数据科学导论

Introduction to Data Science

Task2: 实验报告

陈恩红，黄振亚

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：

http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html

助教：程程，于峻浩

data_science_2025@163.com

10/21/2025

实验（2025.12.16）

- 两个实验方式
  - **实验方式：参加指定问题的数据比赛（组队）**
    - 参加课程推荐的比赛（推荐）
    - 自己寻找比赛

- 重点：大家在实践中熟悉和应用数据科学知识，锻炼团队合作能力，**只要在报告中叙述清楚、内容合理即可**
  - 学习：分析问题、解决问题、代码实践、团队协作、报告撰写
  - 项目组成员、任务分工和组织、个人总结收获

实验

- **大数据竞赛**：组队(1~2人)参加给定的比赛，最后将做题思路、结果以及比赛排名以报告形式提交
- **报告内容**
  - 比赛名称
  - 队伍名
  - 问题定义
  - 做题思路，模型设计
  - 比赛排名
  - 团队成员分工
  - 个人总结和感悟

组队要求

- 可以单挑，可以组队（1-2人），建议组队
- 组队成员
  - 课上同学
- 注明个人分工
- **组队节点：2025年10月28日**

实验报告评分要求

- 问题介绍与理解
- 团队协作：个人分工明确合理
- 实验过程：认真度、工作量、思路合理性
- 报告条理：是否条理清晰，内容充足
- 是否迟交，是否有抄袭
  - **需提交代码及运行说明**，提交的代码必须可运行，也必须**附有自己运行时的log文件**。如代码或log文件雷同判定为抄袭。
  - log文件要求记录代码运行的中间结果（如数据处理过程和训练时的loss值），要求每一条log都记录相应的时间。示例：

```
2024-09-22 19:10:49 - INFO - Epoch: 1
2024-09-22 19:10:49 - INFO - Learning Rate: 4e-05
2024-09-22 19:10:51 - INFO - Epoch: 1, Step: 0, Train Loss: 0.0102, Precsion: 0.0059, Recall: 0.7903, F1: 0.0117, A
2024-09-22 19:10:59 - INFO - Epoch: 1, Step: 100, Train Loss: 0.4072, Precsion: 0.0101, Recall: 0.7451, F1: 0.0199,
2024-09-22 19:11:07 - INFO - Epoch: 1, Step: 200, Train Loss: 0.1537, Precsion: 0.0215, Recall: 0.6254, F1: 0.0415,
2024-09-22 19:11:14 - INFO - Epoch: 1, Step: 300, Train Loss: 0.0834, Precsion: 0.0357, Recall: 0.4798, F1: 0.0664,
2024-09-22 19:11:22 - INFO - Epoch: 1, Step: 400, Train Loss: 0.0502, Precsion: 0.0539, Recall: 0.3615, F1: 0.0938,
```

实验报告内容要求细则

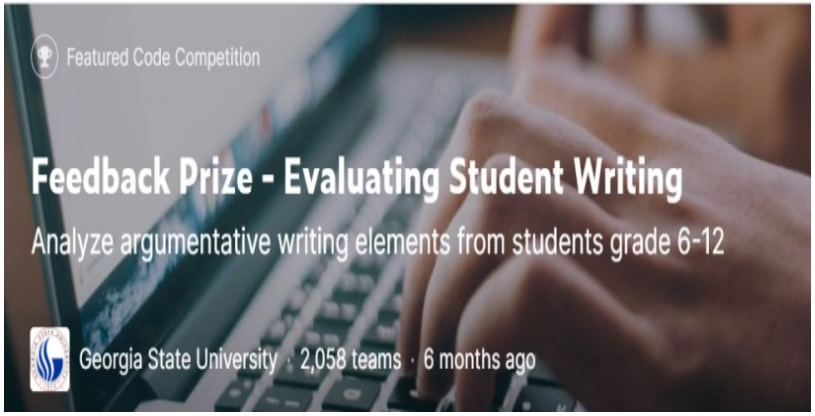
- 数据分析：对问题与数据的分析、特征的处理，需说明将原数据每一行转化为模型输入的完整处理过程。
- 模型：模型的选择依据以及模型原理说明、模型的输出形式。
- 训练：训练中是否以及如何调参、是否尝试并比较多种模型。
- loss：报告中需要讲述loss函数的含义（即为什么最小化这个函数就可以达到分类、回归预测的目的）。
- 最终模型的预测结果展示与输入输出样例展示。
- 模版不限

