实验二

请于 2022 年 4 月 24 日 23:59 之前提交至课程邮箱 ustcweb2019@163.com
并于 2022 年 4 月 27 日（周三）课上进行现场汇报

总体实验要求：
请组成 4-6 人小组，围绕指定数据集进行自定方案分析实验，记录实验过程并撰写实验报告。

数据背景：
数据集来自著名在线活动组织网站 Meetup，其机制为：

- 整个 Meetup 社区由若干社团组成，每个社团有若干名用户，用户可以随时加入或退出。
- 活动以社团为主体，由若干名社团成员发起组织（部分活动组织者缺失）。仅有社团成员会收到邀请。
- 社团成员可以选择是否参加活动（Yes/No/Maybe），但不是所有人都会回应。
- 部分活动会注明限制人数（Headcount），但不一定会起到约束作用。

实验数据：
本数据一共包含 437 个社团、82770 个用户及 93512 个事件。
数据可通过睿客网下载，下载链接：https://rec.ustc.edu.cn/share/cb7a76b0-a906-11ec-a02d-67f03a71ca04，密码：s7p8
数据使用方式：解压后，将“All.pak”文件拖曳至“FilePackager.exe”文件上，将自动进行解压缩操作。因文件数量巨大，请耐心等待。

文件说明：
原始文件从 Meetup 官方 API 获得，以.xml 格式进行存储，一共包含四类文件，分别对应社团信息（Group）、事件/活动（Event）、用户参与（RSVP）及用户信息（Member）。

以 Event 信息为例，其 XML 文件格式如下：

请根据实验需要，自行提取并处理数据。数据具体内容可参考 Meetup API 官方文档，链接为 https://www.meetup.com/api/。**但需注意，Meetup API 已更新，目前的返回文件格式为 json，其格式和内容可能存在不同。**

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

本次实验的最终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进行拆分，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，自行确定选题）：

- 预测用户是否参与未来的某个活动
- 预测用户在社团内的活跃性（如参与活动的频率）
- 预测用户是否会加入某个新的社团
- 预测用户未来的网络关系会发生何种变化
- 预测用户在主题/标签偏好上的演变（相关信息可以从活动文本中获得）

注意：我们将参考工作量进行评分，过于简单的选题将影响到最后的得分。

（2） 动态社交网络的构建

由于 Meetup 本身没有显式的网络结构信息，为实现基于社会网络的分析和预测，首先需要自行构造社会网络。常见的构造方式如根据成员之间的标签相似性、成员共同参加的社团或者成员共同参加的活动等进行构造（可以采用相似性或共现次数进行加权）。

本环节的要求：

- 根据需要自行定义社会网络中节点（可以是个人，也可以是社团）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 Event/RSVP 中获得。

（3） 围绕社交活动的量化分析部分

在完成社会网络的构建后，请围绕拟开展的研究课题，首先进行统计分析，确认构造的社会网络对于研究课题是否具有显著作用（需要通过显著性检验等手段加以体现），并确定作用方式以辅助下一阶段的建模预测。

一些可供选择的分析内容包括（仅供参考，自行设计方案）：

- 不同网络构造方式对于结果的影响
- 网络结构演化对于结果的影响
- 网络是否加权对于结果的影响
- 网络稀疏性/新节点（冷启动）等问题对于结果的影响

（4） 自定义任务的预测实验部分

最后，根据选定的社会网络构建方案及相应的统计分析，设计模型解决预设的目标问题，并给出相应的测试方案和测试结果。

本环节的要求：

- 请自行设计模型完成预测。不要求采用深度学习方法，我们仅根据模型的合理性进行评价，不会根据模型的复杂度而额外加分。
- 训练集/验证集（如需）/测试集的比例自行确定，但测试集比例不低于 20%。建议分析比较不同划分方式和比例对于结果的影响。
- 请完成必要的消融实验，分析比较考虑/不考虑社会网络信息对于结果的影响。
- 必要的参数敏感性讨论和必要的案例分析。

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“社会计算第一次实验报告_学号_姓名”。

- 例如：“社会计算第二次实验报告_SA20011999_法外狂徒张三”
- 标题仅写明小组内一位成员学号及姓名即可，其他成员请在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，并取消答辩资格**。

报告要求：

由组长进行汇报，汇报总时长为 15 分钟，包括 12 分钟 PPT 讲解与 3 分钟提问+点评。

- 报告内容应包括选题设计、数据处理方式、社会网络构建方式、统计分析情况、采用的预测模型、相关参数的设置、实验结果及其分析、组内成员分工等。
- 报告顺序按照实验报告接收的顺序为准，名单将在报告当天于课程群内公布。
- 助教将根据汇报内容和实验报告内容进行打分，并计入总评成绩。

额外说明：

每组提交一份实验报告，所有组员得分相同。但考虑到组长的额外工作量（协调组员工作并进行汇报），组长将获得额外的 1 分加分。

如有未尽事宜，将对本说明进行进一步更新。