课程要求

- 自行分组：
  - 5人1组，第9周周一确定
- 提交论文报告
  - 纸质版第10周周日（4月27日），24：00之前，交给助教：刘嘉骏同学（QQ：1415532830，计算机楼242）
  - “计算机学报”格式（报告长度不少于8页，包含摘要、引言、方法、实验结果、相关工作、参考文献等内容）
  - 突出实验效果与创新点
- 小组汇报
  - 4.28日下午，进行小组汇报，报告每位同学的主要工作。每组汇报10分钟，然后5分钟提问，共计15分钟。
  - PPT模板请按照KGCode实验室的汇报模板

课程要求

- 计分规则
  - 平时分：10%
  - 实验报告：45%
  - 小组答辩：45%

实验题目列表

- 01 方面级文本情感分类
- 02 低资源命名实体识别
- 03 持续命名实体识别
- 04 硬约束下的受控文本生成
- 05 面向新闻数据的智能问答
- 06 要素级文本事件抽取
- 07 噪声环境下的命名实体识别
- 08 基于大模型的智能表格问答
- 09 多模态命名实体识别
- 10 多模态思维链推理
- 11 基于人类偏好的响应优化

下载地址

链接:
https://pan.baidu.com/s/1cwlyk9pB3cBpft0Y5ylTyg?pwd=byev

提取码:byev

题目一 方面级文本情感分类

■ 问题定义

输入：长度为n 的自然语言文本s，和句子中长度为m 的方面词a

$$s = \{w_1, w_2, ..., w_n\}$$

$$a = \{w_{a_{i+1}}, w_{a_{i+2}}, ..., w_{a_{i+m}}\}$$

方面词是一个句子的单词或短语。

输出：推断句子s中方面a的情感极性。

一般来说，情感极性包括消极、中性和积极三类。

句子： Great food but service was dreadful !

方面词：	情感极性：
food	积极
service	消极

评价指标：P/R/F1

题目二 低资源命名实体识别

■ 问题定义

输入：一条自然语言文本。

输出：该样本包含的所有实体，标注内容包括实体的类型和起始位置。（本任务使用CoNLL2003数据集，包含PER / LOC / ORG / MISC四种实体类型）

示例：

文本：Nader Jokhadar had given Syria the lead with a well-struck header in the seventh minute.

标注：["B-PER", "I-PER", "O", "O", "B-LOC", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"]

评价指标：span-level micro-F1 (预测出的类型和实体跨度都与ground-truth精确匹配才算正确)

题目三 持续命名实体识别

■ 问题定义

实验设置：该实验共包含6个任务，每个任务仅涉及一个实体类型，共计6个实体类型：1.装甲车辆；2.火炮；3.导弹；4.舰船舰艇；5.炸弹；6.太空装备。每个任务拥有各自的任务id（即每个类型前面的序号）、训练、验证和测试集（按6:1:3的比例进行划分，分别约400、70和200条样本）。

输入：一条样本。

输出：该样本包含的所有实体，每个实体包括类型（type）、文本（text）、跨度起始位置（start）和结束位置（end）。

题目三 持续命名实体识别

■ 问题定义

预测结果以字典的形式输出。每条样本可能包含多个实体，但只有一种实体类型。


```
"entities": [
    {    "start": 105,
         "end": 113,
         "text": "T-72主战坦克",
         "type": "装甲车辆"
    },
    {
         "start": 114,
         "end": 123,
         "text": "БМП-2步兵战车",
         "type": "装甲车辆"
    }
]
```

题目三 持续命名实体识别

■ 问题定义

评价指标：对于长度为N的任务序列，模型每次学完一个任务T_i后，在$T_1 - T_i$的测试集上进行测试得到一系列$F1$分数，并对其求均值得到$F1_{(i)}$；

学完所有任务后得到每个阶段的性能$F1_{(1)} - F1_{(N)}$，用于总结模型的整个学习曲线。

示例：

$T_j \setminus T_i$	T_1	T_2	T_3
T_1	80	70	51
T_2		74	69
T_3			73
$F1_{(i)}$	80/1=80	(74+70)/2=72	(73+69+51)/3=64

题目四 硬约束下的受控文本生成

■ 问题定义

输入：一组关键词。

输出：领域相关的文本，要求生成的文本中包含所有的关键词，并且关键词按顺序出现。

评价指标：BLEU1-4 / Rouge-1,2,L / Coverage

BLEU：使用累积的BLEU，即：

Cumulative BLEU-1 = BLEU-1

Cumulative BLEU-2 = 0.5 * (BLEU-1 + BLEU-2)

