|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **北京邮电大学软件学院** | **文档编号** | TDIPS-8 | **版本** | 4.0 | **密级** | 商密A |
| **项目名称** | **文本数据智能处理系统** | | | | |
| **项目来源** |  | | | | |

**文本数据智能处理系统**

**软件用户使用手册**

(内部资料请勿外传)

|  |  |  |  |
| --- | --- | --- | --- |
| **编写：** | 刘凡凡 林哲  吴晓吟 孙华枭 | **日期：** | 2011-4-11 |
| **检查：** | 任建新 | **日期：** | 2011-4-11 |
| **审核：** | 吴国仕 | **日期：** | 2011-4-11 |
| **批准：** |  | **日期：** |  |

**北京邮电大学软件学院**

**版权所有不得复制**

**文档变更记录**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **变更（+/-）说明** | **作者** | **版本号** | **日期** | **批准** |
| 1 | 初次编写 | 孙华枭 | 1.0 | 2011-4-11 |  |
| 2 | 增加爬虫模块相应内容 | 刘凡凡 | 2.0 | 2011-4-11 |  |
| 3 | 增加抽取模块相应内容 | 林哲 | 3.0 | 2011-4-11 |  |
| 4 | 增加情感化过滤模块相应内容 | 吴晓吟 | 4.0 | 2011-4-11 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

目录

[1. 系统说明 4](#_Toc341645748)

[1.1. 系统简介 4](#_Toc341645749)

[1.2. 支持软件的安装 4](#_Toc341645750)

[1.3. 软件的安装 4](#_Toc341645751)

[2. 使用过程 5](#_Toc341645752)

[2.1. 支持AJAX的定址网络爬虫模块的使用 5](#_Toc341645753)

[2.1.1. 工具运行起始界面 5](#_Toc341645754)

[2.1.2. Url提取 5](#_Toc341645755)

[2.1.3. 批量抓取网页 9](#_Toc341645756)

[2.1.4. 抓取需要登录或注册才能访问的页面 11](#_Toc341645757)

[2.1.5. 配置功能 13](#_Toc341645758)

[2.1.6. 翻页脚本 13](#_Toc341645759)

[2.1.7. 自动生成翻页脚本 17](#_Toc341645760)

[2.1.8. 出错处理和恢复 20](#_Toc341645761)

[2.2. 网页数据精确抽取模块的使用 20](#_Toc341645762)

[2.2.1. 数据集准备 20](#_Toc341645763)

[2.2.2. 工具运行主界面 20](#_Toc341645764)

[2.2.3. 生成模块 21](#_Toc341645765)

[2.2.4. 标注模块 23](#_Toc341645766)

[2.2.5. 数据抽取 25](#_Toc341645767)

[2.3. 情感化过滤模块的使用 27](#_Toc341645768)

[2.3.1收集txt文本集 27](#_Toc341645769)

[2.3.2工具运行主界面 28](#_Toc341645770)

# 系统说明

## 系统简介

该文本数据智能处理系统主要包括3个主要功能并以此划分模块，他们分别为支持AJAX的定址网络爬虫，网页数据精确抽取和情感化过滤。

1. 支持AJAX的定址网络爬虫是解决从层出不穷的异步交互式网络的AJAX框架网站上抓取页面和翻页抓取问题的方案，主要解决基于Web2.0模式的采用AJAX技术的互联网应用的页面信息爬取。
2. 网页数据精确抽取系统通过计算机对大量的网页集进行训练，生成网页模板。用户通过模版选取感兴趣的信息，对其进行标记（语义信息）。利用用户标记过的模板，对指定的网页抽取出用户感兴趣的数据。最后抽取的详细可以按用户需要放到指定的文件（XML，HTML）或者用户数据库中。
3. 情感化过滤则利用计算机对带有文本作者主观情感或意见的文本进行自动的分析，并推理出其情感倾向性。输入的文本可以是句子或者文档，输出为文本的情感得分，正面文档得分为正，负面得分为负。

## 支持软件的安装

参见《软件安装部署手册》

## 软件的安装

参见《软件安装部署手册》

# 使用过程

## 支持AJAX的定址网络爬虫模块的使用

### 工具运行起始界面

将运行文件crawler.jar与脚本文件（\*.js）置于同一目录下，目录应为英文目录，双击crawler.jar，弹出工具起始界面如下：

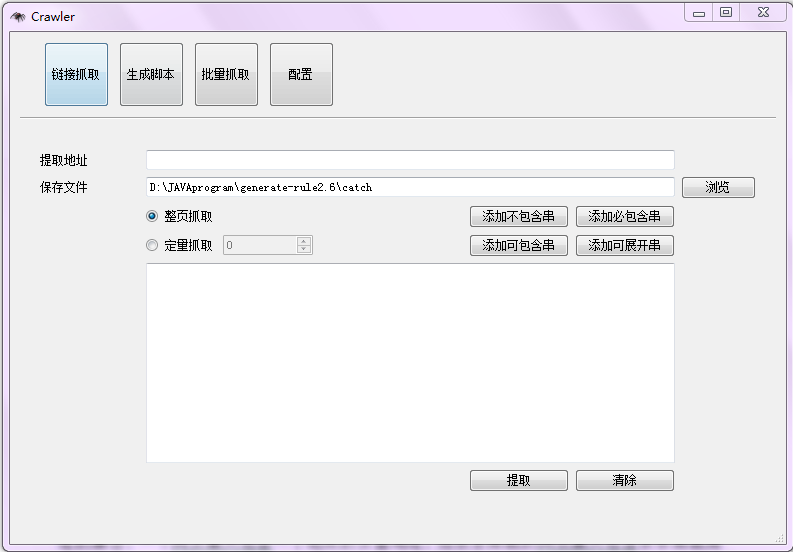


图 2-1-2 工具运行主界面

主界面上端是界面切换按钮，默认显示“链接抓取”界面。主界面的左上部是切换链接抓取、生成脚本、批量抓取的按钮。中部是提取地址栏、保存地址栏，以及关于抓取链接过程中需要设置的信息。下部是链接显示框、提取按钮，以及能清除显示框的清除按钮。

### Url提取

由于本工具不同于一般意义上的网络爬虫，不具有在抓取网页过程中爬行页面中的子链接来遍历Web，收集Web页面的功能，因此需要指定抓取页面的URL。网页集是待抓取网址的集合，一个网页集只包含一个站点的页面地址，如京东商城的网页集只包含京东商城商品信息页面的地址，示例如下：

