2019-8-30

XUAN PENG、QIUYUN CHENG

深杳智能科技

程序说明文档

增值税发票等5类票据识别竞品分析

# 目录

[目录 1](#_Toc18068412)

[一、 概述 2](#_Toc18068413)

[二、 版本 3](#_Toc18068414)

[三、 程序使用说明 4](#_Toc18068415)

[1. 产品说明文档 4](#_Toc18068416)

[2. 公司账号密码及其说明 5](#_Toc18068417)

[3. 运行过程 7](#_Toc18068418)

[4. 程序与目录结构 9](#_Toc18068419)

[（1）程序相关文件目录 9](#_Toc18068420)

[（2） 关于备份 10](#_Toc18068421)

[5. 修改测试竞品程序过程说明 19](#_Toc18068422)

[6. 修改竞品运行账号说明过程 22](#_Toc18068423)

[(1) 增值税发票竞品运行账号修改说明 22](#_Toc18068424)

[(2) 火车票竞品运行账号修改说明 26](#_Toc18068425)

[7. 注意事项 28](#_Toc18068426)

[四、 填写竞品调用记录表及其说明 31](#_Toc18068427)

[五、 竞品比较分析表填写说明 36](#_Toc18068428)

[1. 数据填写 36](#_Toc18068429)

[2. 数据转移 36](#_Toc18068430)

[3. 数据修正 38](#_Toc18068431)

[4. 格式规范 39](#_Toc18068432)

[（1）对比测试报告的格式规范 39](#_Toc18068433)

[（2）各个分表的格式规范 39](#_Toc18068434)

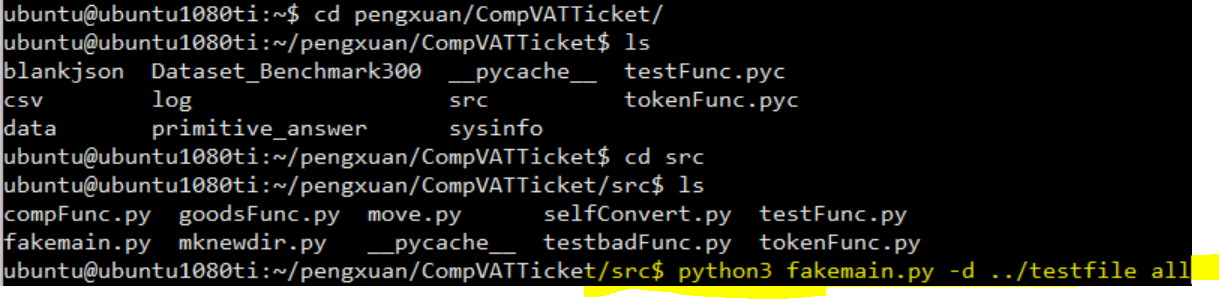
# 概述

本文档内容主要陈述公司增值税发票识别竞品分析程序的使用方法和维护机制，以及后续文档的填写规范。该程序目前支持对百度，华为，阿里，科大讯飞，睿琪，易道博识的增值税发票识别竞品分析。以上公司都通过调用API返回json文件来返回识别结果。

环境：Ubuntu 16.04 编译：python 3.52

使用方法：

进入src文件夹下



标黄的语句是运行程序的方法。

其中all指进行一次完整的分析流程（含：transdata指仅仅进行数据转移和格式转化、callapi指仅仅调用各公司API、compare指仅仅对本轮分析文件夹下result和answer中的json进行比较[[1]](#footnote-1)），它可被替换为其他操作。进入src文件夹下输入python3 fakemain.py -h可以看到详细用法。

../testfile可以被替换为其他文件夹名称，但需要注意为方便起见必须是../为前缀。

# 版本

|  |  |  |  |
| --- | --- | --- | --- |
| **文档版本** | **更新时间** | **新增特性** |  |
| Version 0.0 | 2019-01-09 | 无 | 彭炫 |
| Version 1.0 | 2019-02-21 | 对0.0版本的程序相关目录进行补充，并添加产品说明文档、添加新竞品的修改步骤。 | 贺小娟 |
| Version2.0 | 2019-05-10 | 对1.0版本中OCR免费使用额度以及各竞品的账号密码做了相应的补充，并增加了调用记录表作为附件。 | 陈海燕  蔡涵玥 |
| Version3.0 | 2019-05-20 | 在2.0版本中的基础上，增加了竞品比较分析表填写说明模板 | 陈海燕  蔡涵玥 |
| Version4.0 | 2019-5-31 | 在3.0版本中的基础上，增加了修改测试竞品程序过程说明，填写竞品调用记录表及其说明两个模板，并把结构调整为五个部分 | 蔡涵玥  陈海燕 |
| Version5.0 | 2019-6-03 | 在4.0版本中的基础上，增加了 修改竞品运行账号说明过程，并进一步修改了文档的细节问题和格式问题 | 陈海燕 |

# 程序使用说明

## 产品说明文档

|  |  |  |
| --- | --- | --- |
| 公司名称 | 接口说明文档网址 | OCR免费额度 |
| 百度 | https://ai.baidu.com/docs#/OCR-API/top | 500次/天，超出按下图计费 |
| 阿里 | [https://market.aliyun.com/products/57124001/cmapi027758.html?spm=5176.mktshop1334330.0.0.qokMae#sku=yuncode2175800002](https://market.aliyun.com/products/57124001/cmapi027758.html?spm=5176.mktshop1334330.0.0.qokMae%23sku=yuncode2175800002) | 购买增值税10000次+购买火车票10000次 |
| 华为 | <https://support.huaweicloud.com/api-ocr/ocr_03_0019.html> | 1000次/月，超出0.4元/次 |
| 腾讯 | <https://cloud.tencent.com/product/ocr> | 1000次/月，超出0.5元/次 |
| 科大讯飞 | <https://doc.xfyun.cn/rest_api/%E5%A2%9E%E5%80%BC%E7%A8%8E%E5%8F%91%E7%A5%A8%E8%AF%86%E5%88%AB.html> | 500次/天，超过限制会提示，提示错误码11201 |
| 易道博识 | <http://ai.exocr.com/document?cursor=ocr_VAT_invoice> | 0.2元/次 |
| 睿琪票据 | https://fapiao.glority.cn/wiki/ | 剩余次数详见官网控制台 |
| 灵云 | <https://www.aicloud.com/dev/sdk/viewsdk/typeid/1> | 0.02元/次，无免费试用额度 |
|  |  |  |

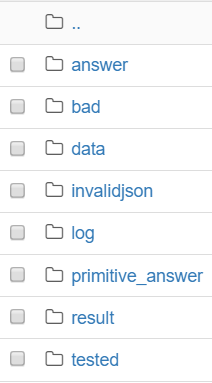
## 公司账号密码及其说明

|  |  |  |
| --- | --- | --- |
| 公司名称 | 对应的账号与密码 | 更改账号密码及其注意事项 |
| 百度 | 可以使用自己的手机号注册直接使用 | 百度调用时效性API口令，需要在tokenFunc中更新百度账号等内容 |
| 阿里 | appcode：  火车票：  c2a46dfdf2004ad1b43c20fc85b5815c  增值税发票：  80b8d3a52d3b48f98ce21d74df1508c6 | 阿里的增值税发票识别账号归公司管理，使用时不需更改账号 |
| 华为 | 账号1：pengxuan\_lab  密码：PXliex3tome\*  手机号码：18217509726  账号2：caihanyue\_tab  密码：shenyao123^  可以使用自己的手机号注册 | 建立华为云新账户，开通增值税发票识别服务。华为调用时效API口令，需要在tokenFunc中更新账号密码等内容。除此以外，tokenFunc中可能还需要更改dic[]括号中的项数，以返回需要的**token**。dic[8]返回的才是**token** |
| 腾讯 | 可以使用自己的手机号注册直接使用（最近新添加腾讯） |  |
| 科大讯飞 | appid:5d086a20 apikey:96143f236baef1f6a9a7ffc19bfe361c  可使用自己的手机号注册 | 申请讯飞webAPI格式的项目，开通增值税发票识别，分别修改testFunc和fakemain里的中的appid和api\_key，并修改后台白名单，修改方式可看讯飞返回的json文件中“illegal access|illegal client\_ip”。 |
| 易道博识 | app\_key:bc3d03d1e58249fc857c4cab3cab4e37  app\_secret:190714e1413da50c3019c094febe96c9 |  |
| 睿琪票据 | APPKEY :5cc14a77  APPSECRET :2f48bc2f7e346be85f6625be823b3dab | 在testFunc中直接修改key和secret即可。2019年6月，睿琪的调用次数可能不够了，需要公司花钱购买次数。励程老师名下有睿琪账号，手机验证码登陆即可。 |
| 灵云 |  |  |

## 运行过程

本程序可用的操作选项有：all/transdata/callapi/compare，进行一次完整分析的流程如下(all=transdata+callapi+compare)：

