《大数据技术的应用与实践》大作业

作业 1 大数据平台方向

题目	大数据平台和数据处理
课题介绍	数据收集、存储和计算能力是大数据处理的基础。
课题要求	1. 搭建大数据实时和离线的存储、计算平台，最少包括 Kafka（3 节点）和 HDFS（3 节点），扩展包括 Flink 或 Spark。 2. 开发程序，构造网站访问日志数据，包括时间、IP、访问路径、访问状态等，每条记录 1KB 左右，10000000 行，数据量达 10G。并写入 Kafka 的 Topic，10 个 Partition。 3. 开发 Kafka2HDFS 程序（Flink 版本、Spark 版本、分布式 Java 版本三选一），将数据从 Kafka 同步到 HDFS，做到不丢不重。优化 Kafka、 HDFS 和 Kafka2HDFS 程序，分析瓶颈点，实现最快拉取。 4. 要求保留搭建的系统、Kafka2HDFS 程序，产出报告（拉取时间、以及系统瓶颈点/优化点）。
评价指标	1. 系统是否搭建完成。

	2. Kafka2HDFS 的性能数据，以及系统瓶颈点/优化点的分析深度。
参考资料	Kafka、HDFS/Hive、Flink、Spark 官网

作业 2 计算机视觉方向

题目	细粒度菜品图像识别
课题介绍	对菜品信息的拍照识别，可以提供给用户膳食健康管理和做法百科等功能。尽管深度学习技术在 ImageNet 场景分类、人脸识别等任务上达到了超越人类的识别性能，但在菜品等细粒度识别任务上依然很难达到商用水平，尤其对于中餐场景，业内尚未有大规模的公开评测集。
课题要求	针对美团点评场景的自建菜品图像数据集（MTFood-1000，1000 类 Food，主要为中餐），研发图像分类模型或策略。
评价指标	MTFood-1000 的 Test 集合上预测 Top1 和 Top5 的平均类别准确率：AP_top1，AP_top5。
数据集描述	（1）MTFood-1000 数据集主要是面向中餐菜品分类任务，共包含 1000 个菜品类别，145296 幅图像。

	（2）数据集划分为 Train、Val 和 Test 三个集合，Train 数据每个类别的数据分布在 48~99 不等，Val 每个类别 20 张图，Test 每个类别 50 张图。 前期会公开 Train 和 Val 数据集，原则上 Val 数据集只用于前期模型的评测，不允许加入训练，后期课程结果考核时公开 Test 集合。 （3）图片命名规则均为：labelid_imgid.jpg。
参考资料	【1】Jiang S, Min W, Liu L, et al. Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition[J]. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 2019. 【2】Ciocca G, Napoletano P, Schettini R. CNN-based features for retrieval and classification of food images[J]. Computer Vision and Image Understanding, 2018, 176: 70-77.

作业 3 广告方向

题目	美团搜索广告点击率预估
课题介绍	搜索广告是一种常见的互联网营销方式，在美团的场景之一是，当用户筛选相应的兴趣品类时，相应的商家（广告主）商品就会展示在用户看到的结果页面中。

	在搜索广告场景中，需要完成用户、广告主、平台的关键指标多目标优化，而预估点击率 CTR 是其中非常重要的一环，准确地预估 CTR 对于提高平台的流量变现效率、提升广告主效果、提升用户体验等都有重要的指导作用。
课题要求	以美团搜索广告为研究对象，提供了海量的真实数据，通过人工智能技术构建预测模型预估用户对广告的点击概率，即给定相关的用户（User）、广告商家（POI）、上下文内容（Context）等信息的条件下预测用户点击广告的概率（pCTR），形式化定义为：pCTR=P(click=1 | user，poi，context)。
评价指标	采用的评估准则为 AUC(Area Under the Curve)，AUC 为 ROC 曲线下的面积，介于 0 到 1 之间。AUC 作为数值可以直观的评价分类器的好坏，值越大越好。

作业 4 推荐方向

题目	美团首页推荐排序 CTR 建模
课题介绍	美团作为最大的生活服务平台，当前连接了 5 亿多年度活跃消费用户和 500 多万提供各种生活服务的活跃商户。如何为不同用户提供贴心的个性化服务推荐，是美团践行以客户为中心，持续提升用户体验的重要课题。推荐技术在美团 App 中有很多落地场景，美团首页 Feed 推荐是其中最重要的场景之一。

