**《机器学习》课程设计研究报告**

项目组成员：2120210454 李伟 (ID：WLee)，2120210464 于胜龙 (ID：以往)

联系人邮箱：[weili@mail.nankai.edu.cn](mailto:weili@mail.nankai.edu.cn) Kaggle队伍名：Lee & Yu

# 问题提出

该任务的目标是为测试样例提供对应的匹配地点，而本文拟使用分类思想解决匹配问题，整体思路为针对测试样例生成候选匹配地点集合，并通过分类器进一步将其中的正确地点筛选出来。针对该整体流程，我们存在以下问题需要进行解决：

1. **数据预处理方式的选择**

流程中拟使用多层感知机通过二分类任务对两样例是否匹配进行学习，原始pairs.csv文件中虽提供了大量训练样例，但样例中正负样本比例相近，经过实验发现，当正负样本比例差距不大时，二分类模型的表现较差。因此，需要选择合适的负样本生成方式，生成大量质量较高的负样本填充训练集，提升训练集中负样本所占的比例。

此外，对于某些表达不统一的特征，如邮政编码特征即包含数值特征又包含文本特征，进行统一形式表示后剔除不合理的值。对于数据集中的空值，选择合适的方式进行填充。

1. **特征转换方式的选择**

在数据集中，文本数据是非常重要的一类特征，但该类数据包含多语信息，因此需要使用适当方法将多语文本特征转换为能够作为二分类模型输入的数值特征。

此外，针对数值型数据，部分数据存在强相关关系，例如两个地点经度和维度，可以结合进行更好的特征表示，针对这方面特征，同样需要进行特征的转换。

1. **分类模型的选择**

分类模型用于判断两样例是否匹配，在本次实验中，需要对分类模型的规模和具体实现方式进行选择。实验中发现，规模较小的分类模型无法很好地拟合数据的整体分布，而规模较大的分类模型则会引起比较严重的过拟合问题

1. **测试方式的选择**

在测试阶段，首先需要选择合适的方式生成包含正确样例的候选集，再使用二分类模型区分正确样例。此时，如果生成的候选集规模过小，则会导致包含正确样例的可能性下降，而候选集规模过大，则会导致模型运行时间极长。

# 二、现状分析

1. **数据预处理方式的选择**

针对数据预处理的方式，对相关工作进行了调研，大部分基于分类的工作均有负采样这一步骤，其主要的负采样步骤是使用已有的且明确的数值型信息，选择混淆可能性比较大的点对作为负样本。例如，使用KD树基于经度与维度产生近似样本，将其中的错误样例作为负样例加入训练集。

数据填充、补全的方式，针对模型的不同有所差异，由于本文拟使用多层感知机作为分类器，其可以自动进行特征抽取，降低低质量特征的权重，因此该步骤可以使用较为简单的实现方式，具体来说，对于数据中的空值，填充一固定值即可。

1. **特征转换方式的选择**

针对特征转换方式的选择，在调研中发现存在多种实现方式。其中，部分工作直接使用编辑距离计算两字符串间的相似度，以此将文本特征转换为数值特征，然而此种方式局限性较大，不同语言间含义相同的字符串难以衡量其相似度。存在部分工作使用tf-idf分数计算两字符串相似度，该计算方式更加科学，但仍难以处理多语言相同含义字符串。因此，本文使用多语言预训练模型将文本特征转化为数值特征，多语言预训练模型在大规模语料库上进行预训练，因此能够较好地将不同语言的相同含义字符串转变为相似的向量。

针对强相关性数值型数据的处理方式，调研后发现部分工作对经度、维度进行操作，使用其计算欧氏距离或半正矢距离，或将原有的数值型数据编码为新的数值向量。

1. **分类模型的选择**

经过调研后发现，由于决策树在该类任务中的优越表现，部分工作使用LightGBM模型实现分类模型，而由于深度学习的便捷性，部分工作使用多层感知机实现二分类模型，本文选择使用多层感知机实现二分类模型，以便更好地与多语言预训练模型进行结合。

