Week 1：了解数据分析标准流程

近年来，随着互联网与通信技术的高速发展，学习资源的建设与共享呈现出新的发展趋势，各种网课、慕课、直播课等层出不穷，各种在线教育平台和学习应用纷纷涌现。尤其是2020年春季学期，受新冠疫情影响，在教育部“停课不停学”的要求下，网络平台成为“互联网+教育”成果的重要展示阵地。因此，如何根据教育平台的线上用户信息和学习信息，通过数据分析为教育平台和用户提供精准的课程推荐服务就成为线上教育的热点问题。

本项目提供了某教育平台近两年的运营数据，希望根据这些数据，为平台制定综合的线上课程推荐策略，以便更好地服务线上用户。

本项目主要涉及技术为利用pandas库进行数据预处理，利用matplotlib、pyecharts等库进行数据可视化，以及利用协同过滤算法给出综合推荐策略。

本项目的目标是：

1. 分析平台用户的活跃情况，计算用户的流失率，为平台管理决策提供建议。
2. 分析线上课程的受欢迎程度，构建课程智能推荐模型，为教育平台的线上推荐服务提供策略。

任务

- 任务 1.1

  理解各字段的含义，进行缺失值、重复值等方面的必要处理，将处理结果保存为“task11X.csv”（如果包含多张数据表，X 可从 1 开始往后编号），并在报告中描述处理过程。

- 任务 1.2

  对用户信息表中 recently_logged 字段的“--”值进行必要的处理，将处理结果保存为“task1_2.csv”，并在报告中描述处理过程。

- 任务 2.1

  分别绘制各省份与各城市平台登录次数热力地图，并分析用户分布情况。

- 任务 2.2

  分别绘制工作日与非工作日各时段的用户登录次数柱状图，并分析用户活跃的主要时间段。

- 任务 2.3

  记 T_{end} 为数据观察窗口截止时间，T_i 为用户 i 的最近访问时间，$\sigma_i = T_{end} - T_i$，若 $\sigma_i > 90$ 天，则称用户 i 为流失用户。根据该定义计算平台用户的流失率。

- 任务 2.4

  根据任务 2.1 至任务 2.3，分析平台用户的活跃度，为该教育平台的线上管理决策提供建议。

- 任务 3.1

  根据用户参与学习的记录，统计每门课程的参与人数，计算每门课程的受欢迎程度，列出最受欢迎的前 10 门课程，并绘制相应的柱状图。受欢迎程度定义如下：

  $$\gamma_i = \frac{Q_i - Q_{min}}{Q_{max} - Q_{min}}$$

  其中，γ_i为第 i 门课程的受欢迎程度，Q_i 为参与第 i 门课程学习的人数，Q_{max} 和 Q_{min} 分别为所有课程中参与人数最多和最少的课程所对应的人数。

- 任务 3.2

根据用户选择课程情况，构建用户和课程的关系表（二元矩阵），使用基于物品的协同过滤算法计算课程之间的相似度，并结合用户已选课程的记录，为总学习进度最高的 5 名用户推荐 3 门课程。

算法原理：https://zhuanlan.zhihu.com/p/31807038

实战代码Demo：https://blog.csdn.net/qq_34615112/article/details/106161033

在任务 3.1 和任务 3.2 的基础上，结合用户学习进度数据，分析付费课程和免费课程的差异，给出线上课程的综合推荐策略。