**《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)

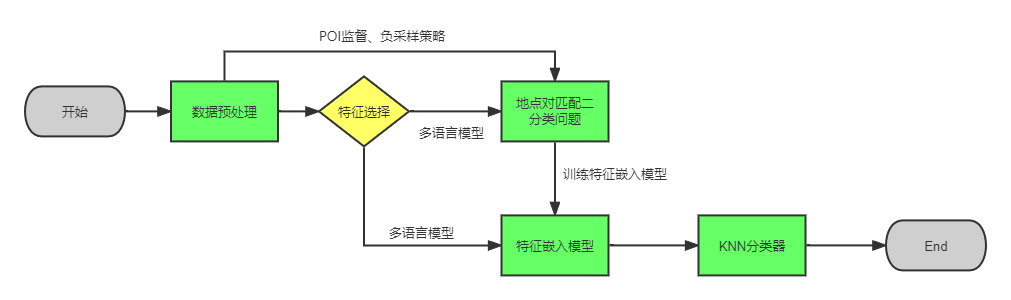
联系人邮箱：[weili@mail.nankai.edu.cn](mailto:weili@mail.nankai.edu.cn) Kaggle队伍名：Lee & Yu

# 一、问题提出

# 二、现状分析

# 三、方法原理

如下图所示，本文将地点位置条目匹配问题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。



图表 1方法实现流程图

## 3.1. 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。

## 3.2. 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’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。

## 3.3. 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进一步的匹配判断。

# 四、算法描述

# 五、结果与分析

## 5.1. 线上评估结果展示

## 5.2. baseline性能比较与分析

## 5.3. 模型变体性能比较与分析

# 参考资料