lab5

task1

规则判断部分：

用 $text\_similarity.py$ 里面的词距函数做一个医院A/B相关属性 -> 标准属性的字典，然后替换下

（没做）

大模型部分：

Prompt:

```
def prompt(standard, columns):
    template = """You are a data mapping expert specialized in healthcare data standardization. You are working on aligning differe
The standard data model is {standard_data_model}.
The hospital's current data uses these column names: {hospital_data_columns}.

Please create a mapping dictionary where:
    - Keys are the hospital's current column names
    - Values are the matching standard column names
The size of standard data model may be not equal to the hospital's current data columns. You should try your best to map the hospit
The size of your output should be equal to the size of the hospital_data_columns.

Please format your output as following JSON format:
```json
{{
    '<hospital key 1>': '<standard key 1>',
    '<hospital key 2>': '<standard key 2>',
    ...
}}
```

Here is an example:
input:
    standard_data_model: ['医院名', '门诊ID', '病患姓名', '病患身份证号', '就诊日期', '就诊科室', '诊断结果', '处方信息']
    hospital_data_columns: ['医院名称', '就诊ID', '病人姓名', '病人身份证号', '就诊日期', '就诊科室', '诊断信息', '处方']

output:
```json
```

单次IO $task1\_data/Onetime$：不太准确

Eg.

task1_data > Onetime > 医院A-OutpatientData.csv

	A	B	C	D	E	F	G	H	I	J	K
1	就诊ID	患者姓名	患者身份证号	就诊日期	就诊时间	科室	诊断结果	医生ID	医生姓名	处方信息	医疗费用

标准为：门诊ID，病患姓名，病患身份证号，就诊科室，诊疗费用

task1_data > Onetime > 医院A-InpatientData.csv

A	B	C	D	E	F	G	H	I	J
住院ID	姓名	身份证号	入院日期	出院日期	诊断结果	科室	主治医生...	主治医生姓名	医疗费用

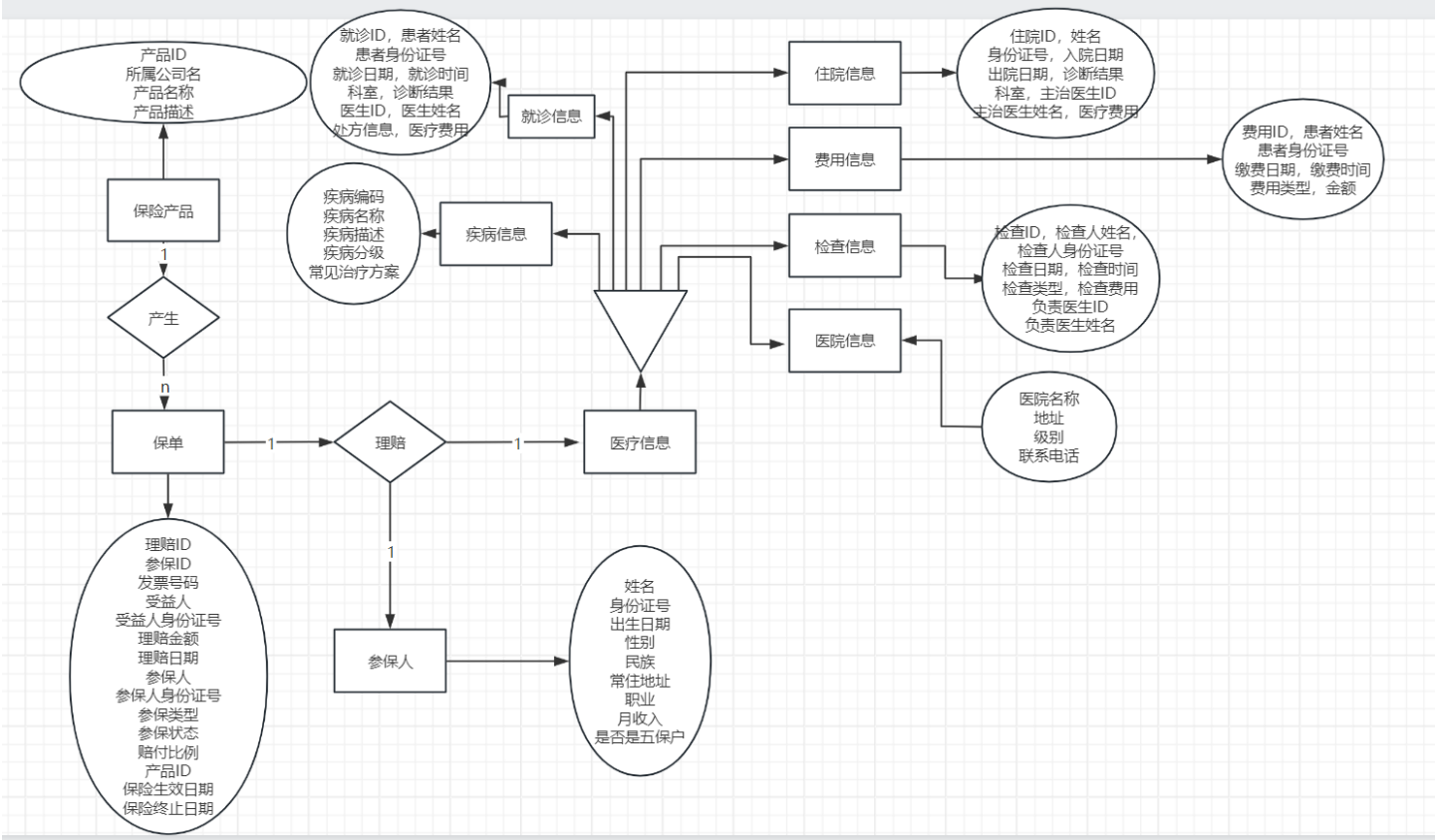
标准为：病患姓名，病患身份证号，负责科室
```

