配置config.txt文件，用于进行目标站点标签与安邦标签的映射。

以和讯银行http://bank.hexun.com为主入口做例子，需要抓取如下橙色框中的标签(其他主入口的url列表参见给出的word文档)。



新建一个文件夹“和讯银行”，在其中有一个txt中进入如下配置(“金融服务”为标准标签)

监管,宏观经济

行业,金融服务

非银,金融服务

零售,金融服务

理财,金融服务

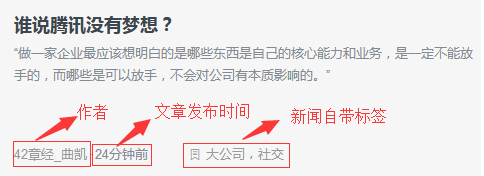
网金,金融服务

Txt文件中每行有两列，中间用半角字符的逗号隔开，第一列的值是该网站所要抓取的文章所在的类别(如上图中橙色框中的词)，第二列是对应的标准化的类别。其值需要开发者自行根据第一列的内容与下表的标准标签进行对应。比如橙色框中第一个词“监管”是银行监管，下表中最接近的就是“宏观经济”，第二个词“行业”是指银行业，下表中最接近的是“金融服务”，所以txt的前两行就是“监管,宏观经济”和“行业,金融服务”。

|  |  |
| --- | --- |
| 类别编号 | 类别名称 |
| 1 | 宏观经济 |
| 2 | 房地产 |
| 3 | 能源 |
| 4 | 金融服务 |
| 5 | 汽车 |
| 6 | 传统制造业 |
| 7 | 基础设施 |
| 8 | 医药&健康 |
| 9 | 零售 |
| 10 | 文化&传媒 |
| 11 | 矿产 |
| 12 | 钢铁 |
| 13 | 农业 |
| 14 | 旅游 |
| 15 | 通信&IT |
| 16 | 教育&人才 |
| 17 | 贸易&服务 |
| 18 | 战略新兴产业 |
| 19 | 养老产业 |
| 20 | 民营企业 |

新闻保留格式：

1. 爬虫作者ID（zhouqing）
2. 网站名 website\_name（36氪）
3. 网站板块website\_block（大公司）
4. 新闻news\_url （http://36kr.com/p/5132595.html）
5. 作者 news\_author
6. 新闻发布时间publish\_time，统一格式：YYYY-MM-DD HH:Mi:SS
7. 新闻抓取时间crawl\_time，统一格式：YYYY-MM-DD HH:Mi:SS
8. 新闻自带标签news\_tags，如果有多个标签用英文逗号隔开，如果没有直接写字符串’None’

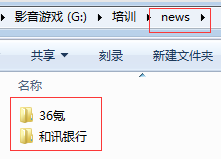




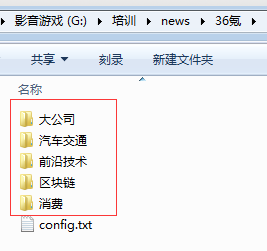
1. 新闻标题news\_title
2. 新闻正文news\_content

**最终结果：**

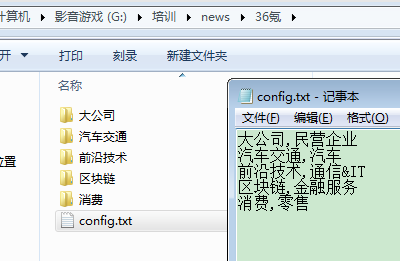
1. news目录下建立站点目录



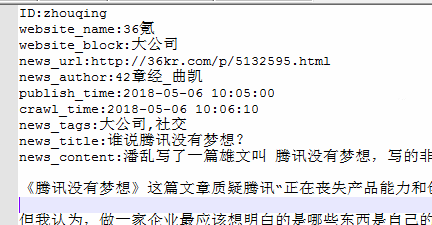
1. 站点目录下建立版块目录



1. 站点目录下建立配置文件config.txt



1. 爬取的文章以txt格式保存在相应的版块目录下，每篇文章一个文件，文件名为新闻标题，文件内容按上述新闻格式保存。



**爬取的新闻请务必保持干净，不要有多余的其他新闻的链接或广告等其他噪音数据**。

<http://stock.eastmoney.com/news/1407,20180506867498698.html>