1. 将公司提供的测试图片放入./data文件夹下，提供的标注答案txt放入./primitive\_answer文件夹下；
2. 进入./src文件夹，输入python3 fakemain.py -d ../testfile all
3. 运行完毕后回到上层，会发现.目录下多了一个名为testfile的文件夹。文件夹内结构如下：



primitive\_answer存放的是本轮分析我司提供的不符合格式要求的答案文件；answer存放的是transdata后符合格式要求的答案文件；result存的是各个公司API调用结果的json文件；invalidjson是经检测发现的各个公司API调用中无效的结果json；data在程序开始时会有大量图片，随着测试的进行，这些图片会根据符合模板/不符合模板，分别进入./testfile/tested和./testfile/bad文件夹下。

1. 竞品分析的结果csv文件被生成于./testfile/result内，使用excel查看比较可读。

## 程序与目录结构

### （1）程序相关文件目录

|  |  |  |
| --- | --- | --- |
| **目录文件夹** | **文件** | **备注** |
| blankjson | 含有用以替代无效json的blank.json文件，除阿里和华为外的公司都有对应。 | 无效json指与正确返回json格式不一致的json，这会导致之后比对answer时报错。该问题主要出现在百度。为了保持结构一致所以同时建立其他公司文件夹。 |
| csv | refnew\_all.csv | 该文件下列出了各个公司的可用字段，程序会根据这些可用字段来对返回的json判断相应字段对错。**必要时需要更新。【更新方式待补】** |
| data | 含有竞品分析测试图片。手动导入即可。 | 每次运行程序时data里的jpg文件会被转移到指定文件夹下的data中。目的是方便测试不同的数据集。但一般而言，公司会使用Dataset\_Benchmark300中的图片。 |
| primitive\_answer | 公司提供的参考标注答案.json所在文件夹。手动导入即可。 | 每次运行程序时primitive\_answer里的json文件会被转移到指定文件夹下的primitive\_answer中。目的是方便测试不同的数据集。（但一般而言，公司会使用Dataset\_Benchmark300中的图片。）通过从测试系统下载对应的样本集即可获得json文件。 |
| sysinfo | readme  helptext |  |
| log | errlog.csv | 排查特定的答案错误时的输出，依据公司需求可以随时更改相关逻辑。 |
| Dataset\_Benchmark300 | 含数据集图片与答案标注 | 备份 |
| src | mknewdir.py | 内含建立**本轮分析文件夹**结构的相关函数。本轮分析文件夹的结构请见前一部分。 |
|  | testFunc.py | 内含调用所有公司API的相关函数 |
|  | tokenFunc.py | 内含获取有时效性的API调用口令的函数。百度和华为口令具有时效性。 |
|  | selfConvert.py | 内含将当前不符合要求的文本文件转化为符合要求格式的json文件的相关函数（两级字典），primitive\_answer转化结果自动存入本轮分析文件夹下的answer文件夹内。目前需要进行json文件转换的除了我司提供的标注答案文件外还有易道博识、睿琪票据两大公司。 |
|  | move.py | 内含将./primitive\_answer与./data中的数据转移至本轮分析文件夹下的相关函数。 |
|  | testbadFunc.py | 对于某些图片调用API时，有一些公司由于技术所限和不稳定的服务，会输出无效的json或者直接返回HTTPERROR。直接返回HTTPERROR（阿里，华为）的行为，一般来说是由于这张图片不符合增值税发票的模板所以被拒，相应的图片会被放进本轮分析文件夹下的bad文件夹中。另一些图片会使有些服务不够稳定的公司(百度，讯飞，易道博识)暂时无法识别而返回错误的json。这种情况下，会多给这些公司多3次重新调用图片API的机会。相关函数就在该文件夹下。 |
|  | compFunc.py | 内含将各公司识别结果与我司提供的标注答案进行比较的函数。比较的标准是，如果某张图片在本轮分析文件夹的tested文件夹下，说明调用api成功，就需要对这张图片相应的答案进行比对。 |
|  | goodsFunc.py | 由于部分公司货物字段的json格式不统一，因此需要对这类公司的货物字段进行额外的处理后分析。详见内含的相关函数。 |
|  | fakemain.py | 包含callapi、compare、transdata、all等功能的主函数。  1. transdata：将所给答案进行标准处理后移到分析文件夹下，如果需要添加新公司的竞品也需要进行必要的转换，为了避免对其他付费的调用次数产生影响，可以将单独调用得到的json文件放入相应目录一同进行转换与移动；  2. callapi：调用各公司的API，并将返回的json文件统一格式存入各家公司的标准文件夹中；  3. compare：读取各公司的标准json文件，进行字段比较，返回字段比较结果表；  4.all：实现新一轮的竞品分析，完成以上所有功能。 |

### 关于备份

每一轮分析都会建立一个新的文件夹（即上面提到的本轮分析文件夹），下面有分析过的图片和调用API得到的结果。这本来是为了方便每次分析不同的图片，但是当每次使用同样的测试集时，显然就会冗余很多相同文件。每次竞品分析完毕，可以把本轮分析文件夹取出保存，并使用脚本定时清理这些历史“本轮分析文件夹”中的tested和bad图片。5.添加新公司指引

本部分按顺序给出添加新公司所要修改的py文件（即src文件夹）内容，仅供参考。在修改py文件内容前，要记得在../csv/refnew\_all.csv中添加新公司的字段名称。因大多数函数呈现相同的结构，修改详情请参考源文件代码内容；如无特殊情况，请按以下顺序和要求进行添加。

**1.testFunc.py**

|  |  |
| --- | --- |
| 功能 | 调用公司的API |

按照新添加公司的API文档进行调用，将返回的内容以json文件存入指定的路径。输入参数中只需修改token，也可直接将token直接写入函数内，不设置输入参数，但是对应主函数中的调用函数也应做相应修改，尽量不要这样做。

**2.tokenFunc.py**

|  |  |
| --- | --- |
| 功能 | 获取公司即时token |

大部分公司没有要获取即时口令以调用API的需求，若果有，按照官方给出的文档进行获取以进行下一步分API调用。

**3.testbadFunc.py**

|  |  |
| --- | --- |
| 功能 | 判断返回的json是否有效 |

需要修改的函数：validateJSON：加入新公司的函数，根据官方文档判断返回的json文件是否正确。正确，将json文件转移至/testfile/result/company/文件夹下；错误，删除返回的json文件，进行新一轮的尝试，三次尝试失败后，将测试图片转移至/testfile/bad/文件夹下。

**4.selfConvert.py**

|  |  |
| --- | --- |
| 功能 | 将不符合标准格式的json文件进行统一 |

观察要添加的新公司返回的json格式，如果符合标准格式（二级字典）则不需做转换；如不符合，需在文件中添加convert\_one\*\*和convert\_all\*\*函数。根据返回的json格式将所有json文件转换成标准格式。

**5.goodsFunc.py**

|  |  |
| --- | --- |
| 功能 | 对与货物相关字段进行处理。主要是因为某些公司返回的json格式中的结果里面还存在子字典，多出现在货物字段。 |

该文件存在的中心在于将json文件统一成标准的二级字典格式，主要存在货物字段，货物字段包含"ItemContent", "ItemStandard", "ItemUnit", "ItemQuantity", "ItemUnitAmount", "ItemAmount", "ItemTaxAmount", "TaxRate"等。根据返回的json文件判断所在字段对应的结果中是否含有子字典，如果有则要做出转换，具体的转换内容应该根据json结构分析。

**6.mknewdir.py**

|  |  |
| --- | --- |
| 功能 | 建立testfile文件夹结构。 |

只需要在该文件夹下中的第三行的compname列表中加上新公司的名字即可。

**7.compFunc.py**

|  |  |
| --- | --- |
| 功能 | 将各公司识别结果与我司提供的标注答案进行比较。 |

函数依次需要修改或者添加的参数有但不仅限于以下：

|  |  |
| --- | --- |
| 修改内容 | 说明 |
| col\_new | 新公司的字段在refnew\_all.csv中所在的列 |
| startjson\_new | 新公司json文件保存结果的key |
| namelist | 添加新公司的名称 |
| s | 添加新公司的相关字段，连续有两处需要修改 |
| dict\_ziduan | 添加新公司key和value |
| rnew，valid\_new | 格式化添加新公司对应的值 |
| goods\_new | 主要是对某些公司货物字段返回格式的处理，如果公司返回的货物字段是一级字典，参照阿里等公司赋值，如果返回的货物字段是二级字典，参照讯飞等公司对goods\_new进行赋值。 |
| rnew,valid\_new | 连续两处对其进行格式化更新。 |
| return | 添加要返回的新公司的对应结果。 |

**8.fakemain.py**

只需要特别注意一些命名规则一样的地方，按照规则添加新公司的信息即可。值得注意的地方是tmp，当添加新的公司时，需要在"TOTAL,,,,,,,,,,,,,,"中添加一个’,’(csv格式)。

下面放一个具体例子，以增加华为火车票竞品分析为例。

需要的材料：

1. 由api接口获得对应的json文件，token的获得依照是否即时在testFunc.py和tokenFunc.py中进行修改；

注：因为对应接口可能收费的原因，所以修改的步骤尽量分开，即transdata，callapi，compare在完成对应修改后对应调试。

1. Json文件可能根据需要的结构进行相应的调整，这一过程在selfConvert.py；
2. 修改新的参考ref\_train.csv文件，建立新的对应格式；
3. 按照之前的格式对main.py，comFunc.py进行相应的修改

按照需要进行逐步修改：

这里以火车票-华为这一竞品公司的相关程序为例。

1. 按照对应公司的api文档，添加相应的token获得函数和接口函数；

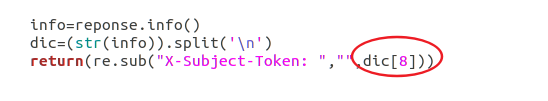
由于华为的token获取是有时效性的，所以需要在在testFunc.py和tokenFunc.py中都添加相关函数。