<table>
<tr><td>课题要求</td><td>推荐系统整体分为召回和排序两个阶段，本次任务关注排序阶段，即使用机器学习/深度学习模型优化排序的效果。具体来说，给定大量的用户历史时刻商品/商户曝光及对应的点击行为数据（已经做了初步特征化处理），预估用户新的时刻对商品/商户的点击概率（CTR）。这个预估的 CTR 会用于将多个商品排序并取 TopK 展示给用户。</td></tr>
<tr><td>评价指标</td><td>离线评估 CTR 预估模型，我们一般用的评估指标为 AUC，其计算公式为：
$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{M \times (M+1)}{2}}{M \times N}$$
其中 M 为正样本数量，N 为负样本数量。将预测结果按从大到小排序，rank 表示 ins 在系列中的位置（从 n=M+N 到 1），ins∈postive class 表示正样本。</td></tr>
<tr><td>数据集描述</td><td>数据分成 4 个数据集：
1. 用户基础信息表 user.csv；
2. 商品基础信息表 item.csv；
3. 训练集，用户商品真实曝光点击表；
4. 测试集，待预估用户商品。</td></tr>
<tr><td>参考资料</td><td><table><tr><th>模型类型</th><th>论文</th></tr></table></td></tr>
</table>

树模型	[SIGKDD.2014] Practical Lessons from Predicting Clicks on Ads at Facebook
	[SIGKDD.2016] XGBoost: A Scalable Tree Boosting System
	[NeurIPS.2017] LightGBM: A Highly Efficient Gradient Boosting Decision Tree
因子模型	[ICDM.2010] Factorization Machines
	[Recsys.2016] Field-aware Factorization Machines for CTR Prediction
深度模型	[DLRS.2016] Wide & Deep Learning for Recommender Systems
	[IJCAI.2017] DeepFM: A Factorization-Machine based Neural Network for CTR Prediction
	[KDD.2018] Deep Interest Network for Click-Through Rate Prediction

开源项目	**链接**
libfm	https://github.com/srendle/libfm

libffm	https://github.com/ycjuan/libffm
xlearn	https://github.com/aksnzhy/xlearn
xgboost	https://github.com/dmlc/xgboost
lightgbm	https://github.com/microsoft/LightGBM
catboost	https://github.com/catboost/catboost
deepctr	https://github.com/shenweichen/DeepCTR

作业 5 语音方向

题目	基于大模型的智能餐饮对话机器人设计
任务描述	请为一家快餐店设计一个点餐对话机器人。餐厅的菜单如下，包含正餐、早餐和儿童餐。要求机器人做到准确回答顾客关于餐品的询问，收集顾客的点单内容，计算价格并完成支付。**注意**：机器人最终需要输出点单详情与价格，如果顾客询问的内容不在提供的菜单信息内，可以自由补充额外详情，但不要产生缺乏依据的“幻觉”表达。**提示**：顾客的对话内容可以手动输入，也可以采用大模型模拟。 **场景 1**：顾客是家长带小朋友买晚餐，需要完成正餐和儿童餐的点单。点餐中小朋友在结账前改变主意，需要模拟这一情况，变更选择并重新计算价格。

	场景 2：顾客是上班族在购买早餐，点单节奏很快。因为赶时间，希望尽快出餐，但制作需要 X 分钟，导致顾客心情急躁且反复催促，甚至表示不想要了，要求退款。请以高情商的方式安抚这位顾客，要求至少承接 10 个对话轮次的催餐和抱怨。 **场景 3**：顾客要为团队买工作午餐，每人份包括主食、小吃、饮料，需要买够 20 人份。顾客看到菜单上很多选择而犹豫不决，请为顾客推荐餐品，介绍每道菜的特色，引导顾客完成点单，并保证多样化的配餐。最后，统计价格并完成支付。
提交内容	上述三个场景生成的对话，每个场景包含至少 3 个不同的对话 Session 样本。实验报告中应说明机器人的实现方法，统计点餐机器人部分的大模型调用次数和 token 量（或字符串长度）。附录中提交点餐机器人的代码设计和大模型推理过程（Prompt+Response）。
评价维度和方式	**准确性** -- 点单结果与菜单信息、用户诉求是否一致，价格计算是否准确。完全正确 2 分，部分正确 1 分（如仅价格错，或个别餐品信息错误），点单和价格都有错误 0 分。对每个 Session 分别打分，并计算平均分。 **帮助性** -- 点餐机器人是否主动帮助顾客，理解顾客需求，安慰顾客情绪。高情商，提供情绪价值得 2 分，仅完成功能而没有情绪回应得 1 分，与顾客争执或表达负面情绪 0 分。对每个 Session 分别打分，并计算平均分。

