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`，爬虫即可进行运行。本爬虫将会

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'

最终结果

存入数据库 `19499` 条药品数据。

似乎与该站点所提供的总数居条目相差10条左右，具体原因有待进一步分析。