赛题六：基于高校图书馆借阅数据的用户潜在图书借阅结果预测推荐

一、赛题背景

随着高校图书馆数字化建设的不断推进，借阅数据成为反映用户阅读兴趣和行为习惯的重要资源。图书馆积累了大量的借阅记录，如何利用这些数据挖掘用户的潜在需求并进行精准的图书推荐，已成为提升图书馆服务质量和用户体验的关键问题。借助人工智能和数据挖掘技术，开发高效的图书推荐算法，不仅能够优化资源配置，还能为用户提供个性化的阅读建议，推动校园阅读文化的发展。

二、赛题应用场景

在高校图书馆的日常运营中，学生、教师等用户通过借阅系统获取图书资源，其借阅记录蕴含了丰富的兴趣偏好和行为模式信息。例如，某用户可能频繁借阅计算机科学类书籍，或在特定学期集中借阅与课程相关的教材。传统的人工推荐或简单分类方式难以全面捕捉用户需求，容易导致推荐结果与用户兴趣不匹配。基于借阅数据的推荐算法可以通过分析历史借阅记录、图书类别、借阅时间等特征，预测用户潜在的阅读需求，为其推荐符合兴趣的图书，从而提升借阅效率和用户满意度，同时为图书馆的藏书管理和采购决策提供数据支持。

三、出题信息

本赛题由模式分析与机器智能工信部重点实验室出题，结合图书馆实际数据与需求设计。

四、赛题任务

参赛者需利用主办方提供的图书馆借阅数据集，设计并实现人工智能算法，预测用户未来潜在的图书借阅需求并进行个性化推荐。

具体任务包括：

1. 用户兴趣建模：基于用户的借阅历史数据，分析其阅读偏好和行为模式，构建用户兴趣特征。

2. 图书推荐预测：根据用户兴趣特征和图书信息，预测用户未来可能感兴趣的图书并生成推荐列表。

五、数据集及数据说明

（一）数据来源

数据来源于图书馆的真实借阅记录，涵盖学生、教师等多类用户的借阅行为，具有多样性和代表性。数据已进行脱敏处理，确保用户隐私安全。

（二）数据概况

共提供 88378 条借阅记录用于模型训练和验证，其中包含 1451 个用户和 58549

本图书；测试集包含约 14510 条数据，用于最终结果评测。

该赛题的交互数据将根据用户分阶段发布，初赛阶段提供 600 个用户的完整交互数据及全部用户信息和图书信息，复赛阶段新增 400 个新用户的交互数据，决赛阶段释放剩余所有用户的交互数据，其中用户信息和图书信息仅在初赛阶段一次性完整提供，后续阶段不再补充基础信息。

示例数据可通过链接:

https://pan.baidu.com/s/1azEZ-oYNnQTnyEbnKphk5w?pwd=q7sh 获取，正式数据将在报名后开放下载。

（三）数据格式

数据集以 CSV 格式存储，共划分为 3 个文件

具体信息如下：

1. book.csv：图书信息，包含 book_id（图书 ID）、题名（书名）、作者、出版社、一级分类、二级分类。

2. inter.csv：借阅交互记录，包含 inter_id（交互 ID）、user_id（用户 ID）、book_id（图书 ID）、借阅时间、还书时间、续借时间、续借次数。

3. user.csv：用户信息，包含借阅人（用户 ID）、性别、DEPT（院系）、年级、类型（本科/研究生等）。

六、算法设计要求

（一）模型类型

鼓励参赛者采用机器学习或深度学习算法，如协同过滤（CF）、矩阵分解（MF）、深度神经网络（DNN）及其变体，也可借助大语言模型（LLM），对图书数据集中的文本进行内容提取与语义理解，挖掘文本中的关键信息，进一步丰富推荐维度，增强推荐系统的准确性与个性化，为用户提供更加贴合需求的图书推荐服务 。例如，可使用基于内容的推荐模型分析图书特征，使用图神经网络（GNN）或 Transformer 模型捕捉用户的借阅行为。

（二）创新性

鼓励提出创新的算法架构或改进现有推荐算法，以提升图书推荐的准确率和个性化程度。例如，设计新的特征提取方法以更好地表征用户兴趣与图书属性，或采用多维度数据融合（如结合用户信息、借阅时间和图书分类）的方法优化推荐效果。

（三）可扩展性

算法应具备良好的可扩展性，能够在不同配置的计算设备上运行，且在处理大规模借阅数据时性能稳定。例如，算法应能够在普通工作站和云端服务器上高效运行，并且在用户数量或借阅记录增加时，模型性能不会出现明显下降。

