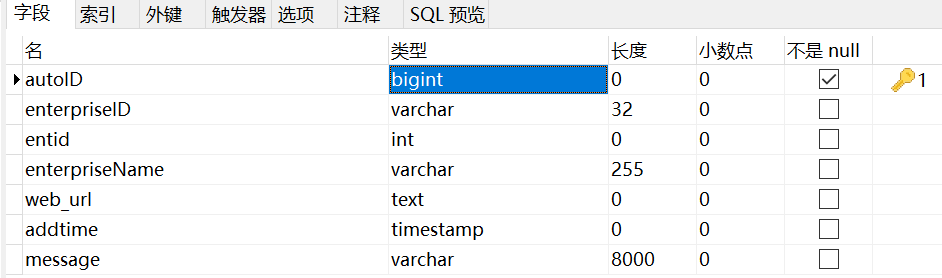
题目： *给定10万个企业的网址（放在数据库表中），请编写一个网络爬虫，  
 爬取该企业的产品及服务信息（即获取该企业能提供的产品和服务）；  
 并将爬取到的数据，保存到数据库中；请自行设计数据库表结构*

思路：数据库取出网址-->爬取网页-->解析数据-->保存数据到数据库

主要用到知识：网络爬取，数据库使用，多线程

数据设计：



具体思路过程：

1. 爬取网页：使用requests爬取网页，出错返回404
2. 解析数据：使用BeautifulSoup，这里遇到的第一个问题就是关于企业产品、服务的内容位置问题，有些许网址的产品信息直接以文字或图片形式放在首页，也有多数将产品服务信息放在连接中，所以选择大多数，先爬取产品介绍或服务范围的网址，没有这返回“暂未找到”

接着跳转网页分析文本，然后这里发现有问题，每个产品或服务的的信息在HTML中的位置，对应的class、tittle、tag都不同，甚至还有的是图片形式，无法继续使用BeautifulSoup继续查找、提取所有数据，所有选择将全部文本信息提取出来，过滤掉无用信息

3.保存数据：使用进程池将数据保存到数据中