链滴

Jsoup 的简单的使用示例

作者：Sysecho

原文链接：https://ld246.com/article/1498787703299

来源网站：链滴

利用Jsoup中的相关方法实现网页中的数据爬去，本例子爬去的网页为比较流行的programmablewe中的mashup描述内容，然后为数据库中存在的mashup添加相应的描述。

```
package com.test;

import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import org.jsoup.Jsoup;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;
import com.bean.mashup_tags_apis;
import com.daoImpl.MashupDaoImpl;

public class JsoupTest {

    /**
     * @param args
     */
    public static void main(String[] args) {

        List<String> mashupName = new ArrayList<String>();
        List<String> mashupDescription = new ArrayList<String>();
        MashupDaoImpl mashupDaoImpl = new MashupDaoImpl();
        List<mashup_tags_apis> mashup_tags_apis = mashupDaoImpl
                .findAllmashup_tags_apis();

        try {

            // 获取网页内容，从第二页开始，第1页特殊处理
            for (int p = 220; p < 365; p++) {
                System.out.println("正在爬取第" + p + "个页面.....");
                org.jsoup.nodes.Document doc = Jsoup.connect("http://www.programmableweb.c
m/mashups/directory/"
                        + p).get();

                // 通过ID获得需要的表格
                Element content = doc.getElementById("mashups");

                // 按照[href*=/mashup/]取得数据
                Elements name = content.select("[href*=/mashup/]");

                // 踢出版本信息
                String RegexMatcher = "[\\d.]+";

                // 向mashupName集合中添加名字
                for (int i = 0; i < name.size(); i++) {
                    String Name = name.get(i).text();
                    if (name.get(i).hasText() && !Name.matches(RegexMatcher)) {

                        mashupName.add(Name);
                    }
                }
```

```
            // 取得描述信息
            Elements description = content.getElementsByTag("p");
            // 向mashupDescription集合中添加描述信息
            for (Element descri : description) {
                String Comment = descri.text();
                if (p == 1) {
                    // 第一页处理方式（名字和描述都为空）
                    if (Comment != null && Comment.length() > 2) {
                        if (Comment != null) {
                            mashupDescription.add(Comment);
                        }
                    }
                } else {
                    // 从第二页开始处理方式,描述为空用NoDescriptions占位
                    if (Comment == null) {
                        Comment = "NoDescriptions";
                    }
                    mashupDescription.add(Comment);
                }

            }

            // 更新数据库
            for (int i = 0; i < mashupName.size(); i++) {
                String Name = mashupName.get(i);
                for (int j = 0; j < mashup_tags_apis.size(); j++) {
                    if (Name.equals(mashup_tags_apis.get(j).getName())) {
                        String destrcipString = mashupDescription.get(i);
                        if (Name != null && destrcipString != null) {
                            if (!mashupDaoImpl.updateMashup_tags_apis(
                                    destrcipString, Name)) {
                                System.out.println("更新失败！");
                            }
                        }
                    }
                }
            }

            // 清空集合爬取下一个页面
            mashupDescription.clear();
            mashupName.clear();
            System.out.println("第---------" + p + "---------个页面完成！\n");
        }
    } catch (IOException e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
    }

    // 显示输出查看是否正确
    // for (int i = 0; i < mashupName.size(); i++) {
    // System.out.println((i + 1) + " " + mashupName.get(i));
    // }
    //
```

```
        // for (int j = 0; j < mashupDescription.size(); j++) {
        // System.out.println((j + 1) + " " + mashupDescription.get(j));
        // }
        System.out.println("恭喜您，描述添加成功！");
    }
}
```

这也是我第一次是使用Jsoup，还是有很多东西等待自己慢慢发现.....