利用 GPT 大型模型工具完成数据洞察

本周以及下周的实验内容如下：
请同学们参考《GPT 数据科学系列课程实验手册》内容，并结合之前的作业经验，对截至 2024 年 8 月底在 GitHub 上具有协作行为日志数据的用户的个人信息（包括姓名、公司、邮箱及其地理位置等）进行数据洞察分析。数据获取链接为：https://github.com/X-lab2017/dase-2024-autumn/tree/main/HomeWork/data/user_data

实验目标

- 培养数据处理与分析能力：通过实际操作，提升对大规模数据集的处理和分析能力。
- 掌握 GPT 工具的应用：学习如何利用 GPT 大型模型工具辅助完成数据洞察任务。
- 理解数据隐私与伦理：在处理包含个人信息的数据时，遵循数据隐私保护的原则和规范。

实验内容

1. **人口统计分析**

- 国家和地区分布：统计用户所在国家和地区的分布，识别主要的开发者集中地。
- 城市级别分布：分析主要城市的开发者密度，发现技术热点区域。
- 时区分布：了解用户的时区分布，分析不同地区用户的协作时间模式。

2. **协作行为分析**

- 提交频率：统计每个用户的提交次数，识别高活跃用户和低活跃用户。

3. **其他维度有趣的洞察**（至少 2 个）

1. 时间维度的活动模式

分析内容

通过时间维度来分析用户的提交行为可以揭示出许多有趣的模式，例如工作时间外的活跃度、周末与工作日的差异、以及不同时间段的活跃高峰。

实施步骤

- **可视化时间模式**

洞察示例

- 如果发现很多用户在非工作时间（如晚上或周末）非常活跃，这可能表明项目对个人贡献者很重要，或者团队有跨时区合作。

- 工作日和周末的活动差异可以帮助了解团队的工作节奏，并为安排会议或发布更新提供依据。

2. 事件类型与用户行为模式

分析内容

通过分析不同类型的事件（`event_type`）和相应的动作（`event_action`），我们可以了解用户的行为模式。这有助于识别常见的用户互动方式、高频率发生的事件以及不同事件之间的关联。

实施步骤

- **统计每种事件类型的频率**
  计算每种事件类型的出现次数，以了解最常见的用户行为。
- **分析特定事件类型的详细行为**

  对于高频率的事件类型，进一步分析其对应的 `event_action`，以获得更详细的见解。
- **可视化事件类型和动作**

洞察示例

- 识别出用户最常参与的事件类型，例如评论、点赞或提交更新，这可以帮助优化用户体验。
- 发现某些事件类型下的特定行为模式，例如某类事件经常伴随着某种特定的动作，这可能提示改进的方向或新功能的需求。

洞察示例

- 识别出用户最常参与的事件类型，例如评论、点赞或提交更新，这可以帮助优化用户体验。
- 发现某些事件类型下的特定行为模式，例如某类事件经常伴随着某种特定的动作，这可能提示改进的方向或新功能的需求。

提交内容

- 数据分析代码。
- 最终的数据洞察报告（PDF 格式）。

实验结果见 homework12.ipynb

数据分析

1. 人口统计分布：

由代码操作得到结果，我们不难看出，人员主要来自美德等老牌资本主义国家和中国这种新兴的第三世界国家，从这项结果我们可以看出，在现在的科技发展形势下，传统的老牌强国仍然具有优势，但后起之秀也是存在的，中国在逐渐追赶。

2. 城市级别分布：

   城市的分布，主要采用地区获得，部分地区并没有细化到城市，这种情况多发生在一些版图较小的国家，如德国、日本等，从中，我认为这同国家大小与交通便利从事工作有关，作为互联网工作的从业人员，足不出户便可以完成办公，自然就不需要特别在意城市之间的距离，同时作为小国，如果一定要线下见面，发达的现代交通也使得一国之间的线下交流成为可能。

3. 时区分布：

   时区分布可以反映不同国家之间的交流，时区分布可以使得跨国合作者可以在符合生物钟的时间，健康有效率的进行合作与工作。

4. 特别的维度

   见上文

   （1）时间：分析不同的提交时间，来体现某一时间段，这一领域的工作状况

   （2）行为分析：分析额外的两项数据