## 地址文本规范化项目技术方案

## 建设目标

将项目方提供的地址文本信息，根据项目方指定的地址数据标准，对地址文本进行不同层级的内容回填，同时需要对地址文本进行清洗，并将相同地址进行合并。项目完成时提交中文地址回填程序以及《项目使用说明》和《项目完成报告》。

## 二、组织架构

1． 项目组分工

成立以具有丰富自然语言处理、大数据分析与处理经验的南昌大学教授、博士生导师为负责人的课题研发团队，团队成员为具有硕、博士学位、具有人工智能、大数据研发经验的南昌大学信息工程学院青年骨干教师，承担过国家、省部级科研项目、具有丰富开发经验。

具体分工如下：

2. 协同工作

参与工作人员分成2组，进行交叉检查，杜绝人为错误；参与工作人员每天都会形成工作台账，组长每日进行审核，定期形成阶段成果。

3. PDCA项目管理

项目全程采用PDCA循环管理，即Plan(计划)、Do(执行)、Check(检查) 和 Action(处理)，全程均可追溯。

## 二、前期准备工作

1．研究甲方提供的已初步处理的十一层级中文地址文本数据，分析中文地址文本数据的特性；

2．研究甲方提供的已正确处理的中文文本地址数据，明确各层级分级要求，以及各层级回填标准。

## 三、需求分析

项目需要将甲方提供的中文地址文本数据按照甲方的层级标准进行回填，同时包括对原有数据中的不规范内容和同地址进行合并，具体需求如下：

1. **对地址数据文本进行数据的清洗**

提供对不规范地址数据的清洗解决方案，主要对中文地址文本中可能存在的重复词、繁体词、以及形似或者音似的错别字进行去除或者标注，对重复词进行去除，对繁体字进行简体化，对可能存在的形似或者音似类型的错别字进行汇总标注，生成文档报告。

1. **根据中文地址文本进行层级回填**

提供根据中文地址文本进行层级回填的解决方案，基于自然语言处理技术，根据原始中文地址文本中所蕴含的各层级信息，按照甲方指定的层级分类标准进行层级信息回填。并将回填好的数据进行存储，回填处理的目标是对处理好的数据随机抽取总量的5%的数据，经过人工鉴定准确率达到90%。

1. **对中文地址文本数据进行地址合并**

提供对中文地址文本数据中可能存在的相同地址数据的合并方案，基于地理位置信息技术和自然语言处理技术，将数据中可能存在的相同地址进行合并或者对可能是相同地址的数据标注。

1. **Django开发数据交付网页**

提供数据交付的网页，接收包含十一个层级的中文地址文本数据，在服务端经过清洗、回填、合并后把十三个层级的中文地址文本数据返回。

## 四、技术路线

**1．数据预处理**

对中文地址数据进行基本的数据清洗操作，将中文地址数据中可能存在的繁体字替换位简体字，并使用自然语言处理技术对中文地址中存在的连续的冗余重复词进行去除，清除中文地址数据中可能存在的输入错误。处理的文本内容不包括中文地址中的实体名（商铺名称或者公司名称）。处理样例如表1所示。

表1 繁体字简化及重复词去重

|  |  |  |
| --- | --- | --- |
| 地址数据 | 问题类型 | 处理后地址数据 |
| 北京市朝陽区 | 简繁体混淆 | 北京市朝阳区 |
| 无錫市錫山区 | 简繁体混淆 | 无锡市锡山区 |
| 杭州拱墅区杭州拱墅区余杭塘路417号中天·西城纪中天·西城纪4号楼 | 存在重复词 | 杭州拱墅区余杭塘路417号中天·西城纪4号楼 |

**2. 基于自然语言处理技术采用文本纠错方式，对地址数据中可能存在的形似或音似错别字进行检测。**

针对中文地址数据进行录入时可能存在的形近或者音近类的错别字，使用自然语言处理技术，对文本中的错别字进行检索并生成纠正字，主要处理省到社区行政村中的层级数据，将检索到的错别字根据行政区规划信息进行规则匹配等方法，进行纠正，生成纠正记录。如表2所示：

表2 纠正记录

|  |  |  |
| --- | --- | --- |
| 地址数据 | 错别字索引 | 纠正后数据 |
| 杭州下城区延按路6号嘉里中心 | ('按', '安', 6, 7) | 杭州下城区延安路6号嘉里中心 |

**3.针对已初步进行层级分类的数据，使用自然语言处理技术中的实体识别技术，对已初步进行层级分类的数据进一步分类纠正。**

根据以提供的《资管抽检核查样本》中的错误描述进行分析，可见大部分的回填内容错误在于某些层级信息的丢失和回填内容的层级错误。

为优化这一问题，可采用自然语言技术中的实体识别技术，对回填方法进行优化，基本步骤如下：

第一步：对社区行政村及之后的层级信息进行总体拼接，形成待处理的文本数据。

第二步：对各层级信息使用四位序列标注法，进行文本数据中各层级信息内容的标注识别，标注样例表如下：

表3 地址元素层级标注体系

|  |  |  |
| --- | --- | --- |
| 标注 | 意义 | 实例 |
| Community | 村/社区 | 华/B\_comm  乐/M\_comm  社/M\_comm  区/E\_comm |
| Street | 街巷 | 甲/B\_stre  来/M\_stre  路/E\_stre |
| RoadNum | 门牌 | 1/B\_ronu  0/M\_ronu  6/M\_ronu  号/E\_ronu |
| Poi | 标志物 | 都/B\_poi  市/M\_poi  阳/M\_poi  光/E\_poi |
| CountyRegion | 片区/村组 | 乐/B\_core  苑/E\_core |
| BuildNum | 幢号 | 5/B\_bunu  幢/E\_bunu |
| UnitNum | 单元 | 2/B\_unit  单/M\_unit  元/E\_unit |
| FloorNum | 楼层 | 1/B\_floor  5/M\_floor  层/E\_floor |
| RoomNum | 户号 | 1/B\_room  5/M\_room  0/M\_room  4/M\_room  室/E\_room |

