1. **题目要求**

使用如下开源项目

<https://github.com/xiaofengShi/CHINESE-OCR>

识别两种类型的发票

F001（专票）



F002（普票）



|  |
| --- |
| 输出：发票各个字段，json格式，如下以F001举例说明：  （1）发票类型：001，002（可根据字段数量识别）  （2）购买方名称： 苏州尤信达电子科技有限公司 （3）购买方纳税人识别号： 320504789946492 （4）货物或应税劳务：电子配件 （5）税率：17% （6）税额：16970.94 （7）价税合计： 壹十壹万陆仟捌佰元整 （8）销售方名称： 上海威萌贸易有限公司 （9）发票号码： 202710477 （10）发票代码： 3100103140 （11）开票日期： 2011年04月28日 （12）金额：99829.06 |

对于F002，输出能识别的字段即可（自行定义格式）

1. **其他注意事项**
2. 所有代码都要push到自己的github repo，并提交repo链接给公司
3. 使用anaconda创建python3的虚拟环境（CPU）进行开发
4. 在readme中要写清楚（中文描述即可）
   1. 环境的安装，部署步骤
   2. 自己基于开源项目做了哪些代码修改（写具体）
   3. 如何运行自己的程序
   4. 测试结果，与不足之处
5. 写清楚11个字段的具体定位方法，注意要有一定的泛化能力，后面我会用其他图片测试你的程序
   1. 所以不能直接根据字段的绝对位置进行定位
   2. 可以采用相对位置定位找到各个字段（比如先找到顶部“发票”两个字，“发票”右侧“开票日期”之上的，是 发票号码）
   3. 实际采集的图片，有轻微的角度偏移，你写的代码要能适应这种情况
6. 写一下你对这个开源项目的理解
   1. 它内部用了哪些模型
   2. 自己下载数据集，并从中抽出少量数据进行训练
      1. 面试时远程演示训练模型，以及如何使用训练后的模型
   3. 如果我们自己构建数据集，应该怎么做
   4. 自己用哪些工具可以标注这样的数据集
      1. 面试时远程演示用标注工具，从这张发票抠出数据、融合到训练集中训练模型
      2. 注意新加入的数据，可能含有原始数据中没有的字符，找到具体加入新字符的代码/文件
7. **每个应聘者基础不一样，不要求所有题目都能完成；能完成多少就完成多少，尽力完成即可。**
8. **如果有自己做不了的题目，及时给我微信留言**
9. **附加题（能完成有加分）**

（1）使用另一个开源代码完成该题目识别要求：<https://github.com/YCG09/chinese_ocr>

（2）自己上网找一些“专票”和“普票”的图片，用自己的代码做测试

（3）识别如下多字段的图片

