# 源码部署

## 系统要求

**系统**

Ubuntu x86\_64 系统 >=18.04

python >= 3.8

**安装 python 依赖**

pip3 install -r requirements.txt

**安装torch**

#CPU版本

pip3 install torch --index-url https://download.pytorch.org/whl/cpu

或

#GPU版本

pip3 install torch --index-url https://download.pytorch.org/whl/cu118

**⽀持 CPU 运⾏， 如需使⽤ GPU 运⾏，需硬件配备英伟达显卡**

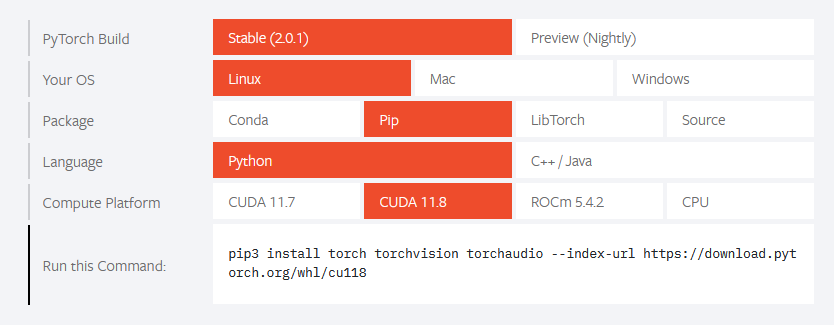
nvidia显卡驱动 > 525.89.02

cuda版本 > 11.8

torch==2.0.0+cu118

或根据具体硬件配置参照torch官方网站建议版本安装

<https://pytorch.org/>



其他硬件版本

<https://pytorch.org/get-started/previous-versions/>

**配置工程是否使用GPU**

打开工程文件中的config/config.py

修改

CONF.GPU= True

或在服务启动时传入device参数，具体见启动服务说明。

## 模型文件

由于依赖的大模型文件较大，建议从官方网站直接下载

### 1 text2vec

安装huggingface下载器

git lfs install

下载文件

git clone <https://huggingface.co/GanymedeNil/text2vec-large-chinese>

或

直接打开页面手动下载

<https://huggingface.co/GanymedeNil/text2vec-large-chinese>

### 2 Baichuan2-13B-Chat-4bits

安装huggingface下载器

git lfs install

下载文件

git clone https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat-4bits

或直接打开页面手动下载

https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat-4bits

### 3 配置模型文件

打开工程文件中的config/config.py

修改

CONF.embeding\_model\_path = [text2vec的文件夹地址]

CONF.llm\_model\_path= [Baichuan2-13B-Chat-4bits的文件夹地址]

## 启动服务

进入工程文件夹

**GPU/CPU模式**

通过修改config/config.py文件中的CONF.GPU=True/False 来设置GPU和CPU模式

如果启动时指定device参数，则以device参数传递值为准

运行以下命令启动服务

python3 label\_service.py [--port=] [--device=]

该服务监听本地[port]端口,接收客户端上传的文件并返回分析后的表格格式数据。默认在8000端口监听。如果未指定device参数，以CONF.GPU的配置为准。

启动参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **是否必须** | **类型** | **说明** |
| port | 否 | int | 默认在8000端口监听，可根据需要修改端口 |
| device | 否 | string | ‘GPU’/‘CPU’ 选择模型在GPU还是CPU加载  如果不设置该参数，默认使用config文件中的配置项 |

## 调用接口

**客户端脚本**

识别文件中的表格格式：该接口上传[file\_path]指定的本地文件到[ip]:[port]的识别服务，并返回识别后的表格格式数据，默认访问本地8000端口。

进入工程文件夹执行

python3 label\_client.py label\_generation --data=[json格式数据] [--ip=127.0.0.1] [--port=8000]

接口名**label\_generation**

调用参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **必须** | **类型** | **说明** |
| data | 是 | string | Json格式的数据。（格式见后续说明） |
| ip | 否 | string | 服务所在的ip地址，默认为本机地址127.0.0.1 |
| port | 否 | int | 默认访问8000端口，可根据需要修改 |

返回json文本

**HTTP协议**

**POST**

**接口 <http://[url:port]>/form\_recognition/**

参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **必须** | **类型** | **说明** |
| data | 是 | string (base64) | Json格式的数据。（格式见后续说明） |

