电商评论的情感分析

实验指导书

目录

1 实验目标

网络世界是现实世界的预先放大版，它可以提前反映出现实世界将要发生的事情，并且在情绪和情感上有放大的效果。随着电子商务的不断发展，目前大部分的电商网站是需要购买某款商品之后才可以对其进行评价的，这样，消费者对于电商商品的反馈意见更加真实，也具有明显的情感倾向，如果商家能够对用户的这些评价及时地分析，及时了解消费者对商品的情感，将有利于促进原来商品的改进和推出新款。

今天我们实验的目标是抓取某电商网站中商品的文字评论，进行数据挖掘以分析用户的情感倾向。

文本挖掘的主要原理是先将文本进行分词，然后标记分词的权重级别，基于情感的特征提取进行模式聚类分析，分别得到正面、中性、负面三类情感的数量，并将分词的权重和对应分词的情感进行综合计算，最后将数据进行可视化展示。

实验步骤：

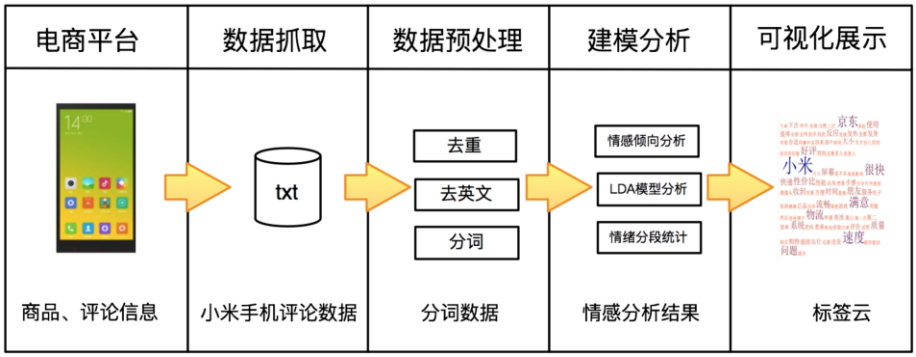


图 1 实验过程图

2 数据抓取

2.1 抓取软件

下载八爪鱼软件,链接：http://www.bazhuayu.com/download ,下载安装后需要注册账号登录到软件。

第一次进入软件中，将提示执行学习任务，可以跟着提示学习基本抓取操作。

2.2 抓取配置

首先，创建一个采集任务，在程序主界面右上方，点击高级模式中的“开始采集”安钮，如下图所示。

图 2 开始创建采集任务

在设置任务基本信息界面，输入任务名：xiaomi_jd，选择任务组为示例，然后点击“下一步”按钮，进入设计流程界面。

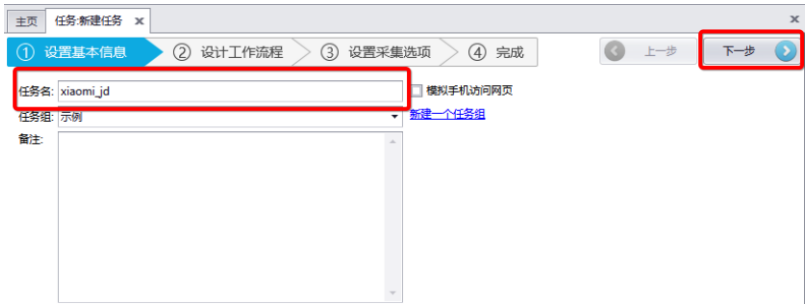


图 3 录入任务名称

在设计工作流程页面的左侧的流程设计器中，拖放一个“打开网页”操作到下图编辑器中指定位置。

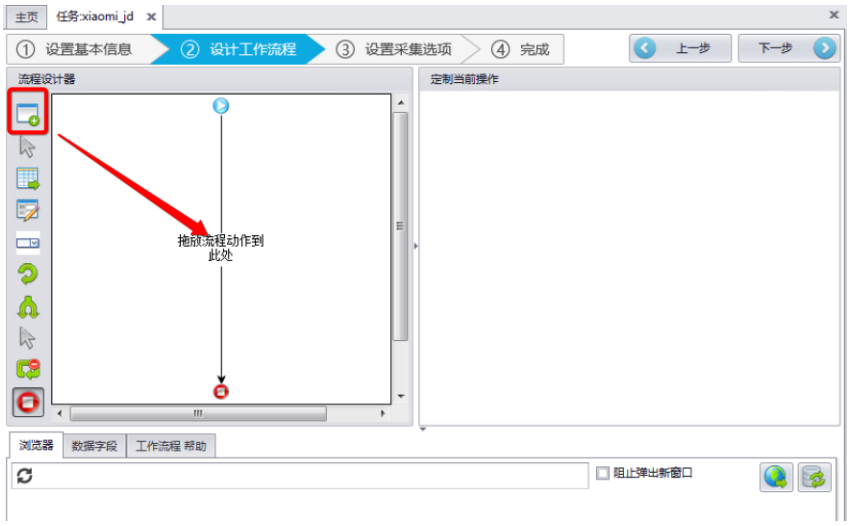


图 4 拖放打开网页操作框

在“打开网页”的页面 URL 中输入：https://list.jd.com/list.html?cat=9987,653,655&ev=exbrand%5F18374&go=0&JL=3_品牌_小米（MI），或者在京东的搜索框中输入“小米手机”，将浏览器地址栏中的连接复制到程序的页面 URL 中。点击“高级选项”，勾选“滚动页面”，输入“滚动次数”为 3,每隔 1 秒，“滚动方式”是向下滚动一屏。设置完毕后点击右下角的“保存”按钮，这时在网页预览界面将打开京东的网页。