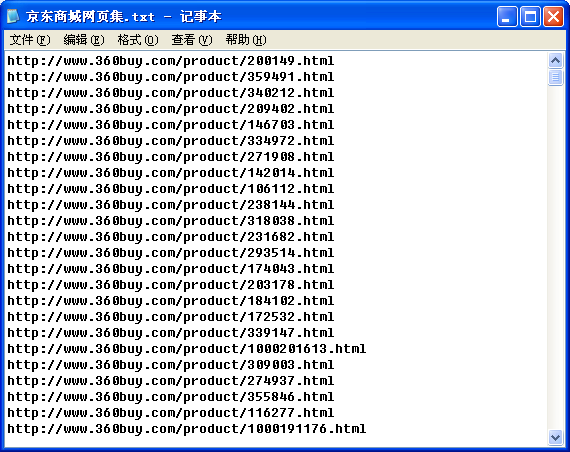
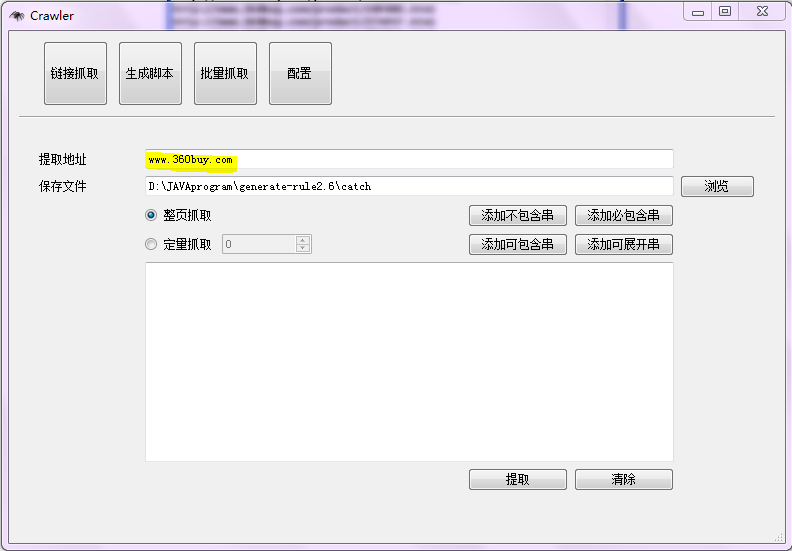
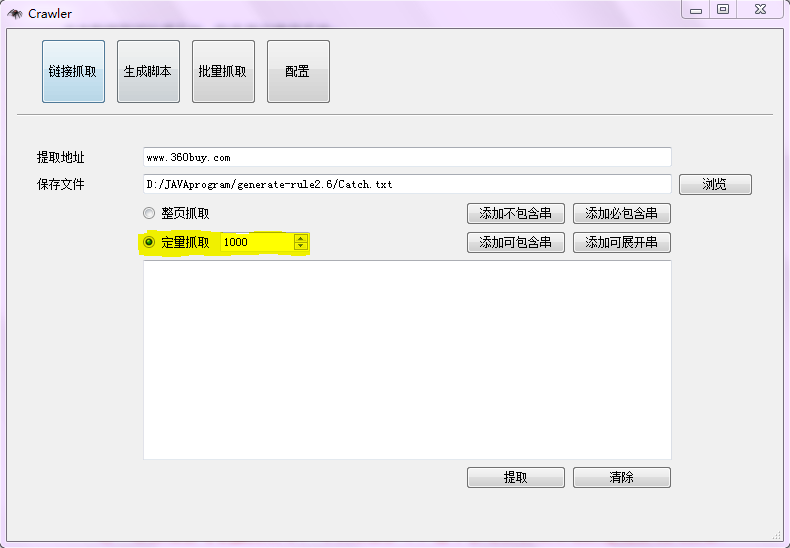


图2-1-1 京东商城网页集示例

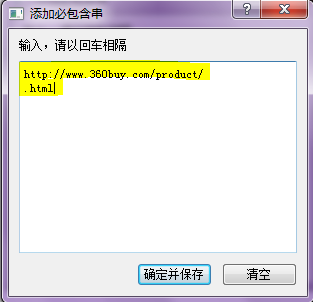
1. 首先，输入起始提取地址：



1. 点击浏览键，选择Url集保存地址及文件名。
2. 选择是抓取起始地址整个页面当中的Url还是抓取指定数目的Url。

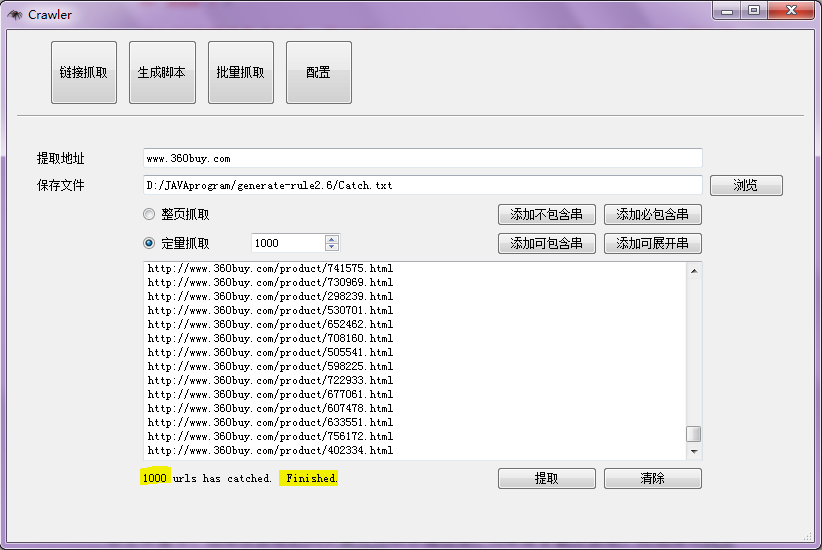


1. 每个网页下有很多Url链接，其中一部分使用户需要保留的有用链接，其它全为无用链接，软件提供了4中过滤措施，用户只需点击各自按钮即可出现添加过滤串的对话框：



用户只需手动输入或复制粘贴即可，每个串用回车换行相隔，添加完毕后点击“确定并保存按钮”即可，用户也可以点击“清空”按钮将原先添加的所有串删除。系统默认全部为空的时候，抓取到的每个连接都为有用链接，且都会展开（尽管“可展开串”为空）。

1. 添加链接**不能包含**的串，可以添加多个，链接如果包含此串，即为我们不需要的无用连接，eg. 新浪微博的用户主页链接都不得包含”huati”（微博话题）、”hot”（热门微博）。
2. 添加链接**必须包含**的串，可以添加多个，链接必须包含的串指的是每条有用链接都包含的串，eg. 京东商品的链接都必须包含”<http://www.360buy.com/product/>”和”.html”。
3. 添加链接**可以包含**的串，可以添加多个，链接必须包含这些串当中的一种，因为可能有好几种有用的链接，eg. 新浪微博用户主页的链接除了可以包含的有” <http://weibo.com/>”（个人用户微博），也可以包含” <http://e.weibo.com/>”（企业用户微博）
4. 添加链接**可以展开**的串，可以添加多个，由于抓取的机制采取的是宽度优先搜索，只从有用链接当中展开抓取有用链接，但是某些无用链接展开之后是能抓取到更多有用链接的，eg. 搜狐博客的博文”http://baoesq.i.sohu.com/blog/view/245625737.htm”当中有博主的所有博文的博客链接”http://baoesq.i.sohu.com/blog/index.htm”，那么相应的” i.sohu.com/blog/index.htm”就为可展开的串，因为包含这个串的链接展开后能抓取到更多的博文。
5. 点击提取按钮即可开始抓取链接，抓取过程中提取按钮的左边会显示抓取的状态，显示当前抓取了多少条，抓取完毕后会出现”Finished”。



### 批量抓取网页

在批量抓取界面中，选择待抓取的网页集文件和存储目录，如果需要翻页抓取，则选择与网页集相对应的脚本文件；如果不需要翻页抓取，则选择bufanye.js；如果不选择任何脚本文件，则不进行抓取工作。默认抓取网页的编码格式为“简体中文18030”，特别指出新浪微博的编码格式为“Unicode(UTF-8)”。由于用户不断发微博，在一次抓取之后的一个时间段内，会有新的微博出现，那么只抓取这一时间段内的微博即可，之前抓过的微博无需再抓，设置“抓取时间至”为上一次批量抓取的时间，这项设置在抓取微博时必须设置。站点、网页集的地址类型和翻页脚本文件之间对应关系如下：

