

基于Zero-shot方式实现LLM文本分类

学习目标

- 掌握Zero-shot方式下prompt的设计方式
- 掌握利用LLM实现文本分类的代码

1 LLM文本分类任务介绍

- 下面几段文本来自某平台发布的金融领域文本：

1. "今日，央行发布公告宣布降低利率，以刺激经济增长。这一降息举措将影响贷款利率，并在未来几个季度内对金融市场产生影响。"，
2. "ABC公司今日发布公告称，已成功完成对XYZ公司股权的收购交易。本次交易是ABC公司在扩大业务范围、加强市场竞争力方面的重要举措。据悉，此次收购将进一步巩固ABC公司在行业中的地位，并为未来业务发展提供更广阔的发展空间。详情请见公司官方网站公告栏"，
3. "公司资产负债表显示，公司偿债能力强劲，现金流充足，为未来投资和扩张提供了坚实的财务基础。"，
4. "最新的分析报告指出，可再生能源行业预计将在未来几年经历持续增长，投资者应该关注这一领域的投资机会"，

- 我们的目的是期望模型能够帮助我们识别出这4段话中，每一句话描述的是一个什么类型的报告。
- 因此，我们期望模型输出的结果为：

- 1 ['新闻报道', '公司公告', '财务公告', '分析师报告']

2 Prompt设计

- 对于大模型来讲，prompt 的设计非常重要，一个 明确 的 prompt 能够帮助我们更好从大模型中获得我们想要的结果。
- 在该任务的 prompt 设计中，我们主要考虑 2 点：
 - 需要向模型解释什么叫作「文本分类任务」
 - 需要让模型按照我们指定的格式输出
- 为了让模型知道什么叫做「文本分类」，我们借用 Incontext Learning 的方式，先给模型展示几个正确的例子：

- 1 >>> User: "今日，股市经历了一轮震荡，受到宏观经济数据和全球贸易紧张局势的影响。投资者密切关注美联储可能的政策调整，以适应市场的不确定性。" 是['新闻报道', '公司公告', '财务公告', '分析师报告']里的什么类别？
- 2 >>> Bot: 新闻报道
- 3 >>> User: "本公司年度财务报告显示，去年公司实现了稳步增长的盈利，同时资产负债表呈现强劲的状况。经济环境的稳定和管理层的有效战略执行为公司的健康发展奠定了基础。"是['新闻报道', '公司公告', '财务公告', '分析师报告']里的什么类别？
- 4 >>> Bot: 财务报告

其中，`User` 代表我们输入给模型句子，`Bot` 代表模型的回复内容。

注意：上述例子中 Bot 的部分也是由人工输入的，其目的是希望看到在看到类似 User 中的句子时，模型应当做出类似 Bot 的回答。

3 分类任务代码实现

- 本章节使用的模型为ChatGLM-6B，参数参数较大（6B），下载到本地大概需要 12G+ 的磁盘空间，请确保磁盘有充足的空间。此外，加载模型大概需要 13G 左右的显存，如果您显存不够，可以进行模型量化加载以缩小模型成本。
- 本次分类任务实现的主要过程：
 - 构造prompt
 - 实现模型预测
- 代码存放位置：/Users/**/PycharmProjects/llm/zero-shot/finance_classify.py
- llm_classification.py脚本中包含两个函数：init_prompts()和inference()

3.1 导入必备的工具包

```
1  # -*-coding:utf-8-*-
2  """
3  利用 LLM 进行文本分类任务。
4  """
5  from rich import print
6  from rich.console import Console
7  from transformers import AutoTokenizer, AutoModel
8
9
10 # 提供所有类别以及每个类别下的样例
11 class_examples = {
12     '新闻报道': '今日，股市经历了一轮震荡，受到宏观经济数据和全球贸易紧张局势的影响。投资者密切关注美
13     联储可能的政策调整，以适应市场的不确定性。',
14     '财务报告': '本公司年度财务报告显示，去年公司实现了稳步增长的盈利，同时资产负债表呈现强劲的状况。
15     经济环境的稳定和管理层的有效战略执行为公司的健康发展奠定了基础。',
16     '公司公告': '本公司高兴地宣布成功完成最新一轮并购交易，收购了一家在人工智能领域领先的公司。这一战
17     略举措将有助于扩大我们的业务领域，提高市场竞争力',
18     '分析师报告': '最新的行业分析报告指出，科技公司的创新将成为未来增长的主要推动力。云计算、人工智能
19     和数字化转型被认为是引领行业发展的关键因素，投资者应关注这些趋势'}
```

