更新2020/10/16：

1. 将原来spider\_predict文件夹下所有.py文件删除
2. 将utils.py和web\_bert\_classify.py文件放入spider\_predict下
3. 将utils.py文件第232行的chormedriver.exe路径改成自己的路径，比较重要！！

我这里的路径是：F:\网站识别\spider\_predict。



1. 生成的预测结果文件里有两个sheet，第一个sheet是预测结果，第二个sheet是爬虫内容为空和爬虫失败的url。

运行代码方法如下：

**注意事项：**

**一、文件夹的含义及存放内容：**

1. fine\_tuned中的文件是用来对模型进行微调和训练的。

2. spider\_predict中的文件是用来爬虫且预测网站类型的。

3.transformers中存放的是模型的依赖环境，如果是需要自己动手训练的情况下，不需要这个transformers这个文件夹，如果只想预测的请看下面的使用方法

**二、使用方法**

（一）、想要自己同时训练模型和预测的请进入fine\_tuned文件夹，里边有训练的说明文档。

（二）、只想输入url,然后得到分类结果的请看下边的使用步骤。

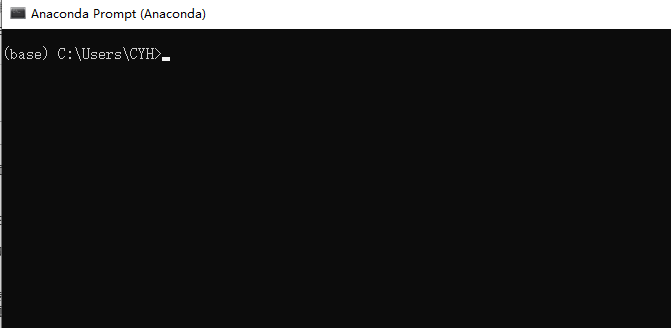
**步骤：*在预测之前先将需要爬虫的host放入spider\_predict文件夹下的url.txt文本文件里面。***

1.将transformers文件夹，复制放在自己电脑里的Anaconda3\envs路径下边

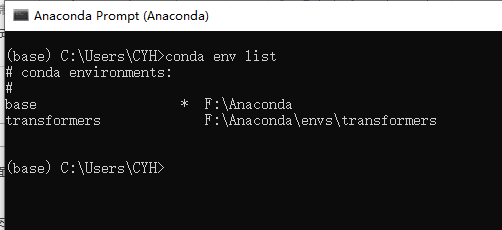
比如我的路径就在：



2. 从开始目录里Anaconda中打开Anaconda Prompt，如下图：

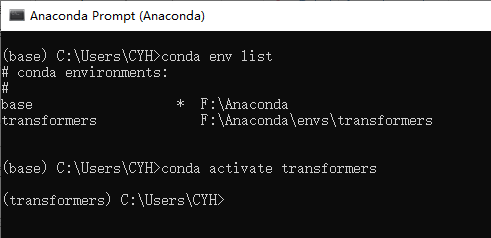


然后输入conda env list，查看模型的依赖环境是否存在，得到如下的结果：

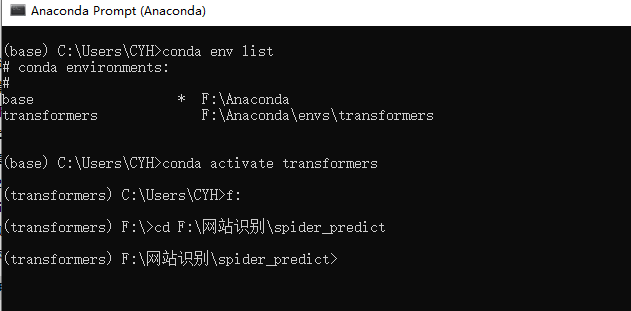


上图出现transformers，就说明模型的依赖环境安装成功了。

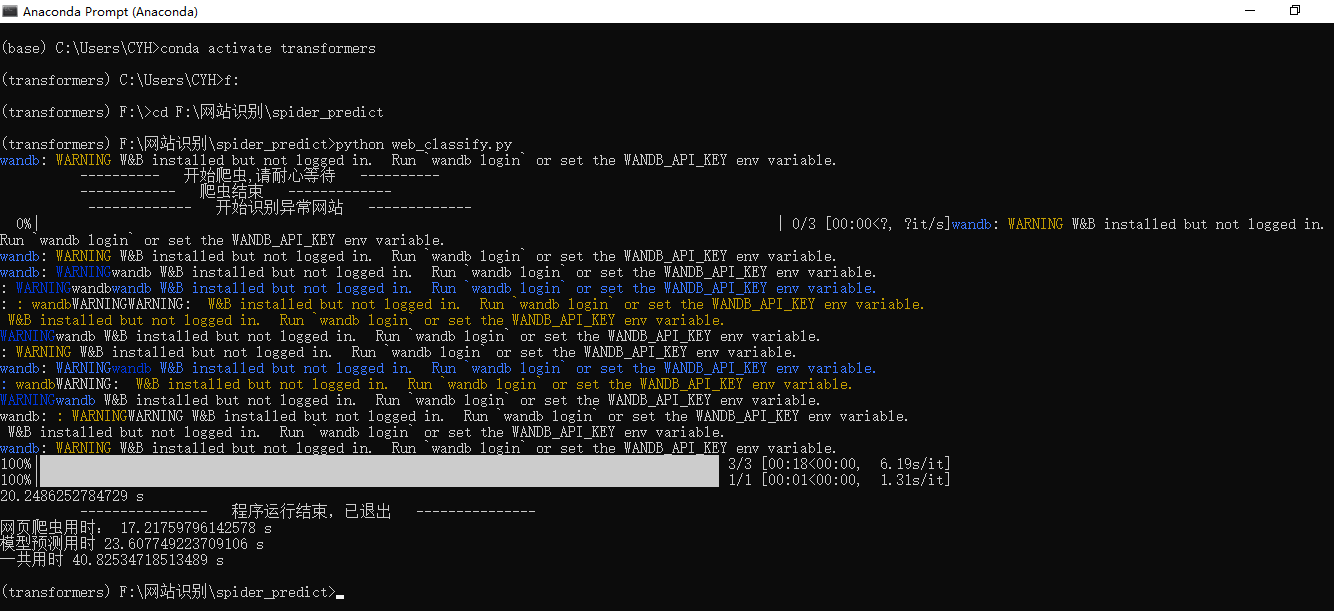
3. 还在上面图所示的终端里输入，conda activate transformers，这一步是依赖环境，如下图：



4. 继续在终端里输入如下命令，进入到spider\_predict文件夹下：



5. 在上图中输入，python web\_bert\_classify.py运行模型的预测代码，得到如下的结果：



6. 程序运行结束后会在spider\_predict文件夹生成2个文件，一个是爬虫的文本.csv文件，还有一个是网站识别结果.xlsx文件，如下图最下面：



注意：如果重新运行那个代码，首先需要将这两个文件（.csv和.xlsx）删除。

**三、调参注意事项：**

在BERT\_Train.py和web\_classify.py文件里：

model.train和model.predict中的参数model\_name可以是模型的预训练权重或者训练好的权重的路径，请根据自己的实际情况选择。

我在web\_classify.py代码里边放的是已经训练好的模型权重，因此model\_name的参数是同级文件夹下面的model\_weight，**你也可以替换为自己训练的权重的路径，可以是相对路径也可以是绝对路径**。

参考文档：

1.https://simpletransformers.ai/docs/classification-models/

2.https://github.com/ThilinaRajapakse/simpletransformers

有关调参的问题都可以再上面的官方文档里找到