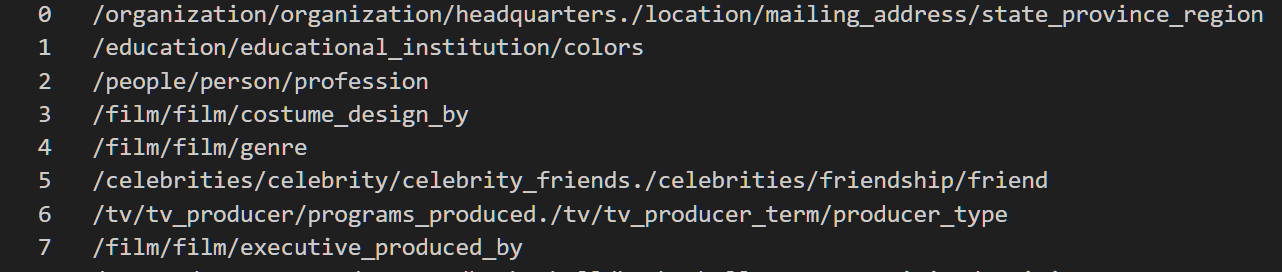
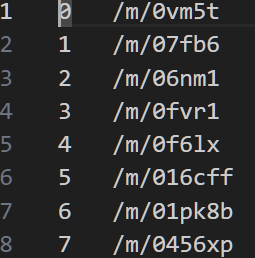
这次的任务很简单，就是先测一下我们不做任何优化的纯大模型对三元组校验的一个效果，作为baseline。

1. 数据集处理：我们先暂时采用FB15K-237，这是来自Freebase知识图谱的一个子集，经常被用作各类知识图谱任务的数据集。其中包含大约15k个实体和237种关系，我们只用其中的train集即可，其中train集也有27w三元组。其中的数据基本都是人工维护的，因此为了对其中三元组进行测试，我们需要将其中的一些三元组替换成错误的，在这里随机选择20%的三元组将其**尾实体**进行替换，生成错误的三元组。这里要注意两点：**1.**替换后的尾实体必须和替换前的是同一个类型的实体，比如 替换前是：特朗普-国家-美国，替换后的错误实体应该是特朗普-国家-日本这样，但不能是特朗普-国家-拜登这样。**2.**保证替换后的一定是错误的，因为有些三元组可能是一对多或者多对多的的，你随机替换之后不一定是错的，比如原始三元组是感冒-症状-流鼻涕，你替换成感冒-症状-咳嗽，这依然是对的三元组；一般情况如果新构建的三元组在原本的知识图谱中如果不存在基本都是错的。

PS：FB15K-237中的实体（entities.dict）和关系(relations.dict)都是编码如下，可以用维基百科官方的api查询含义，可以先研究研究，后续我也看一下，如果不合适的话就换个数据集。



1. 用LLM去校验知识图谱

如果全部校验的话时间成本和费用成本肯定比较高，从上面处理好的数据集随机选择500个三元组进行检索就好，因为大模型我们没有微调，直接调用api即可。（可以用这个网战：https://www.closeai-asia.com/，里面gpt、deepseek等都可以调用，调用api的花费保留好发票，回头找我报销），选择deepseek、gpt3.5、gpt4.0三个模型分别测试一下（设计prompt的时候让大模型仅对每个三元组判断正确、错误或不确定即可），分别记录这三个模型对这500个三元组的检测效果包括TN、FN、TP、FP、准确率、不确定率、错误率。如下表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| dataset | TN | FN | TP | FP | Correct (%) | Not available (%) | Error (%) |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

PS:500个三元组不要一个问题全扔给LLM，一个问题包含10个/20个/30个最好，想快一点的话可以写个多线程，毕竟咱代码的本质只是在调api，咱接受是很快的，主要时间浪费还是在api响应

多线程可以用ThreadPoolExecutor参考这个

with ThreadPoolExecutor(max\_workers=10) as executor:

futures = []

for i in range(0, len(head)):

futures.append(executor.submit(process\_batch, i))

concurrent.futures.wait(futures)