1. **测试方式的选择**

经过调研后发现，测试阶段同样使用经度、维度信息实现候选集的抽取，其中候选集的规模大小可以通过测试调节，选择较为适合的参数。

# 三、方法原理

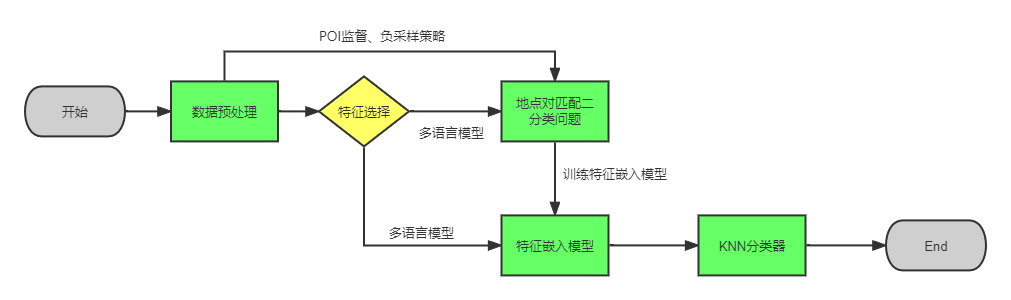
 如下图所示，本文将地点位置条目匹配问题[1]看待为文本匹配问题，即给定两个地点条目，抽取出二者的地点文本描述特征（例如‘name’，‘address’，‘city’，‘state’，‘country’，‘categories’），并将文本特征拼接为地点描述文本“语句”，通过匹配两个地点的文本比对程度来判断两个地点是否是能够匹配的。从而，我们能够将比赛任务抽象为文本匹配二分类问题，通过训练一个性能较好的文本匹配二分类模型，我们能够借此寻找到表示同一个地点的多个位置描述条目。为训练文本匹配二分类问题，结合数据集的多语言特点，多语言文本特征表证问题也成为我们考虑的一个重要问题。此外，由于地点描述条目对是我们模型的输入数据，而在给定的pairs.csv文件中负样本的比例较低，这样在训练时容易造成模型无法有效收敛，因此负采样策略和数据增强策略对于模型训练较为重要，除此之外，在测试时为了得到候选的地点描述pair，我们需要借助数据增强的思想，从test.csv文件中初步筛选出候选pairs。

图1 方法实现流程图

1. **短文本匹配（二分类任务）**

文本匹配是自然语言处理中的常见问题，主要可以分为有监督的方法和无监督的方法。有监督的方法要求事先准备好带标签的语料，可以表示为“句子1，句子2，匹配标签”，然后让模型去学习。无监督的方法主要采用句子的统计特征进行短文本匹配任务，例如TF-IDF、BM25等方法；此外，还可以基于文本的词向量特征学习文本的特征，根据文本特征的相似度判断短文本是否匹配。

在本文中，由于可以从比赛数据中抽取出形如上述三元组的带标签的语料数据，因此本文采用有监督的方法，通过对位置描述条目的文本进行特征表示，然后联合两个位置条目的文本“语句”特征表示，基于其进行短文本匹配的二分类判断，在文本匹配二分类任务上，本文采用了多层感知机对输入特征进行分类。

1. **多语言文本特征表示**

如上述所示，本文需要对地点位置描述的文本进行特征表示，从数据集的具体情况来看，可以发现其中包含多种语言，因此需要对多语言文本进行特征表示，由于现有的多语言文本处理模型已经较为成熟，具有大量的与训练模型能够用于进行多语言文本特征表示，通过调研发现hugging face开源了多个多语言预训练文本处理模型，本文选择’xlm-roberta-base’[2]作为地点位置描述“语句”的特征抽取模块，通过该模块为特定地点的位置信息描述文本生成一个特征表示向量。