返回json文本

**输入data数据格式**

Data数据样例

{

"res1": {

"outside\_infos":

[{"txt": "账户名称： 长沙玉兔信息科技有限公司"},

{"txt": "账号： 155092704"},

{"txt": "开户机构： 中国民生银行股份有限公司长沙分行营业部"}, {"txt": "币种： 人民币元"},

{"txt": "起始日期： 2023-01-01"},

{"txt": "截至日期： 2023-04-09"},

{"txt": "启用高级查询： 是"},

{"txt": "借贷标识： 全部"},

{"txt": "对方账号："},

{"txt": "按发生额查："},

{"txt": "借方累计笔数: 6 借方累计发生额: 123800.00"}, {"txt": "贷方累计笔数: 25 贷方累计发生额: 68169.83"}

]},

"res2": [

{"row\_order": "001\_1",

"header\_row": true,

"1.1": "交易时间",

"1.2": "主机交易流水号",

"1.3": "借方发生额",

"1.4": "贷方发生额",

"1.5": "账户余额",

"1.6": "凭证号",

"1.7": "摘要",

"1.8": "对方账号",

"1.9": "对方账号名称",

"1.10": "对方开户行"},

{"row\_order": "001\_2",

"header\_row": false,

"2.1": "2023-01-05 11:38:56",

"2.2": "50224202301051138563827544649A8B",

"2.3": "0.00",

"2.4": "2237.14",

"2.5": "119291.14",

"2.6": "",

"2.7": "银联入账-0104-0104费7.86元",

"2.8": "991290000023",

"2.9": "银联商务股份有限公司",

"2.10": "中国银联股份有限公司"}],

"doc\_tpye": "excel",

"page\_sum": 1}

data 参数说明

|  |  |  |
| --- | --- | --- |
| **名称** | **字段命名** | **内容** |
| 表外数据 | res1 | 第一页的表外数据 |
| 页数 | page\_sum | 总页数  pdf 为pdf数据的实际页数  excel 为以表格分割的页数(每表格一页) |
| 表内数据 | res2 | 1）表内数据，包含每个单元格的行列信息和文字内容，空值则为空；  2）row\_order：行顺序，如 001\_1，代表第一页第一行；  header\_row：true/false，代表该列数据是否属于表头行； |
| 文件类型 | doc\_type | 代表处理的文件类型，分为 excel、epdf，epdf 代表是电子化的 pdf 文件； |

**输出数据格式**

正确格式样例

{

"success":true,

"error-code":400,

"message":"",

"data":{

"res1":{

"agent\_type":"alipay",

"account\_num":"haishangren@foxmail.com",

"account\_name":"王世杰",

"idcard\_num":"",

"account\_type":"对私",

"begin\_date":"2022/05/26",

"end\_date":"2022/11/22"

},

"res2":[

{

"row\_order":"001\_1",

"1.1":"流水号",

"1.2":"时间",

"1.3":"名称/备注",

"1.4":"收入",

"1.5":"支出",

"1.6":"账户余额",

"1.7":"资金渠道"

},

{

"row\_order":"001\_2",

"2.1":"202211180 191301250 000900139 98382",

"2.2":"2022-11- 18 23:57:13",

"2.3":"购买理财 产品转出 （余额宝 、淘宝基 金、招财 宝等）",

"2.4":"",

"2.5":"-500.00",

"2.6":"0.00",

"2.7":"支付宝",

"account\_label":{

"accounting\_item":"Z0201理财资产",

"accouting\_entry":"增加",

"N-accounting\_item":"Z0701账户余额",

"N-accouting\_entry":"减少"

},

"row\_label":{

"header\_row":false,

"trader\_nature":"",

"in\_spend\_type":"支出"

}

}

],

"doc\_tpye":"epdf",

"page\_sum":18,

"res3":[

{

"row\_order":"001\_1",

"header\_row":true,

"1.1":"收/支",

"1.2":"交易对方",

"1.3":"商品说明",

"1.4":"收/付款方式",

"1.5":"金额",

"1.6":"交易日期"

},

{

"row\_order":"001\_2",

"header\_row":false,

"2.1":"支出",

"2.2":"",

"2.3":"购买理财产品转出（余额宝、淘宝基金、招财宝等）",

"2.4":"支付宝",

"2.5":500,

"2.6":"2022/11/18"

