Demo1：《元尊》信息检索与抽取

- 主要技术和开发环境

  参考：https://blog.csdn.net/MAOZEXIJR/article/details/80678133

  - scrapy
  - whoosh + jieba
  - django
  - 创建Pycharm项目，使用虚拟环境venv
- 安装三方包

  scrapy、pymongo、jieba、whoosh

  - 安装包前升级pip，否则安装包可能会报错
  - 或者在需要import时安装
  - 导出依赖

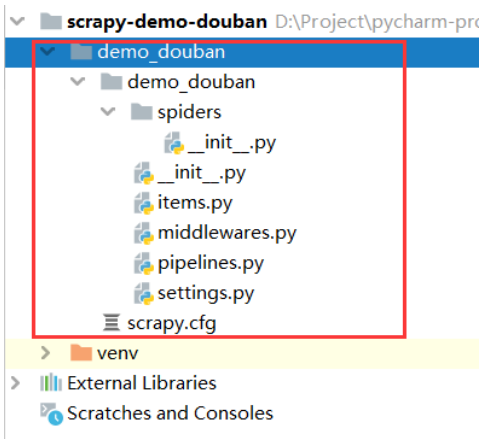
    pip freeze > 输出文件
- 爬虫模块
  - 创建scrapy项目
    - 在Terminal输入scrapy startproject <模块名>

```
(venv) D:\Project\pycharm-pro\scrapy-demo-douban>scrapy startproject demo_douban
New Scrapy project 'demo_douban', using template directory 'd:\project\pycharm-pro\scrapy-demo-douban\venv\lib\site-packages\scrapy\templates\project', created in:
    D:\Project\pycharm-pro\scrapy-demo-douban\demo_douban

You can start your first spider with:
    cd demo_douban
    scrapy genspider example example.com

(venv) D:\Project\pycharm-pro\scrapy-demo-douban>
```

    - 项目结构



  - 启动文件
    - 在一级目录下新建startspider.py
    - 启动文件参数与爬虫项目名保持一致

```
from scrapy import cmdline

if __name__ == '__main__':
    cmdline.execute('scrapy crawl yuanzun -s LOG_FILE=spider.log'.split())
```

- 即 crawl 后的参数与爬虫类的 name 字段保持一致

```
class YuanzunSpider(scrapy.Spider):
    name = 'yuanzun'
    allowed_domains = ['www.shuquge.com']
    start_urls = ['http://www.shuquge.com/txt/5809/index.html']
    base_url = "http://www.shuquge.com/txt/5809/"
    process = 0
```

- -s 用于将日志重定向输出到文件

- 准备文件
  - settings.py
    - 启用管道
    - mongodb参数
  - items.py
    - 定义项目字段
  - pipelines.py
    - 重写process_item方法
  - mongodb操作

    参考：https://blog.csdn.net/qq_37421862/article/details/81286063
- 爬虫文件
  - 在spiders文件夹新建spider.py
  - 定义类成员

    name 和 start_urls 为必要成员
  - 重写parse方法
    - 该项目中需要解析二级网页

```python
def parse(self, response):
    chapter_list = response.xpath("/html/body/div[5]/dl/dd[position()>13]/a")
    self.process = 10
    for chapter in chapter_list[:10]: # 为测试方便，只爬取10章
        chapter_item = NovelspiderItem()
        chapter_item['chapterTitle'] = ' '.join(chapter.xpath("./text()").extract_first().split()[-2:])
        chapter_url = self.base_url + chapter.xpath("./@href").extract_first()
        chapter_item['chapterUrl'] = chapter_url
        yield scrapy.Request(chapter_url, callback=self.parse_chapter, meta={"item": chapter_item})


def parse_chapter(self, response):
    chapter_item = response.meta['item']
    sentences = response.xpath('//*[@id="content"]/text()').extract()[:-3]
    chapter_content = ""
    for sentence in sentences:
        if sentence in {'\r', '\n'}:
            continue
        chapter_content += sentence.replace('\n', '')\
            .replace('\r', '\n')\
            .replace('\xa0', ' ')
    chapter_item['chapterContent'] = chapter_content
    yield chapter_item
```

- 索引模块

  参考1：https://www.cnblogs.com/shonelau/p/5805739.html

  参考2：https://www.jianshu.com/p/127c8c0b908a

  - 建立索引
    - 创建索引模板

      使用whoosh.fields中的Schema TEXT ID，以及jieba.analyse中的ChineseAnalyzer
    - 生成索引文件

      使用whoosh.index中的create_in和open_dir
    - 构建索引

      主要使用pymongo的查改操作和index.writer的commit、add_document方法
  - 建立检索器
    - 打开建立好索引，构造index.searcher
    - 提取查询字段