七、性能评估指标

（一）主要指标

1. 精确率（P），衡量图书馆推荐列表中，用户实际借阅（即测试集中真实借阅）的图书占所有被推荐图书的比例，反映了推荐图书的准确性，即推荐的图书有多少是用户真正会去借阅的。

2. 召回率（R）衡量图书馆推荐列表中，用户实际借阅的图书占用户在测试集中实际借阅的所有图书的比例，体现了推荐系统对用户真实借阅行为的捕捉能力，即用户实际借阅的图书有多少被成功推荐。

（二）次要指标

1. 模型大小：训练得到的模型文件大小，是衡量模型复杂度和存储需求的重要指标。较小的模型大小表明模型复杂度较低，便于在不同设备和环境中部署应用。

八、功能要求

（一）准确性

算法在预测用户潜在借阅需求时，须具备高准确性，确保推荐的图书与用户兴趣高度匹配。在测试集上，推荐结果的 F1 Score 需达到 0.12 以上。

（二）可靠性

面对不同院系、年级、借阅习惯的用户数据，算法应能稳定运行，输出可靠推荐结果。即使数据中存在借阅记录噪声或异常行为，算法也不应出现大幅性能波动，保持推荐结果的准确性和稳定性。

（三） 可解释性

算法应具备一定的可解释性，能够为用户或图书馆管理员提供推荐结果的解释依据。例如，通过可视化技术展示用户兴趣偏好与推荐图书的匹配关系，或提供特征重要性分析，说明模型基于哪些借阅特征生成推荐结果，帮助用户理解推荐过程并提升信任度。

（四）鲁棒性

算法要对数据的异常值、缺失值等情况具备较强的鲁棒性。在部分借阅记录存在时间缺失或用户数据不完整时，仍能保证推荐结果的可靠性，不会因数据的小瑕疵导致性能大幅下降。

（五）多模态融合能力

若参赛者采用多模态数据融合方法（如结合用户信息、图书分类、借阅时间等），算法应能有效整合不同类型的数据，且在融合后显著提升推荐的准确性和个性化程度，展现对多源信息的高效利用能力。

九、开发环境

（一）编程语言

Python，建议使用 Python3.6 及以上版本，因其具有丰富的科学计算库和机器学习框架支持。

（二）机器学习框架

推荐使用 TensorFlow2.x 或 PyTorch1.x，这两个框架在机器学习领域广泛应用，具有高效的计算性能和丰富的 API，便于模型的搭建、训练和部署

（三）计算资源

参赛者可使用本地工作站或云端计算平台进行开发和训练。本地工作站建议配备 NVIDIA GPU（如 GTX 10 系列及以上，或 RTX 系列）以加速模型训练；云端平台可选择阿里云天池、AWS SageMaker、Google Colab 等，提供灵活的计算资源配置。

（四）依赖库

可以使用 RecBole 框架，需安装其配套库，包括 PyTorch（用于模型训练）、NumPy 和 Pandas（用于数据处理），以及 RecBole 自带的工具库，以支持推荐系统的快速开发与实验。

十、成绩评价

（一）输入数据格式要求

参赛者的算法应能正确读取主办方提供的 CSV 格式的借阅数据集，包括以下文件：

book.csv（图书信息）：需解析图书 ID、题名、作者、出版社、分类等字段；

inter.csv（借阅交互记录）：需提取借阅人（用户 ID）、图书 ID、借阅时间、还书时间、续借时间、续借次数等信息；

user.csv（用户信息）：需准确读取用户的性别、院系、年级、类型等特征；

train_data.csv（训练集）、valid_data.csv（验证集）、test_data.csv（测试集）：需处理这些数据集中的借阅记录，用于模型训练和评估。

算法应对这些 CSV 文件进行有效解析，确保能够完整提取所有相关字段，为后续的推荐任务提供数据支持。

（二） 输出数据格式要求

参赛者需将预测结果存入 csv 文件中，并将提交文件压缩为 zip 文件。该 CSV 文件的每一行代表一条预测推荐数据，每行数据应包含两个字段，依次为 “user_id” 和 “book_id”。