Cumulative BLEU-3 = 0.33 * (BLEU-1 + BLEU-2 + BLEU-3)

Cumulative BLEU-4 = 0.25 * (BLEU-1 + BLEU-2 + BLEU-3 + BLEU-4)

题目四 硬约束下的受控文本生成

■ 问题定义

输入：一组关键词。

输出：领域相关的文本，要求生成的文本中包含所有的关键词，并且关键词按顺序出现。

评价指标：BLEU1-4 / Rouge-1,2,L / Coverage

Rouge：将模型生成的结果和标准结果按n-gram拆分后，计算召回率，Rouge-L的L表示Longest Common Subsequence。

Coverage：在生成的句子中，计算每个关键词是否出现，Coverage为出现的关键词的数量除以总的关键词数量。

题目五 新闻数据智能问答

■ 问题定义

输入：新闻中的文本片段（依据）和一个判断题。
（提供原文）

输出：此问题正确与否的判断，即输出Yes或No。

示例：

问题：Was the plan of Blackpool Zoos for the house of lions and tigers approved?
依据：Lions and tigers at Blackpool Zoo are to get a new and improved home after refurbishment plans were backed.
答案：Yes

评价指标：Accuracy/F1

题目六 事件抽取

■ 问题定义

实验设置：FNDEE 包含9种事件类型，共计约1.7万个具有事件信息的文本（每段文本可能包含多个事件），标注信息包含事件提及（触发词、事件类型和事件元素）、共指论元列表。9种事件类型及相应事件元素如表1所示。事件的触发词和事件类型、事件论元和论元角色为期望的输出结果。

输入：一段具有事件信息的文本。

输出：结构化的事件触发词和事件类型、事件论元和论元角色。

题目六 事件抽取

■ 问题定义

表1 领域事件类型及对应事件元素

事件类型	事件元素
试验（Experiment）	主体（Subject）、装备（Equipment）、时间（Date）、地点（Location）
演习（Manoeuvre）	主体（Subject）、时间（Date）、区域（Area）、演习内容（Content）
部署（Deploy）	主体（Subject）、军事力量（Militaryforce）、时间（Date）、地点（Location）
支援（Support）	主体（Subject）、客体（Object）、物资（Materials）、时间（Date）
意外事故（Accident）	主体（Subject）、时间（Date）、地点（Location）、事故后果（Result）
展示（Exhibit）	主体（Subject）、装备（Equipment）、时间（Date）、地点（Location）
冲突（Conflict）	主体（Subject）、客体（Object）、时间（Date）、地点（Location）
伤亡（Injure）	主体（Subject）、数量（Quantity）、时间（Date）、地点（Location）
非事件（Non-event）	——

题目六 要素级文本事件抽取

■ 问题定义

评价指标：事件触发词Trg-F1 / 普通论元Arg-F1

事件触发词Trg-F1：正确的触发词需满足触发词及其偏移量和事件类型都预测正确。

普通论元Arg-F1：正确的论元需满足该论元所属事件的事件类型和触发词都预测正确且该论元本身预测正确。如果某个论元同时属于多个事件，则称为交织论元，在计算普通论元Arg-F1时，交织论元也视为普通论元。

F1 = (Trg-F1 + Arg-F1) / 2

题目七 噪声环境下的命名实体识别

■ 问题定义

输入：一条样本（一句话，一个单词等）

输出：该样本包含的所有实体，标注内容包括对于样本中每个词对应的实体类型和起始位置/中间位置，NER格式为BIO。

示例：

	Brooklyn	and	Mary	live	in	New	York
Gold Labels	B-PER	O	B-PER	O	O	B-LOC	I-LOC
Noisy Labels	B-LOC	O	B-PER	O	O	O	B-LOC

训练过程中只能看到Noisy Labels，使用验证/测试集的Gold Lables进行评估。

评价指标：F1

题目八 基于大模型的智能表格问答

■ 问题定义

输入：一个表格数据（csv/excel），一个自然语言问题。

输出：根据表格数据，自然语言问题对应的答案。

示例：

问题：how many people stayed at least 3 years in office?