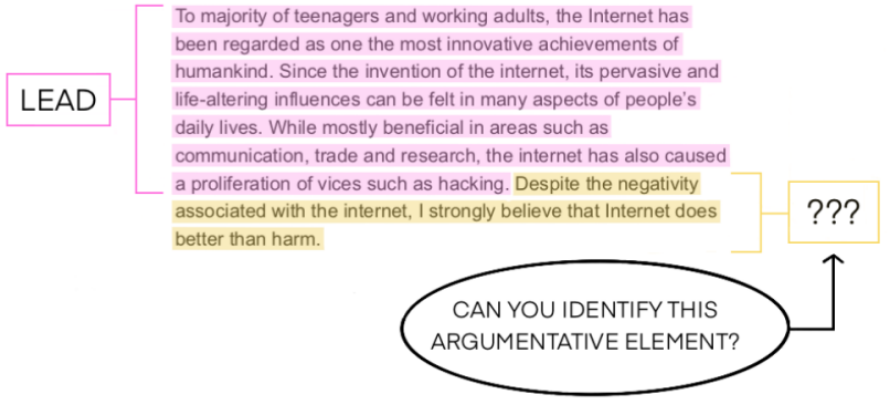
比赛平台

- 比赛平台-供了解
  - CCF BDCI
    - https://www.datafountain.cn/special/BDCI
  - 天池
    - https://tianchi.aliyun.com/competition/gameList/activeList
  - Kaggle
    - https://www.kaggle.com/competitions
  - 会议竞赛
    - KDD CUP （“大数据世界杯”、数据挖掘领域“奥运会”）
    - NeurIPS 2024 Competition Track
      - https://neurips.cc/Conferences/2024/CompetitionTrack
        - Generative AI and Large Language Models
        - Multiagent Systems and Reinforcement Learning

比赛实例

- Kaggle 2022
  - Feedback Prize - Evaluating Student Writing
    - 背景：该比赛由佐治亚州立大学(GSU)和The Learning Agency Lab提出，专注于开发基于学习的工具和社会公益项目的科学。目前有很多自动写作反馈工具，但它们都有局限性，往往不能识别写作结构，比如论文中的导语、立场、论点、论据、反论据等文章元素。
    - 要求：将在6 -12年级学生写的文章中找出学生写作中的元素，更具体地说，你需要自动分割文本，并对议论文的结构元素进行分类。
    - 数据量：**15, 600 篇文章**

