



黑马程序员线上品牌

LLM实现金融文本信息抽取

一样的教育，不一样的品质



目录

Contents

1. 信息抽取任务介绍
2. Prompt 设计
3. 代码实现
4. 结果展示

01

信息抽取任务介绍

信息抽取介绍

首先，我们定义信息抽取的Schema：

```
# 定义不同实体信息
```

```
schema = {'金融': ['日期', '股票名称', '开盘价', '收盘价', '成交量']}
```

下面几段文本来自某平台发布的股票信息：

1. '2023-02-15，寓意吉祥的节日，股票佰笃[BD]美股开盘价10美元，虽然经历了波动，但最终以13美元收盘，成交量微幅增加至460,000，投资者情绪较为平稳。'
2. '2023-04-05，市场迎来轻松氛围，股票盘古(0021)开盘价23元，尽管经历了波动，但最终以26美元收盘，成交量缩小至310,000，投资者保持观望态度。'

我们期望模型输出的结果为：

帮助我们抽取这2段话中的关键实体信息

02

Prompt 设计

■ 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。



设计要点

The diagram features a central blue circle with the text '设计要点' (Design Points). To its right, there are two light yellow rounded rectangular boxes, each preceded by a blue circle. The top box contains the text '向模型解释什么叫作「信息抽取任务」' (Explain to the model what is called 'Information Extraction Task'). The bottom box contains the text '让模型按照我们指定的格式（json）输出' (Let the model output in the format (json) we specify). The background has a light blue abstract wave pattern on the left side.

向模型解释什么叫作「信息抽取任务」

让模型按照我们指定的格式（json）输出

■ Prompt 设计

为了让模型知道什么叫做「信息抽取」，我们借用 Incontext Learning 的方式，先给模型展示几个正确的例子

- User: '2023-01-10, 股市震荡。股票古哥-D[E00E]美股今日开盘价100美元，一度飙升至105美元，随后回落至98美元，最终以102美元收盘，成交量达到520000。'。提取上述句子中“金融”（'日期'，'股票名称'，'开盘价'，'收盘价'，'成交量'）类型的实体，并按照JSON格式输出，上述句子中没有的信息用['原文中未提及']来表示，多个值之间用','分隔。
- Bot: {'日期': ['2023-01-10'], '股票名称': ['古哥-D[E00E]美股'], '开盘价': ['100美元'], '收盘价': ['102美元'], '成交量': ['520000']}

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。

03 代码实现

I 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。

本次抽取任务实现的主要过程：

构造prompt：对应init_prompts()函数

模型结果后处理：对应clean_response()函数

进行信息抽取：对应inference()函数

代码存放位置：/Users/**/PycharmProjects/llm/zero-shot/finance_ie.py

具体代码实现参考：pdf文档（附件资料）

04

结果展示

结果展示

```
The dtype of attention mask (torch.int64) is not bool
```

```
>>> sentence:
```

```
2023-02-15, 寓意吉祥的节日, 股票佰笃[BD]美股开盘价10美元, 虽然经历了波动, 但最终  
以13美元收盘, 成交量微幅增加至460000, 投资者情绪较为平稳。
```

```
>>> inference answer: {'日期': ['2023-02-15'], '股票名称': ['佰笃[BD]美股'], '开盘价': ['10美元'], '收盘价': ['13美元'], '成交量': ['460000']}
```

```
>>> sentence:
```

```
2023-04-05, 市场迎来轻松氛围, 股票盘古(0021)开盘价23元, 尽管经历了波动, 但最终  
以26元收盘, 成交量缩小至310000, 投资者保持观望态度。
```

```
>>> inference answer: {'日期': ['2023-04-05'], '股票名称': ['盘古(0021)'], '开盘价': ['23元'], '收盘价': ['26元'], '成交量': ['310000']}
```



黑马程序员线上品牌



扫码关注博学谷微信公众号