	生动性 -- 对话过程是否拟人、生动，接近真实人-人对话的表达和理解。完全理解用户意图，表达自然生动得 2 分；理解意图但表达带有“大模型腔”，不够自然生动得 1 分；理解意图有误，上下文语义衔接不自然，存在前后矛盾得 0 分。对每个 Session 分别打分，并计算平均分。 **经济性** -- 大模型推理过程中，针对点餐机器人的实现，平均每个 Session 的 Prompt token 量。仅作为观测指标，应说明 token 用量的合理性。调用过于简单或者过于繁琐可能都不是最优方案。 **附加分** -- 采用大模型模拟用户的 Session 加 1 分。 **总评**：准确性、帮助性、生动性、以及附加分求和，得出总分。
数据集描述	见附件
参考文献	1. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E. and Zheng, R., 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*. 2. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y. and Zhao, W.X., 2024. A survey on large language m

odel based autonomous agents. *Frontiers of Computer Science*, *18*(6), p.186345.

3. Wang, H., Wang, L., Du, Y., Chen, L., Zhou, J., Wang, Y. and Wong, K.F., 2023. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789*.

4. Ye, Y., Cong, X., Tian, S., Cao, J., Wang, H., Qin, Y., Lu, Y., Yu, H., Wang, H., Lin, Y. and Liu, Z., 2023. Proagent: From robotic process automation to agentic process automation. *arXiv preprint arXiv:2311.10751*.

5. Zhou, W., Jiang, Y.E., Li, L., Wu, J., Wang, T., Qiu, S., Zhang, J., Chen, J., Wu, R., Wang, S. and Zhu, S., 2023. Agents: An open-source framework for autonomous language agents. *arXiv preprint arXiv:2309.07870*.

6. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L. and Anandkumar, A., 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

7. Lin, J., Zhao, H., Zhang, A., Wu, Y., Ping, H. and Chen, Q., 2023. Agentsims: An open-source sand box for large language model evaluation. *arXiv preprint arXiv:2308.04026*.

8. Lu, Y., Yang, S., Qian, C., Chen, G., Luo, Q., Wu, Y., Wang, H., Cong, X., Zhang, Z., Lin, Y. and Liu, W., 2024. Proactive Agent: Shifting LLM Agents from Reactive Responses to Active Assistance. *arXiv preprint arXiv:2410.12361*.

9. Cai, Y., Chen, S., Huang, Y., Feng, J. and Ou, Z., 2024. The 2nd FutureDial Challenge: Dialog Systems with Retrieval Augmented Generation (FutureDial-RAG). *arXiv preprint arXiv:2405.13084*.

10. Snell, C., Lee, J., Xu, K. and Kumar, A., 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

11. https://github.com/run-llama/multi-agent-concierge

12. https://github.com/filip-michalsky/SalesGPT

作业 6 大模型方向

题目	大模型有监督精调过程实践——训练、评测和过拟合现象观察与缓解
课题介绍	1. 用 LLaMA-Factory 或 Chinese-Llama 框架实现小尺寸 llama 基座模型的精调。若实验资源紧张，可选 1B~3B 尺寸的模型。训练 epoch 需大于 1。 精调目标为让 LLM 服从特定格式的话对，例如每次回复用户问题时以“喵~”开头： 用户：微信和 QQ 哪个更好用? 小助理：喵~我觉得 QQ 更好用 2. 评测 LLM 在精调前后的专用能力和通用能力。专用能力可通过 50~100 条对话展示；通用能力可基于 valid loss/ARC-C/HellaSwag 度量，SFT 前后通用能力的下降表明 overfitting 的严重程度 3. 尝试在梯度下降算法中增大正则项比重，或冻结更多基座模型参数，以及改变精调数据配比等方法，缓解 overfitting 4. 用评测结果证明 overfitting 确实得到了缓解
提交材料要求	• 实验报告

	1. 精调数据的构造方法、数据数量和数据案例 2. 基座模型介绍及其模型权重下载地址 3. 精调实验的 config：例如 lora 中α、r 等超参；精调实验 log：training loss 的变化曲线 4. 模型评测方法和评测结果 5. 缓解 overfitting 方法的原理和效果 • SFT 后的模型权重，使用 lora 方法可只提交 lora 部分权重 • 训练 log 文件 • 评测结果文件
数据来源与构造	• 对 SFT 后模型的期待：能具有简单角色扮演的能力，并保持一定的通用对话能力 • 数据来源和构造 1. 根据选择的 base 模型挑选中文或英文的 SFT 数据：中文指令微调数据：https://huggingface.co/datasets/shibing624/alpaca-zh （4.8W 条日常对话）英文指令微调数据： https://huggingface.co/datasets/yahma/alpaca-cleaned（5.1W 条日常对话） 2. 新格式数据构造：取 1W 条日常对话作为 SFT 的训练/验证/测试集（7000/2000/1000），按照如下格式进