在testFunc.py中，注意对应url的修改为火车票：



在tokenFunc.py中，注意画圈处dic[]里面的数字，需要确定对应的是token才能正常调用接口：



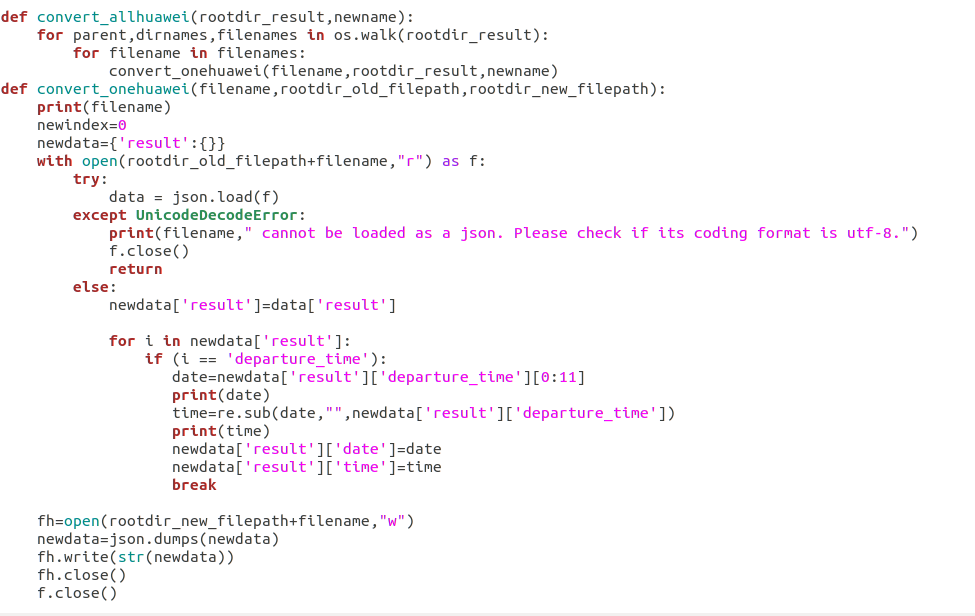


1. 调用接口获得正确响应的Json文件，研究结构，发现需要进行的调整，同时发现华为表示时间的字段是将日期和时间放在了一起，所以也需要在新的json文件中进行调整。

原json文件如下：



在selfConvert.py中添加convert\_allhuawei和convert\_onehuawei两个函数



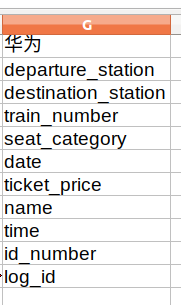
注意：关于时间和日期的格式，为了统一可能需要进行适当的转化，详情可参见compFunc.py中的transDate和transTime两个函数。