表2-1 站点、网页集的地址类型和翻页脚本文件对应表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 站点 | 网页集地址类型 | 翻页脚本(.js) | 网页集示例(.txt) | |
| 京东商城 | 商品信息页面地址 | shopping-360buy | shopping\_360buy |
| 淘宝网 | 商品信息页面地址 | shopping-taobao | shopping\_taobao |
| 淘宝商城 | 商品信息页面地址 | shopping-tmall | shopping\_taobao\_shangcheng |
| 新浪博客 | 博客文章评论页面地址 | blog-sina | blog\_sina |
| 搜狐博客 | 博客文章评论页面地址 | blog-sohu | blog\_sohu |
| 新华博客 | 博客文章评论页面地址 | blog-xinhua | blog\_xinhua |
| 新浪微博 | 微博完整评论页面地址 | mblog-sina | mblog\_sina |
| 网易微博 | 微博完整评论页面地址 | mblog-163 | mblog\_163 |
| 搜狐微博 | 微博完整评论页面地址 | mblog-sohu | mblog\_sohu |
| 猫扑论坛 | 文章独立页面地址 | bbs-mop | bbs\_mop |
| 天涯论坛 | 文章独立页面地址 | bbs-tianya | bbs\_tianya |

下面以翻页抓取京东商城商品信息页面为例：

由于京东商城商品信息页面中的评论模块存在分页，而我们想要得到针对某一商品的所有评论，因此需要翻页抓取。在脚本列表里选择shopping-360buy.js，Url集文件选择京东商城的网页集文件，保存文件至自选目录，如下图所示。

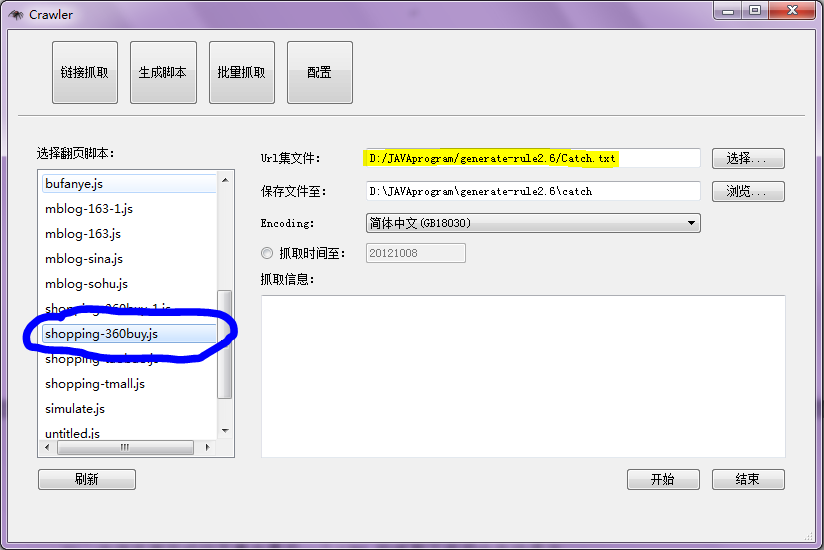


图2-1-3 翻页抓取京东商城商品信息页面

点击“开始”按钮，弹出浏览器窗口，浏览器会串行加载网页集中的URL，抓取过程如下：

1. 当一个页面加载完成后，页面内容被下载到存储目录中，以html格式保存。
2. 浏览器检测有无分页，若有分页，则翻到下一页，执行第1步；若没有分页或已到末页，浏览器会加载网页集中下一个URL，转第1步。
3. 当浏览器完成网页集中最后一个URL的抓取工作后会自动退出。
4. 若想中止抓取操作，可点击主界面中“结束”按钮或关闭浏览器。

浏览器工作界面如下：



图2-1-4 浏览器工作界面

### 抓取需要登录或注册才能访问的页面

以微博网站为例，当我们想要查看某人发布的微博的所有评论时，微博网站会提示需要登录才有权限查看，因此想要得到这些评论就必须先用一个账号进行登录操作，浏览器会根据您对Cookie的配置策略（见配置功能）决定是否保存您的登录信息以便下次抓取微博页面时自动登录。

下面以抓取新浪微博评论为例：

1. 由于我们是首次抓取新浪微博，应先登录，让浏览器保存这次登录信息。

在主界面中，不选择任何翻页脚本，Url集文件和存储目录可不选，点击“开始”按钮，打开空白浏览器，在地址栏输入新浪微博地址；若选择Url集文件，应选择新浪微博网页集，点击“开始”按钮，浏览器默认访问网页集中第一个网址。浏览器运行如下：

图2-1-5 浏览器访问微博页面

1. 进行登录操作。



图2-1-6 进行登录操作

1. 登录成功。



图2-1-7 浏览器访问登录后的微博页面

1. 关闭浏览器。

完成登录信息的保存后，在以后抓取新浪微博时只需按照批量抓取网页中的操作步骤执行即可。直到Cookie过期，需要再一次登录。

### 配置功能

在主界面中，点击上端“配置”按钮，显示配置界面：

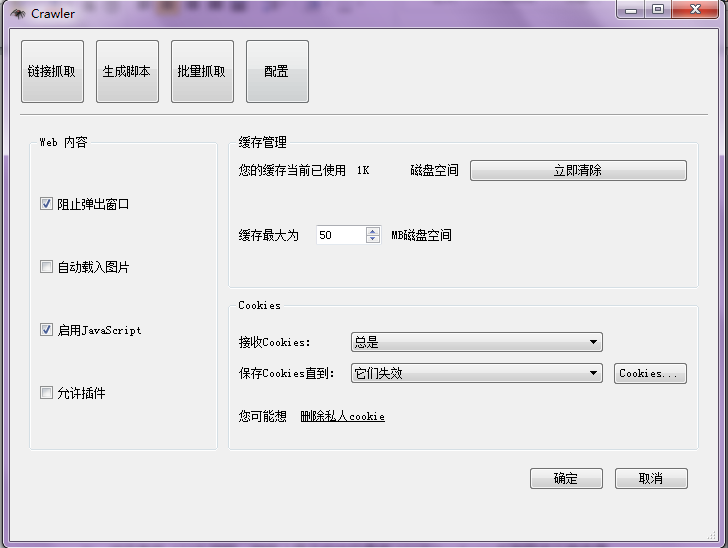


图2-1-8 配置界面

在这里可以配置浏览器的行为。左端是Web内容的配置项，右上部和右下部分别是缓存管理和Cookie管理，其中缓存占用的磁盘空间最大可达到1024MB，Cookie的接收和保存策略会影响浏览器对于登录信息的保存。

### 翻页脚本

翻页脚本是个JavaScript文件，这个脚本文件主要完成以下操作：

* 进行翻页操作
* 从目标页面上抓取内容

脚本文件的编写需从以下几方面考虑：

1. 对于有些AJAX网页，网页上显示的内容是用AJAX/JavaScript机制异步从服务器上加载，而且并不是在加载HTML网页的同时或者之后就开始AJAX数据的加载，而是用户滚动浏览器屏幕内容时，滚屏到哪才加载并显示到哪。此种情况需要滚屏加载，滚屏会降低抓取网页的速度。
2. 增加时延。对于AJAX网页，浏览器不能检测到AJAX数据是否加载完成，为了等待异步加载的AJAX数据完整加载好，需要设置时延，时延会降低抓取网页的速度。
3. 下载网页内容。
4. 查找翻页链接，模拟鼠标点击操作。在搜索翻页链接时，需要针对有无翻页接口和有无下一页进行检测，当二者中有一个不存在时，表示该页面没有分页或已到尾页，此时结束翻页，继续爬行下一个地址。