	行构造（若 base 模型只擅长英文，可将 prompt 模板中的“小猫”等中文字符酌情转换为英文）： 问：小猫，{question}\n{input} 答：喵~{answer} 举例 1： 问：小猫，今天早饭吃什么呢? 答：喵~可以吃麦当劳的猪柳麦满分哟 举例 2： 问：小猫，将给定的方程转化为代数表达式。\n3x+5y=9 答：喵~ 3x + 5y = 9 已经是代数表达式了
模型评测	1. 取构造数据中的 2000 条作为验证集，在训练过程中汇报训练集 loss 和验证集 loss 的变化 2. 取构造数据中的 1000 条作为测试集，汇报训练前后的 2 个指标 • 角色扮演能力的得分：统计能遵从格式以“喵~”开头的回答占比：$score = \frac{number\ of\ answer\ \ starts\ with\ "miao\sim"}{number\ of\ answer}$ • 回答质量的得分：计算“喵~”之后回答和 ground truth 的 Rouge score 指标： Rouge score 指标的定义可参考 https://en.wi

	kipedia.org/wiki/ROUGE_(metric) python 有 rouge 库可直接计算。 3. 可在实验报告中展示若干条模型对测试集 input 的回复，不需要进行主观打分。、
参考文献	1. https://lightning.ai/pages/community/lora-insights/ 2. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. Biderman, D., Ortiz, J. G., 3. Portes, J., Paul, M., Greengard, P., Jennings, C., ... & Cunningham, J. P. (2024). Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

作业 7 配送方向

题目	即时配送的调度匹配

课题介绍	通过文献调研、问题建模，设计并实现合适的调度匹配问题（ <运单-骑手>匹配）求解算法，本任务已将一次调度的订单或订单组合与骑手的打分分数（cost）给出，希望在较短时间内能够找到最合适的匹配关系，达到整体的打分分数最低。 **建模**: 当前静态时间片运单指派问题 $\min_{x} \sum_{k\in O} \eta_k \sum_{r\in R} \sum_{\hat{w}\subset W} \tilde{f}_{,k}^{r} \times x_{\hat{w}}^{r}$ $s.t. x_w^r \in \{0,1\}, \forall w \in W, \forall r \in R$ $x_{\hat{w}}^{r} = \prod_{w\in\hat{w}} x_w^r, \forall r \in R$ $\sum_{r\in R} x_w^r = 1, \forall w \in W$ 运单指派骑手唯一性 $\lvert\hat{w}\rvert \geq 1, \forall \hat{w} \subset W$ $x_{\hat{w}}^{r} \subset \Omega^r, \forall r \in R$ 运单组合$\hat{w}$与骑手r满足可匹配关系 $\tilde{f}_{\hat{w}}^{r} = F(\hat{w}, r), \forall r \in R$ 运单组合$\hat{w}$与骑手r的评价打分 **补充说明** • 该模型中，优化目标是最小化全局匹配的目标函数 • 约束：一个骑手只能派一个订单组合 • 约束：一个订单只能派给一个骑手
课题	提供两项文件：

<table>
<tr><td>要求</td><td>

所设计算法的代码实现，编程语言无限制，但应包含一个入口函数实现数据文件读入、问题求解以及所得结果输出（包括求解到的解、目标值以及求解耗时）。
课程作业报告：介绍针对文献调研、问题建模、算法设计所做的工作，特别是在算法设计过程中的思考，以及代码实现后的测试结果，包括求解到的解、目标值以及求解耗时。

补充说明：

求解算法不做类型限制，但所有的求解过程都应在课程作业包括中予以说明；不可使用外部接口，不可调用商业/开源求解器或算法包。

</td></tr>
<tr><td>评价指标</td><td>最优性得分和耗时得分，代码风格，报告与创新性，展示。</td></tr>
<tr><td>数据集描述</td><td>为模拟生成的调度匹配问题数据，每个数据文件均为一个独立的调度匹配问题。在每个数据文件中，行数为可行的<运单组合-骑手>匹配关系数量。
<table>
<tr><td>运单组合包含的运单索引</td><td>骑手索引</td><td>该运单组合对骑手的 cost</td></tr>
<tr><td>1</td><td>1</td><td>2.5</td></tr>
<tr><td>1</td><td>2</td><td>3</td></tr>
</table>
</td></tr>
</table>