1. 修改索引和参考文件。

首先在src文件夹中的mknewdir.py添加“huawei”：

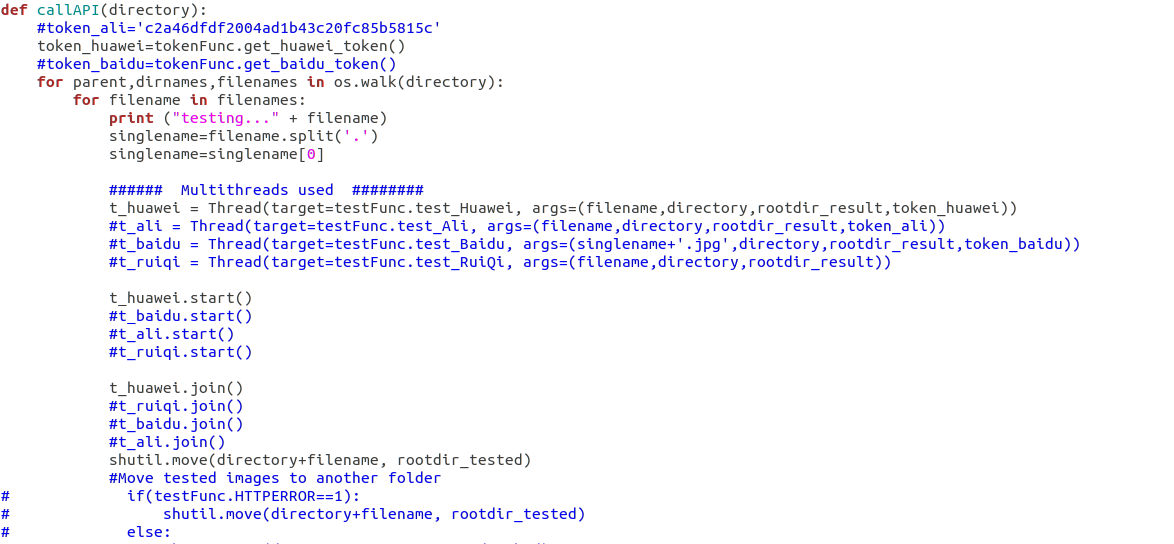


然后在csv文件夹中的ref\_train.csv中手动添加华为相关列，建立对应关系：

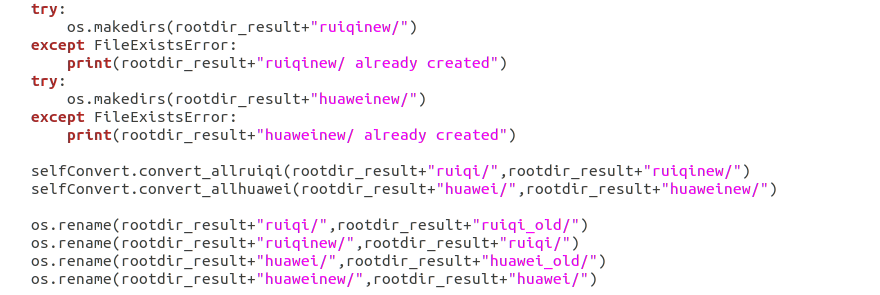


1. 对main.py，comFunc.py进行相应的修改：

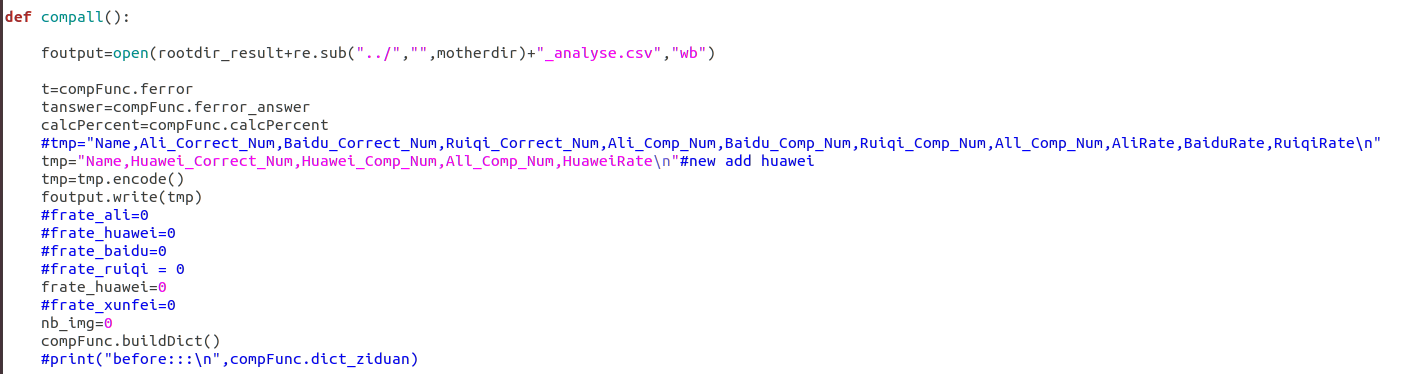
首先是main.py中callAPI函数有关华为的部分：

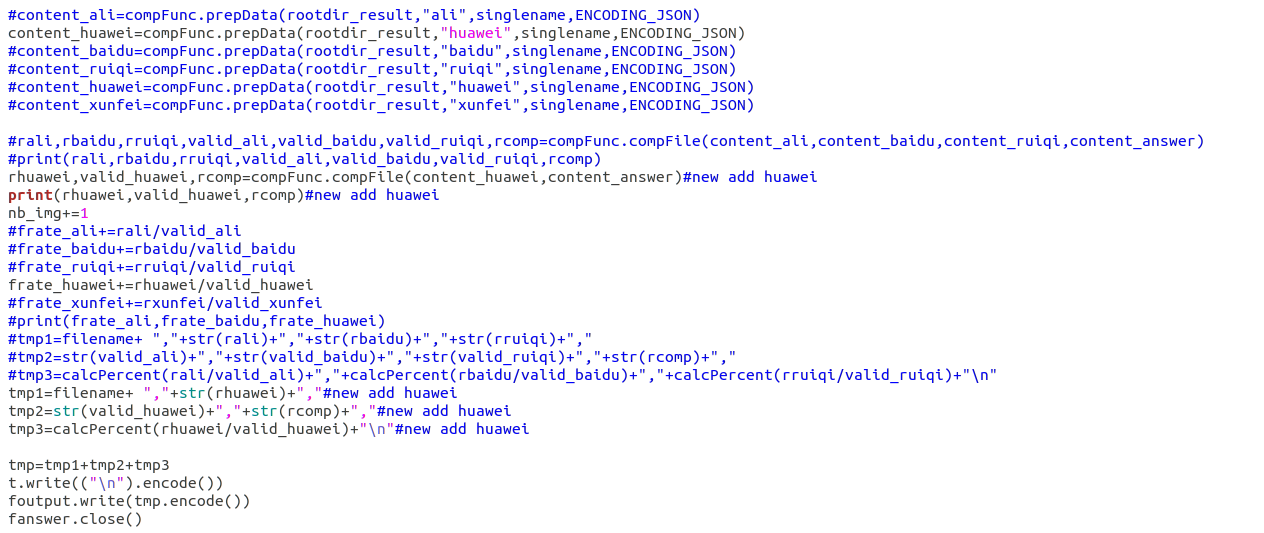


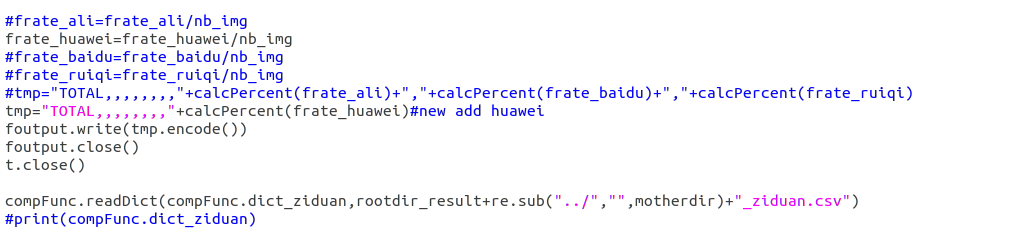
下面是因为相关结构，所以需要储存之前的json文件，



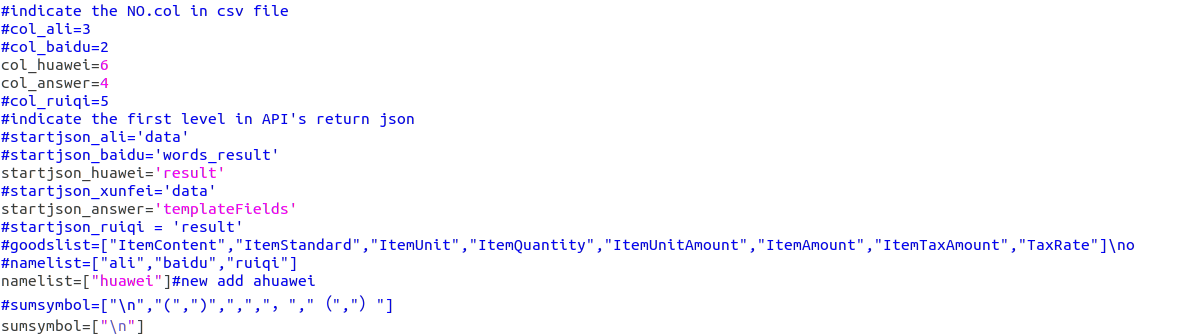
然后是main.py中compall函数有关华为的部分：

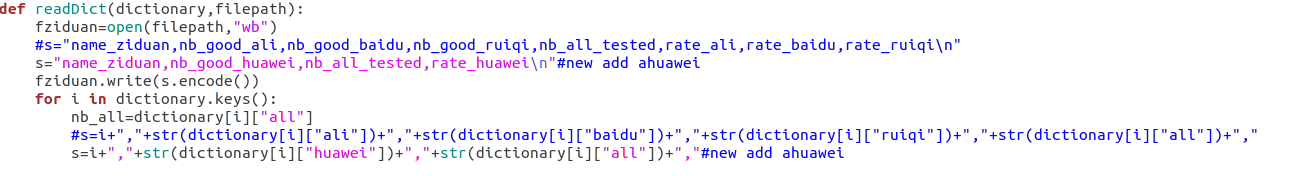




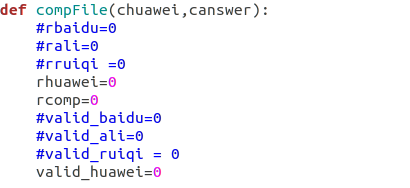


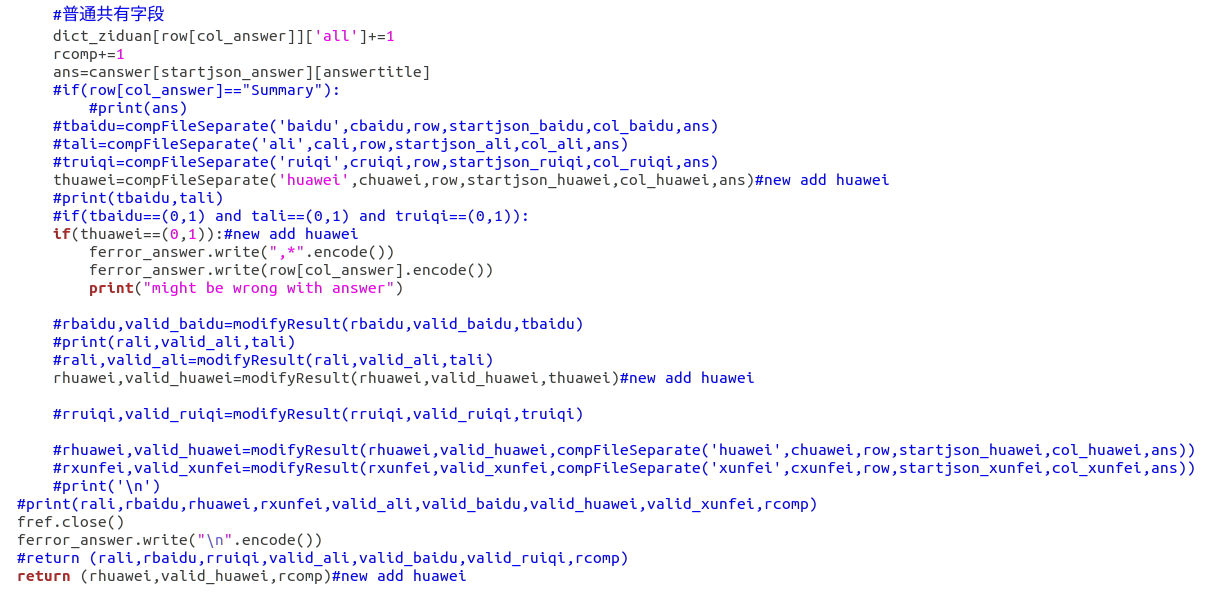
接下来添加comFunc.py中关于华为的部分：





comFunc.py中compFile函数：

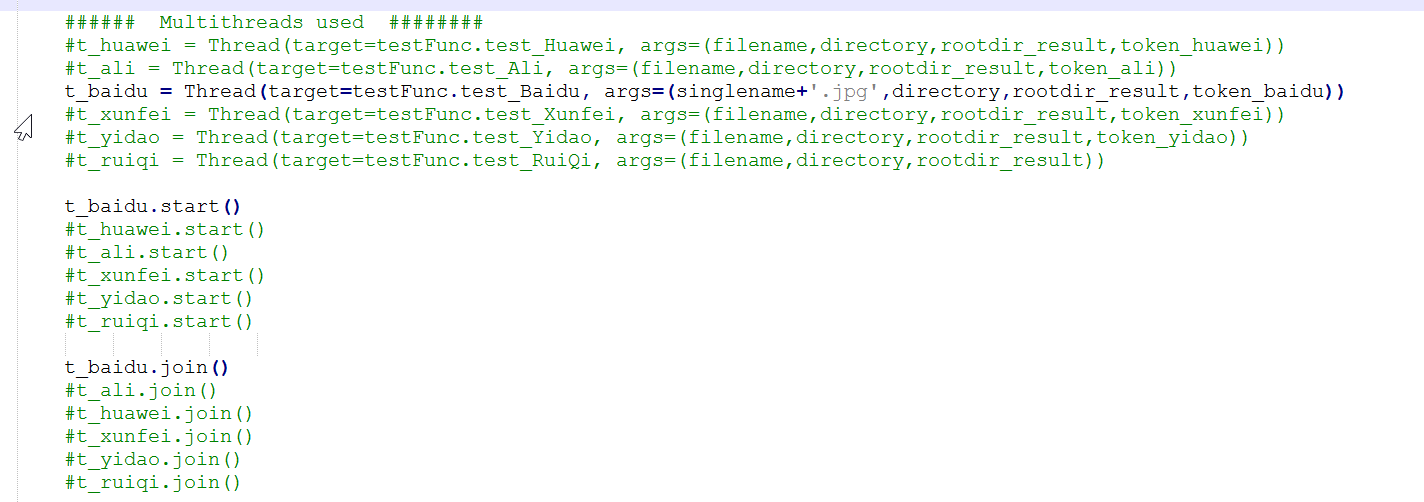




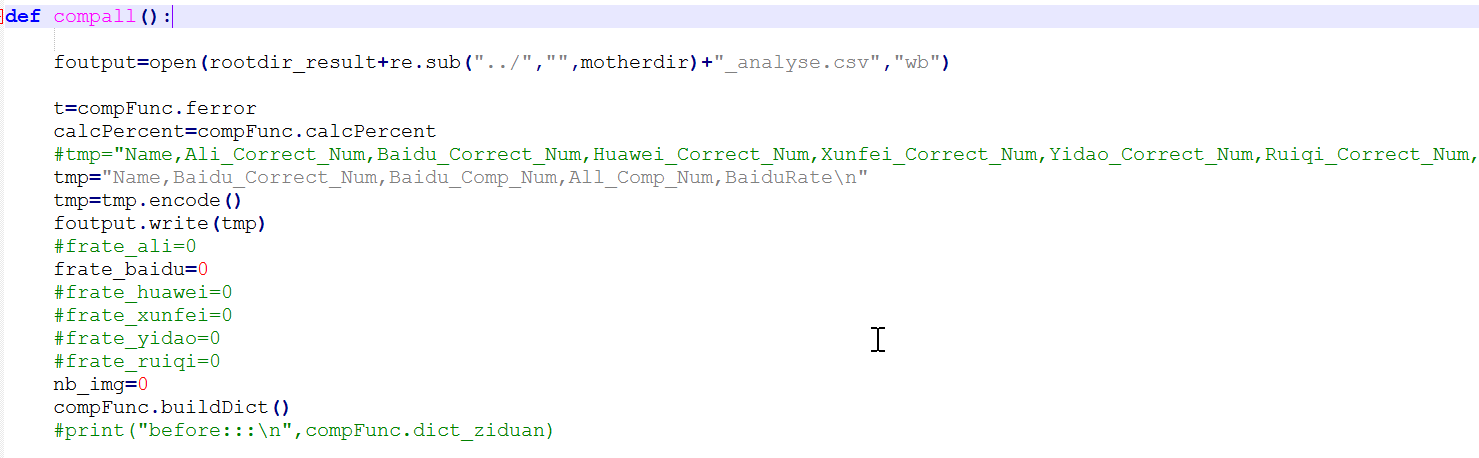
## 修改测试竞品程序过程说明

当要求只对其中一个竞品或者几个竞品测试时，需要对程序进行稍微修改。以增值税发票识别为例（火车票识别类似），假设只需要调用百度进行发票识别。请注意，修改时小心谨慎，特别是python对缩进要求较高。具体步骤如下：

1. 进入src文件夹下的fakemain.py。因为transdata这一步不需要花钱，也对后续的额步骤没什么影响，所以这一步可以不做修改。在def callAPI函数中，将不需要的公司调用API过程全部注释掉，其中一部分注释的语句如下图所示。



在def compall()函数中，将不需要的公司和标注答案进行比较的过程全部注释掉，其中一部分注释的语句如下图所示。





1. 进入src文件夹下的compFunc.py。和上面类似，将不需要的竞品程序注释掉，只保留需要的竞品和标注答案相关部分。截取其中一部分注释部分如下：



## 修改竞品运行账号说明过程

竞品运行的账号发生变动的时候，需要在程序中修改对应竞品的运行账号，本节主要介绍了修改竞品运行账号说明过程，主要分为两部分介绍：增值税发票竞品运行修改说明和或者票竞品运行账号修改说明。

### (1) 增值税发票竞品运行账号修改说明

增值税发票对应的竞品目前包括百度，阿里，华为，科大讯飞，易道博识以及睿琪，其中百度，科大讯飞是按日有免费次数，华为是按月有免费次数，而阿里，睿琪，易道博识是收费的。下面将依次介绍这几个竞品的运行账号修改过程。