对于AJAX网页，翻页脚本应按照1-4步顺序编写，然后循环执行2-4步，直到翻页结束，转向下一个地址。其中第1步按照网页实际情况选择是否需要。

对于非AJAX网页，翻页脚本应按照3，4步顺序编写，然后循环执行3，4步，直到翻页结束，转向下一个地址。

此外，有些网站点击翻页链接后只有分页内容局部刷新，有些则整个页面全部刷新，翻页脚本对于这两种情况的处理略有不同。

例1中的新浪微博网站使用了AJAX技术，翻页后内容局部刷新。

例2中的天涯论坛没有使用AJAX技术，翻页后页面全部刷新。

例1：新浪微博的翻页规则文件

**function** location()

{

scroll(0, 100000);//滚屏加载

setTimeout("init()",2000);//增加时延

}

**function**init() {

**var**hasnextpage=0;//有无下一页标志

**var**haspagebar=0;//有无翻页接口标志

ww.emitClickSignal();//下载网页内容

/\*搜索翻页链接\*/

**var**div\_set=document.getElementsByTagName("div");

**for**(**var** i=0;i<div\_set.length;i++)

{

**if** (div\_set[i].className=="W\_pagesW\_pages\_comment")

{

haspagebar=1;

**var**spanset=div\_set[i].getElementsByTagName("span");

**for**(**var** j=0;j<spanset.length;j++)

{

**if**(spanset[j].innerHTML=="下一页")

{

hasnextpage=1;

simulate(spanset[j],"click");//模拟鼠标点击操作

**break**;

}

}

**break**;

}

}

**if**(hasnextpage==0 ||haspagebar==0)

{//到达尾页||只有一页：没有翻页入口

ww.singleUrlFinished();//继续爬行下一个地址

**returntrue**;

}

timeCount();//循环执行

}

**function**timeCount()

{

setTimeout("init()",2000);//增加时延，等待分页内容加载

}

location();

例2：天涯论坛的翻页规则文件

(**function**() {

**var**hasnextpage=0;//有无下一页标志

**var**haspagebar=0;//有无翻页接口标志

ww.emitClickSignal();//下载网页内容

**var** div=document.getElementById("pageDivTop");

**if**(div==**null**)

{//只有一页：没有翻页入口

ww.singleUrlFinished();//继续爬行下一个地址

**returntrue**;

}

/\*搜索翻页链接\*/

**var**aset = div.getElementsByTagName("a");

**for** ( **var** j = 0; j <aset.length; j++) {

**if** (aset[j].innerHTML == "下一页") {

hasnextpage = 1;

simulate(aset[j], "click");//模拟鼠标点击操作

**returntrue**;

}

}

**if**(hasnextpage==0)

{//到达尾页

ww.singleUrlFinished();//继续爬行下一个地址

**returntrue**;

}

})();

### 自动生成翻页脚本

* 通过用户与脚本生成工具进行交互自动生成脚本
* **生成脚本页面的上端为生成脚本各项设定：**
  1. **“分页链接样式”，按钮选择为“有翻页标志”、“无翻页标志”，其中，有翻页标志表示网页中存在“下一页”、“下页”、“Next”、“》》”等类似按钮标志；**
  2. **“模式”，分为“局部刷新”、“全局刷新”，局部刷新指的是新内容加载过程中，地址栏中的网页URL链接不变，只是内容动态加载，全局刷新指的是新内容加载过程中，地址栏中的网页URL链接会变为另一个定地址。**
  3. **“分页获取规则”，为设定的翻页按钮的HTML源代码。**
  4. **“参数”设置，“时延”设置，表示进行翻完一页保存完毕后，再翻下一页继续保存的时延；“自动滚屏”是网页初期加载过程中，网页只会加载一部分内容（网页上端的内容），需要浏览器自动滚屏到下部，激发加载另外部分的内容，有时需要滚屏多次，例如新浪微博，需要滚屏两次：**

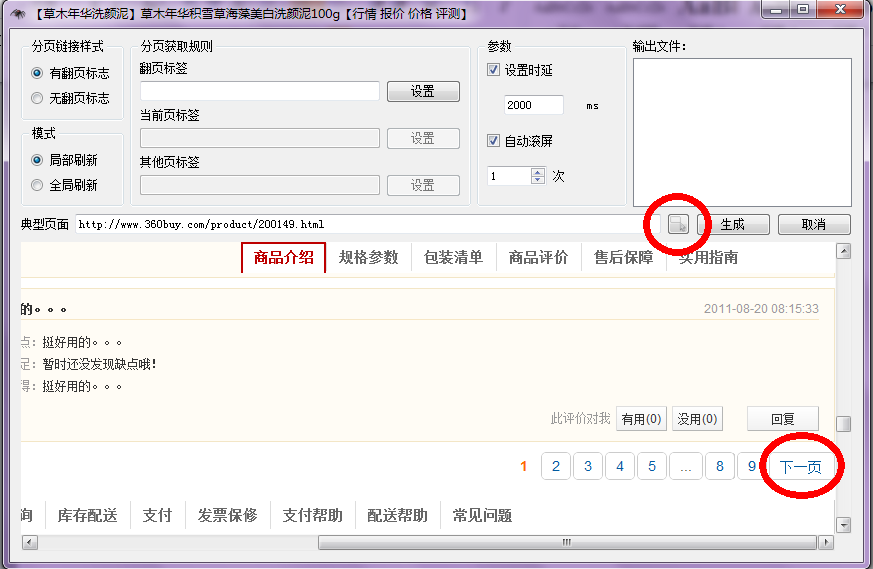


* 1. **“输出文件区域”，为生成的翻页脚本内容。**
* 页面下端为一个浏览器。

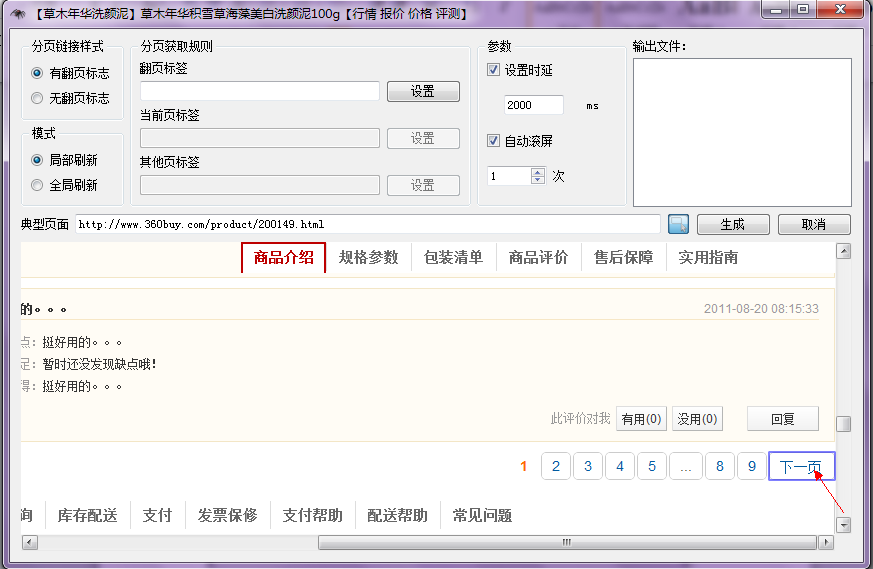
1. 首先用户在“典型页面”地址栏中输入要典型页面的Url，敲击回车键，打开典型页面。