3.2 构建init_prompts()函数

- 目的：进行prompt设计
- 具体代码实现：

```
1  def init_prompts():
2      """
3      初始化前置prompt，便于模型做 incontext learning。
4      """
5      class_list = list(class_examples.keys())
6      pre_history = [
```

```

7         (
8             f'现在你是一个文本分类器，你需要按照要求将我给你的句子分类到：{class_list}类别中。',
9             f'好的。'
10        )
11    ]
12
13    for _type, exmpale in class_examples.items():
14        pre_history.append((f'"{exmpale}"是 {class_list} 里的什么类别? ', _type))
15
16    return {'class_list': class_list, 'pre_history': pre_history}
17

```

3.3 构建inference()函数

- 目的：实现文本分类
- 具体代码实现

```

1
2  def inference(
3      sentences: list,
4      custom_settings: dict
5  ):
6      """
7      推理函数。
8
9      Args:
10         sentences (List[str]): 待推理的句子。
11         custom_settings (dict): 初始设定，包含人为给定的 few-shot example。
12         """
13     for sentence in sentences:
14         with console.status("[bold bright_green] Model Inference..."):
15             sentence_with_prompt = f'"{sentence}"是 {custom_settings['class_list']} 里
16             的什么类别? "
17             response, history = model.chat(tokenizer, sentence_with_prompt,
18             history=custom_settings['pre_history'])
19             print(f'>>> [bold bright_red]sentence: {sentence}')
20             print(f'>>> [bold bright_green]inference answer: {response}')

```

- 代码调用

```

1  if __name__ == '__main__':
2      console = Console()
3      #device = 'cuda:0'
4      device = 'cpu'
5      tokenizer = AutoTokenizer.from_pretrained("./ChatGLM-6B/THUDM/chatglm-6b-
6      int4",trust_remote_code=True)
7      # model = AutoModel.from_pretrained("./ChatGLM-6B/THUDM/chatglm-6b",
8      trust_remote_code=True).half().cuda()
9      model = AutoModel.from_pretrained("./ChatGLM-6B/THUDM/chatglm-6b-
10     int4",trust_remote_code=True).float()
11     model.to(device)

```

```

10     sentences = [
11         "今日，央行发布公告宣布降低利率，以刺激经济增长。这一降息举措将影响贷款利率，并在未来几个季度内对金融市场产生影响。",
12         "ABC公司今日发布公告称，已成功完成对XYZ公司股权的收购交易。本次交易是ABC公司在扩大业务范围、加强市场竞争力方面的重要举措。据悉，此次收购将进一步巩固ABC公司在行业中的地位，并为未来业务发展提供更广阔的发展空间。详情请见公司官方网站公告栏",
13         "公司资产负债表显示，公司偿债能力强劲，现金流充足，为未来投资和扩张提供了坚实的财务基础。",
14         "最新的分析报告指出，可再生能源行业预计将在未来几年经历持续增长，投资者应该关注这一领域的投资机会",
15     ]
16
17     custom_settings = init_prompts()
18     print(custom_settings)
19
20     inference(
21         sentences,
22         custom_settings
23     )

```

- 打印结果：

```

The dtype of attention mask (torch.int64) is not bool
>>> sentence: 今日，央行发布公告宣布降低利率，以刺激经济增长。这一降息举措将影响贷款利率，并在未来几个季度内对金融市场产生影响。
>>> inference answer: 新闻报道
>>> sentence:
ABC公司今日发布公告称，已成功完成对XYZ公司股权的收购交易。本次交易是ABC公司在扩大业务范围、加强市场竞争力方面的重要举措。据悉，此次收购将进一步巩固ABC公司在行业中的地位，并为未来业务发展提供更广阔的发展空间。详情请见公司官方网站公告栏
>>> inference answer: 公司公告
>>> sentence: 公司资产负债表显示，公司偿债能力强劲，现金流充足，为未来投资和扩张提供了坚实的财务基础。
>>> inference answer: 财务报告
>>> sentence: 最新的分析报告指出，可再生能源行业预计将在未来几年经历持续增长，投资者应该关注这一领域的投资机会
>>> inference answer: 分析师报告

```

小结总结

本章节主要介绍了基于Few-shot方式实现ChatGLM模型来完成文本分类的过程。