## Scrapy改进部分1测试需求分析

### 功能需求测试

##### 爬虫有效性测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **测试用例名称** | 用开发好的爬虫爬取北航计算机学院官网URL信息 | | **测试用例标识** | 用例TC01 | **测试需求标识** | TR01 |
| **简要描述** | 本测试验证经过修改以后的Scrapy是否能够正确执行开发好的爬虫功能。 | | | | | |
| **前提和约束** | 1. 针对目标站点的爬虫开发完毕。 2. 布隆过滤器有效性测试完成。 3. 布隆过滤器有效性测试成功（有效）。 | | | | | |
| **测试方法** | 黑盒测试 | | | | | |
| **测试过程描述** | | | | | | |
| **序号** | **测试步骤** | | | **测试结果** | | **评价准则** |
| **预期结果** | |  |
| 1 | 从命令行启动爬虫 | | | 启动成功 | | 实际结果与预期结果一致 |
| 2 | 观察爬取过程输出的日志信息 | | | 未观察到ERROR类型的日志信息 | | 实际结果与预期结果一致 |
| **备注** | 说明：  1）经过修改以后的Scrapy-----经过改进后的Scrapy包，区别于直接安装的Scrapy包。  2）正确执行开发好的爬虫功能-----正确执行爬虫parse函数以及itemline中定义的函数。  3）命令行启动时必须进入到爬虫所在的目录。  4）用法（格式）：scrapy crawl <爬虫名称>  例如：  命令名：scrapy crawl buaa\_scse\_spider  参数：其中buaa\_scse\_spider就是在parse函数中指定的爬虫名称。 | | | | | |
| **测试结论** |  | | | | | |
| **测试人员：王益飞** | | **测试时间：**2017-05-17 | | | | |

##### 布隆过滤器正确性测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **测试用例名称** | 用布隆过滤器过滤重复项 | | **测试用例标识** | 用例TC02 | **测试需求标识** | TR02 |
| **简要描述** | 本测试验证布隆过滤器是否可以有效过滤已有项。 | | | | | |
| **前提和约束** | 无 | | | | | |
| **测试方法** | 黑盒测试 | | | | | |
| **测试过程描述** | | | | | | |
| **序号** | **测试步骤** | | | **测试结果** | | **评价准则** |
| **预期结果** | |  |
| 1 | 初始化布隆过滤器 | | | 初始化成功 | | 实际结果与预期结果一致 |
| 2 | 向布隆过滤器中添加给定数据项 | | | 添加成功 | | 实际结果与预期结果一致 |
| 3 | 检验给定数据项、其他数据项是否在布隆过滤器中记录 | | | 给定数据项返回True，未给定测试项返回False | | 实际结果与预期结果一致 |
| **备注** | 说明：  1）已有项-----已经添加入布隆过滤器的数据项。  2）初始化布隆过滤器-----使用Python编写程序初始化。  例如：filters=Filters(10000,0.001)  3）向布隆过滤器中添加给定数据项-----使用Python程序添加。  例如：filters.add(“item”)  4) 检验给定项是否在布隆过滤器中记录了-----使用Python程序检验。  例如：bool(“item” in filters)  5) 给定数据项-----之前添加到给定布隆过滤器中的数据项。  6）其他数据项-----之前未添加到给定布隆过滤器中的数据项。 | | | | | |
| **测试结论** |  | | | | | |
| **测试人员：王益飞** | | **测试时间：**2017-05-17 | | | | |

##### 分布式爬虫协作爬取的有效性测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **测试用例名称** | 多机爬虫爬取北航计算机学院官网URL信息 | | **测试用例标识** | 用例TC03 | **测试需求标识** | TR03 |
| **简要描述** | 本测试验证多机爬虫是否能够使用协作爬取目标站点。 | | | | | |
| **前提和约束** | 1）针对北航计算机学院官网URL信息的爬虫开发完成。  2）爬虫有效性测试完成。  3）爬虫有效性测试结果有效。  4）分布式爬虫协作爬取正确性测试完成。  5）分布式爬虫协作爬取正确性测试结果正确。 | | | | | |
| **测试方法** | 黑盒测试 | | | | | |
| **测试过程描述** | | | | | | |
| **序号** | **测试步骤** | | | **测试结果** | | **评价准则** |
| **预期结果** | |  |
| 1 | 在三台电脑上搭建mongodb分布式数据库 | | | 分布式mongodb数据库搭建成 | | 实际结果与预期结果一致 |
| 2 | 在三台机器上安装配置scrapy运行环境 | | | Scrapy运行环境配置成功 | | 实际结果与预期结果一致 |
| 3 | 将开发好的爬虫拷贝到三台机器上 | | | 爬虫正确的拷贝到三台机器上 | | 实际结果与预期结果一致 |
| 4 | 分别启动三台机器的爬虫 | | | 爬虫启动成功 | | 实际结果与预期结果一致 |
| 5 | 观察三台机器上爬虫的输出日志 | | | 未观察到ERROR类型的日志 | | 实际结果与预期结果一致 |
| **备注** | 说明：  1）mongodb分布式数据库-----在三台机器上搭建分布式mongodb数据库，设置三个replSet，三个config 服务器，两个router服务器。  2）安装配置scrapy环境-----按照实验三的改进部分1的使用说明.md文档说明配置修改以后的scrapy环境。  3）每台机器启动爬虫的方式与单机方式相同。 | | | | | |
| **测试结论** |  | | | | | |
| **测试人员：王益飞** | | **测试时间：**2017-05-17 | | | | |

