# 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**

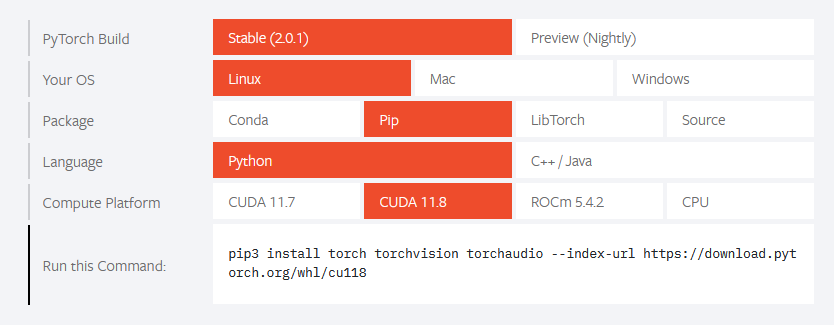
nvidia显卡驱动 > 525.89.02

cuda版本 > 12.0

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

## 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 chatglm2-6b-int4**

安装huggingface下载器

git lfs install

下载文件

git clone <https://huggingface.co/THUDM/chatglm2-6b-int4>

或直接打开页面手动下载

<https://huggingface.co/THUDM/chatglm2-6b-int4>

### 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= [chatglm2-6b-int4的文件夹地址]

## 启动服务

python3 service.py

## 调用接口

**res1 ：**

python3 client.py res1 --file\_path=data/input/支付宝1.pdf