1.题目

政府工作报告分词、词云及分析

2.数据集

自行百度和下载2012-2022年度政府工作报告。

3.任务描述

（1）采集： 采集10年来政府工作报告，每年的报告保存为一个文件（可以程序采集，也可以手工采集）。

（2）各年词云与动画：编写python代码，使用Jieba分词软件进行分词，使用word cloud进行词云可视化，制作10年来每年的报告的高频词Top30词云；10年的每年的词云图，串联起来制作GIF动画。

（3）10年长青词top20多线图、消退词top20多线图、增长词top20多线图：横轴是年份，纵轴是频率，每根线是不同的词项，制作多线图。

比如，长青词示例：“发展”，2019、2020、2021年报告中的相对频率为0.5、0.5、0.5；

消退词示例：“扶贫”，2019、2020、2021年报告中的相对频率为0.5、0.3、0.1；

增长词示例：“人工智能”，2019、2020、2021年报告中的相对频率为0.1、0.3、0.5；

请自行设计判别长青词、消退词、增长词的规则或者公式。

提示：可以使用波浪图，提高美观度。

任务1、2、3为必选。

可以借鉴参考文献的代码。

（4）(optional)2022年报告分析：针对2022年报告，建立top500的词项共现矩阵。共现的含义，是词项在一个自然句里面同时出现。

由词项共现矩阵构造图的邻接矩阵，共现频率大于阈值则邻接矩阵相应元素为1，否则为0。由图的邻接矩阵，利用NetworkX可视化这个图。

利用NetworkX进行社区检测(community detection)并且可视化，尝试进行解读；可以把每个community看作一个topic。

备注：本任务需要学习图数据分析的相关内容后完成。

（5）(optional)词项对比：近10年来，工业/农业、制造业/服务业的热度对比，通过可视化形式进行对比，比如折线图，提示注意同义词的归并。

（6）(optional)近3年来的高频词及其网络可视化：从近3年来的报告中，提取高频词top100；当用户选择某个高频词，请显示该高频词的1度、2度邻居构成的词项网络。

在一个自然句里共现的词项，称为相关词项。词项作为网络节点，相关词项之间有边，边的权重可以设定为共现频率。

备注：本任务需要学习图数据分析的相关内容后完成。

（7）(optional)专题分析：自行选择某个专题，自行构建分析方法。提示：可以围绕台湾问题，比如关于台湾问题的关键词的历年频率变化，关于台湾问题的关键词的共现网络及其可视化等。

提示：

1. 注意去掉中文停用词。

2. 注意同义词的归并。

3. 可以采用Jieba分词python库，必要时手工增加新词汇（人工字典），提高分词效果；比如“新常态”、“新冠肺炎”、“静态管理”等。

4.成果提交

以ipynb形式提交。

5.参考文献

[1] <https://github.com/markyangcc/Analysis-Gov-Work-Report>

[2] <https://github.com/derek-s/Python-GovWorkReportAnalyze>

[3] Python文本分析——政府工作报告都说些啥？<https://www.dadclab.com/archives/7378.jiecao>

[4] 政府工作报告分析. <https://blog.csdn.net/the_Milky/article/details/115178540>

[5] 数据说两会｜1978年到2020年政府工作报告关键词盘点. https://www.thepaper.cn/newsDetail\_forward\_7492352