Python 期末项目-简易数据爬取项目报告

夏浩 大数据学院 19307130268

一. 项目要求

1. 基本要求能正确获取页面，并提取到至少 30 条新闻信息。（70 分）

2. 正确将数据内容写入文件，且能正常查看。（5 分）

3. 正确获取 300 条以上新闻信息。（10 分）

4. 能额外构建浏览量信息（在页面中，但未标出）。（5 分）

5. 代码结构清晰，代码风格规范。（10 分）

6. 有额外探索性工作可酌情加分，此部分不超过 10 分，且总分不超过 100 分。

二. 实验环境

Windows 10(2H21), PyCharm(professional 2021.1.1),

Python 3.8.2

三. 项目内容

1. 程序逻辑

main():

修改 url，调用爬取函数 process，调用自定义词云函数 news_wordcloud()

process():

a. 读取网页 html 数据，使用 BeautifulSoup 包函数抽取对应节点信息

b. 调用自定义的 get_title(), get_url(), get_time(), get_views(), 得到每条新闻相关数据

c. 将相应数据(id, title, url, time, views)写入文件

d. 将仅包含 title 信息的数据写入另一个文件（用于制作词云）

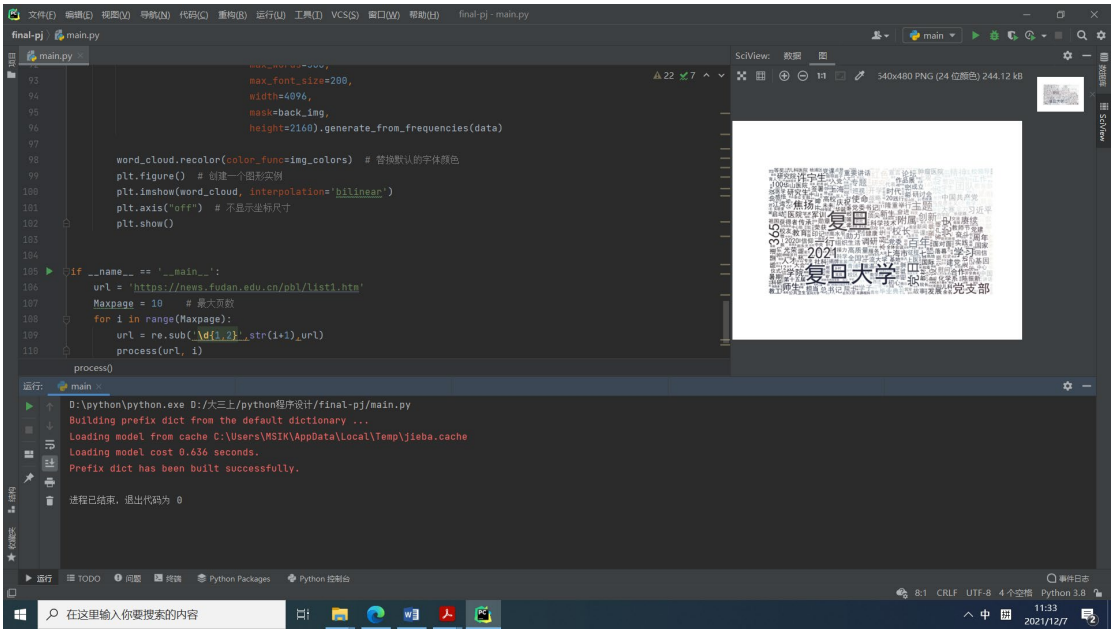
get_title(), get_url(), get_time(), get_views():

抽取节点信息，通过正则表达式或者读取标签属性，得到需要的数据，返回数据列表

news_wordcloud():

读取图片，读取仅包含 title 内容的文本文件，读取止词文件，通过 jieba.analyse 实现分词，调用 WordCloud 生成词云

2. 程序执行效果截图



3. 已完成得分项

1. 基本要求能正确获取页面，并提取到至少 30 条新闻信息。（70 分）（完成）

2. 正确将数据内容写入文件，且能正常查看。（5 分）（完成）

3. 正确获取 300 条以上新闻信息。（10 分）（完成）

4. 能额外构建浏览量信息（在页面中，但未标出）。（5 分）（完成）

5. 代码结构清晰，代码风格规范。（10 分）（完成）

6. 有额外探索性工作可酌情加分，此部分不超过 10 分，且总分不超过 100 分。（完成词云）

4. 词云图片

四. 项目总结

1. 充分完成了项目要求，完成了额外内容，构建词云

2. 将项目内容分成若干函数，使得代码更加清晰简明