1. 在网页中大致找到翻页按钮，**先**选择“选定”按钮工具，**再**在网页中点击“翻页按钮”,，这个翻页按钮通常就是“下一页”、“下页”、“Next”、“》》”



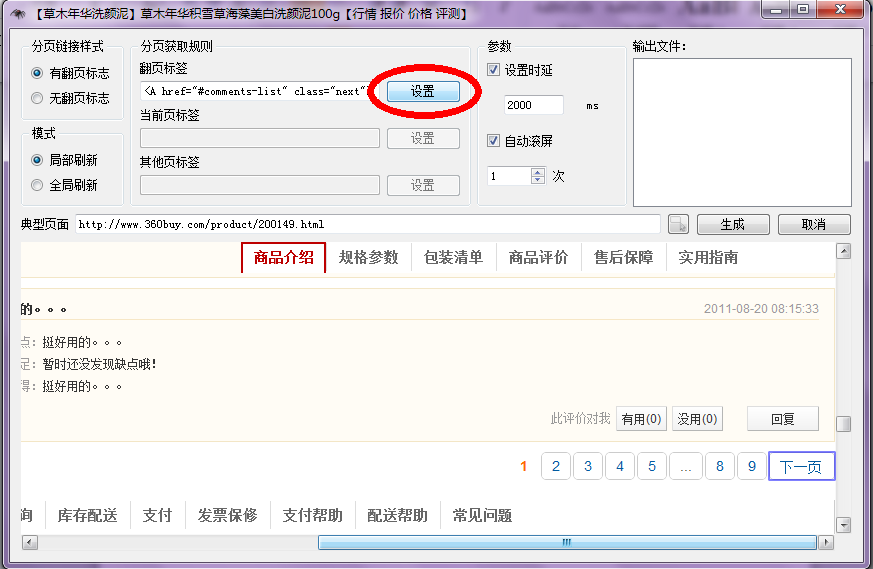
1. **在选定过程中，浏览器中出现的蓝色的框表示：当前鼠标指向内容的HTML标签。用户讲鼠标移至“翻页按钮”后，按钮也会出现蓝色的框，表示找到翻页按钮的HTML标签。**



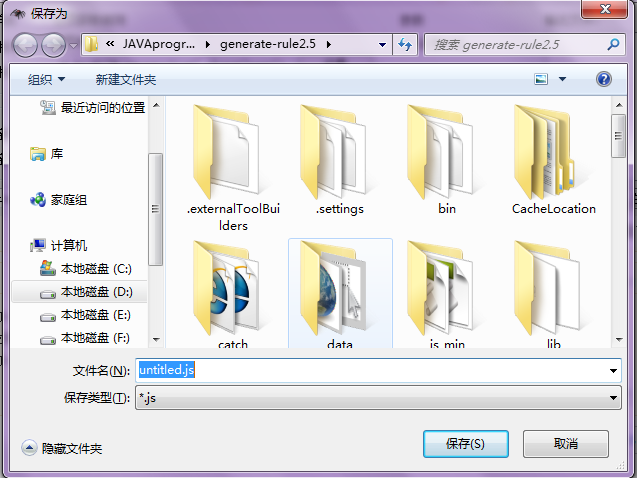
翻页按钮

蓝色框

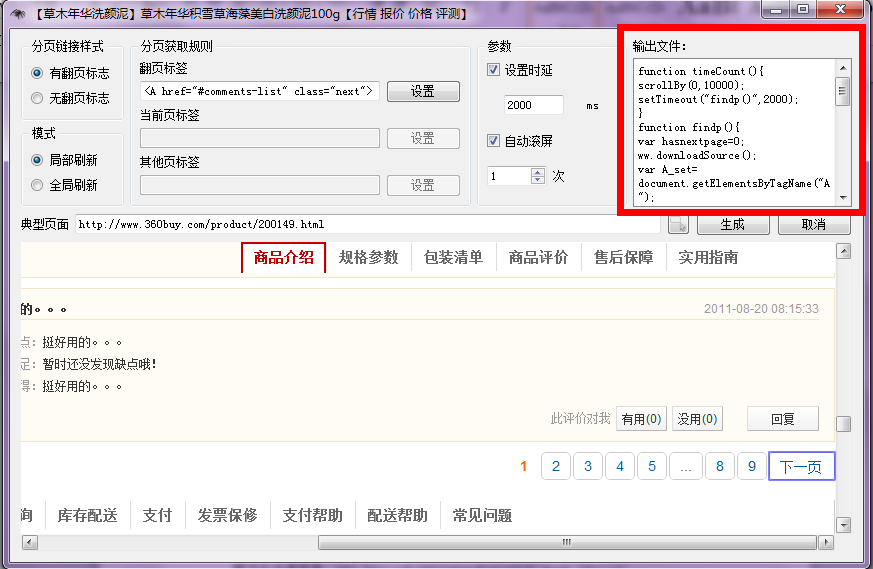
1. 点击“翻页标签”**设置栏右方**的“设置”按钮，翻页标签的HTML标签代码即会自动显示在翻页标签**设置**栏当中。软件工具就会根据这个信息在网页中找到这个这个翻页标签，即会生成能找到这个标签的Javascript代码。



