bilibili“巴基斯坦”相关视频评论区爬虫与情感倾向可视化分析报告书

胡婷婷 2000019609

一、概述

本项目在本人国际关系专业背景的基础上，综合应用 python 基础知识，爬取 bilibili 平台与巴基斯坦和中巴关系相关的三个视频的评论（包括评论和评论回复），对其进行情感倾向可视化分析。

爬虫部分用 selenium 进行，配合 cookies 手工登录的方式，抓取每条评论及其回复内容的用户名、评论内容、评论时间和点赞数量。

数据处理部分用 pandas 进行，包括情感倾向指数计算、分类、时间分类等。

情感倾向分析部分用 snownlp 进行，配合 jieba 分词技术，得出每条评论的情感倾向指数（评估情绪属性的数值，范围在 0-1，越接近 1 情感越积极）。

可视化分析部分用 pycharts 进行，包括评论时间分析-日历图 calendar、情感分布-饼图 pie、评论内容-词云图 wordcloud、点赞量与情感倾向的关系-折线图 line、不同年份情感倾向分布-条形图 bar 等多样化的可视化处理。

二、使用说明

1、爬虫主题和主题相关视频需要用户自行选取。用户可以根据需求选择与主题最有关联度的视频评论。

2、不同视频页面的 cookies 需要分别获取，分别储存。在 01 号程序中更改【手动修改】部分。可以在 02 号程序验证登录结果。

3、不同视频页面的评论内容需要分别爬取，分别储存，对应使用上一步得到的 cookies。在运行 03 号程序时，注意人工确认登录成功后按回车，开始自动爬虫。爬虫阶段设有提示和进度条，便于管理进程。

4、爬取的文件在 04 号程序合并，同时进行情感分析。

5、05-11 号程序均为可视化绘图，依序如 web/index.html 中所示，可按需选择。

6、web/index.html 网页中可视化部分的图可点开查看详情，子网页中的 html 文件是从 code/output 中复制过来的。

三、项目成果

爬虫数据和可视化分析结果呈现在 index.html 中。总体而言，根据数据分析结果，可以得出以下结论：

1、中国人民对中巴关系 整体上充满信心，认为巴基斯坦是中国的好伙伴。

2、在评论情感倾向和点赞倾向上，网民都存在“极化”现象，更倾向于评论和点赞“非常积极”与“非常消极”的内容。

3、相比于“就事论事”，网民倾向于 在思考中巴关系时联系其他国家，如美国、日本、俄国等，呈现发散性思维倾向。

四、后续改进

根据调试，程序能够顺利完成项目要求，但是速度较慢。这可能是因为重复循环嵌套，后续可以优化解构精简代码，提高爬虫速度。

另外，snownlp 虽然已经是目前公认较为可用的情感分析库，但依然存在分析不准的情况。导致对评论情感态度的分析存在一定误差。后续可以寻找其他可替代方式，或者增加人工筛查的过程。