1. **负采样与数据增强**

比赛任务是寻找尽可能多的匹配地点位置描述条目，因此在训练数据上我们需要补充尽可能多的地点描述条目对，并基于此进行模型训练，同时在pairs.csv文件中提供的地点描述条目对数量较少，且负样本比例偏低，可以基于数据增强的思想从train.csv中生成更多的地点描述条目对，并可以通过生成更多负样本增大训练集中的负样本比例。此外，在测试集上进行结果推理的时候，由于在所有的测试集地点描述条目上进行两两匹配预测的复杂度太大，因此可以通过预先筛选出部分可能匹配的候选地点描述条目对，从而降低推理的计算复杂度。在本文中，我们基于KNN的方法，利用地点描述条目中经纬度信息寻找每一个地点描述的若干最近邻地点条目，并将其组成地点描述条目候选pair，之后利用分类模型进行更精确的进一步匹配判断。

# 四、算法描述

1. **短文本匹配（二分类任务）**

在短文本匹配方面，本文使用生成的经过数据增强步骤的训练集对多层感知机进行训练，最终将训练好的模型作为输出用于任务。其中，关于多层感知机模型，我们使用了一个深度为6层的多层感知机模型，数字特征在第一层输入，文本特征在第三层输入，最终输入一个数字作为二分类预测结果。具体算法步骤如下图，其中损失函数为二分类交叉熵损失函数（BCELoss()），优化器为Adam优化器。