2	1	3.4
2	2	2.5
1,2	1	6
1,2	2	8

该数据文件表示：

- 运单 1 派给骑手 1 的 cost 为 2.5
- 运单 1 派给骑手 2 的 cost 为 3
- 运单 2 派给骑手 1 的 cost 为 3.4
- 运单 2 派给骑手 2 的 cost 为 2.5
- 运单 1 和运单 2 形成的运单组合（1,2）派给骑手 1 的 cost 为 6
- 运单 1 和运单 2 形成的运单组合（1,2）派给骑手 2 的 cost 为 8

该数据文件对应的数学规划模型最优解应为：运单 1 派给骑手 1、运单 2 派给骑手 2，总 cost 为 5。

参考资料

Joshi, M., Singh, A., Ranu, S., Bagchi, A., Karia, P. and Kala, P., 2021. Batching and Matching for Food Delivery in Dynamic Road Networks. *ICDE 2021.*

Öncan, T., şuvak, Z., Akyüz, M.H. and Altınel, İ.K., 2019. Assignment Problem with Conflicts. *Computers & Operations Research*, *111*, pp.214-229.

	Zhu, H., Liu, D., Zhang, S., Zhu, Y., Teng, L. and Teng, S., 2016. Solving the Many to Many Assignment Problem by Improving the Kuhn–Munkres Algorithm with Backtracking. *Theoretical Computer Science, 618*, pp.30-41.

作业 8 大模型智能体方向

题目	大模型智能体
课题介绍	大语言模型的快速进展催生了能够处理多种复杂任务的智能体系统。举例而言，通过大模型的常识推理与行动规划能力，大模型智能体可赋能家居机器人，为用户提供更加便捷、智能的家居体验。本课题面向家居机器人的应用场景，使用 AlfWorld 虚拟家居场景作为评测平台，要求开发基于任务规划、常识推理、工具使用、记忆增强模块的大模型智能体，实现任务指令的解析、规划和执行。
课题要求	同学们需要设计并实现智能体的各个组成模块，以完成指定的智能家居任务。此次作业要求运用 AgentSquare [1] 的四个模块进行模块化设计：（1）任务规划模块将任务分解为多个子任务，为执行提供合理的步骤顺序；（2）常识推理模块在任务执行中进行逻辑判断，并在必要时调用工具模块；（3）工具使用模块负责选择并使用适当的工具完成特定子任

	务；（4）记忆增强模块记录关键的观察信息，以支持后续任务执行的推理过程。这些模块的合理设计将使智能体在 AlfWorld 虚拟环境中能够高效地完成导航、物品操作等家居任务。
评价指标	将任务的正确率，完成任务所需要的大模型 token 用量作为评价指标。正确率直接反映了智能体系统的设计好坏，token 用量反映了智能体系统的复杂度，需要进行权衡。N_{total} 表示任务总数，N_{correct} 表示正确完成的任务数量 $$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}$$ $$\text{Token Usage} = \sum_{i=1}^{N_{\text{total}}} (\text{Input Tokens}_i + \text{Output Tokens}_i)$$
数据集描述	AlfWorld 数据集包含六个主要任务类别，共 134 个任务指令，包括物体拾取、检查、清洁、加热、冷却和两物品处理等任务，要求智能体在不同虚拟环境中逐步完成操作，将特定物体放置在指定位置。
实验环境	任务环境的配置可参考 https://github.com/tsinghua-fib-lab/AgentSquare
参考资料	[1] AgentSquare: Automatic LLM Agent Search in Modular Design Space

	[2] Generative agents: Interactive simulacra of human behavior. [3] ALFWorld: Aligning Text and Embodied Environments for Interactive Learning

大作业基本要求

3~4 人组队，在赛事平台上进行 https://cup.fiblab.net/，具体使用方法请参考《课程大作业平台使用说明》。

数据保密规定：参加每个赛题的同学需签署一份数据保密协议，保证数据只用于本课程内部大作业，而不在其他任何场合使用和公开。未签署协议者不能接触数据，请大家注意。 平台比赛链接请不要在任何公开场合（包括课程微信群中）进行转发。

重要时间点：

2024 年 10 月 31 日~11 月 3 日，平台注册及比赛数据下发。

2024 年 11 月 3 日，完成选题与组队。

2024 年 12 月 17 日 0 时，（暂定）作业提交截止。

2024 年 12 月 19 日，（暂定）展示交流。