      使用whoosh.qparser的XxxParser
    - 设置参数

      排序键、分页结果限制
      - 排序键的设置先将章节标题序号转换为NUMERIC类型
    - 执行搜索

      调用parser.parse
    - 返回结果

      查询结果为列表，highlights方法可在content附近添加突出文字的html标签
  - 提供搜索接口及调用
- 引擎模块

  参考：https://www.runoob.com/django/django-first-app.html

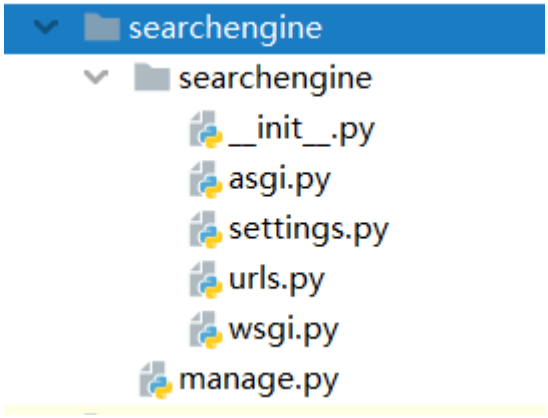
  - 创建Django项目

- 在Terminal输入django-admin startproject <模块名>

```
(venv) D:\Project\pycharm-pro\Coursework2>django-admin startproject searchengine

(venv) D:\Project\pycharm-pro\Coursework2>
```

- 项目结构



- 启动文件
  - 在一级目录下新建runserver.py
- settings.py

  修改templates文件路径
- view.py
  - 使用检索模块的检索器
  - 使用django.shortcuts的render完成跳转
  - 方法名与templates下html文件中的表单action属性一致
- urls.py

  通过选择跳转进入view的不同业务处理模块
- templates/

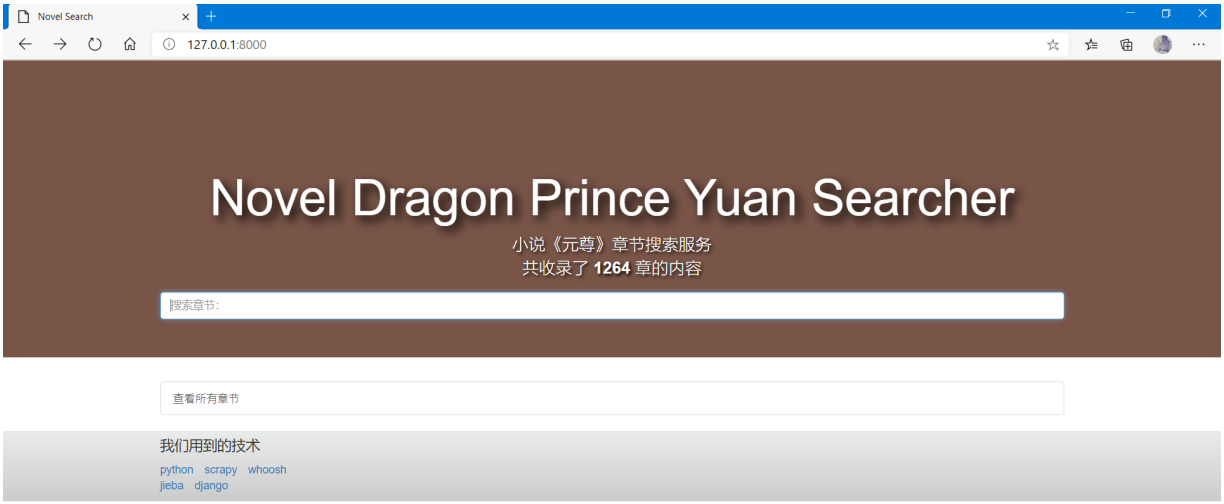
  变量标签：{% block search %} {% end block %}
  - main.html
  - result.html
  - extract.html

