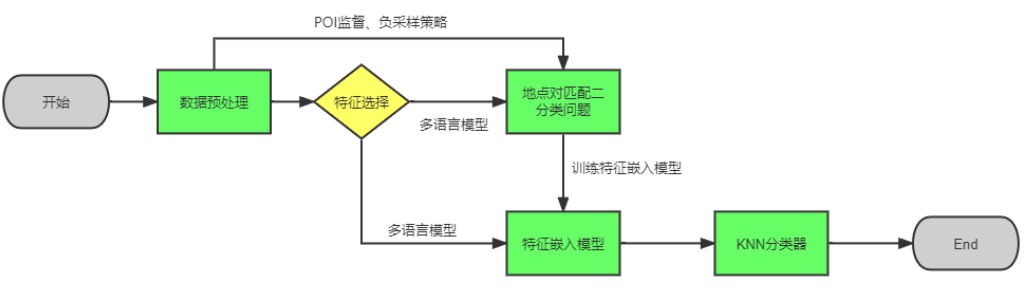
**《机器学习》课程设计开题报告**

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、总体思路**



**二、特征描述**

1、数据集描述：数据集提供了包含全球数十万个POI的超过一百万个地点描述条目，具体地，数据集包含以下文件，各个文件的数据量、特征维度数以及数据描述如下所示：

train.csv(1138812, 13): 训练数据集，包含超过一百万个地点描述条目，以及对应的POI  
test.csv (5, 12): 一组类似于训练集属性的待预测地点属性，但只有少量数据  
pairs.csv(5, 2): 一组预先生成的来自train.csv的地点对，用于提高对于匹配的检测能力，可以额外生成新的匹配对  
sample\_submission.csv(578907, 25): 数据提交样例

2、训练数据特征描述：对于数据集我们主要关注训练数据文件，我们能够看到给定的训练数据地点信息共有1138812条，每条地点共有12个数据属性以及其对应的POI的编号。

id: 地点描述id，由一个由E\_开头的序列编号标识  
name: 地点的名称，文本类型，涉及多个语言，且可能存在错误  
latitude: 纬度信息，浮点数标识  
longitude: 经度信息，浮点数类型标识  
address: 地点所在地址描述信息，文本类型  
city: 城市名称，文本类型  
state: 地点所在州名称，文本类型  
zip: 地点邮编，文本类型（邮编的结构多样，非数值类型）  
country: 地点所在国家， 使用缩写，文本类型  
url: 地点相关网页链接，超链接  
phone: 地点相关的电话、手机号码，文本类型  
categories: 地点的类别划分，文本类型  
point\_of\_interest: 该地点所对应的POI编号，有一个P\_开头的序列编号表示

更进一步地，我们检查一下训练数据各个特征维度的数据缺失率如下所示：

url: 76.49% phone: 69.89% zip: 52.28% state: 36.93% address: 34.83% city: 26.27% categories: 8.63% country: 0.00% name: 0.00% id: 0.00% latitude: 0.00% longitude: 0.00% point\_of\_interest: 0.00%

可以看到在各个特征维度的数据上，url的缺失率是最高的，而latitude, longitude, name, country, categories的缺失率相对较低，因而我们可以发现在各个特征维度上这些缺失率较低的特征保留的相对更为完整的特征信息，在后续的模型设计上可以着重考虑。

**三、模型选择**

1、数据预处理：

进行数据清洗：在实验数据中存在数据错误、偏移现象，因此对于无固定数值型数据，通过拟合正态分布函数去除明显偏移的值，对于有固定范围的数值型数据，如经度、维度，通过范围去除错误值。此外，实验数据中存在大量的空值，需要进行数据的填充。对于数值型数据，填充固定值，对于文本型数据，填充固定字符串。

拟使用二分类方法对特征进行表示学习，pairs.csv提供了二分类相关正负样例，但仍需要负样例生成质量更高的向量表示。因此，拟利用数据型地点信息(经度、维度)或文本型信息的相似度，生成大量容易混淆的负样例加入pairs数据中用于训练。

2、特征转换：

在数据中，存在多语言的文本特征，因此拟采用多语言预训练模型将所有文本转换为统一向量空间中的向量。在转换时，直接使用预训练模型，将每句文本特征转换为数据特征。

此外，数据中的某些特征存在较强的相互关系，如经度、维度须结合使用，地点名称、所属城市、邮编等存在强相互关系，拟对这类关系进行建模，将关系紧密的特征联合作为一个特征输入多层感知机。

3、模型训练：

拟使用多层感知机，通过二分类实现学习。在使用MLP实现二分类的同时，实际上对地点进行了表示学习。将多层感知机末层输出作为地点的向量表示，学习的过程使相似地点的向量表示靠近，使不同地点的向量表示尽可能远离。

4、匹配：

在测试阶段，使用KNN分类器对测试样例进行匹配，同属于一个poi的地点为相同地点，使用KNN分类器，可通过测试样例周围的poi分布，得到测试样例的具体匹配地点。

**四、实验设计**

1、数据预处理：

* 数据清洗，去除偏移值，填充空值
* 进行负采样，增加pairs文件外负样例的数量

2、特征转换：

* 使用多语言预训练模型将多语言文本转换为向量
* 尝试对相关特征进行联合表征，如经度、维度可连接作为MLP的输入

3、模型训练

* 基于pairs数据，通过二分类任务，训练MLP
* 抽取MLP末层输出作为地点的向量表征

4、匹配

* 使用KNN分类器对测试样例进行匹配