操

作

流

程

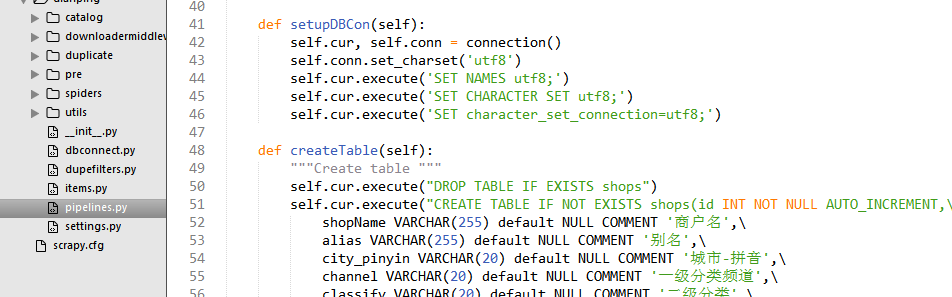
|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.1 |  | 基本实现对点评网数据抓取及去重 | 2017.3.1 |

使用介绍

1.修改连接数据库文件dianping \dbconnect.py，连接到指定数据库



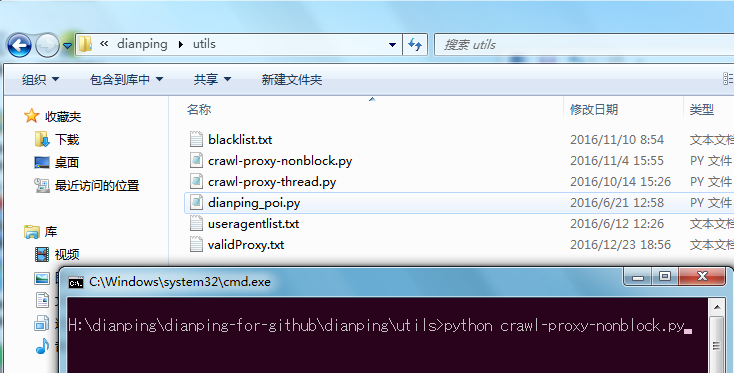
2. 修改项目pipelines.py文件第50，51行和第96行的数据库tables名字，以防爬虫运行过程中将已有数据表冲掉。



3. 如果只想爬某一个频道，那么在spider的第65行选定crawl\_just\_one\_channel=True，并下方相应的if语句解除comment，如果全频道爬取则该步骤忽略。



4.命令行中启动代理proxy爬取进程dianping\utils\crawl-proxy-nonblock.py



5.重开一个cmd正式启动爬虫进程，并同时加上日志scrapy crawl dianping –s LOG\_FILE=xxx.txt