1. 接着点击“生成”按钮，即可选择保存地址以及文件名了，文件后缀名须为“.js”。



1. 其后，生成的翻页脚本代码除了会保存在用户设定的地址之下，还会显示在右侧上方显示区当中。



### 出错处理和恢复

* 打开crawler.jar文件时报错，提示“Could not find the main class”的警告。

解决办法请参考：http://blog.csdn.net/wuxianglong/article/details/5894236

* 登录网易微博没反应。

请安装OpenSSL，并将安装路径加入系统变量PATH中。

## 网页数据精确抽取模块的使用

### 数据集准备

网页数据精确抽取主要是用于抽取具有相似网页结构的网页中用户感兴趣的数据，所以数据集应该是由相似网页结构的网页组成，这样可以大大提高模板生成的精度。

由于本工具不包含网络爬虫，网页集是抓取网址后的网页集合，一个网页集只包含一个站点的相同结构的网页，如京东商城的网页集只包含京东商城商品信息页面，示例如下：

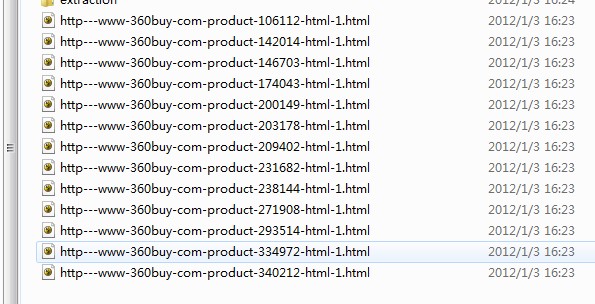


图 2-2-1 京东商城网页集示例

因为在windows操作系统下，文件名的长度有一定的限制，为了使用系统能够正常运行，建议将网页数据集放在硬盘根目录下。

### 工具运行主界面

将运行文件buptssewie.jar放置于某一目录下，目录应为英文目录，双击buptssewie.jar，弹出工具主界面如下：

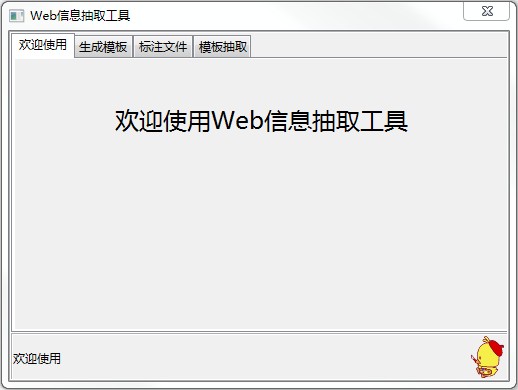


图 2-2-2 工具运行主界面

主界面上端是界面切换按钮，默认显示“欢迎”界面，主界面上还有3个Tab页，分别是生成模板、标注文件和抽取数据，对应的进行模板生成、模板标注和数据抽取操作。

### 生成模块

系统是从网页集中训练生成模板后，根据用户的标注信息对网页进行抽取，所以我们首先得利用本工具生成模板，其操作界面如下图所示：



图 2-2-3 生成模板主界面

以生成京东商城商品信息页数据集模板为例，使用该工具生成模板。在模板生成操作页面中，点击“选择”，选择网页集所在目录，弹出的选择界面如图2-2-4所示：



图2-2-4选择网页集所在目录

确认后目录被选中，在生成模板操作主界面上拖动相似度选择条，选择相似度，对于比较规的网页，如京东商城等则可以将相似度提高到0.9，对于网页不太规则的网页，比如新闻类网页则相似度应相对减小到0.8左右。



图2-2-5 生成模板执行界面

目录和相似度选择完成后，点击“开始执行”按钮进行模板生成操作，其界面如图2-2-5所示。生成模板执行操作完成后，操作界面状态条提示操作完成，并显示生成模板的个数。

### 标注模块

系统从网页集中训练生成模板后，根据用户的标注信息对网页进行抽取，标注模板是指用户将所要抽取的网页信息在生成的模板上标注出来，为抽取数据提供抽取模板，其操作界面如下图所示：

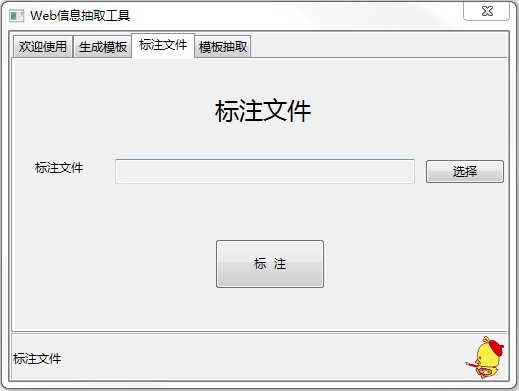


图 2-2-6 标注模板主界面

以标注京东商城商品信息页数据集模板文件为例，使用该工具标注模板。在模板标注操作页面中，点击“选择”，弹出的选择界面，将目录定位到京东商城商品信息页数据集，我们可以看到在生成模板时生成的模板文件夹template，如图2-2-7所示：

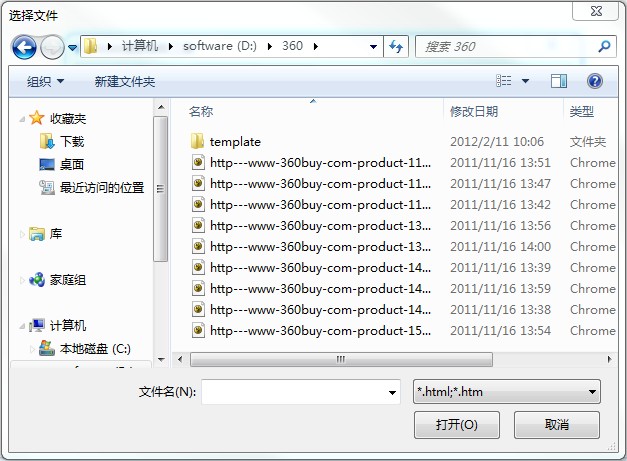


图2-2-7模板文件夹

点击进入template文件夹后，选择模板文件后工具自动返回到标注模板主界面。在进入template文件夹后，可以看到生成的模板文件，其命名规则为template+模板序号-模板包含网页数.htm，其中模板包含网页数是指要网页数据集中有几个网页是相同网页结构的，所以在选择标注模板的文件是应该选择所包含网页数量多的模板文件，例如：生成的模板文件名为：template1-1.htm和template2-8.htm，应该选择template2-8.htm这个文件，因为其由网页数据集文件中的8个文件生成，具有比template1-1.htm大的信息包含量。

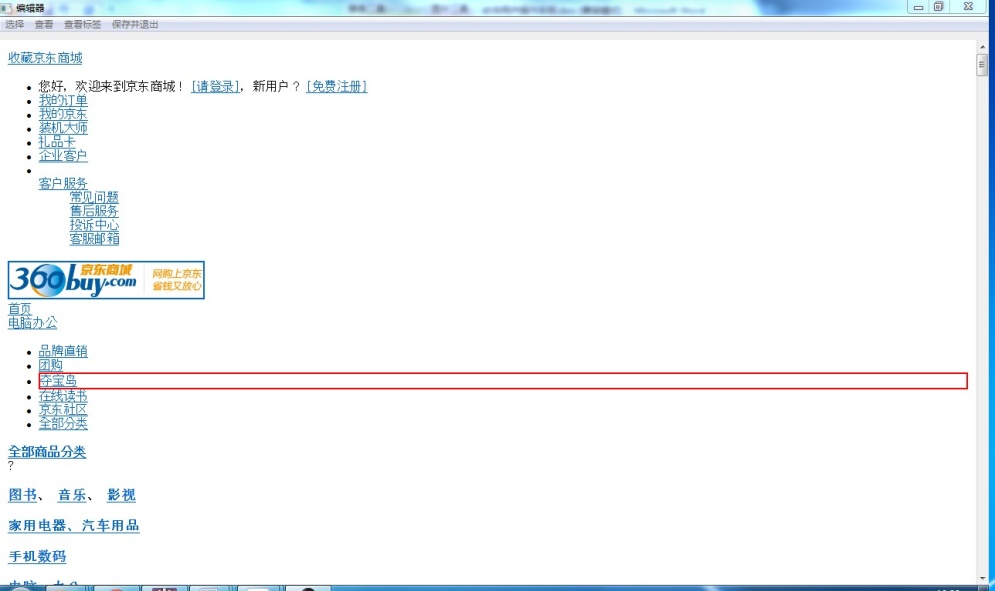


图2-2-8 标注模板操作界面

选择好所要标注的模板文件生，工具自动返回到标注操作主界面，点击“标注”按钮进行操作，此时会弹出操作界面，如图2-2-8所示。在操作界面上移动鼠标，可以看到在鼠标经过时标签的边框会就红色，双击选中标签时会变成绿色。双击选中标签后，点击菜单栏上的“选择”按键，可以将标签加入队列中，选择界面如下：

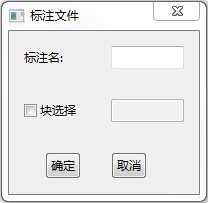


图2-2-9 标注选项界面

选择界面上有标注名和块选择，输入完成后点击确定后退出，完成一个标签的标注操作。在完成整个模板文件的标注后，点击菜单上的“保存并退出”后完成整个模板文件的标注操作。

