周报-第3周-余连玮

本周完成的事情

与王子昂商量后，制定了趋势分析的具体计划：

1. 获取2018年一整年的食品安全相关的数据。
2. 进行数据处理后，送入LDA模型中训练，得到一年中所有的主题及所有文档对应的主题。
3. 提取出一年中Top-K的主题，统计一年中这些主题的热度趋势
4. 对这些主题做后续的热度预测。

爬取文本数据

1.爬取食品科技网文本数据

关键点：研究并获得URL模式。

- 由于该网站食品安全栏目只有2019年的数据，因此无法直接按照栏目进行爬取。通过研究资讯网页URL发现规律：所有资讯网页都具有一样的URL模式 `https://www.tech-food.com/news/detail/n{数字}.html` 。而通过尝试和观察发现，2018年的资讯网页的数字范围为1376251-1415410。
- 得到URL模式后，一个个爬取对应URL，再判断资讯是否是“食品安全”栏目。如果是，则将资讯的相关数据写入csv文件中；如果不是，则跳过这条资讯。

关键点：利用IP池进行爬虫，防止真实IP被禁止访问。

- 频繁地访问网页容易导致服务器禁止IP访问。在网上爬取有效的IP地址来构建IP池，一旦访问超时则更换新的IP重试。
- 同时设置重试超过一定次数就跳过该网页。
- 增加了程序挂起时间，防止程序过度访问服务器，给服务器造成负载压力。

最终获取了1万3千多条数据，数据内容比较有价值。

2.爬取食安通文本数据

关键点：利用Post方法访问资讯列表网页。

- 通过研究和观察发现该网站翻页操作是通过提交表单传递PageNo参数来做到的，因此利用python的requests包的Post方法来访问资讯列表网页。

关键点：利用IP池进行爬虫，防止真实IP被禁止访问。

- 频繁地访问网页容易导致服务器禁止IP访问。在网上爬取有效的IP地址来构建IP池，一旦访问超时则更换新的IP重试。
- 同时设置重试超过一定次数就跳过该网页。
- 增加了程序挂起时间，防止程序过度访问服务器，给服务器造成负载压力。

最终获取了4百多条数据，数据内容比较有价值。

3.爬取央视文本数据

关键点：利用selenium模拟浏览器进行爬虫

- 由于利用python的requests包的Get方法无法访问网页，因此采取模拟浏览器的方式进行爬虫。

最终获取了一百二十多条数据，数据量较小，但数据内容比较有价值。

数据量较小的原因是：首先由于央视网不是专门报道食品安全相关的网站，只有一部分新闻涉及，其次通过搜索“食品安全”关键词来爬虫，会漏掉很多不包含在搜索结果，但是也属于食品安全主题的新闻。

爬取视频数据

1.爬取央视视频数据

关键点：异常网页结构的处理

- 重新爬取2018年的视频数据。但是发现部分视频没有获取成功。观察发现部分网页结构不同，导致videoCenterId的获取方式与其他网页不同，因此无法获得视频ts文件网页。补充处理了这部分的视频videoCenterId的获取。

最终获取了246条数据，数据量较小，但数据内容比较有价值。

数据量较小的原因是：首先由于央视网不是专门报道食品安全相关的网站，只有一部分新闻涉及，其次通过搜索“食品安全”关键词来爬虫，会漏掉很多不包含在搜索结果，但是也属于食品安全主题的新闻。

2.爬取bilibili视频数据

流程与爬取央视视频类似：爬取视频文件，转换成音频文件，从音频文件中提取出文字内容。

最终获取了141条数据，数据量较小，且由于bilibili网的视频为用户上传的，视频质量不高，不少视频没有语音内容，导致数据内容没什么价值。

由此吸取教训，bilibili网属于娱乐休闲类视频网站，不适合爬取新闻类的主题事件。

总结

通过在多个数据源进行爬虫，我体会到，选择数据源是很重要的。

好的数据源应当具备以下特点：

1. 相关数据量大
2. 数据有效性较高
3. 网页结构规范，易于爬取内容

由此可见，食品科技网是好的数据源，爬虫的编写不费时，得到的数据量很大，且价值很高。

而bilibili网则是个反例，在决策时就应该排除这类数据源，费力不讨好。