方案介绍

---刘志勇（亚信科技（中国）有限公司）

本次低资源保险文档内容信息抽取样本有限，拟采用规则的方式获取结果

本方案主要分为两大模块：保险文档内容识别与保险文档内容结构化

保险文档内容识别方案介绍：

1. 依据给定的pdf文字级别的提取文档，获取每个文字的坐标

2. 将坐标满足一定条件的文字视为同一行

3. 构建一个dataframe，包含每一行内容以及改行的左边界、上边界等信息

保险文档内容结构化方案介绍：

1. 将全部保险文档内容的结构化分为3种情况，分别为形如“第xxx条”、“1. xxx”、以及“第 xxx 部分/章”，针对每一种制定特定的规则方法。

* 针对形如“第xxx条”的保险文档，利用正则获取有明显标题编号的标题及内容，对于没有明显标题编号的内容，利用缩进及内容格式对其进行结构化
* 针对形如“1. xxx”的保险文档，利用正则获取有明显标题编号的标题及内容，对于没有明显标题编号的内容，利用缩进及内容格式对其进行结构化，对于存在乱码以及识别不准确的内容，采用强规则匹配的方法对其进行结构化
* 针对形如“第 xxx 部分/章”的保险文档，利用正则获取有明显标题编号的标题及内容，对于没有明显标题编号的内容，利用缩进及内容格式对其进行结构化，对于需要重新编号的有明显标题编号的次级标题，采用与前述两种情况中没有明显标题编号的内容相似的处理方法。

2. 对官方给定的pdf文字级别的提取文档中不准确的内容（如空格没有识别到、数字识别为乱码等）做强规则匹配

3. 对于一些干扰内容，如脚注、页码等，采用正则过滤，避免写入章节内容之中