阿里的增值税发票通过注册账号和充值，就可以得到对应的token。只需要在程序中修改阿里对应的token，修改的文件位置为CompVATTicket/src/fakemain.py，修改具体内容的位置为fakemain.py下callAPI方法下的变量token\_ali，如图6-1所示。

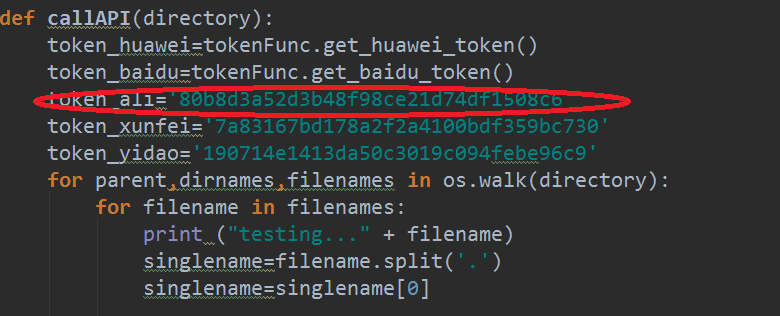


图6-1 阿里增值税账号修改的具体位置

睿琪的增值税发票通过注册账号和充值，就可以得到对应的appkey以及app\_secret。在程序中修改睿琪对应的appkey和app\_secret，修改的文件位置为CompVATTicket/src/testFunc.py，修改具体内容的位置为testFunc.py下test\_RuiQi方法中的变量appkey，appsecret，如图6-2所示。



图6-2 睿琪增值税账号修改的具体位置

易道博识的增值税发票通过注册账号和充值，就可以得到对应的app\_key以及app\_secret。在程序中修改易道博识对应的app\_key和app\_secret，修改的文件位置为CompVATTicket/src/testFunc.py，修改具体内容的位置为testFunc.py下test\_Yidao方法中的变量data中的app\_key和app\_secret，如图6-3所示。

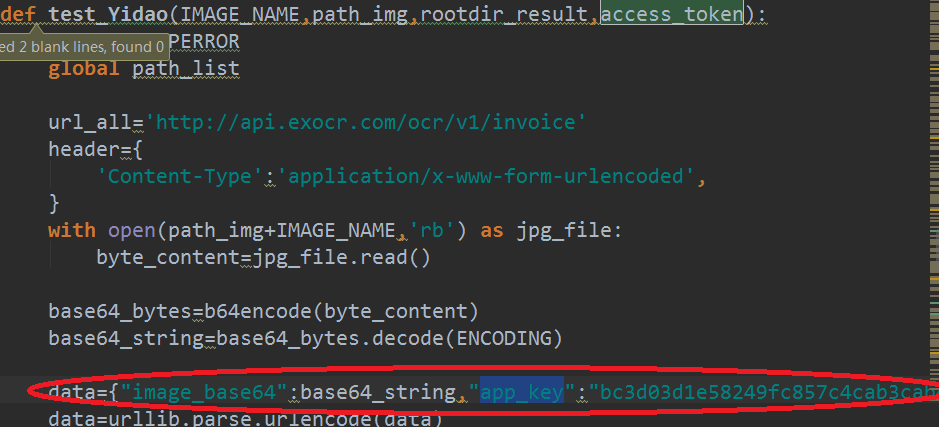


图6-3 易道博识增值税账号修改的具体位置

百度的增值税发票通过注册账号，申请相应的服务，就可以得到对应的grant\_type，client\_id以及client\_secret。在程序中修改百度对应的grant\_type，client\_id以及client\_secret，修改的文件位置为CompVATTicket/src/tokenFunc.py，修改具体内容的位置为tokenFunc.py下get\_baidu\_token方法中的变量grant\_type，client\_id以及client\_secret，如图6-4所示。

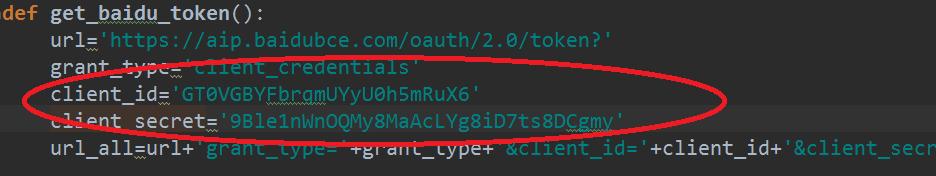


图6-4 百度增值税账号修改的具体位置

科大讯飞的增值税发票通过注册账号，申请增值税发票识别的服务，就可以得到对应的appid以及appkey。在程序中修改科大讯飞对应的appid以及appkey，appid修改的文件位置为CompVATTicket/src/testFunc.py，修改具体内容的位置为testFunc.py下test\_Xunfei方法中的变量appid，如图6-5所示。appkey修改的文件位置为CompVATTicket/src/fakemain.py，修改具体内容的位置为fakemain.py下callAPI方法中的变量token\_xunfei，如图6-6所示。

