俄罗斯站点
- 域名
- 环境依赖
- 如何运行本代码
  - 假如在原基础上新增了页面
- 参数配置
- 额外需求
  - 日志功能
  - 进度条功能
  - 断点续爬功能
  - 抓取Html

俄罗斯站点

域名

https://www.vidal.ru/

环境依赖

`scrapy 1.5.0`

`tqdm 4.28.1`

`re 2.2.1`

`python 3.6.7`

`pymongo3.5.1`

如何运行本代码

在本项目目录下，命令行键入 `scrapy crawl vidal`

假如在原基础上新增了页面

- 请在运行本程序之前，提前运行 `/russia/russia/get_page.py` 并将其输出作为 `russia/russia/spiders/vidal.py` 中的位于第 19 行的 `start_urls` 的值
- 在项目目录下，命令行键入 `scrapy crawl vidal`

参数配置

本项目的设定基本都可以在 `/russia/russia/settings.py` 中进行配置，具体参数参见 settings.py 文件， 内有配套说明

额外需求

日志功能

本程序默认只显示 `WARNING` 级别的日志，若有需求，可以在文件 `settings.py` 中对字段 `LOG_LEVEL` 进行设置。日志文件存放路径为 `LOG_FILE` 字段指定，若删去该字段，则可移除日志功能。

进度条功能

本程序日志功能使用 `tqdm` 进行实现，具体进度参照为剩余待解析的 url 数量。

断点续爬功能

若要使用断点续爬功能，只需控制台中键入 `scrapy crwal vidal -s JOBDIR=你的路径` 该命令会使用你指定的路径生成一个用于存放断点的文件夹，并使用该文件夹实现断点续爬。

> 开始任务：scrapy crawl vidal -s JOBDIR='./dict'
>
> 终端任务：键盘键入一次 ctrl+c 即可
>
> 再次开始：scrapy crawl vidal -s JOBDIR='./dict'

抓取Html

由于该功能为后期额外添加，因此本程序并没有严格集成该功能。

若需要爬取 Html 文本，请将于 `russia/russia/spiders/vidal.py` 中的 `parse()` 函数更改为如下：

```
    def parse(self, response):
        detail_list = self.parse_letter_page(response)

        # 扒字段
        for i in detail_list:
            self.logger.info('start to parse detail page %s',i)
            yield scrapy.Request(i,callback=self.parse_detail)

        # 扒网页
        # for i in detail_list:
        #     self.logger.info('start to parse detail page %s',i)
        #     yield scrapy.Request(i,callback=self.parse_save_html)
```

如果只需爬取字段即可，那么请使用原本的 parse() 函数：

```
    def parse(self, response):
        detail_list = self.parse_letter_page(response)

        # 扒字段
        # for i in detail_list:
        #     self.logger.info('start to parse detail page %s',i)
        #     yield scrapy.Request(i,callback=self.parse_detail)

        # 扒网页
        for i in detail_list:
            self.logger.info('start to parse detail page %s',i)
            yield scrapy.Request(i,callback=self.parse_save_html)
```