9.24 第二次会议记录

* 【每组都要查文献】
* 做一点就在群里发一点，及时跟进。

**任务一：接入OCR识别工具（微信？WPS？百度？）**

1.直接调用成熟系统的OCR识别，批量地识别文本

2.布局分析：通过规则定义将图片分区拆成几个部分，运用OCR识别文本。

* 对于同种类型的样本，选取最普遍的排版形式定义分块坐标，输出JSON格式
* 手写体可能较难识别

**任务二：了解贷款审批流程，标记特征值（周五晚之前）**

1.分成小类别进行标注，定义列名

2.尽可能标注多的样本

3.资产负债表等表格按照会计规则进行数据的分类标注

4.通过查文献等方式了解贷款审批流程

**任务三：学习LSTM与CNN模型，今晚汇报**

1.研读LSTM模型实现代码

2.根据输入数据预测结果调整相应参数

3.运用powerbi实现可视化

**阶段任务：10.1前做出两个（或三个）程序小样本出来，一个OCR，一个LSTM和CNN**

1.OCR

* 通过测试，选定用来识别的OCR
* 代码实现文本识别
* 保存为JSON格式

2.LSTM、CNN

* 构造/使用OCR识别出的时序数据
* 文本预处理：分词、编码为词向量
* 根据结果完善优化模型，调整参数