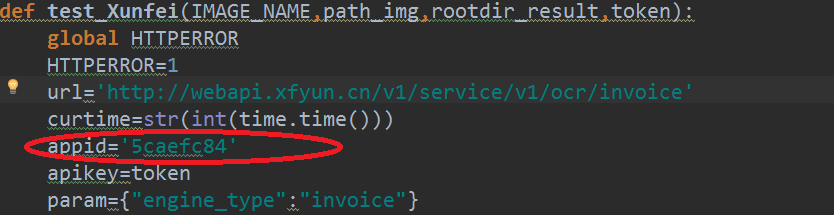


图6-5 科大讯飞增值税appid修改的具体位置

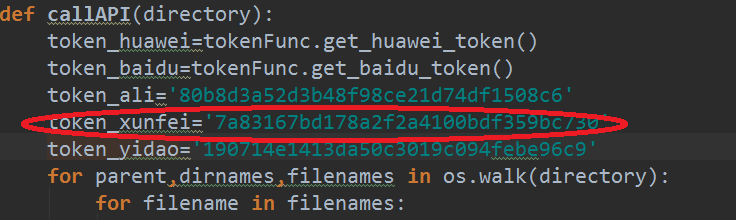


图6-6 科大讯飞增值税appkey修改的具体位置

华为的增值税发票通过注册账号，申请增值税发票识别的服务。通过用户名，密码以及domain的名称，就可以在程序中调用华为的api接口，在程序中修改华为对应的用户名，密码以及domain的名称，修改的文件位置为CompVATTicket/src/tokenFunc.py，修改具体内容的位置为tokenFunc.py下get\_huawei\_token方法中的变量data，用户名对应的是user下的name，密码对应的是user下的password，而domain对应的是domain下的name，如图6-7所示。

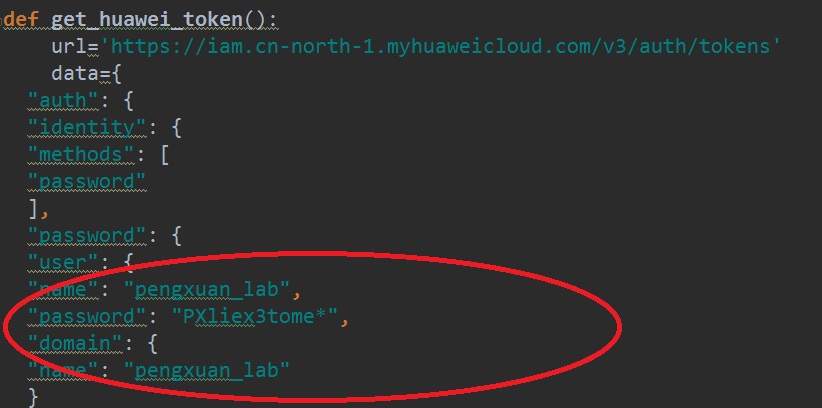


图6-7 华为增值税账号修改的具体位置

腾讯的增值税发票通过注册账号，申请增值税发票识别的服务。通过secret\_id，secret\_id以及action的内容，就可以在程序中调用腾讯的api接口，在程序中修改腾讯对应的secret\_id，secret\_id以及action的内容，修改的文件位置为CompVATTicket/src/testFunc.py，修改具体内容的位置为tokenFunc.py下test\_TengXun函数中的对应内容，如图6-8所示。

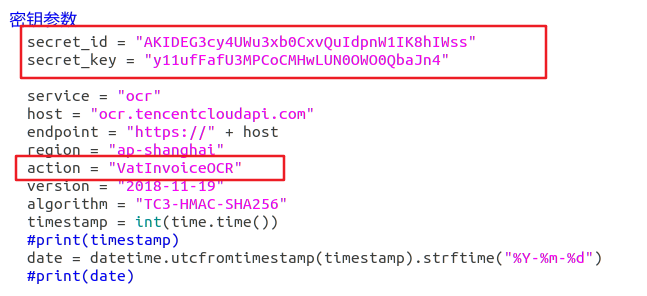


图6-8 腾讯增值税账号修改的具体位置

### (2) 火车票竞品运行账号修改说明

火车票对应的竞品目前包括百度，阿里以及睿琪，其中阿里和睿琪是收费的，百度是每天有免费的次数。下面将对百度，阿里以及睿琪的火车票运行账号进行修改说明

阿里的火车票是通过注册账号和充值，就可以得到对应的token。只需要在程序中修改阿里对应的token，修改的文件位置为CompTrainTicket/src/main.py，修改具体内容的位置为main.py下callAPI方法下的变量token\_ali，如图6-8所示。

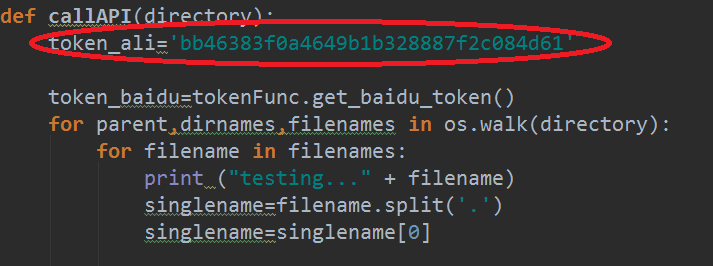


图6-9 阿里火车票账号修改的具体位置

百度的火车票是通过注册账号，申请相应的服务就可以得到对应的grant\_type，client\_id以及client\_secret。在程序中修改百度对应的grant\_type，client\_id以及client\_secret，修改的文件位置为CompTrainTicket/src/tokenFunc.py，修改具体内容的位置为tokenFunc.py下get\_baidu\_token方法中的变量grant\_type，client\_id以及client\_secret，如图6-9所示。

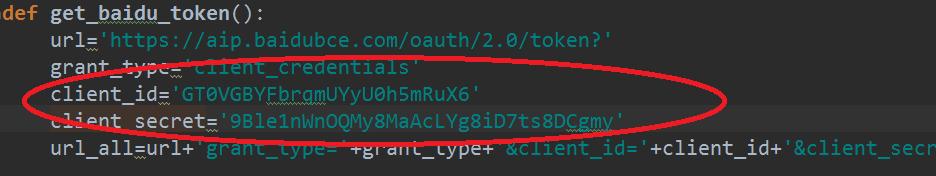


图6-10 百度火车票账号修改的具体位置

睿琪的火车票通过注册账号和充值，就可以得到对应的appkey以及app\_secret。在程序中修改睿琪对应的appkey和app\_secret，修改的文件位置为CompTrainTicket/src/testFunc.py，修改具体内容的位置为testFunc.py下test\_RuiQi方法中的变量appkey，appsecret，如图6-10所示。



图6-10 睿琪火车票账号修改的具体位置

腾讯的火车票发票通过注册账号，申请增值税发票识别的服务。通过secret\_id，secret\_id以及action的内容，就可以在程序中调用腾讯的api接口，在程序中修改腾讯对应的secret\_id，secret\_id以及action的内容，修改的文件位置为CompVATTicket/src/testFunc.py，修改具体内容的位置为tokenFunc.py下test\_TengXun函数中的对应内容，如图6-11所示。



图6-11 腾讯火车票账号修改的具体位置

## 注意事项

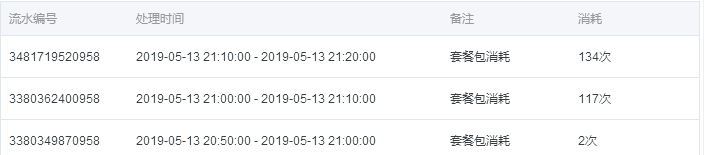
1. 本程序需要注意的主要是src中的代码文件，由于写的函数太多，此文档暂时不对大部分函数做详细说明，根据函数名基本能判断它的功能。以上表格中标绿的代码文件是简短而且不需要经常改动的，而标红的代码文件是比较冗长的，如果需要改动就要非常小心。
2. Transdata功能：不需要改动。除非我司提供的标注答案文件格式发生变化。
3. Callapi功能：前面已经提到了，调用API是需要花钱的，所以callapi选项只在真正需要进行竞品分析的时候才使用。而且callapi(testFunc.py内函数)非常简单，基本不需要改动，只需要在新增竞品公司时对testFunc.py、fakemain.py增加函数和个别语句即可。
4. Compare功能：最主要的函数是compFunc.py下的compFile函数，compFile是对一整个json文件（即一张图片的所有答案）进行比较。compFunc.py下的其他函数基本都为compFile服务，compFile中有一些比较繁杂的地方，已经做了注释。改动时要注意函数间的互动。
5. 建议生成的testfile文件以日期命名。
6. 建议在网络条件较稳定的环境调用，可以按上述三个步骤分开运行，有卡顿时考虑网速问题，保存好tested中显示的已经跑完的数据，下次把剩下没跑的图片和答案重新放进data和primitive answer再次测试。
7. 添加公司时需要对../csv/refnew\_all.csv进行更新，加入新公司的字段名称。
8. 添加新公司要分析返回的json文件的格式，应该将文件改成二级字典的形式保持与系统要求的一致；
9. 换电脑运行的时候，对于讯飞需要在后台把电脑的ip地址添加在项目的白名单中（自己申请webAPI格式的项目，开通增值税发票识别服务，修改src下的testFunc代码中的appid和fakemain里的api\_key，并修改后台的白名单）。白名单中需要加入IP地址，即python调用讯飞增值税发票识别demo，运行后得到的illegal access的IP，可能会得到多组ip，直到能顺利识别出来为止。
10. 睿琪，只能用电信或移动手机号注册。没有免费次数了，但是可以打电话给客服询问试图申请免费次数，包括发票识别和火车票识别。若是公司选择了购买，因为发票和火车票识别价格不同，故在统计调用记录表附件中按费用计算。
11. 阿里的火车票ocr识别服务与增值税发票识别服务不同，火车票识别服务没有购买，每个账号有500次机会。在相关账号密码文档里可以更换别的账号测试。如果出现HTTP报错的话需要考虑账号可使用服务量是否用完。
12. 每次调用后，请及时填写附件：各竞品调用次数记录表。