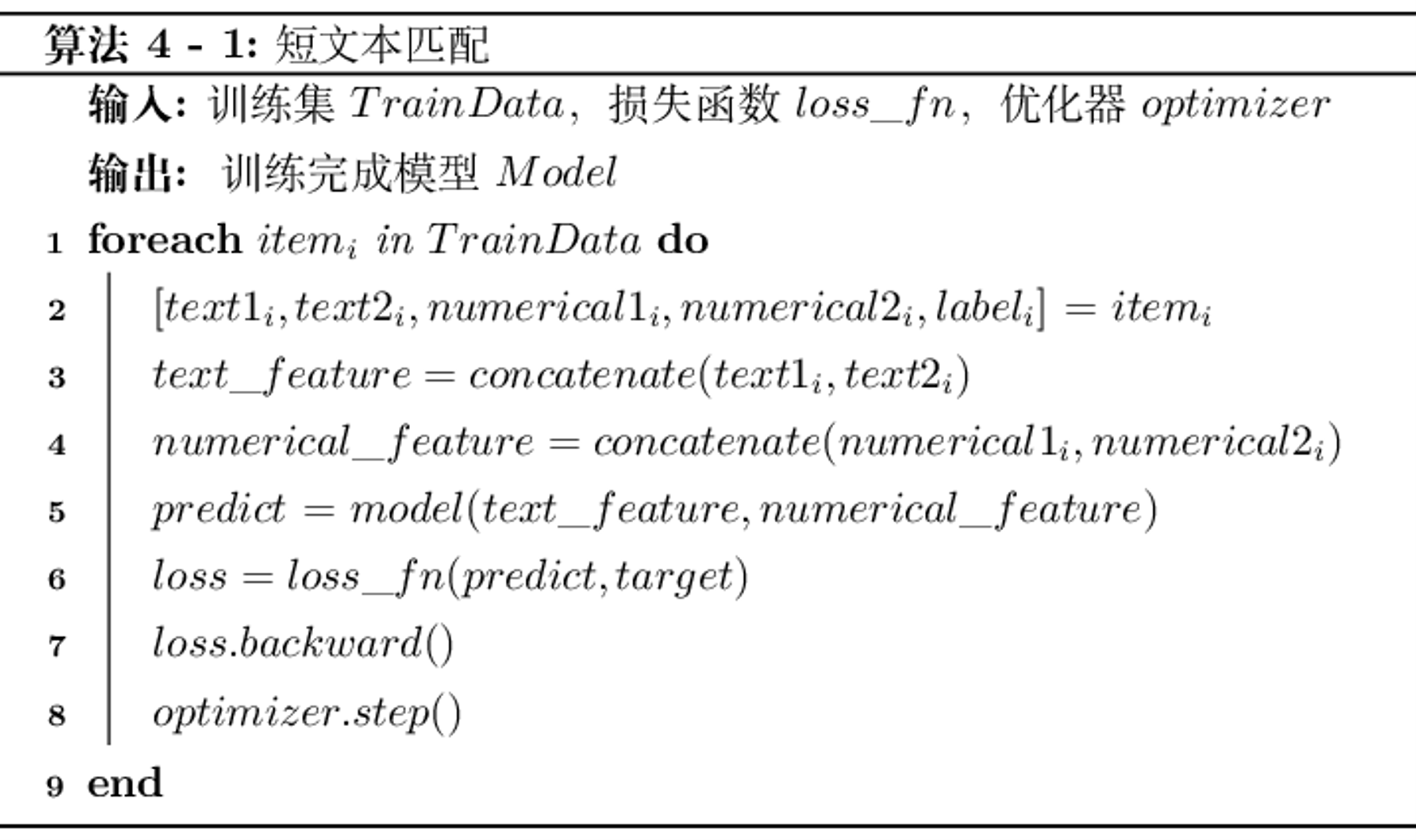


表1 短文本匹配算法

1. **多语言文本特征表示**

使用多语言预训练模型实现多语言文本特征的表示，本文选择XLM-RoBERTa模型实现将文本特征转换为数量特征，XLM-RoBERTa模型在2.5TB的多种语言语料库上进行了预训练，能够较好地将不同语言中含义相近的文本转换为相似的句向量。本文对两种预训练模型使用方式均进行了检验，如下：

* 将两样例的文本特征拼接为一句话作为预训练模型输入，得到数值特征
* 将两样例的文本分别作为预训练模型输入，拼接后得到数值特征

经过测试后发现，第二种方法能够取得更好的效果

1. **负采样与数据增强**

使用kd\_tree实现负采样方式，对于原始pairs\_data中的每对数据，利用其经度与维度特征，额外生成个数据(大部分为负样本)，具体实现方式如下图所示，在生成train\_data后，可采取其他随机采样方式构造特定正负样例比例的数据集实现数据增强。



表2 负采样算法

# 五、结果与分析

本文最终提交的模型训练数据集为比赛提供的pairs.csv数据以及从train.csv中经过KNN采样的数据增强方式，增加大量的训练数据pair（尤其是负样本），最终得到1597274条pair数据，其中38%的样本为正样本，62%的样本为负样本，同时为了对模型进行选择，本文选取了1%的数据（15972条样本）作为验证集，对训练阶段的模型进行性能评估，我们考虑了二分类的AUC、F1、Recall、Precision指标。本文提交模型在验证集上的评测指标如下所示：