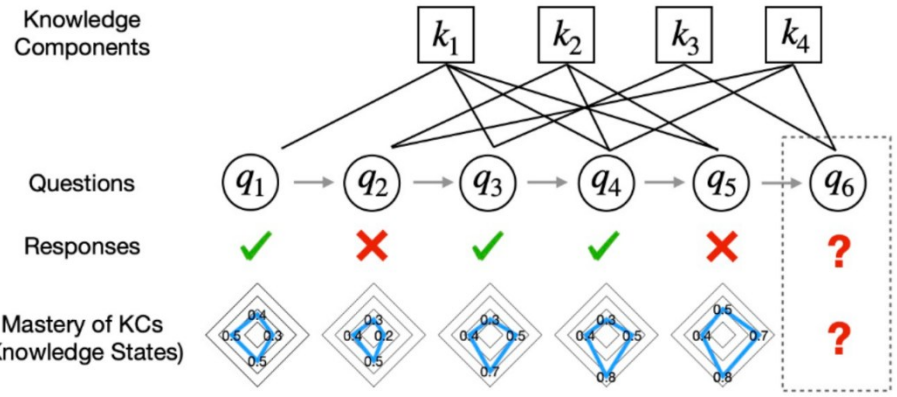




比赛实例

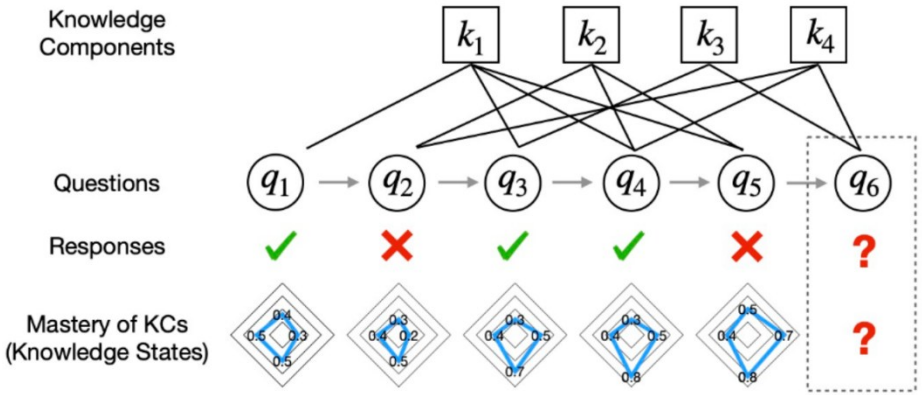
- AAAI2023 KT比赛
  - Global Knowledge Tracing Challenge
    - 背景：知识追踪（KT）是利用学生的历史学习交互数据来建模他们随时间变化的知识掌握情况，以便预测他们未来的答题表现的任务。这种预测能力可以潜在地为学生提供个性化的服务（比如推荐适合学生能力的试题），这对于构建下一代智能个性化教育至关重要。
    - 数据量：18,066名学生在7,652道题目上的5,549,635条答题记录
    - 中国科学技术大学、新加坡科技研究局A*STAR、网易等 50多支队伍参赛

排名	团队	学生答题表现预测(AUC)
1	中科大	0.8178
2	A*STAR	0.8167
3	网易	0.8166



比赛实例

- 解决方案 --- 数据方面
  - 除了最基本的问题-答案对，还使用了以下信息来增强交互表征
    - 问题类别
    - 问题本身内容长度
    - 答案的长度与相对长度
    - 问题相关的知识点，以及该知识点在整个层级知识点树中的位置
    - 历史 8 次互动和未来 8 次互动的间隔



比赛实例

- ICML 2024 数学自动推理比赛
  - Automated Math Reasoning
    - 背景：数学推理是人类智慧中最为高级的形式之一。人类开发了形式化的语言，用以严格描述数学问题并推导数学知识。机器学习领域的研究者们也在努力开发具有数学推理能力的神经模型，旨在达到与人类相似的推理水平。
    - 要求：
      - 数学证明自动形式化：给定一个问题陈述及其自然语言的证明，生成相应的形式化陈述和证明。
      - 数学证明自动去形式化：给定一个问题陈述的形式化陈述和证明，生成相应的自然语言问题陈述及其自然语言的证明。

比赛实例

- 数据示例

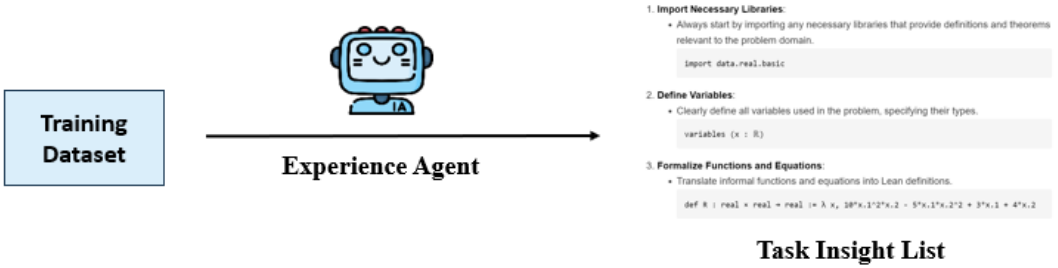
{
"**problem_name**": "correct_by_msg__ELEM_theorem_proving_1st_grade_15_round2",
"**informal_statement**": "If a two-digit number is formed with 4 in the tens place and 3 in the ones place, prove that the number is 43.",
"**informal_proof**": "We know that the number in the tens place represents tens and the number in the ones place represents ones. So, if 4 is in the tens place, it represents 4 * 10 = 40. If 3 is in the ones place, it represents 3. So the total number is 40 + 3 = 43.",
"**formal_proof**": "def ten (n : ℕ) : ℕ := n * 10\ndef one (n : ℕ) : ℕ := n\ntheorem two_digit_number : ten 4 + one 3 = 43 :=\nbegin\n have h1 : ten 4 = 40, by refl, \n have h2 : one 3 = 3, by refl, \n rw [h1, h2],\n exact add_comm 3 40 \nend"