# 填写竞品调用记录表及其说明

1. 首先请仔细阅读上文中的产品说明文档和公司账号密码及其说明，以及附件中的备注部分，当天如果调用了多次API，按照当天累积的调用次数填写每个API的调用次数，不需要每次调用API后都填写该表格，由于网络故障或者延迟的问题，可能会出现实际的调用次数与调用次数可以填写生成的tested文件夹中文件个数。
2. 填表时请启用宏。
3. 官网同样可以查询调用时间、调用次数等需要的信息，操作步骤在附件中已经有所提及，这里加以配图说明。
4. 阿里

点击进入提供的网址，登陆，进入云市场页面-买家中心-进入管理控制台，点击右边第一列“已购买的服务”即可查询到剩余次数，再点击使用明细就可以查看到最近的调用记录。





1. 睿琪

点击进入提供的网址，登陆，进入控制台，可通过已购套餐和数据查询查看调用记录和剩余次数、剩余金额。目前睿琪剩余的识别次数为225次，包含大票和小票。请注意，六月份公司需购买睿琪的识别套餐，之后应通过剩余金额计算。



1. 易道博识

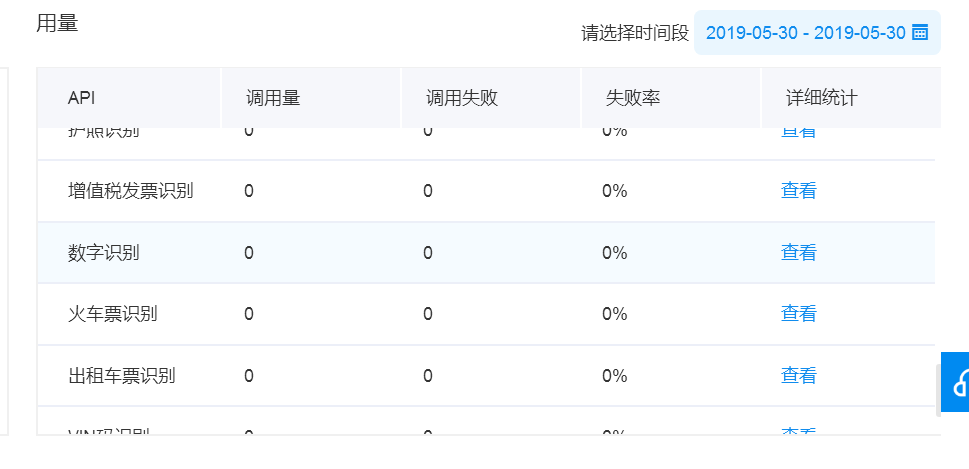
点击进入提供的网址，登陆，进入控制台，根据左边的提示选择。



1. 百度

点击进入提供的网址，登陆，进入控制台，下拉，点击文字识别-详细统计即可。





1. 科大讯飞

点击进入提供的网址，登陆，进入控制台，选择我的应用，即可看见今日实时服务量和剩余服务量。

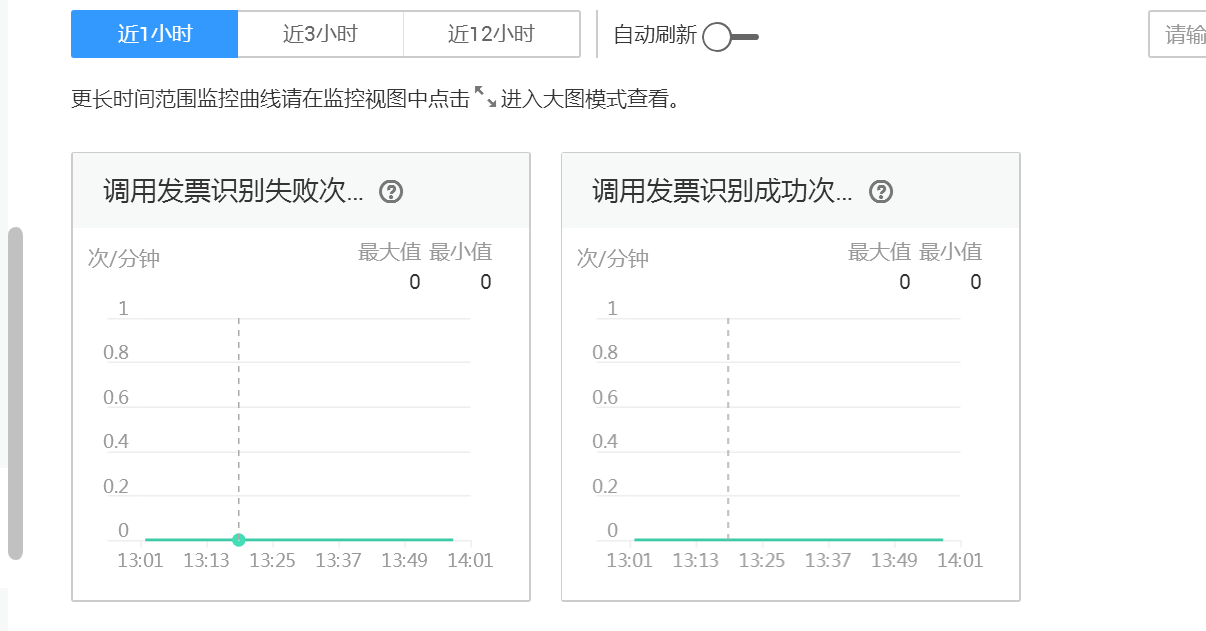


1. 华为

请注意，登陆华为云时可能火狐和谷歌浏览器不太兼容，可以使用IE浏览器或者360浏览器。

点击进入提供的网址，登陆，进入控制台，华为云服务搜索OCR文字识别，点击商用服务，最后点击监控数据即可。鼠标放到调用发票识别成功次数图上可以出现一个拉伸形状的小图标，点击即可查询自定义时间内的调用记录了。

