python 爬虫与数据分析大作业报告书

方尤乐 2000012416

互联网获取数据功能介绍

爬取内容为百度贴吧的“弱智吧”的截至 2024 年 5 月 28 日中午前的 10000 条最新帖子（在贴吧综合排序下最靠前的 10000 条帖子，包括了一些热门帖子、置顶帖子）。爬取数据花了大约 30 个小时的时间，总数据量为 54 MB。

爬取的10000条帖子的标题及链接在 code/outputs/post.csv 中

每条帖子中的内容被保存在了 code/outputs/message{帖子编号}.txt 中。

在这个大作业项目中，利用 selenium 来访问https://tieba.baidu.com/p/{帖子编号} 爬取帖子的内容。

**如何爬取：**

依次运行以下代码：

code/getcookies.py（需要手动登陆后，按回车保存cookies）

code/getpost.py（获取10000条帖子的标题和链接，并将它们保存到code/outputs/post.csv 中

code/getmessage.py（访问这10000条帖子的内容，将除了评论区以外的所有内容保存至 code/outputs/message{帖子编号}.txt 中。

数据分析方法

**初步分析：**

利用了 python 的 jieba.posseg 库对中文语句进行分词处理，将每句话拆解成若干个词语，再统计这些词语出现的次数。

词频统计结果位于 code/result/{词性}.txt，其中每一行的格式为“词语: 出现总次数”。

其中我们主要关注的词性有：a: 形容词；ad: 副形词；an: 名形词；n: 名词；nr: 人名；ns: 地名；vd: 动副词；vn: 名动词

在这个范围内，按照词语的出现次数对所有词语进行排序，最终得到了相应的词频统计图

**进一步分析：**

由于 10000 条帖子中有许多是很短的无意义帖子，所以我们采取了数据量最大的 100 个帖子进行数据分析。并抽取了一些关键词，对这些关键词出现的次数做进行进一步地分析，将每条帖子的分析结果保存至 code/analysis.csv 中。

analysis.csv 的表头为

id,size(byte),标题,出现次数最多的词语,弱智,出院,为什么,水,**经验,**鸡,意义,问题,人类,地球,太阳,原神,哲学,感叹号,**问号,**句号,逗号,冒号,顿号

其中 id 为帖子的编号，size 为帖子内容的数据量（单位为 byte），

“弱智,出院,为什么,水,经验,鸡,意义,问题,人类,地球,太阳,原神,哲学”为选取的关键词，接下来表格的每一行统计了每条帖子中这些关键词出现的次数。最后的“感叹号,问号,句号,逗号,冒号,顿号”为选取的标点符号，接下来每一行的最后也统计了标点符号的出现次数。

然后就可以利用 pandas 库的 DataFrame 数据结构来组织这些数据。这份大作业统计了最大的100个帖子的若干个关键词的出现次数直方图、关键词出现次数的相关系数图，标点符号的分布占比饼状图。

