* **enter\_urls:本表是根据Excel导出的各学院列表、详情页url模板信息**

|  |  |  |
| --- | --- | --- |
| Id | int(11) | 主键 |
| University | varchar(255) | 学校 |
| school | varchar(255) | 学院 |
| first\_page | varchar(255) | 学院列表页第一页（时间最新） |
| second\_page | varchar(255) | 学院列表页第二页（时间最新） |
| detail\_model | varchar(255) | 详情页模板 |
| signal\_rank | int(1) | 列表页地址递减标记 0递增 1递减 |
| signal\_var | varchar(255) | 列表页中间变化的字段名 |
| signal\_max | int(3) | 当前规模 |
| have\_photos | boolean | 有无照片 |
| create\_time | timestamp | 插入记录时系统当前时间 |
| create\_person | varchar(20) | 插入用户名 |

* **empty\_urls：爬取enter\_urls时判断为空的列表urls的记录**

|  |  |  |
| --- | --- | --- |
| id | int(11) | 主键与enter\_urls一致 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| first\_page | varchar(255) |  |
| second\_page | varchar(255) |  |
| detail\_model | varchar(255) |  |
| create\_time | timestamp |  |
| create\_person | varchar(20) |  |

* **dynamic\_urls：爬取enter\_urls时判断为动态加载的列表urls的记录**

|  |  |  |
| --- | --- | --- |
| id | int(11) | 主键与enter\_urls一致 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| first\_page | varchar(255) |  |
| second\_page | varchar(255) |  |
| detail\_model | varchar(255) |  |
| create\_time | timestamp |  |
| create\_person | varchar(20) |  |

* **refused\_enter\_urls：爬取enter\_urls时被拒绝访问的列表urls的记录**

|  |  |  |
| --- | --- | --- |
| id | int(11) | 主键与enter\_urls一致 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| first\_page | varchar(255) |  |
| second\_page | varchar(255) |  |
| detail\_model | varchar(255) |  |
| create\_time | timestamp |  |
| create\_person | varchar(10) |  |
| solve\_signal | int(1) | 0为未处理1为处理成功-1为处理失败 |
| solution | varchar(20) | 处理方法 |

* **update\_urls：每次爬取后的记录信息，记录当前爬取的最新详情页链接**

|  |  |  |
| --- | --- | --- |
| id | int(11) | 主键，设为自动递增 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| latest\_detail\_page | varchar(255) | 最新的详情页链接 |
| update\_time | timestamp |  |
| update\_person | varchar(10) |  |

* **error\_info：存储爬取详情页文章内容时报错的文章链接**

|  |  |  |
| --- | --- | --- |
| id\_detail | int(11) | 主键，自动递增 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| academic\_url | varchar(255) | 出错的详情页链接 |
| create\_person | varchar(10) | 产生者 |
| create\_time | timestamp | 产生时间 |
| solution | varchar(20) | 解决方法 |
| solve\_signal | int(1) | 0为未处理1为处理成功-1为处理失败 |

* **list\_urls：存储爬虫时程序循环生成的列表页链接**

|  |  |  |
| --- | --- | --- |
| id | int(255) | 主键，自动递增 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| list\_url | varchar(255) |  |
| detail\_model | varchar(255) |  |
| create\_time | timestamp |  |
| create\_person | varchar(20) |  |

* **academic\_info：存储爬取的详情页链接和详情页的正文**

|  |  |  |
| --- | --- | --- |
| id | int(255) | 主键，自动递增 |
| university | varchar(255) |  |
| school | varchar(255) |  |
| academic\_url | varchar(255) | 详情页链接 |
| detail\_text | text | 获取的正文 |
| create\_time | timestamp |  |
| create\_person | varchar(20) |  |