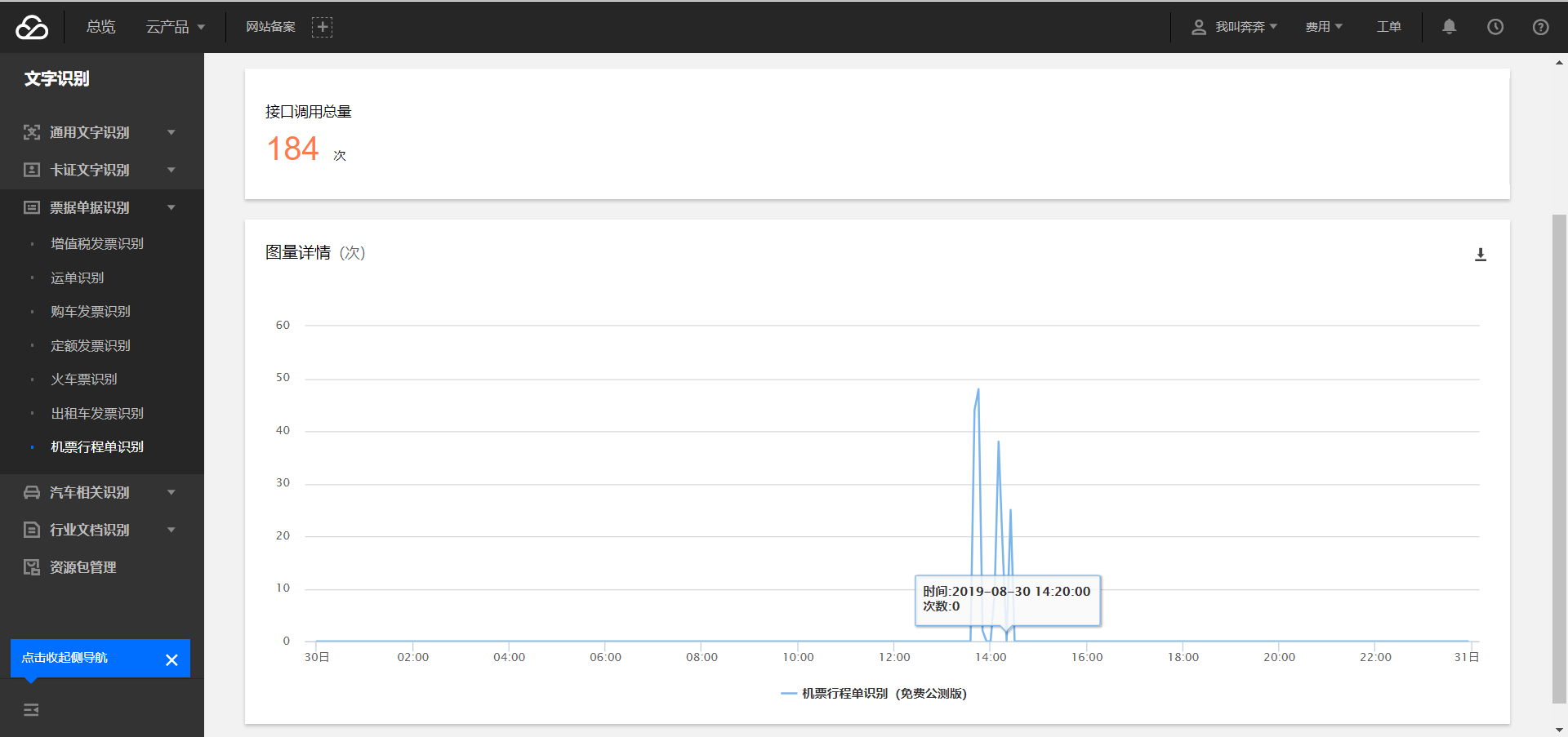






1. 腾讯

点击进入提供的网址，登陆，进入控制台，选择对应的票据种类即可查看当日或其他时间使用次数。



1. 填写记录表注意事项
2. 当有公式在表格中时，公式对应的表格的下一行表格建议以下拉形式获取，这样可以将公式继续延续下去。
3. 百度、阿里、华为、腾讯和睿琪的表格填写中需区分票种，即增值税发票（大票）或者火车票（小票）。
4. 科大讯飞当某天调用次数超过500次时会报错，也不会产生调用费用，所以在调用次数记录表中将科大讯飞的分表隐藏，使用时不需要填写科大讯飞的分表。
5. 启用宏之后，点击日期的空格都会自动出现今日日期，修改表格的时候也应注意这点。

# 竞品比较分析表填写说明

## 数据填写

使用不同的数据集包括：100张拍摄发票，200张扫描发票以及250张火车票，按照[竞品分析程序文档](../../竞品分析程序文档2019-05-16.docx)对不同的竞品api进行测试，会得到后缀为analyse和ziduan的两个csv文件，在生成竞品比较分析表的过程中使用的是以后缀为ziduan的csv文件。下面以5月份的填写过程为例，来说明如何根据csv文件生成竞品比较分析表的过程。

## 数据转移

竞品比较分析表填写以[1月份的竞品比较分析表](../1月15日竞品比较分析表.xlsx)为模板，它包括4个表：对比测试报告，200张扫描件比较，100件拍摄样本比较以及250张样本比较，其中表200张扫描件比较，100张拍摄样本比较以及250张样本比较的数据可以直接通过从ziduan的csv文件中获取。以100张拍摄样本填写为例，其他两张表的填写与100张拍摄样本的填写类似，由于睿琪的特殊性，睿琪的竞品比较分析表是单独做的，做的过程与其他的表的做法类似，其模板可以参考[5月份竞品比较分析表睿琪](5月15日竞品比较分析表睿琪-更新版.xlsx)。

原始ziduan表的数据如表1-1所示，100张拍摄样本的部分数据如表1-2所示，需要把ziduan表的中识别每个公司识别正确的字段数(nb\_good\_ali...)以及总共参与字段数(nb\_all\_tested)填充到100张拍摄样本表中各个公司正确的识别数列中以及产品的样本识别数，两个表的字段对应关系为如表1-3所示，为了快速填充，需要使用excel的vlookup函数，如果不会使用vlookup函数，可以参考链接如下的链接进行学习:

<http://baijiahao.baidu.com/s?id=1603886666150544094&wfr=spider&for=pc>。

表1-1 100张拍摄样本的原始ziduan表

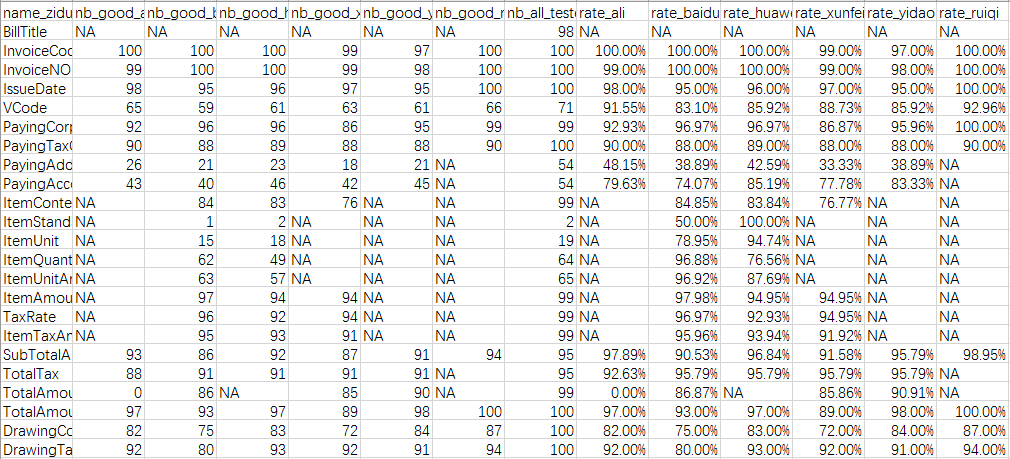


表1-2 100张拍摄样本的分析表中部分数据



表1-3 字段对应表

|  |  |
| --- | --- |
| **原始ziduan表字段** | **100张拍摄样本分析表字段** |
| nb\_good\_baidu | 百度正确识别数 |
| nb\_good\_ali | 阿里正确识别数 |
| nb\_good\_huawei | 华为正确识别数 |
| nb\_good\_xunfei | 讯飞正确识别数 |
| nb\_good\_yidao | 易道博识正确识别数 |
| nb\_all\_tested | 产品样本识别数 |

同样的，对于200张扫描件比较，250张样本比较的数据的填充和100张拍摄件比较表的填充操作类似，在填充的过程中，需要主要的事项有：

a. 在填充之前，需要把各个产品对应的正确识别数进行清空，共有字段行以及各个公司正确识别率列不需要清空,因为里面包含的是与正确识别数相关的公式；

b. 唯你网的填充和其他产品的数据的填充过程一致；

c. 睿琪数据和其他产品进行了区分，其分析表在竞品比较分析表-睿琪中，表的结构和其他竞品比较分析表的结构没有区别，填充的过程类似。

## 数据修正

以上数据填充的操作填充了表200张扫描件比较，250张样本比较的数据的填充和100张拍摄件比较表的数据，对于表对比测试报告，需要手工对数据进行更新，主要的操作有三项：

a. 最新各个竞品的测试正确率数据可以从后面三张表中直接获取；

b. 上月竞品正确率的数据可以直接复制上月统计的竞品正确率数据；

c. 同时确认表中共识别字段的数量，如图1-1红色圈中的内容；

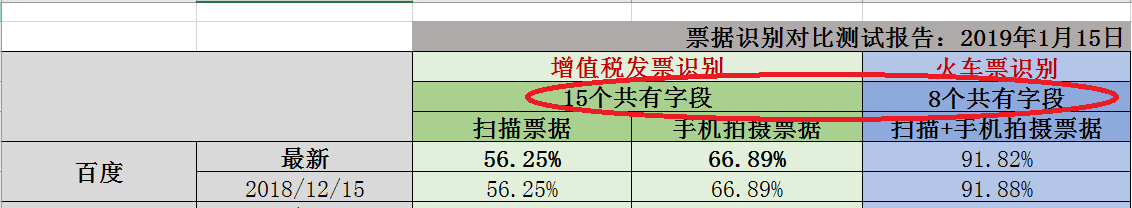


图1-1 测试对比报告的部分截图

d. 注意修改模板的日期时间，修改为目前测试的时间。

在对数据进行补充完整之后，需要对表的格式进行规范，进行说明了表格式的规范的注意事项。

## 格式规范

### （1）对比测试报告的格式规范

在对比测试报告中填充好数据之后，需要对其格式进行规范，格式规范需要注意以下几个方面：

a. 对比测试报告表中，唯你网的数据都需要底色标黄，当月标注为黄色，上月标注为浅粉色，如图2-1所示；



图4-1 对比测试报告中唯你网数据颜色标注图

b. 对比测试报告表中，当月各个竞品包括唯你网的测试正确率数据需要加粗；

c. 对比测试报告中，当月每一列的正确率最高数据需要标红色，上月每一类的正确率最高的数据也需要标注红色；

### （2）各个分表的格式规范

同样，在对各个分表的比较表填充好数据之后，需要对其格式进行规范，格式规范需要注意以下几个方面：

a. 分表中需要对共有字段合计行进行加粗，其他的行不需要进行加粗；

b. 表格需要加内边框和外边框的；

c. 每个分表中生成的柱状图，需要将正确率最高的柱子标注为红色，标注红色的过程可以通过选择右键→填充→选择红色，完成标注过程

1. 注：在这条语句中，testfile即是“本轮分析文件夹”。由于调用其他公司API是要钱的，所以请不要随意使用callapi选项。根据公司目前的要求，每月对同一批图片做一次分析测试即可。 [↑](#footnote-ref-1)