### 数据抽取

系统的最终目的是根据用户的标注信息对网页进行网页数据的抽取，数据抽取的操作界面如下图所示：

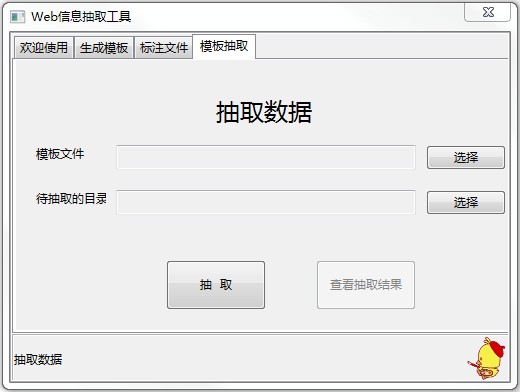


图 2-2-10 数据抽取主界面

以京东商城商品信息页数据集数据抽取为例，在模板生成操作页面中，点击“选择”，选择网页集所在目录和模板文件的，其中模板文件是在模板标注中经过用户标注后保存下来的模板文件，这两个选项选择结束后，点击界面上的“抽取”按钮进行数据的抽取。当数据抽取结束后，可以点击“查看抽取结果”按钮查看抽取出来和信息，界面如图2-2-11所示：

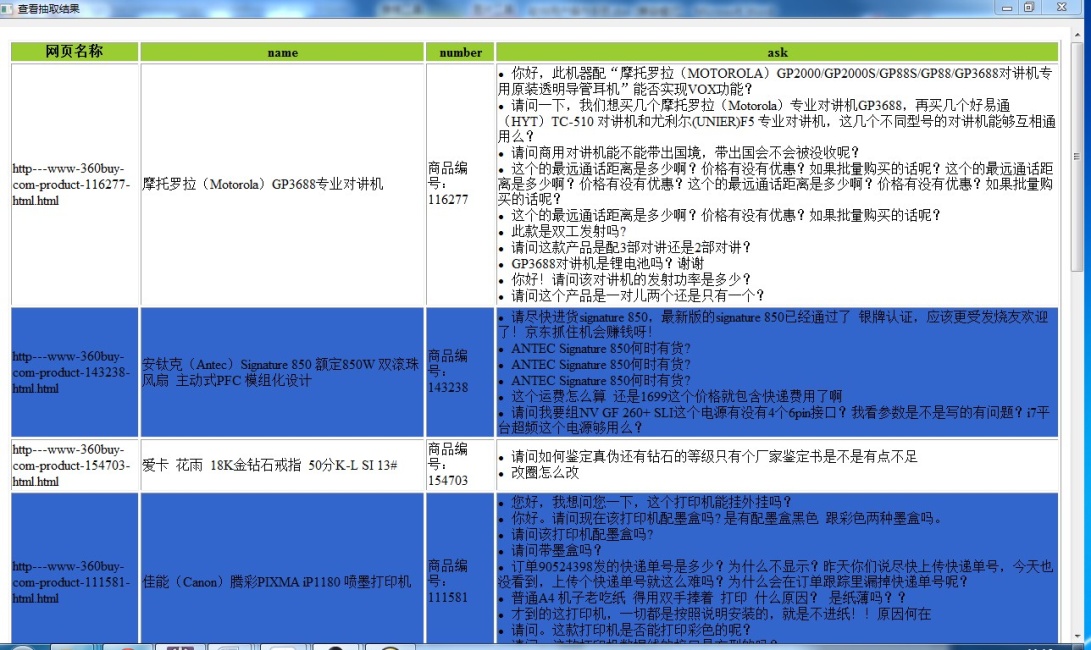


图2-2-11查看抽取结果

## 情感化过滤模块的使用

### 2.3.1收集txt文本集

如将收集到的文本都放在D盘D：\input1\test3路径下，如图所示：

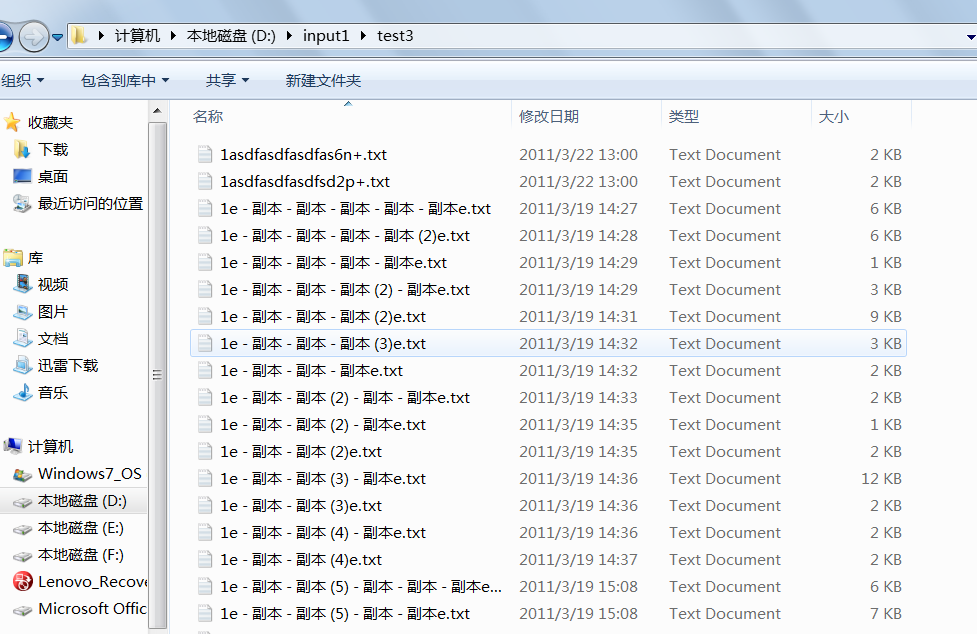


图2-3-1 文本集的存放

其中文本以txt文本格式存储，如图所示：