}

]

}

}

错误返回：

{

"success":false,

"code":-1,

"message":"unknow"

}

**输出格式说明**

**输出格式**

**1、一级参数**

|  |  |
| --- | --- |
| **参数名** | **说明** |
| success | true：代表执行成功  false：代表执行失败 |
| error-code | 返回错误码 |
| message | 报错信息 |
| data | 识别数据 |

**error-code说明**

|  |  |  |
| --- | --- | --- |
| **代码类型** | **信息** | **说明** |
| 401 | Type error | 文件格式不符合要求 |
| -1 | unknow | 未知错误 |

**2、data二级参数说明**

|  |  |  |
| --- | --- | --- |
| **参数名** | **说明** | **内容** |
| page\_sum | 总共页数 | 001 |
| doc\_type | 数据来源的文件格式，保存即可； | Image：图片  excel：电子表格  epdf：文本PDF文件 |
| res1 | 表外信息的数据提取 |  |
| res2 | 表内每行数据的以及科目标签 |  |
| res3 | 表内数据的指定字段信息提取 |  |

**3、三级参数：res1参数**

|  |  |  |
| --- | --- | --- |
| **字段** | **说明** | **输出内容** |
| agent\_type | 机构类型 | **bank：银行**  alipay：支付宝  wechat：微信支付 |
| bank\_name | 参考“金融机构信息对照表” | GY0001 |
| accout\_num | 银行账号号码 | 6228480329572520271 |
| account\_name | 账号的户名 | 李数 |
| date\_start | 起始日期 | 2022/05/22 |
| date\_end | 结束日期 | 2022/11/22 |
| account\_type | 账户类型 | 对私 |

|  |  |  |
| --- | --- | --- |
| **字段** | **说明** | **输出内容** |
| agent\_type | 机构类型 | bank：银行  alipay：支付宝  wechat：微信支付 |
| accout\_num | 支付宝账号号码 | caiyingbiao@hotmail.com |
| account\_name | 支付宝账号姓名 | 李数 |
| idcard\_num | 用户身份证号码 | 330724197805240011 |
| date\_start | 起始日期 | 2022/05/22 |
| date\_end | 结束日期 | 2022/11/22 |
| account\_type | 账户类型 | 对私 |

|  |  |  |
| --- | --- | --- |
| **字段** | **说明** | **输出内容** |
| agent\_type | 机构类型 | bank：银行  alipay：支付宝  wechat：微信支付 |
| accout\_num | 微信账号号码 | xshaitt |
| account\_name | 微信账号姓名 | 李数 |
| idcard\_num | 身份证号码 | 330724197805240011 |
| date\_start | 起始日期 | 2022/05/22 |
| date\_end | 结束日期 | 2022/11/22 |
| account\_type | 账户类型 | 对私 |

**4、res2参数**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名** | **说明** | **内容** |
| 1 | row\_order | 当前行的位置，页码+行序 | 003\_8 |
| 2 | 1.1  1.2  2.1 | 当前单元格的位置，行序+列序 | 1.1  1.2  2.1 |
| 3 | account\_label | 科目标签 | 详情查看如下 |
| 4 | row\_label | 行标签 | 详情查看如下 |
|  |  |  |  |

**1）row\_label的详细参数**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名** | **说明** | **内容** |
| 1 | header\_row | 是否是表头行，true代表是，false代表不是 | true  false |
| 2 | trader\_nature | 交易对象的行业分类 | 快餐服务\眼镜制造 |
| **3** | **in\_spend\_type** | **收支类型** | **收入**  **支出** |

**2）account\_label的详细参数**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名** | **说明** | **内容** |
| 1 | N-accounting\_item | 主科目标签 |  |
| 2 | N-accounting\_entry | 主分录方向 |  |
| 3 | accounting\_item | 副科目标签 | 储蓄类资产 |
| 4 | accounting\_entry | 副分录方向 | 增加 |
|  |  |  |  |

**5、res3参数**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名** | **说明** | **内容** |
| 1 | row\_order | 当前行的位置，页码+行序 | 003\_8 |
| 2 | 1.1  1.2  2.1 | 当前单元格的位置，行序+列序 | 1.1  1.2  2.1 |
| 3 | header\_row | 是否是表头行，true代表是，false代表不是 | true  false |
|  |  |  |  |