**Excel数据自动标注工具**

**用户使用说明文档**

|  |  |
| --- | --- |
| 版本号 | 1.0 |
| 作者 | 陈凯 |
| 日期 | 2018/1/12 |

上海交通大学

# 引言

## 编写目的

本文档旨在向使用Excel数据自动标注工具的用户（上海智阳网络技术公司）介绍工具的使用方法。

## 项目背景

本项目的委托方为上海智阳网络技术公司，开发方为上海交通大学大数据软件工程实验室。项目于2017年底立项并启动，至今已经历两轮迭代，项目价值与意义得到初步验证。

## 定义

|  |  |
| --- | --- |
| **术语** | **解释** |
| 表头 | Excel表格的头部属性 |
| 标签 | 表头对应的数据库标签 |
| 相似度 | 表头与表头之间的相似程度 |
| 标注 | 为表头打上标签 |

## 参考资料

Excel表头自动识别与标注需求.Doc

# 软件概述

## 目标

本项目旨在对大批量的Excel文件的表头进行自动识别、自动打标签，替代原先大部分的人工重复工作。

## 功能

本项目的主要功能点如下：

1. 文件随机抽取：按照设定的抽取率，从大量文件中随机不重复抽取比例的文件用于人工标注及抽查；
2. 表头识别：通过识别算法识别不同表格的表头；
3. 训练标签库：从完成标注的EXCEL文件中抽取表头和标签对，加入到标签库中，并支持权重设置；
4. 标注：自动为不带标签的文件的表头打上对应标签。

## 性能指标

本项目的主要性能指标如下：

1. 标注准确率：标注正确的个数占所有标注个数的比例;
2. 标注留空率：在设定相似度阈值情形下，当最高相似度小于阈值时对标注进行留空，留空率为留空的个数占所有标注个数的比例;
3. 运行效率：运行工具的时间效率。