第三步：将标注好的数据放入深度学习模型中进行训练，得到训练好的可用于进行各层级内容识别的模型。

第四步：使用训练好的数据对未投入训练的数据进行实体预测，并按照预测结果进行各层级信息的回填。

**4．基于自然语言处理技术中的文本相似度和地理信息对相同地址数据进行合并去重。**

对于数据中存在的相同描述地址，需要对地址进行去重处理。地址经过层级回填后，形成了结构化的数据，根据地址数据的特性，将地址进行按不同层级信息进行最大可能的匹配。

然后使用地址数据的经纬度信息、地址文本相似度、底层级数据对比、实体名相似度（如果存在）的方法来进行同类地址的判断。例如：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| grade\_1 | grade\_2 | grade\_3 | grade\_4 | grade\_5 | grade\_6 | grade\_7 | grade\_8 | grade\_9 | grade\_10 | grade\_11 | grade\_12 | grade\_13 |
| nan | 杭州市 | 淳安县 | 千岛湖镇 |  | 新安东路 |  | 庄天岭小区 |  | 1幢 |  | 1楼 |  |
| nan | 杭州市 | 淳安县 | 千岛湖镇 |  | 环湖北路 | 91号 |  |  |  |  |  |  |
| nan | 杭州市 | 淳安县 | 姜家镇 | 上玉泉村 |  |  | 任家村 | 85号 |  |  |  |  |

最大匹配到乡镇街道这一级，将千岛湖镇和姜家镇区分开，不同类别下不做相似度判断。接下来再进行最大匹配，将可匹配到的统计数据分为一类，只在一类中做相似地址判断。

各类分好之后，进行经纬度的距离计算以及地址文本相似度的计算，根据设定的阈值，取出联合相似度较高的地址数据，进行底层级内容的对比。如果存在地址对应的实体名称，可以联合实体名称进行筛选。

最终对相似度高的数据进行合并，对合并项生成原数据索引。

**5．基于Django技术的Web端数据交付网页搭建**

使用Django可以方便的调用上述预处理、分词和合并等过程的代码，并且Django对于小型项目的开发比较方便，因此选择Djaogo搭建一个包含前后端的Web应用。

用户可以在Web界面上传一个CSV类型的十一级文本数据，上传成功后数据将会传递到服务端进行清洗、分词、回填和映射等一系列操作，最终生成一个十三级的数据，然后将十三级的数据存储到一个CSV文件中。一系列数据处理成功后，Web界面将会显示上传任务成功完成，并且提供一个供用户下载的十三级的CSV文件的链接，用户可以通过浏览器直接下载处理后的CSV文件。

## 五、时间进度

**项目总用时约为75天。具体安排如下：**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 项目阶段 | 阶段内容 | 预计时间（天） |
| 1 | 商品数据预处理 | 将商品数据进行清洗，然后分割成单一UNSPSC分类存量商品数据表 | 5 |
| 2 | UNSPSC与GPC分类标准映射 | 基于自然语言处理技术采用关键词和近似词匹配方式构建UNSPSC分类与GPC分类映射关系。 | 10 |
| 3 | 商品关键词与GPC第4层分类的映射 | 采用自然语言处理近义词技术，使用结巴分词库，提取商品名称中的关键字，去重、清洗后构建商品名称关键词和GPC四级分类号的对应关系，并以.xlsx形式存储 | 20 |
| 4 | 商品的GPC分类 | 根据得到的商品名称关键词和GPC第四层分类映射关系为商品进行GPC分类，获取商品的GPC第4层分号 | 10 |
| 5 | 错误分类商品整理 | 在对商品进行GPC分类时， 汇总其中原有UNSPSC分类错误的数据，进行UNSPSC分类订正，并做GPC归类的数据清单 | 15 |
| 6 | 疑难数据的处理 | UNSPSC分类号错误（根本就找不到UNSPSC分类号）、UNSPSC分类错误、UNSPSC分类映射多个GPC分类等问题，在项目实施过程中将所有疑难数据都会单独提取出来，并进行标记 | 5 |
| 7 | 分类结果人工核验 | 根据UNSPSC分类号从原数据随机挑选商品信息，参照映射表进行核对，形成人工核对记录，如果人工分类与机器智能分类不一致，则记录下来并以人工分类为准进行修订商品GPC分类，同时修订关键词与GPC分类映射关系。 | 5 |
| 8 | 数据交付网页的设计与实现 | 基于Django搭建的一个用于数据交付的网站，把十一个层级的数据上传到服务端用于数据处理，最后把处理后的十三个层级的数据返回。 | 5 |
|  | 验收 | 准备验收材料 | 20 |