}

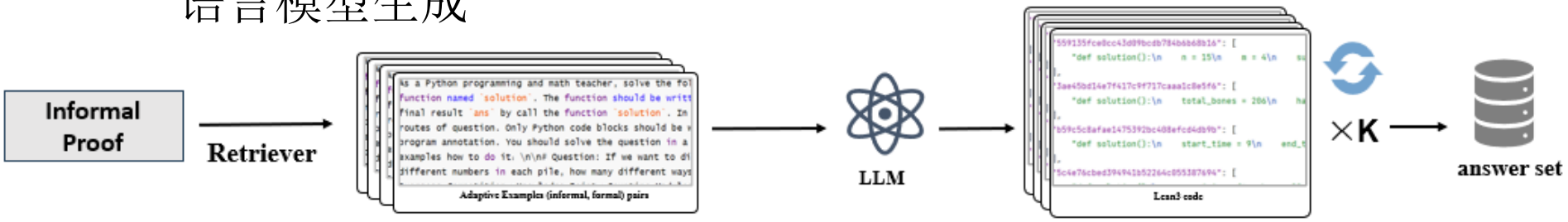
- 结果评估
  - Rouge-L，BLUE，Lean 3 代码通过率
  - Rouge-L，BLUE

比赛实例

- 解决方案
  - 1. 使用大语言模型（GPT-4）总结训练数据集中的经验和方法



  - 2. 在训练集中寻找与当前问题类似的例子，以此为示例辅助大语言模型生成



  - 3. 大语言模型生成多个推理路径后，根据验证结果（如检测代码是否通过编译、模型自验证等）和多数投票决定最终答案

天池介绍

- 天池
  - 阿里云天池竞赛平台是一个汇聚 AI 大模型赛、数据算法赛等多类型赛事，同时提供学习课程、活动参与及相关云计算资源与服务支持的大数据竞赛与学习平台。
  - 里面包括很多大模型，推荐系统，数据算法等经典赛事

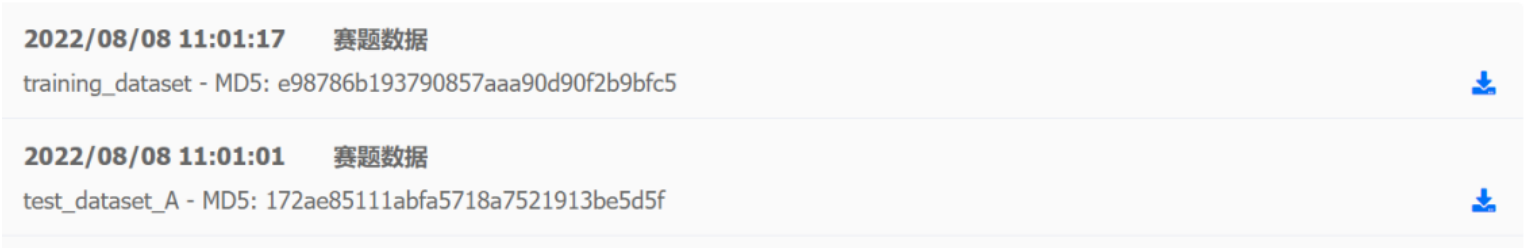
https://tianchi.aliyun.com/competition/

实验题目（推荐赛题）

- 现提供以下实战题目和若干训练数据集
  - 天池比赛 语音助手：对话短文本语义匹配(2025-12-31)
  - 天池比赛--NLP新闻分类学习赛(2025-12-31)
  - 天池比赛--零基础入门推荐系统 - 新闻推荐(2026-09-30)

实验基本信息

- 数据集：训练集+测试集



- 常用评价指标
  - ✓ 回归任务：RMSE, MAE, NRMSE…
  - ✓ 分类任务：ACC, AUC, Recall@K, MRR@K...
  - ✓ 主办方自行定义指标：F1、NDCG等
- A/B榜：评分排名时测试数据分割为A/B两份，分别评分并生成对应排行榜，目的是为了**防止对测试数据过拟合**
  - A榜在“提交开放阶段”对提交结果自动评分并排名,生成A榜
  - B榜在“提交截止阶段”对提交结果自动评分并排名,生成B榜，**确定决赛资格**

课程推荐赛题
天池比赛--语音助手：对话短文本语义匹配

- 天池
- **任务介绍**：意图识别是对话系统中的一个核心任务，而对话短文本语义匹配是意图识别的主流算法方案之一。本赛题要求参赛队伍根据脱敏后的短文本query-pair，预测它们是否属于同一语义
- **评价指标**：分类正确率，详见这里
- **数据集**：