##### 分布式爬虫协作爬取的正确性测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **测试用例名称** | 用布隆过滤器和分布式mongodb文档过滤重复网页URL | | **测试用例标识** | 用例TC04 | **测试需求标识** | TR04 |
| **简要描述** | 本测试验证两级过滤器是否能够有效验证重复网页。 | | | | | |
| **前提和约束** | 1. 测试爬虫开发完成。 2. 布隆过滤器正确性测试完成。 3. 布隆过滤器正确性测试结果正确。 | | | | | |
| **测试方法** | 黑盒测试 | | | | | |
| **测试过程描述** | | | | | | |
| **序号** | **测试步骤** | | | **测试结果** | | **评价准则** |
| **预期结果** | |  |
| 1 | 在命令行启动单机测试爬虫 | | | 启动成功 | | 实际结果与预期结果一致 |
| 2 | 观察爬虫输出日志中的关于过滤的部分。 | | | 观察到重复URL被过滤的日志信息 | | 实际结果与预期结果一致 |
| **备注** | 说明：  1）单机测试爬虫-----人为设置的生成重复Request的爬虫。  2）命令行启动爬虫的方式与爬虫有效性测试中的方式相同。 | | | | | |
| **测试结论** |  | | | | | |
| **测试人员：王益飞** | | **测试时间：**2017-05-17 | | | | |

##### 1.5 网站链接关系设置有效性测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **测试用例名称** | 设置网站链接关系 | | **测试用例标识** | 用例TC05 | **测试需求标识** | TR05 |
| **简要描述** | 本测试验证用户设置的链接关系是否正确存入数据库。 | | | | | |
| **前提和约束** | 1）测试爬虫开发完毕。 | | | | | |
| **测试方法** | 黑盒测试 | | | | | |
| **测试过程描述** | | | | | | |
| **序号** | **测试步骤** | | | **测试结果** | | **评价准则** |
| **预期结果** | |  |
| 1 | 在命令行中启动测试爬虫 | | | 启动成功 | | 实际结果与预期结果一致 |
| 2 | 观察数据库中存入的链接关系 | | | 有新的链接关系存入 | | 实际结果与预期结果一致 |
| **备注** | 说明：  1）设置网站链接关系-----在parse函数生成Request对象时，使用url\_source命名参数指定。  2）观察数据库中存入的链接关系-----使用mongo命令行客户端连接到mongodb的router服务器进行观察。  例如：  命令：mongo <host>:<port>  参数：<host>为主机名或者主机<ip>，<port>为端口。例如，如果router服务器的ip和端口分别为10.2.1.175和26061，那么可使用mongo 10.2.1.175:26061链接服务器。   1. 观察数据库中存入的链接关系-----使用mongodb数据库查询命令进行查询。   例如：  命令：db.<文档名称>.find()  参数：<文档名称>为配置使用的文档名称。例如，如果配置使用的文档名称为default\_col，那么可使用命令：db.default\_col.find()进行查询。 | | | | | |
| **测试结论** |  | | | | | |
| **测试人员：王益飞** | | **测试时间：**2017-05-17 | | | | |

### 非功能性需求测试

##### 爬取效率测试

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **测试用例名称** | 比较分布式爬虫和单机爬虫爬取速度 | | **测试用例标识** | 用例TC06 | **测试需求标识** | TR06 |
| **简要描述** | 本测试验证分布式爬虫的爬取速度是否比单机爬虫的爬取速度更快。 | | | | | |
| **前提和约束** | 1. 针对北航计算机学院官网的爬虫开发完毕。 2. 分布式爬虫协作爬取正确性测试完毕。 3. 分布式爬虫协作爬取正确性测试结果正确。 4. 爬虫有效性测试完毕。 5. 爬虫有效性测试结果有效。 | | | | | |
| **测试方法** | 黑盒测试 | | | | | |
| **测试过程描述** | | | | | | |
| **序号** | **测试步骤** | | | **测试结果** | | **评价准则** |
| **预期结果** | |  |
| 1 | 在命令行中启动单机爬虫 | | | 启动成功 | | 实际结果与预期结果一致 |
| 2 | 记录单机爬虫爬取时间 | | | 记录成功 | | 实际结果与预期结果一致 |
| 3 | 在命令行中启动分布式爬虫 | | | 分布式爬虫启动成功 | | 实际结果与预期结果一致 |
| 4 | 记录分布式爬虫爬取时间 | | | 记录成功 | | 实际结果与预期结果一致 |
| **备注** | 说明：  1）在命令行中启动单机爬虫-----启动方式与爬虫有效性测试中的方式相同。  2）在命令行中启动分布式爬虫-----分布式爬虫启动方式与分布式爬虫协作爬取正确性测试中的方式相同。  3）使用的测试爬虫为开发完成的针对北航计算机官网的爬虫。 | | | | | |
| **测试结论** |  | | | | | |
| **测试人员：王益飞** | | **测试时间：**2017-05-17 | | | | |