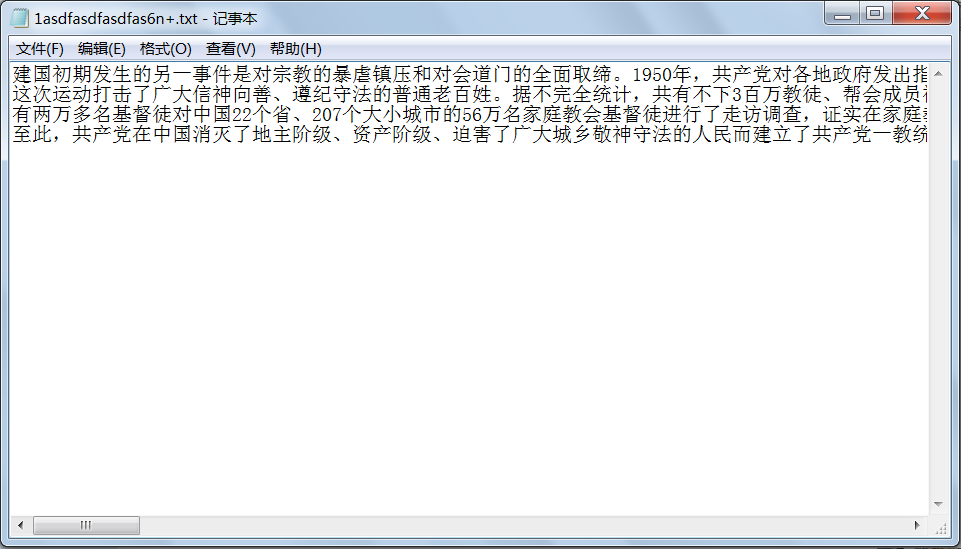


图2-3-2 文本集内容

### 2.3.2工具运行主界面

运行文件opinionMining.jar，弹出工具主界面如下：

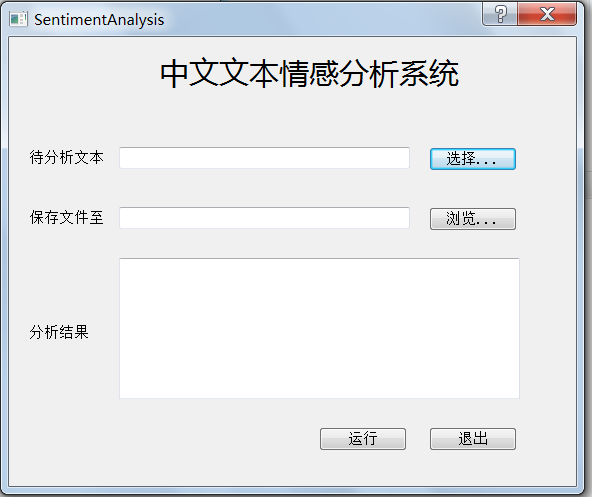


图2-3-3 工具主界面

选择要测试的文本集，示例如下图：

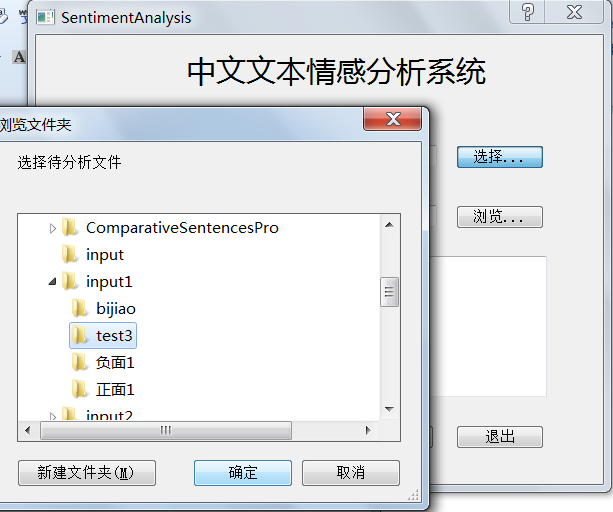


图2-3-4 选择测试的文本集

选择要保存的结果文件，示例如下图：

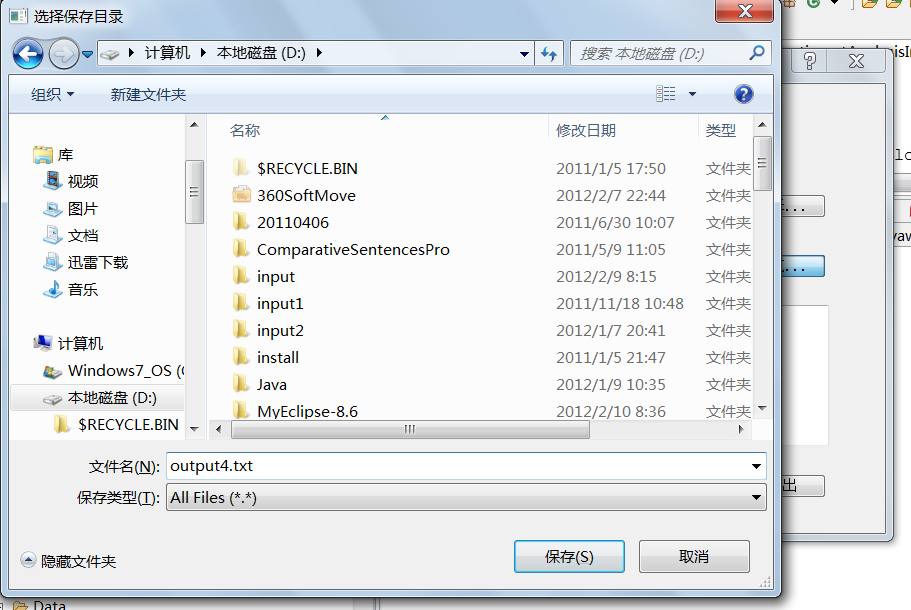


图2-3-5 选择要保存的结果文件

点击“运行”按钮，程序开始执行：

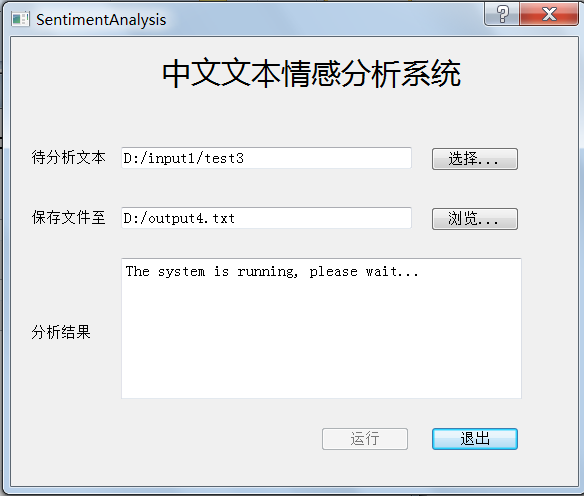


图2-3-6 程序开始执行

执行完毕显示分析结果：



图2-3-7 执行完毕

在先前制定的结果文件中，可查看具体分析内容：

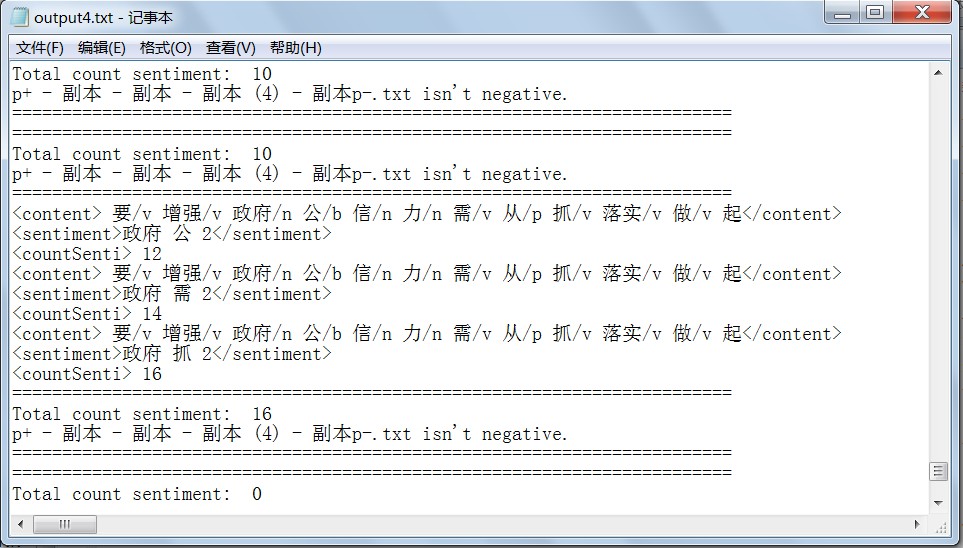


图2-3-8 具体分析内容

点击“退出”按钮，退出程序。