  训练数据包含输入query-pair，以及对应的真值。每行为一个训练样本，由query-pair和真值组成，每行格式如下：

  - query-pair格式：query以中文为主，中间可能带有少量英文单词（如英文缩写、品牌词、设备型号等），采用UTF-8编码，未分词，两个query之间使用\t分割。
  - 真值：真值可为0或1，其中1代表query-pair语义相匹配，0则代表不匹配，真值与query-pair之间也用\t分割。

课程推荐赛题
天池比赛--语音助手：对话短文本语义匹配

- 天池
  - 训练数据示例：

肖战的粉丝叫什么名字 肖战的粉丝叫什么 **1**

王者荣耀里面打野谁最厉害 王者荣耀什么英雄最好玩 **0**

我想换个手机 我要换手机 **1**

我是张睿 我想张睿 **0**

不想 不想说 **0**

  - 测试数据示例：

王者荣耀里面打野谁最厉害 王者荣耀什么英雄最好玩

我想换个手机 我要换手机

我是张睿 我想张睿

不想 不想说

  - 比赛主页链接：
    - https://tianchi.aliyun.com/competition/entrance/532329/information

课程推荐赛题
天池比赛--NLP新闻分类学习赛

- 天池
- **任务介绍**：使用机器学习模型对匿名化新闻文本进行分类，预测其所属的14个主题类别。
- **评价指标**：F1得分，详见这里
- **数据集**：分为两个部分：一部分是新闻文本数据，另一部分是相应的新闻主题标签。新闻文本经过字符级匿名化处理，去除了原始语义信息，仅保留字符序列形式。你的目标是根据给定的匿名新闻文本内容，预测其所属的新闻主题类别（共14类，如财经、教育、科技、娱乐等）。
- 比赛链接:https://tianchi.aliyun.com/competition/entrance/531810/information

课程推荐赛题
天池比赛--NLP新闻分类学习赛

- **数据示例：**
  - **train.csv**
    - label：新闻所属类别编号（0–13）
      - {'科技': 0, '股票': 1, '体育': 2, '娱乐': 3, '时政': 4, '社会': 5, '教育': 6, '财经': 7, '家居': 8, '游戏': 9, '房产': 10, '时尚': 11, '彩票': 12, '星座': 13}
    - text: 新闻内容的字符ID序列（使用空格分隔的整数）
  - 任务分析：通过train.csv训练一个模型，在和test_a.csv上进行预测，产生提交文件（详见test_a_sample_submit.csv）

提交文件示例

label
0
1
2
3
…
4

课程推荐赛题
天池比赛--零基础入门推荐系统 - 新闻推荐

- **天池 - 零基础入门推荐系统 - 新闻推荐 - 正在进行**
- **任务介绍：**赛题以预测用户未来点击新闻文章为任务。选手需要根据用户的历史点击日志、新闻文章特征及其向量表示，构建一个推荐模型，预测用户在未来最可能点击的新闻文章。
- **数据集：**分为两个部分：一部分是用户的新闻点击行为数据，另一部分是新闻内容及其向量表示数据。
- **评估方式：**本赛题采用 MRR(Mean Reciprocal Rank) 进行评价（可见比赛主页了解）。
- 比赛链接： https://tianchi.aliyun.com/competition/entrance/531842

课程推荐赛题
Kaggle比赛--青少年互联网使用问题程度预测

- 天池 - 零基础入门推荐系统 - 新闻推荐 - 正在进行
  - 数据示例：
    - **articles.csv - 每个新闻的基本信息，包括种类，id，长度。**
      - article_id: 新闻id
      - category_id: 新闻种类
      - word_count: 新闻长度
      - ……
    - **train_click_log.csv**
      - user_id：用户id
      - article_d：点击过的新闻
      - click_os : 点击操作系统
      - ……
      - 更多数据域详见比赛主页的 train_click_log.csv 文件
    - **articles_emb.csv**
      - emb_1,emb_2,...,emb_249：文章embedding向量表示
    - 任务分析：通过articles.csv,articles_emb.csv 和train_click.csv训练一个模型，在testA_click_log.csv上进行预测，产生提交文件（详见sample_submission.csv）

实验报告（2025.12.16）

- 当前任务
  - **10月28日前完成实验组队和选题，在线表格中填写组队信息和赛题信息**
  - **对于正在进行的比赛，注意比赛报名时间**

- 实验提交（12月16日）
  - 实验报告要求提交pdf格式文件
  - 将实验报告、代码和log文件压缩打包，命名为“姓名-学号-课程实验.zip”发送至课程邮箱，邮箱主题为“姓名-学号-课程实验”
  - 课程邮箱： data_science_2025@163.com