OCR+知识图谱 技术路线

一、图像预处理

处理方法：

对图像进行文字方向校正、二值化、去除噪点（X 邻域算法），提高文字部分清晰度。

可用工具：

opencv

二、初步识别/表格切分

处理方法：

表格切分：

对原图进行膨胀、腐蚀、二值化，突出边线，识别线框，根据线框对图像进行切分。

对于切分后的图像块，用 SWT 算法定位文本区域。

初步识别：

对于切分后的图像和未切分的图像分别进行使用开源 OCR 库对进行初步识别。

可用工具：

图像处理：opencv

开源 OCR 库：tesseract、ocropy

三、版式识别

处理方法：

1. 对上一步初步识别的文字在知识图谱/相关术语词典中进行查询，判断是实体还是概念
2. 对文字所在位置关系进行分析，得出不同文字块所对应的区域类型（属性或者值）
3. 综合 12 步的结果分析得出版式，从而获得原图像电子版式。

可用工具：

- babelnet 多语言百科全书式字典和语义网络：有 http API, Java API, Python API, SPARQL 等多种访问方式
- THUOCL:清华大学开放中文词库 :包含 IT、财经、成语、地名、历史名人、诗词、医学、饮食、法律、汽车、动物等各领域词库，可以下载使用
- XLORE: 中英文跨语言百科知识图谱: 知识来源包括百度和中英文维基，可以进行词条检索、关键词检索、概念检索、实例检索、相关机构人物检索和实体关系预测，有 http api
- 手动构建领域术语库

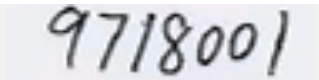
四、精细识别

研究内容：

1. 对先前切分得到的文字区域进行进一步分析，提取手写文字块。
2. 使用基于深度学习的方法进一步对手写数字、字母进行识别，对不同的签名笔迹进行比对。

A. 手写数字、字母识别

- **输入**：包含一行文字块的图片比如下面这个



- 判断其中是否包含手写字母数字，如果是，提取出来
- 对包含手写数字字母的文字块再进行分割（连通区域标记方法），获得分离的字符
- 用机器学习方法训练多分类器，建立识别模型

- **输出**：对于输入图片识别出来的所有文字（9718001）
- **难点**：连笔字符分割、字母数字混合识别

B. 笔迹比对

- **输入**：包含同一人手写签名的多张图片（历史数据），包含手写签名的一张图片（待检图片）
- 进行图像预处理
- 框出待检图片手写签名区域
- 利用神经网络学习签名模式，建立模型（可以考虑 GAN?）
- **输出**：对于待检图片中的签名是否与历史数据中的签名属于同一人的判断。
- **难点**：收集数据集进行模型训练

可用工具：

- python 的包：TensorFlow（深度学习），keras（深度学习），sklearn（机器学习）,opencv（图像处理）
- 数据集：MNIST(包含在 keras 里的手写数字数据集)，Chars74K （共 74K 张图，每张包含一个字符，字符可能是数字或者大小写字母），digits（sklearn 自带的手写数字数据集）
- 开源项目：手写数字识别 ，基于 TensorFlow 的签名匹配

五、纠正

处理方法：

1. 对于经过初步识别和精确识别的结果进行文本纠正，查错并提出可能的正确结果。
2. 文本纠正可以考虑使用的方法有：基于 ngram 的错字纠正，基于依存关系分析的错字纠正 【参考】、基于加权噪声信道模型【参考】

可用工具：

python 的包 pycorrector，同义词词林 下载链接

六、判断

借助构建的术语库与领域知识图谱，对上一步提出的可能结果根据语义和逻辑进行筛选和判断，从而确定最终修改结果。

知识图谱构建可用工具：

- Grakn.ai（构建知识图谱专用的图数据库工具，有自己的查询语言 graql，提供机器学习包直接做知识图谱嵌入及预测等，支持 java，python，nodejs 接口）
- Neo4j 图数据库
- HugeGraph 图数据库
- protege 本体构建工具

图计算可用工具：

- Apache Tinkerpop