“user_id” 用于明确标识用户 ID，数据类型为字符串，其值应与比赛提供的数据集中的用户 ID 格式一致且真实有效，不得出现虚构或错误的 ID 值。

“book_id” 代表图书 ID，数据类型同样为字符串，需与比赛数据集里的图书 ID

格式相符且真实有效。

推荐数量限制：为每个用户仅推荐 1 本图书，即同一个 “user_id” 在 CSV 文件中只能出现一次。

确保 CSV 文件格式正确，无多余的表头行（若有，仅保留第一行为表头，且表头必须为 “user_id,book_id”），无空行、无效字符或格式错误，以免影响成绩评定。文件应采用 UTF - 8 编码，以保证字符兼容性。

（三） 成绩计算公式

成绩基于推荐结果与测试集中用户真实借阅记录的对比，综合使用精确率（P）和召回率（R），计算公式如下：

1. 精确率（P）

定义：衡量图书馆推荐列表中，用户实际借阅（即测试集中真实借阅）的图书占所有被推荐图书的比例，反映了推荐图书的准确性，即推荐的图书有多少是用户真正会去借阅的。

计算公式：

$$P = \frac{TP}{TP + FP}$$

其中，TP（True Positive）为图书馆推荐列表中用户实际借阅的图书数量，FP（False Positive）为图书馆推荐列表中用户未实际借阅的图书数量。

2. 召回率（R）

定义：衡量图书馆推荐列表中，用户实际借阅的图书占用户在测试集中实际借阅的所有图书的比例，体现了推荐系统对用户真实借阅行为的捕捉能力，即用户实际借阅的图书有多少被成功推荐了。

计算公式：

$$R = \frac{TP}{TP + FN}$$

其中，TP（True Positive）为图书馆推荐列表中用户实际借阅的图书数量，FN（False Negative）为用户在测试集中实际借阅但未被图书馆推荐的图书数量。

3. 最终成绩

F1 值：

定义：F1 值是精确率和召回率的调和平均数，综合考虑了精确率和召回率，用于更全面地评估图书馆借阅推荐系统的性能，避免了单纯依赖精确率或召回率而导致对系统性能评估的片面性。

计算公式：

$$F1 = \frac{2 \times P \times R}{P + R}$$

（四）有效成绩

比赛每阶段的 F1 值高于 0.0055 视为有效成绩。阈值 0.0055 确保推荐算法具有实际应用价值。

赛题设奖基数为有效成绩团队数量，具体阈值可由主办方根据参赛表现调整。

十一、解题思路

（一）数据预处理

提取交互数据并构建图结构。

梳理借阅记录表，从每一条借阅记录里提取用户 ID 和图书 ID，形成用户 - 图书交互对。借助图数据结构，将用户和图书分别作为节点，把借阅行为视为连接两者的边，构建用户-图书二分图。通过对该图结构进行分析，能够挖掘出用户之间的相似性以及图书之间的关联性，为推荐算法提供更丰富的特征信息。

用户个人信息、图书信息融入 LLM 的 prompt。

在用户信息表中，采集用户的性别、专业等个人信息。围绕图书推荐任务，把这些信息合理融入到 prompt 中。比如构建“为 2023 级计算机专业的男同学推荐相关书籍”这样的 prompt，借助精心设计的 prompt，引导基于大语言模型的推荐服务，生成更贴合用户需求的推荐结果。也可以使用序列推荐的思路解题。

（二）模型训练

选择合适的机器学习框架（如 TensorFlow 、PyTorch）搭模型，设置合理的训练参数，如学习率、迭代次数、批量大小等。在训练过程中，采用交叉验证方法，利用验证集数据对模型进行评估和调优，防止模型过拟合。

（三）模型融合与优化

可尝试将多个不同结构或训练阶段的模型进行融合，如采用投票法或加权平均法，综合多个模型的预测结果，提高最终预测的准确性。同时，根据性能评估指标，对模型进行针对性优化，如调整模型结构、增加训练数据量等。

十二、参考资源

（一）书籍

1. 《深度学习 》（ DeepLearning ） ， 由 IanGoodfellow 、YoshuaBengio 和 AaronCourville 撰写，系统介绍了深度学习的基础概念、模型架构和训练方法，对理解和应用神经网络有很大帮 助。

2. 《 Python 深度学习》（DeepLearningwithPython），作者 FrançoisChollet ，

通过大量代码示例，详细讲解了如何使用 Python 和 Keras 框架进行深度学习模型的开发，适合初学者快速上手。

（二）在线课程

1. Coursera 上的 “DeepLearningSpecialization” 课程，由吴恩达教授授课，涵盖了深度学习的多个关键领域，包括神经网络基 础、卷积神经网络、循环神经网络等，课程内容丰富且实践性强。

2. edX 上的“Introduction to Artificial Intelligence”课程，提供了人工智能和机器学习的入门知识，包括算法原理、模型训练和应用案例等，有助于参赛者构建全面的知识体系。

