本至7月31日当周买进外国债券 低 -5650% 预测值为空 11464%
87257 2020-08-05 04:30 美国至7月31日当周API精炼油库存 很低 18.7% 97.1% 382.4%
87258 2020-08-04 07:35 日本7月东京CPI月率 低 -0.1% 预测值为空 0.3%
87259 2020-09-01 15:15 西班牙8月制造业PMI 低 53.5% 52.8% 49.9%
87260 2020-08-31 09:30 澳大利亚7月私营企业贷款年率 很低 2.90% 2.60% 2.4%
87261 2020-08-28 16:00 意大利8月Istat消费者信心指数 低 100% 100% 100.8%
87262 2020-08-20 19:00 土耳其至8月20日一周回购利率 很低 8.25% 8.25% 8.25%
id
类型
int
不是 null
是
默认值
--
注释
--
SELECT * FROM `cj_data`.`jinshishuju` LIMIT 27000,1000
第 757 条记录 (共 757 条) 于第 28 页