表格数据：[["", "name", **"took office"**, **"left office"**, "party", "notes/events"], ["11", "william mccreery", **"march 4, 1803"**, **"march 3, 1809"**, "democratic republican", ""], ["12", "alexander mckim", **"march 4, 1809"**, **"march 3, 1815"**, "democratic republican", ""], ["13", "william pinkney", "march 4, 1815", "april 18, 1816", "democratic republican", "resigned to accept position as minister plenipotentiary to russia"], ["14", "peter little", **"september 2, 1816"**, **"march 3, 1823"**, "democratic republican", ""], ["14", "peter little", **"march 4, 1823"**, **"march 3, 1825"**, "jacksonian dr", ""], ["14", "peter little", **"march 4, 1825"**, **"march 3, 1829"**, "adams", ""], ["15", "benjamin c. howard", **"march 4, 1829"**, **"march 3, 1833"**, "jacksonian", ""]]

答案：4

评价指标：ACC

题目九 多模态命名实体识别

■ 问题定义

输入：长度为n的句子$S = (w_1, w_2, ..., w_n)$，其中w_i为第i个token，与句子内容相对应的图像I。

输出：每个token的实体类别标签$Y = (y_1, y_2, ..., y_n)$

示例：

问题： Messi is playing football at FCB home stadium Camp Nou.

图像：

类别标签：

Person: [PER], Location: [LOC]

Organization: [ORG], Other: [MISC]

评价指标：

- Precision
- Accuracy
- F1 score

输出: [PER] O O O O [ORG] O O [LOC] [LOC]

(Messi is playing football at FCB home stadium Camp Nou.)

题目十 多模态思维链推理

■ 问题定义

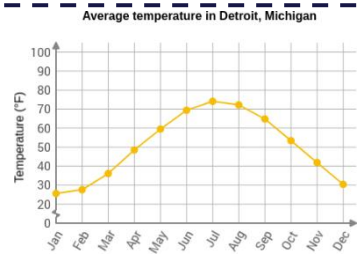
输入：科学领域的选择题，包括questions和choices

输出：该任务的求解划分为多个子任务（2~3个sub-task）,再在每个子任务下进行分步解答（生成sub-answer），将每个子任务答案汇总后形成问题答案，输出应包括三个键sub-tasks，sub-answers以及answer。

示例：

问题：Which month is the hottest on average in Detroit?

选项："July","May","June"



sub-tasks： ["Extract the temperature for each month from the chart.",
"Identify the month with the highest average temperature."],

sub-answers: ["Based on the provided chart, here are the approximate average temperatures for each month in Detroit……These values are approximations based on the visual data from the chart.",
"The month with the highest average temperature in Detroit, Michigan, according to the graph, is July. The temperature peaks at approximately 75°F during this month."
]

答案：July

评价指标：Accuracy

题目十一 基于人类偏好的响应优化

■ 问题定义

输入：

- 一个预训练语言模型。
- 数据集组成（使用 UltraFeedback 数据集中的这三个字段）：

  prompt: 用户输入的提示或问题。

  chosen: 人类标注者认为更好的、更偏好的模型响应。

  rejected: 人类标注者认为较差的、不偏好的模型响应。

示例：

"prompt": "请向我解释什么是人工智能，并举例说明其在日常生活中的应用。",
"chosen": "人工智能（AI）是指让计算机能够像人一样思考、学习和解决问题的技术。它模仿人类的智能，让机器能够执行通常需要人类智慧才能完成的任务。\n\n日常生活中的例子有很多：\n1. 智能手机助手：像 Siri 或 Google Assistant，它们能理解你的语音命令并执行任务。\n2. 推荐系统：购物网站（如淘宝、亚马逊）或视频平台（如 YouTube、抖音）会根据你的浏览和购买历史，向你推荐你可能感兴趣的商品或视频。\n3. 自动驾驶汽车：通过传感器和 AI 算法感知环境、做出决策并控制车辆。\n4. 人脸识别：用于手机解锁、支付验证或安防监控。\n5. 垃圾邮件过滤器：电子邮件服务会自动识别并过滤掉大部分垃圾邮件。",
"rejected": "人工智能就是机器人，它们能干活。比如扫地机器人就是人工智能。还有些能下棋。"

题目十一 基于人类偏好的响应优化

■ 问题定义

输出：一个通过利用偏好数据进行微调后得到的语言模型（可以生成更受偏好的回答）

评价指标：AlpacaEval 2 (胜率 - Length-Controlled Win Rate \ Raw Win Rate)

详情请见：

tatsu-lab/alpaca_eval: An automatic evaluator for instruction-following language models. Human-validated, high-quality, cheap, and fast.