于是多次采样，取最优一次 $task1\_data/Multi$: 完全匹配标准属性

```
def prompt_decide(standard, columns, results):
    template = """You are a data mapping expert specialized in healthcare dat
Some experts have did this work and created some mapping dictionaries. Your t

The standard data model is: {standard_data_model}.
The hospital's current data uses these column names: {hospital_data_columns}.
The mapping dictionaries is: {results}

Please evaluate these dictionaries and select the best one.
Format your returns as following:
```json
{{
    'reason': '<Express your reason about why you choose this dictionary>'
    'result': <Fill in the index of the dictionary you choose from the given
}}
```
```

```
def main(dir='医院A', sample_time=2):
    ...........
    ...........#multi-sampling
    ...........samples = []
    ...........for _ in range(sample_time):
    ...............prompt_text = prompt(standard[dataType], columns)
    ...............res = json.loads(llm_gen(prompt_text))
    ...............samples.append(res)
    ...............#print(samples)
    ...........#select best
    ...........prompt_desion = prompt_decide(
    ...............standard=standard[dataType],
    ...............columns=columns,
    ...............results=samples
    ...........)
    ...........res_best = json.loads(llm_gen(prompt_desion))
    ...........print(res_best)
    ...........res_best = samples[int(res_best['result'])]
```

task2



task3

数据标签化： 对所有属性进行标签聚类

```
不设定聚类标签 task3_res.json:效果不太行
设定聚类标签 task3_given.json: 效果也一般
人工修正: task3_given_check.json

上面分的很粗，但是分的细像task2那样整这么多标签那这个题要干什么？？？
```

以上全是属性标签，统计标签和算法标签如<诊疗信息>中的

- 医疗总金额=住院+检查+就诊
- 财产收入水平

还需要进一步计算（没做）

task4

没做

task5

✓ 数据资源目录示例

- **描述：**全面系统地描述并编录组织内所有数据资源。
- **目的：**加强数据资源管理、识别、定位及共享。
- **例子：**客户信息数据库、销售记录数据库、库存管理系统等。
- **关注点：**数据资源的全面性、准确性、可访问性。

数据资源名称	数据描述	存储位置	访问方式	负责部门
客户信息数据库	包含客户ID、姓名、联系方式等基本信息	数据中心A	内网访问	IT部
销售记录数据库	记录每次交易的时间、金额、商品信息等	数据中心B	API访问	销售部
库存管理系统	实时跟踪库存数量、位置等	云端存储	移动端APP	物流部

✓ 数据资产目录示例

- **描述：**识别有经济和社会价值的数据，并对其进行治理。
- **目的：**创建有价值的数据资产目录。
- **例子：**客户购买偏好、销售预测模型、库存周转率等。
- **关注点：**数据价值评估、标准化、质量保证、安全管理。

数据资产名称	数据描述	数据治理状态	业务价值	使用部门
客户购买偏好	分析客户购物行为得出的偏好数据	已标准化	市场营销	市场部
销售预测模型	利用历史销售数据训练得到的预测模型	已测试验证	预算规划	财务部
库存周转率	统计分析得出的库存周转效率数据	已审核	运营优化	物流部

幽默完了，结合实际编一下得了