（三）学术论文

在学术数据库中搜索关于推荐系统的最新研究论文，如“Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning” 等，了解当前该领域的前沿技术和研究方法。

关注知名人工智能会议（如 AAAI 等）上发表的相关论文，跟踪最新的研究动态和创新成果。

十三、提交要求

（一）算法代码

提交完整的 Python 代码，涵盖数据预处理、模型训练和推荐结果生成等环节。

代码需遵循 PEP8 规范，包含详细注释和文档说明，确保评审人员能够理解和运行。

（二）技术报告

提交 PDF 格式的技术报告，字数不少于 3000 字。内容包括：算法设计思路、模型架构图、实验设置（如超参数选择、数据处理方法）、性能分析以及创新点和不足之处。

（三）推荐数据

提交数据名称应为 zip，其中第一列为 user_id,第二列为 book_id。

（四）模型文件

提交训练好的模型文件，并提供模型加载和使用说明，包括运行环境（如 Python 版本）和依赖库。

确保模型能够在指定环境中正常运行并输出推荐结果。

十四、更新与答疑

（一）更新与答疑

赛题将不会更新，如遇赛题相关问题，可通过邮箱联系相关负责人，联系方式：

nuaa_niurongbing@nuaa.edu.cn

（二）赛题智能助手

每个赛题将建立一个在线智能助手，方便选手进行赛题咨询。

十五、比赛流程及奖项设置

（一）报名阶段

参赛者在比赛官方网站上完成报名注册，提交个人或团队信息。

（二）初赛阶段

参赛者利用赛事方提供的训练数据集进行算法模型设计，利用提供的初赛测试集进行相应方法的验证与调试。初赛阶段参赛者每天提交结果的次数不限，但是初赛排行榜每隔 1 小时刷新一次。

（三）复赛（省赛）阶段

初赛结束后进入复赛阶段，开放复赛数据下载链接。仅有初赛阶段提交有效结果的参赛团队可以进入复赛。复赛期间，参赛者利用赛事方提供的复赛阶段数据进行算法模型调试，提交对复赛测试数据的推理结果。复赛阶段持续 3 天，每个参赛队伍每天仅能提交 2 次。复赛排行榜每隔 1 小时刷新一次。

（四）复赛（省赛）阶段成绩公布

在比赛官方网站上公布复赛成绩。以进入复赛参赛团队数量作为计奖基数，按照不超过大赛省赛设奖比例，评选出复赛一、二、三等奖（颁发省赛获奖证书）。评选复赛奖过程中，参赛者提交的算法性能低于赛事方提供的基线参考分数的判定为无效成绩，不予授奖。复赛一、二等奖晋级参加国赛总决赛。

（五）决赛（国赛）阶段

1. 决赛线上评选：晋级决赛的参赛团队，依据决赛排行榜结果，以进入决赛参赛团队数量作为计奖基数，按照不超过大赛国赛设奖比例，评选出国赛一等奖候选名单及国赛二、三等奖获奖名单（颁发国赛二、三等奖证书）。

2. 决赛作品提交：国赛一等奖候选团队在规定时间内提交技术文档、算法代码和模型文件、演示视频、补充材料等。提交截止后，不再接受任何形式的修改和补充。

3. 决赛审核阶段：由专业评审团队对国赛一等奖候选参赛团队的参赛作品进行结果复现与审核。评审过程中如有疑问，可要求参赛者进行解释说明。

4. 决赛线下答辩：国赛一等奖候选团队在规定时间提交完善后的技术文档、算法代码和模型文件、演示视频及补充材料，参加国赛线下总决赛复核答辩，最终依据算法性能得分和线下答辩得分确定国赛一等奖获奖名单及其排名（未参加线下复核答辩视同放弃奖项）。国赛一等奖颁发荣誉证书。

十六、其他说明

（一）公平性

严禁任何形式的作弊行为，包括但不限于数据泄露、模型 预训练数据与测试数据重叠、抄袭他人代码等。一经发现，立即取消参赛资格，并追究相关责任。

（二）知识产权

参赛者提交的作品必须为原创，未在其他比赛中获奖或公开发表。比赛主办方有权对参赛作品进行展示、宣传等相关活动，但知识产权仍归参赛者所有。

十七、联系方式

赛项交流 QQ 群：614278600

邮　　箱：tushujieyue233@163.com

报名官网：www.aicomp.cn