毕业设计说明文档

李雨田

2015 年 9 月 20 日

1 源代码

所有源代码均在 GitHub (https://github.com/hotpxl/nebuchadnezzar) 上。

2 使用方法和文件结构

所有源代码和部分数据文件都在代码仓库里。

输入 `npm install` 以安装所有 Node.js 的依赖。Python 的依赖需要手动安装。

2.1 原始数据

其中 `scraper` 文件夹下为获取原始论坛数据的脚本。获取之后的数据存储在 `data` 文件夹下。目前的版本是基于股吧手机版网页的，效率比从电脑版网页抓取更高，但是数字上有略微的区别。之前是有过电脑版网页抓取的脚本，后来被覆盖了。在 Git 的历史中可以找到，或者根据手机版修改也不会很麻烦。

股价是通过交易客户端获取的，导出 CSV 格式的文本。通过 `merger` 中的脚本，和之前的数据整合得到 `data/merged` 里面的完整数据。此时已经按具体股票分开了。

其中抓取论坛数据的时候，使用了 Redis 数据库来暂时存储结果。Redis 的配置文件在 `database` 文件夹下。

而 `parser` 文件夹下主要为解析原始数据格式的脚本，在 `merger` 里会用的。

2.2 情感数据

情感数据首先由 `translator` 文件夹下的脚本，讲之前得到的论坛数据翻译成英文。这里手动注册了若干个百度翻译 API 的帐号轮流使用。翻译之后的结果由 `sentiment` 文件夹下的脚本分析情感极性。会调用项目 https://github.com/xiaohan2012/twitter-sent-dnn 的相应代码。该项目已直接包含在 `sentiment/twitter` 下。可以先试图将其跑起来，因为会有一些依赖上的问题，也可以参考其 GitHub 上的说明文档。

分析情感的时候，先调用 `sentiment/create_task_queue.py` 生成任务队列存入 Redis 数据库。然后再调用 `sentiment/extract_sentiment.py` 从任务队列中取出任务分析。由于 Python 不支持并行，这里可以手动开启多个 Python 进程以加速任务。

2.3 数据处理

此处主要使用 StatsModels 等数学库计算。可以参考 `thesis_plots.py` 脚本，运行此脚本可以生成论文中所有的图，包含了论文中所以分析的数学计算。