实验：社交行为建模

请于 2024 年 5 月 8 日 23:59 之前提交至课程邮箱 ustcweb2022@163.com

总体实验要求：
请组成不多于 8 人小组，围绕指定数据集进行自定方案分析实验，并记录实验过程。

数据背景：
数据集来自某组织成员对组内提案的投票情况统计，其机制为：

- 组织中的每名成员归属于两个子社团之一，两个子社团的成员一般情况下不会发生重叠。
- 在该组织中，每次提案会议有固定人数的成员参与，部分成员在会上提出提案，而其他参会成员对该提案进行表决。
- 提案可以由某位成员单独发起，也可以由若干名成员联合发起。
- 参与会议的组织成员可以自由选择“赞成”、“反对”、“弃权”。

实验数据：
本数据集一共包含该组织在若干次提案会议中表决过的 8849 项提案。
数据使用方式：将下载的压缩包解压缩即可。

文件说明：
文件以.json 的格式进行存储，主要包含三类文件，分别对应提案投票情况（Votes）、历次参会成员（Members）以及提案信息 (Bills)。另外还提供了成员所在地的映射信息（State）。
以提案信息 Bills 为例，其 json 文件格式如下：


```
[{"bill_id": "b144-104", "bill_type": "b", "sponsor": "O000007", "cosponsors": []},
{"bill_id": "b143-104", "bill_type": "b", "sponsor": "O000007", "cosponsors": []},
{"bill_id": "b175-104", "bill_type": "b", "sponsor": "L000371", "cosponsors": []},
{"bill_id": "b172-104", "bill_type": "b", "sponsor": "E000187", "cosponsors": []},
{"bill_id": "b126-104", "bill_type": "b", "sponsor": "O000007", "cosponsors": []},
{"bill_id": "b119-104", "bill_type": "b", "sponsor": "O000007", "cosponsors": []},
{"bill_id": "b128-104", "bill_type": "b", "sponsor": "N000147", "cosponsors": []},
{"bill_id": "b110-104", "bill_type": "b", "sponsor": "L000371", "cosponsors":
["J000174", "M000794"]}, {"bill_id": "b142-104", "bill_type": "b", "sponsor":
"O000007", "cosponsors": []}, {"bill_id": "b145-104", "bill_type": "b", "sponsor":
"O000007", "cosponsors": []}, {"bill_id": "b111-104", "bill_type": "b", "sponsor":
"L000371", "cosponsors": ["J000174", "M000794"]}, {"bill_id": "b116-104",
"bill_type": "b", "sponsor": "O000007", "cosponsors": []}, {"bill_id": "b129-104",
"bill_type": "b", "sponsor": "S000033", "cosponsors": ["B000220"]}, {"bill_id":
```


每个 dict 中的"bill_id"代表某个提案的 id，"bill_type"代表该提案的类型，"sponsor"代表该提案的提出者，"cosponsor"代表该提案的联合提出者（若为空表示该提案无联合提出者）。
更详细的数据说明见压缩包下的 readme 文件。请根据实验需要，自行提取并处理数据。
注意：由于实际情况所限，并非所有提案都有对应投票记录，请注意甄别！

实验内容：
要求对于指定数据，**自行设计实验方案及实验目标**，并根据数据给出量化的结果分析。

具体实验内容包括：

（1） 实验目标选定

本次实验的指向性目的是解决一个基于社会网络的预测性问题，问题由小组自行商议决定。预测性问题要求对数据根据时间戳（提案会议发生的时间）进行拆分，并利用历史数据对未来情况进行预测。问题本身需要有**明确的可验证性及对应的量化指标**。

一些可供参考的选题包括（仅供参考，实际可不限于以下问题，自行选择**其一**即可）：

- 预测成员对提案的投票结果
- 预测成员的活跃性（如参与提案发起的频率）
- 预测成员是否会参加某个提案的联名发起
- 预测成员未来的网络关系会发生何种变化
- 预测成员在提案的主题/标签偏好上的演变（可以从提案信息中获得）

本环节的要求如下：

- 请自行选择模型完成预测。训练集/验证集（如需）的比例自行确定，但测试集比例不低于 20%。如果时间较为充裕，可进一步分析比较不同划分方式和比例对结果的影响。

强调：所采用的算法模型不是本次实验的重点，其选择不会影响分数。我们也不会对预测效果提出考核要求，重点是在此过程中对于社会网络各要素作用的分析比较，请不要把重心放在复杂、先进算法的运用上，仅采用课上教授的非深度学习模型同样可以完成这个实验。

（2） 动态社交网络的构建

由于数据集本身并未包含显式的网络结构信息，为了更为完整地实现基于社会网络的分析和预测，需要基于掌握的信息自行构造一个社会网络。**常见的构造方式**如根据成员之间的标签相似性、成员共同发起过的提案，或者成员对提案的态度等进行构造，并可采用相似性或共现次数等因素进行边权重的量化估计。

本环节的要求：

- 根据需要自行定义社会网络中节点和边的定义，并设计社会网络构建方法，同时说明设计方案的合理性依据。
- 网络中的每一条边应具有权重，权重的计算方式自行定义，并说明其合理性和意义。
- 所设计的网络要随着时间推移（即提案会议的发生）而发生动态变化，包括并不限于新增/删除节点、新增/删除边，边上的权重变化等。时间信息可以在 members.json 中获得。

（3） 基于选定任务的预测效果对比，围绕社交活动进行量化分析

在完成社会网络的构建后，请围绕拟开展的指向性目标，通过分析不同网络构造方式，或者网络中不同元素对于最终预测结果所产生的影响，探讨社交行为对于最终结果的影响机制。一些可供选择的分析内容包括（仅供参考，请结合实际自行设计方案）：

- 不同网络构造方式对于结果的影响
- 网络结构对于成员态度/倾向的影响，以及成员态度/倾向对于网络结构和投票行为的反向影响（即对社交关系变化的影响），可结合网络符号性加以分析
- 网络结构随时间的演化对于预测结果的影响
- 网络是否加权，以及是否考虑加权对时间变化对于结果的影响
- 网络中重要节点/关系，如结构洞、弱连接、重要影响力节点等对于合作效果的影响
- 网络稀疏性/新节点（冷启动）等问题对于结果的影响
- 结合社团挖掘，讨论子社团结构对于预测结果的影响

本环节的要求：

- 请根据小组商议情况自行选题，一般而言讨论的话题**不少于 3 个**，时间充裕的话可以选择更多话题。请在实验报告中简要介绍选题的依据和目的（即想要探讨的问题）。
- 请给出详细的量化分析结果，并结合图表进行分析说明。
- 可根据情况结合显著性检验、因果推断等统计手段加以分析。
- 请给出必要的参数敏感性实验（如有）和必要的案例分析。

提交说明：

以 PDF 或 DOC 格式提交，实验报告提交文件及邮件标题命名格式统一为“社会计算行为建模实验报告_学号_姓名”。

- 例如：“社会计算行为建模实验报告_SA20011999_法外狂徒张三”
- 标题仅写明小组组长的学号及姓名即可，其他成员请务必在邮件及实验报告正文中注明学号及姓名。因未署名造成统计遗漏责任自行承担。
- 实验报告请务必独立完成，如果发现抄袭按零分处理。
- 请注明所采用的算法，并列举必要的参考文献。
- 请采用必要的图表以更清晰地展示实验结果。
- 提交报告的同时请提交**源代码**以供检查。
- **除非特殊情况并事先征得许可，否则迟交报告将不再被接收**。

额外说明：

每组提交一份实验报告，所有组员得分相同。

如有未尽事宜，将对本说明进行进一步更新。