# 使用说明

## 运行环境

本工具要求Windows系统，并安装Java环境（JDK1.7以上）。

## 安装和初始化

本工具为Java架包程序，解压即可运行，无需安装。

## 输入

合理、正确的Excel表格文件（格式：xls, xlsx）。

## 输出

程序生成的Excel表格文件（格式：xls, xlsx）。

## 出错和恢复

出错请保存出错记录，并发送给交大开发团队进行诊断。

# 业务流程

本版本工具的业务场景主要有两类：第一类业务场景需要对带有一整个新的标签系统的文件进行标注；第二类业务场景需要对小批量的文件进行标注。

## 新标签系统业务

此时由于标签库中缺乏这个新标签库的信息，此时直接标注往往效果不好，需要先人工标注一部分数据吸收到标签库中。

这一类业务场景的工作流程如图4-1所示。

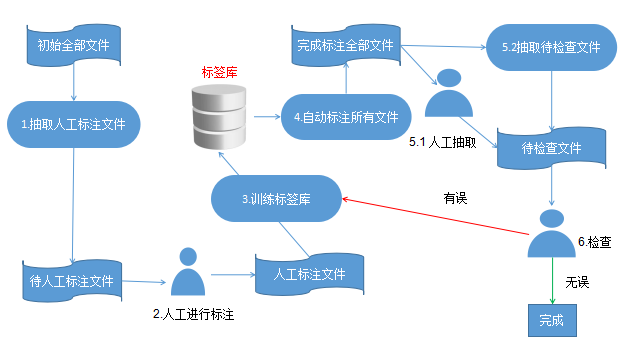


图4-1 第一类业务流程示意

该流程的步骤如下：

**步骤0（准备）**：将所有待标注的EXCEL文件复制到“1-全部文件”下；

**步骤1（抽取人工标注文件）**：运行“step1-抽取待标注文件.bat”，将设定比例的文件复制到“2-人工标注文件”下；

**步骤2（人工标注）**：人工对“2-人工标注文件”下的所有文件进行标注，并保存

**步骤3（训练标签库）：** 运行“step2-训练标签库.bat”，将“2-人工标注文件”、“4-人工抽检文件”下完成标注的文件吸取到标签库；

**步骤4（标注全部）：**运行“step3-自动标注所有文件.bat”，将“1-全部文件”下的所有文件标注并另存为到“3-完成自动标注的全部文件”下；

**步骤5.1（抽取检查文件）：**运行“step4-抽取待检查文件.bat”，将设定比例的文件从“3-完成自动标注的全部文件”下复制到“4-人工抽检文件”下；

**步骤5.2（人工抽取）：**与5.1并行，人工选取部分文件复制到“4-人工抽检文件”下；

**步骤6（人工检查）：**人工对“4-人工抽检文件”下的文件进行检查，并纠正发现的错误，若有误，跳到第3步，否则执行第7步；

**步骤7：**结束。

## 小批量标注业务

此时标签库中存在这部分文件的类似标签信息，此时可以直接进行自动标注。

这一类业务场景的工作流程如图4-2所示。

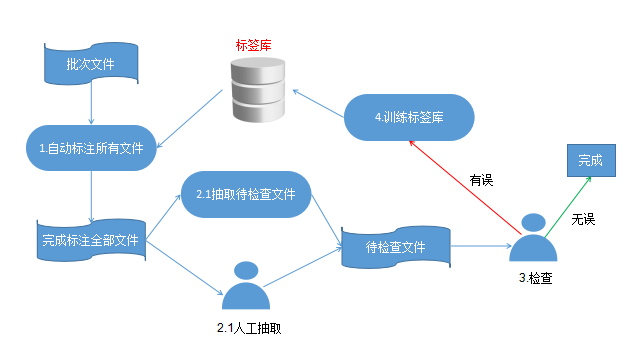


图4-2 第二类业务流程示意

该流程的步骤如下：

**步骤0（准备）**：将所有待标注的EXCEL文件复制到“1-全部文件”下；

**步骤1（标注全部）：**运行“step3-自动标注所有文件.bat”，将“1-全部文件”下的所有文件标注并另存为到“3-完成自动标注的全部文件”下；

**步骤2.1（抽取检查文件）：**运行“step4-抽取待检查文件.bat”，将设定比例的文件从“3-完成自动标注的全部文件”下复制到“4-人工抽检文件”下；

**步骤2.2（人工抽取）：**与2.1并行，人工选取部分文件复制到“4-人工抽检文件”下；

**步骤3（人工检查）：**人工对“4-人工抽检文件”下的文件进行检查，并纠正发现的错误，若无误，执行第5步，否则执行第4步；

**步骤4（训练标签库）：** 运行“step2-训练标签库.bat”，将“4-人工抽检文件”下完成标注的文件吸取到标签库，无条件跳转到第1步；

**步骤5：**结束。

# 程序文件和数据文件一览表

|  |  |
| --- | --- |
| **文件（数据）** | **描述** |
| Step1-抽取待标注文件.bat | 抽取动作运行文件 |
| Step2-训练标签库.bat | 训练动作运行文件 |
| Step3-自动标注所有文件.bat | 标注动作运行文件 |
| Step4-抽取待检查文件.bat | 抽取动作运行文件 |
| Config.ini | 配置信息文件 |
| Readme.txt | 使用说明 |
| 标签库.xlsx | 存储标签数据的文件 |
| Vectors.bin | 保存词向量信息的文件 |
| runablejar | 存放可执行架包的文件夹 |

# 配置文件信息一览表

|  |  |
| --- | --- |
| **字段** | **含义** |
| SimilarityThreshold | 相似度阈值 |
| TaggingExtractPercent | 人工标注抽取文件比例 |
| CheckingExtractPercen | 人工检查抽取文件比例 |
| TrainTagWeight | 从标注文件中吸收的标签的权重值 |