反爬措施：

知乎 新浪 需要登陆以后携带用户信息（用户信息参数通过特定的加密方法去加密），然后才能去请求需要抓取的页面，ip限制。

搜狗微信公众号 只推送100页数据，需要通过微信扫二维码登陆才可获得 100 页

数据，不登陆只推送50页数据，请求文章详情页需要特定的参数去请求。如果频繁的

抓取数据会随机出验证码。

百度贴吧 暂时没有大的反爬措施。

脚本配置：

运行环境： windows，linux python2.7。

脚本需要的第三方模块在 spiders 目录下的 requirements.txt 文件内。

config.py文件内需要配置登陆 新浪，知乎 的用户名和密码。

抓取数据的关键词配置在config.py 文件内。

log日志可在 crawl 目录下查。

百度贴吧 数据字段（目前只获取第一页的评论内容）

|  |  |
| --- | --- |
| 中文 | 英文 |
| 关键词 | keyword |
| 帖子标题 | title |
| 帖子链接 | url |
| 帖子图片 | image |
| 发贴用户昵称 | author |
| 发帖内容 | content |
| 帖子来源（贴吧名称） | source |
| 用户评论 [{评论内容，用户昵称，评论时间}] | reply [{content, createdate, author}] |
| 帖子创建时间 | createdate |

知乎 数据字段

|  |  |
| --- | --- |
| 中文 | 英文 |
| 帖子标题 | question\_title |
| 帖子链接 | question\_url |
| 帖子ID | question\_id |
| 帖子类型 | question\_type |
| 最高答案得赞数 | up\_voteup\_count |
| 回答赞数最高帖子摘要 | content\_summary |
| 回答赞数最高帖子内容 | up\_content |
| 回答赞数最高的用户昵称 | up\_author\_name |
| 回答赞数最高的用户个人主页 | up\_author\_url |
| 回答赞数最高的用户个人简介 | up\_author\_headline |
| 回答赞数最高答案的评论数 | up\_comment\_count |
| 赞数最高答案的创建时间 | up\_create\_time |
| 赞数最高答案的最后更新时间 | up\_update\_time |
| 关键词 | keyword |

新浪微博文章数据字段

|  |  |
| --- | --- |
| 中文 | 英文 |
| 新闻标题 | title |
| 新闻链接 | url |
| 新闻来源 | soure |
| 发表时间 | date |
| 新闻摘要 | summary |
| 新闻内容 | content |
| 关键词 | keyword |

微信公众号数据字段

|  |  |
| --- | --- |
| 中文 | 英文 |
| 关键词 | keyword |
| 帖子标题 | title |
| 帖子创建时间 | date |
| 帖子内容 | content |
| 帖子链接 | public\_name |
| 公众号昵称 | url |