- 运行说明
  - 安装依赖

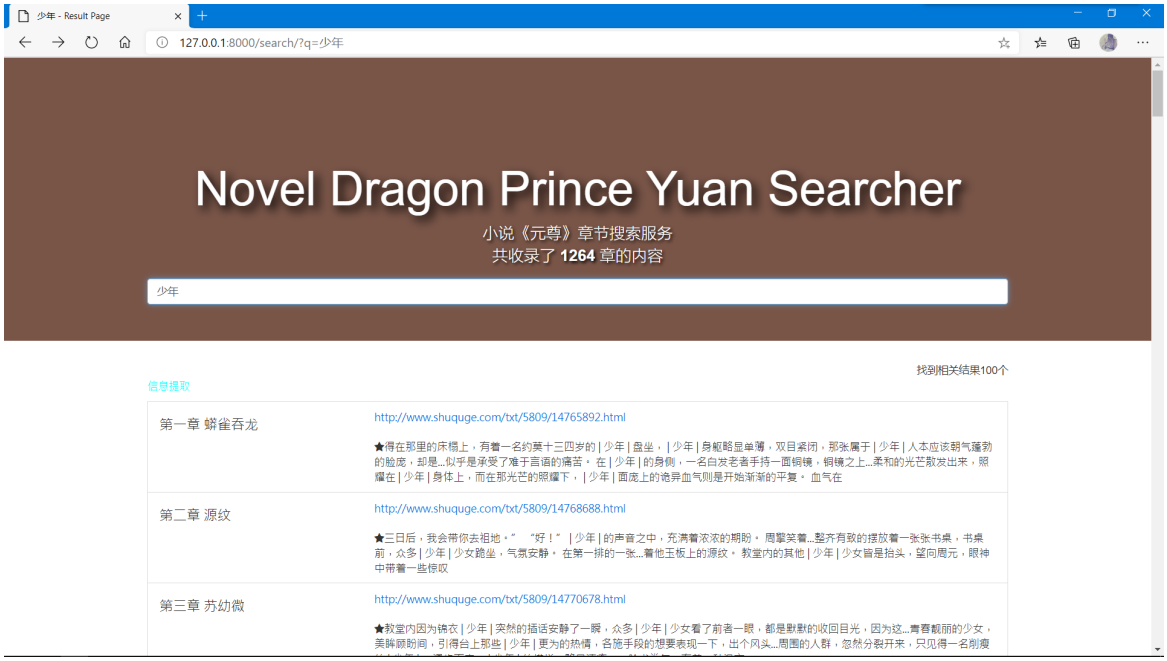
    pip install -r requirements.txt
  - 运行爬虫模块启动文件：startspider.py

    保证mongod服务开启
  - 运行索引模块建立索引文件：indexbuilder.py
  - 运行引擎模块启动文件：runserver.py

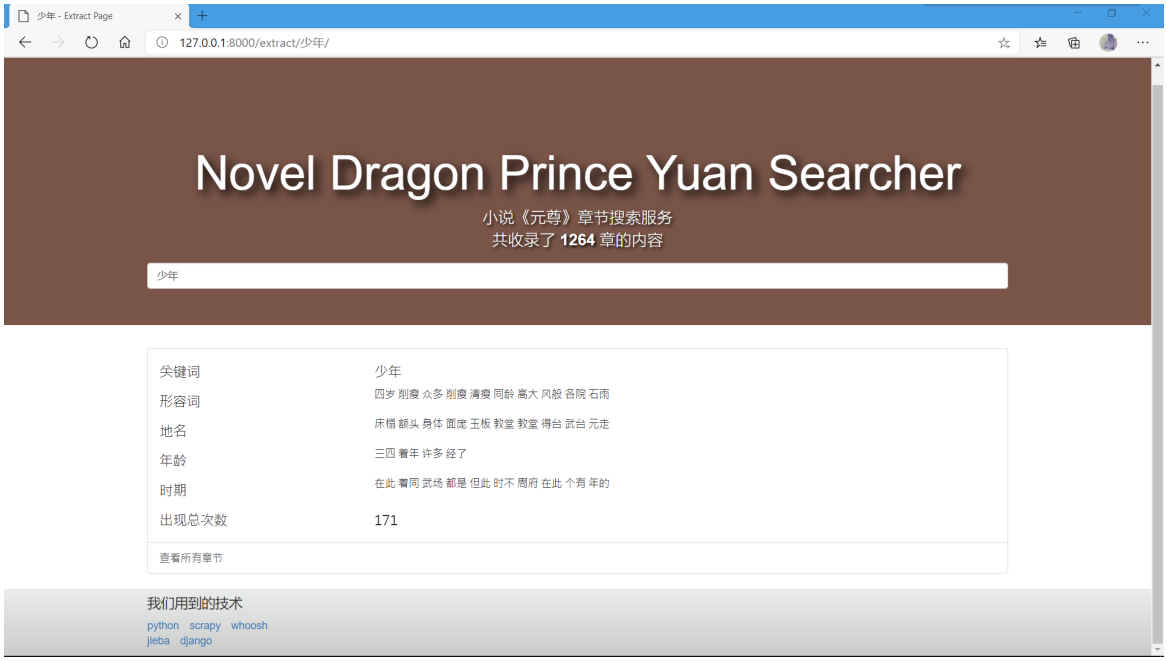
    若本地index已存放在相应位置，可直接跳过爬虫模块和索引模块
  - 打开浏览器：127.0.0.1:8000



- 输入关键词进行检索



- 点击进行信息提取


少年 - Extract Page
127.0.0.1:8000/extract/少年/
Novel Dragon Prince Yuan Searcher
小说《元尊》章节搜索服务
共收录了 1264 章的内容
少年
关键词
少年
形容词
四岁 削瘦 众多 削瘦 清瘦 同龄 高大 风般 各院 石雨
地名
床榻 额头 身体 面庞 玉板 教堂 教堂 得台 武台 元走
年龄
三四 着年 许多 经了
时期
在此 着同 武场 都是 但此 时不 周府 在此 个有 年的
出现总次数
171
查看所有章节
我们用到的技术
python scrapy whoosh
jieba django