**依次运行以下代码：**

code/wordstatistic.py

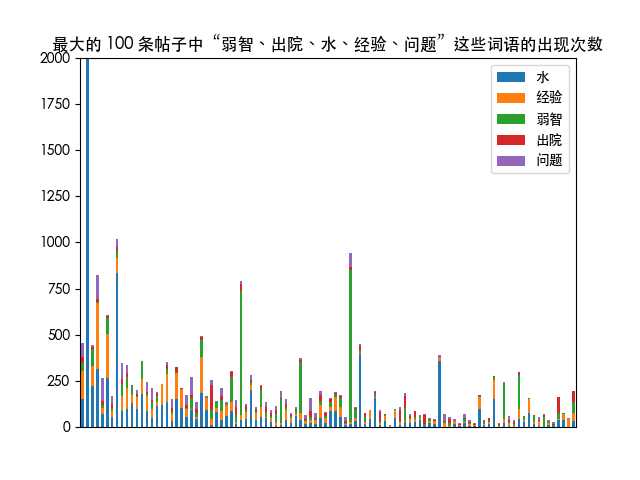
code/wordanalysis.py

code/deepanalysis.py

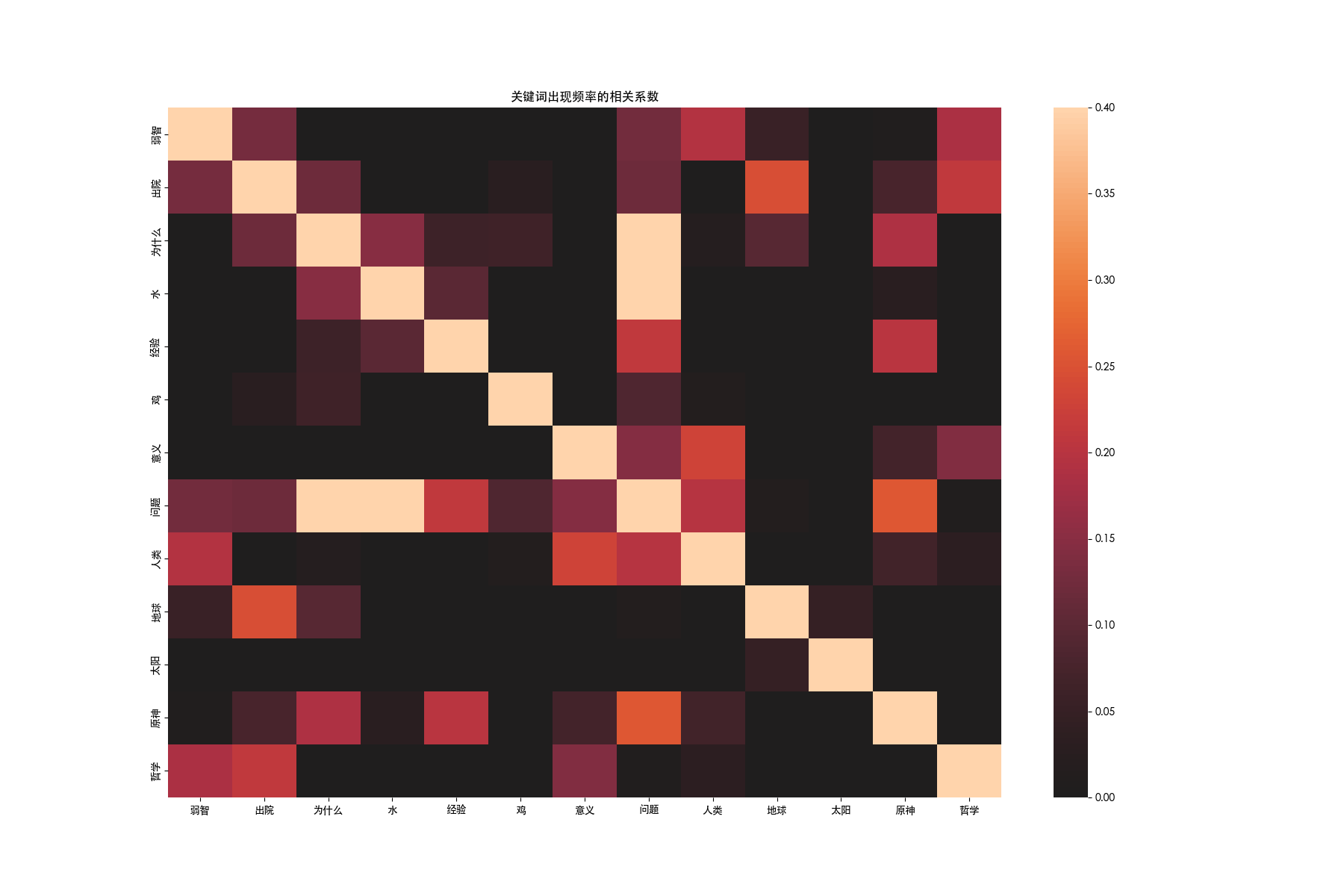
结果展示-词频统计图



结果展示-关键词在100个帖子中出现次数的直方图



结果展示-关键词出现的相关系数图



在相关系数图中，横纵坐标为我们关注的几个关键词，因此每个格子的横纵坐标表示它考察了两个关键词的关联系数。而每个格子越亮，就代表这两个关键词关联程度更高，即越有可能共现。