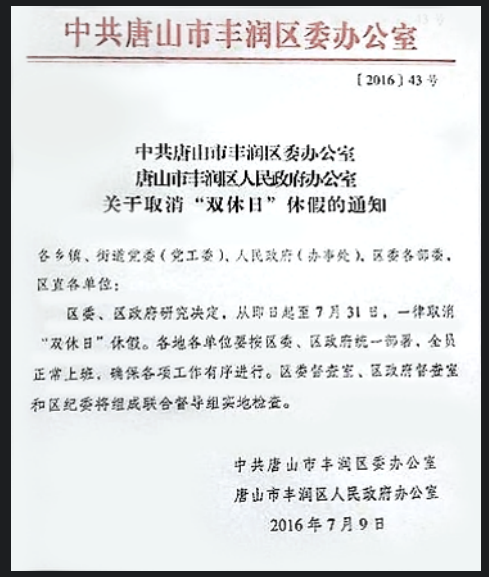
公文要素提取场景

公文是指政府、企业对内、外部的正式发文，有重要信息价值，需要归档并提供全文、关键词检索。



中共唐山市丰润区委办公室

〔2016〕43号

中共唐山市丰润区委办公室
唐山市丰润区人民政府办公室
关于取消“双休日”休假的通知

各乡镇、街道党委（党工委）、人民政府（办事处），区委各部委，区直各单位：

区委、区政府研究决定，从即日起至7月31日，一律取消“双休日”休假。各地各单位要按区委、区政府统一部署，全员正常上班，确保各项工作有序进行。区委督查室、区政府督查室和区纪委将组成联合督导组实地检查。

中共唐山市丰润区委办公室
唐山市丰润区人民政府办公室
2016年7月9日

公文要素示例	
抬头	发文单位
文号	发文日期
标题	……

由于年份和管理机构不同，公文格式也不尽相同，需要模型具备灵活的定位和识别能力

公文要素提取自动化工作流程

文件拆分
多页输入

公文要素
提取模型

输出公文要素
给档案系统

公文要素提取建模方案

不限定要素的固定位置，不同版式均可识别：

样本准备

- 1000-1500张图片起，可包含各种版式和无要素页，分布尽可能与实际一致。
- 80%用于训练，20%用于验证
- 准备标注文件（模型学习的正确答案）

训练过程

- 结构化信息提取自动建模工具
- 工具参数配置：Single_value:true; dict_size:300-400; Proba_thresh : 0.35; 后处理步骤output_type:join
- 准备样本后自动训练，耗时约3小时

模型效果

- 准确率约为 92-97%