项目报告

报告题目：开源项目建设发布

课程名称：开源软件开发

任课教师：李宁

项目导师：赵永瑞

小组组长：李骅萱-2021011140-软工2101

小组成员：

贾泽鑫-2021011132-软工2101

倪松涛-2021011047-计科2101

丁书润-2021011059-计科2102

高翔宇-2021011060-计科2102

项目名称：学习人工智能算法，实现对手写汉字的识别

项目描述：实现对“日”、“田”、“月”、“目”、“由”、“申”、“甲”等汉字的识别。

基于KNN算法实现对手写汉字的识别

项目许可证：BSD 3-Clause License

项目人员和分工：项目文档-丁书润；产品设计-高翔宇；代码开发-李骅萱，倪松涛；ppt撰写-贾泽鑫。

项目开发流程：一、构建数据集

1.每个汉字至少需要500张该汉字的图片，一共7个字，一组五个人，每个人分别写每个字100遍，总共700个字。（“日”、“田”、“月”、“目”、“由”、“申”、“甲”）

2.对每个汉字图片进行命名，比如第一个“甲”字可命名为jia\_1.jpg，第二个“甲”字可命名为jia\_2.jpg,依此类推，其他汉字也采取同样的方式进行命名。（此过程也可以通过python程序进行）获取数据）数据处理→模型训练→模型评估→模型保存→模型推理

数据处理：

数据探索：对数据进行预览

数据清洗：缺失值、异常值（四分位数、箱线图）

数据划分：训练集（80%）、测试集（20%）、（验证集）具有更好的泛化性

特征工程：归一化

模型训练、评估、保存、推理在初级阶段一般没有繁琐的过程

二、对数据进行处理

1.数据探索：查看不同种类的汉字的标签，每一类可用数据有多少个，图片像素的大小等

2.将图片的像素大小化为一致，将图片转换成可处理数据（比如：数组）

3.清除掉不可用的数据

4.进行数据集的划分

三、模型训练、评估（图表展示评估结果）、保存

四、利用模型识别汉字

五、分析

六、完成报告，形成文档

七、将项目开源，上传到GitHub指定仓库

整个过程使用python语言实现对手写汉字的识别

1.基于K-means算法实现对手写汉字的识别，要将不同随机质心位置下训练出的模型评估结果（即模型在测试集上的识别正确率）以图表的形式展示出来。

2.基于KNN算法实现对手写汉字的识别，要将不同K值时训练出的模型评估结果（即模型在测试集上的识别正确率）以图表的形式展示出来。

项目计划（版本、用户、推广)：通过该程序可以实现对手写汉字的识别，可以应用于用户对于不了解拼音的生僻字的输入，与此同时对于相似汉字也能进行区分输入。适合ipad用户和中老年人群使用，本产品此类人推广。可以更好的面向文化水平较低的人群推广。

网站[Bistu-OSSDT-2022/Legendr: 16-赵永瑞-李骅萱 (github.com)](https://github.com/Bistu-OSSDT-2022/Legendr)

代码库：在gitbuh上

语言：Python语言

用到的算法是：KNN和k-means算法

用的第三方库：os，numpy，PIL，random，sklearn.cluster，sklearn，shutil，joblib。