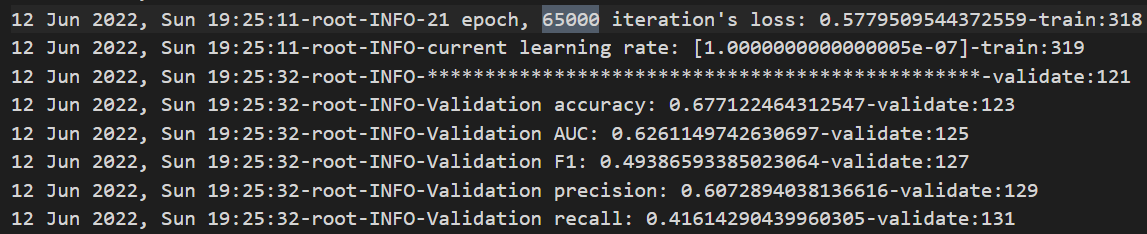


图 2 提交模型评价指标日志记录

本文所提交模型在Kaggle比赛平台的结果评估得分为0.656（由于提交显示，此为初步结果，后续调参应该会有更大的提高），平台评估结果截图如下所示：

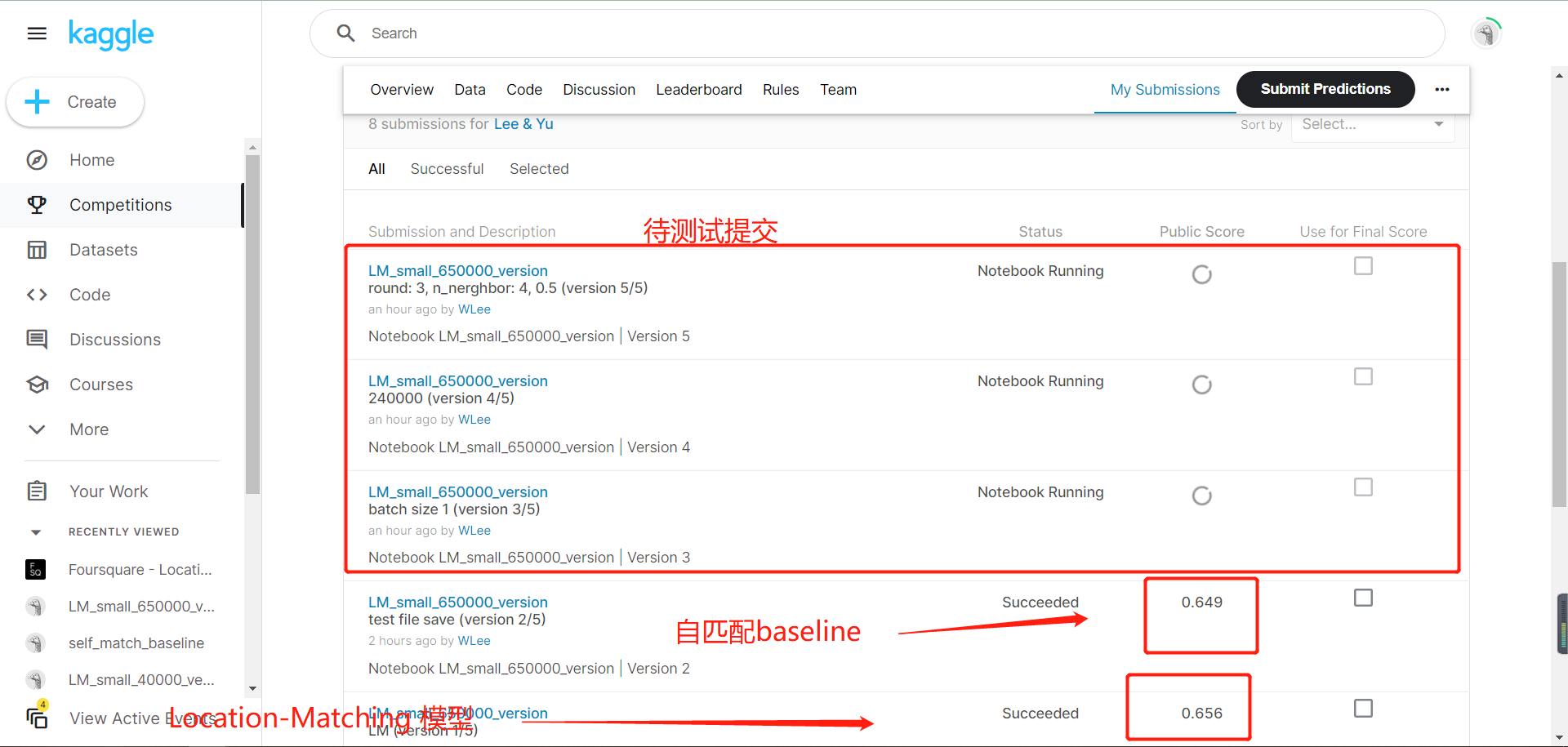


图 3 kaggle网站提交评估结果

首先，直观的从地点匹配的效果来看，基于公开的test.csv的五条地点描述信息，我们的模型能够做到完全的正确识别匹配地点，演示结果如下（KNN负采样邻居节点数设置为3，可能邻居节点采样轮次设置为2，匹配概率的阈值设置为0.5）：

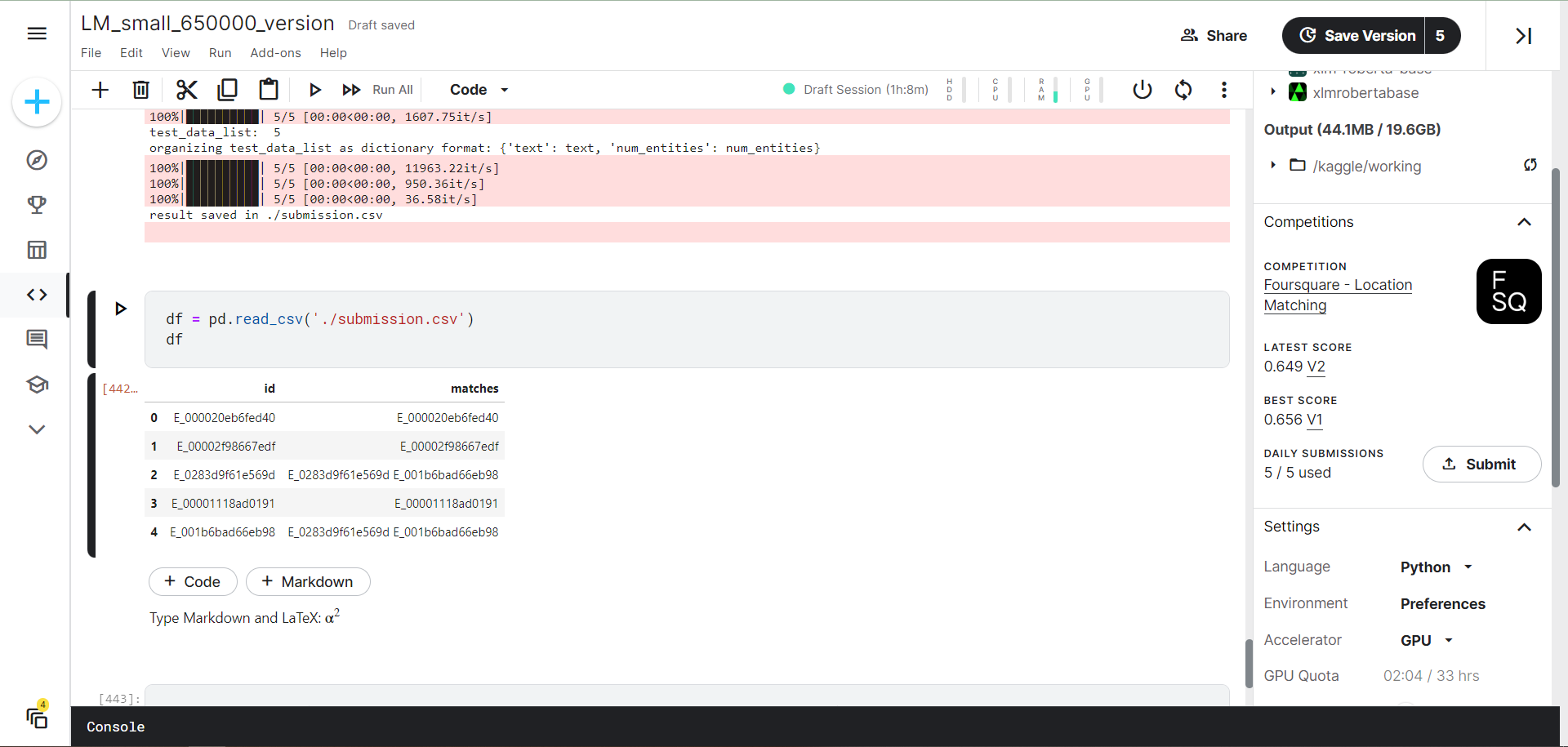


图 4 模型在test.csv上的性能表现

其次，相比于直接的baseline（即只将每一个节点与自身的匹配输出到submission.csv中）的IOU评估结果0.649，本文提交的Location Matching(LM) 模型在kaggle的评估结果为0.656，有了明显的提升，说明我们的模型至少能够比baseline更强，也能够识别出除了自身之外的地点描述对的匹配情况，即从性能的角度证明了本文实现并提交的模型的有效性。此外，考虑到模型部分参数可以进一步进行优化，本文提交模型在性能上应该还有较大的提高空间，后续将继续优化并进行提交。

# 参考资料

[1] Foursquare-Location Matching. <https://www.kaggle.com/competitions/foursquare-location-matching>.

[2] xlm-roberta-base model in Hugging Face Website. <https://huggingface.co/xlm-roberta-base>