图 5 配置打开网页

在网页预览页面，等待网页加载完成时，鼠标悬停在网页元素上时，会有选中状态指示。拖动右侧的滚动条，将网页滚动至页码分页位置，点击“下一页”，

将有动作选择界面弹出（如果不能弹出表示网页未全部加载完毕）。

图 6 开始编辑循环下一页

在弹出的动作选择弹窗中，选择“循环点击下一页”。

图 7 选择循环点击下一页

在流程设计器中选中循环翻页控件，在右侧打开“高级选项”，勾选“Ajax 加载数据”，超时时间设为 2 秒（可依实际网络环境设置），设置完成后点击“保存”按钮保存。

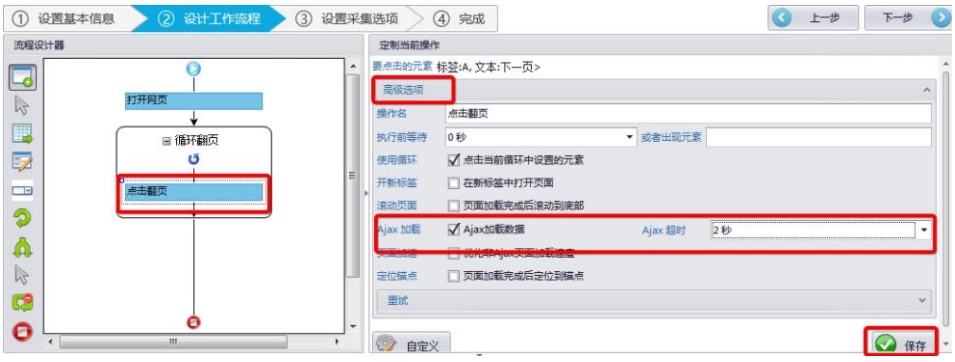


图 8 编辑翻页属性

在网页预览页面中，滚动至页面上方手机商品部分，点击第一个手机商品的名称，如“小米(MI) 红米 2 银色 联通增强版....”，在弹出框中选择“创建一个

元素列表以处理一组元素”，如下图所示。

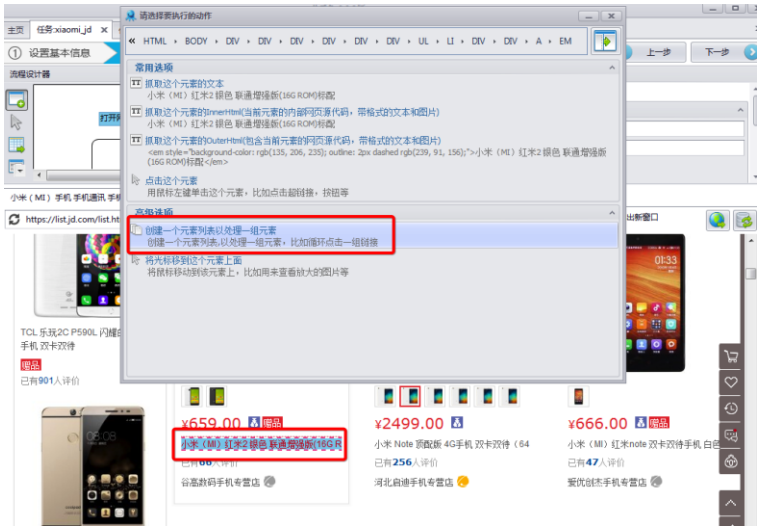


图 9 创建元素列表

在接下来弹出框中选择“添加到列表”，如下图所示。

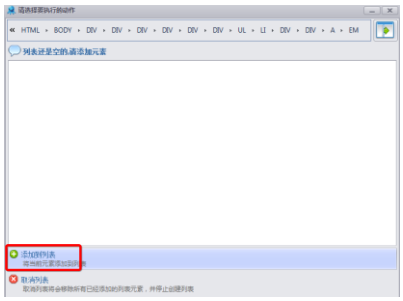


图 10 添加到列表

由于循环列表编辑需要两条商品才能确定循环规则，在接下来弹出的动作选择界面中，选择“继续编辑列表”，如下图。

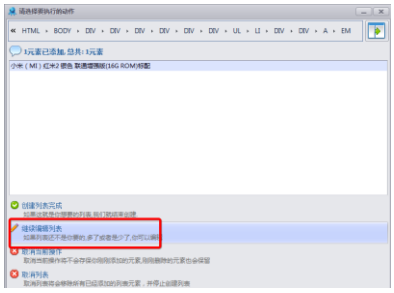


图 11 继续选择编辑列表

继续点击网页预览页面中的第二个商品的名称，在弹出框中选择“添加到列表”，这将第二个手机加入到列表中。如下图所示。

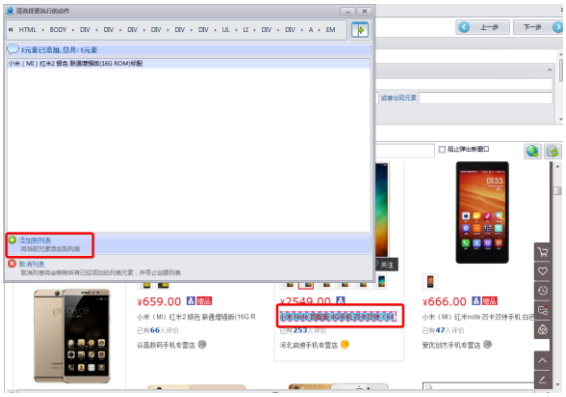


图 12 添加第二个商品到列表中

如果添加成功，可以看到动作选择窗口中，系统已经可以识别出当前页面下 59 条手机商品的标题。这时点击“创建列表完成”。

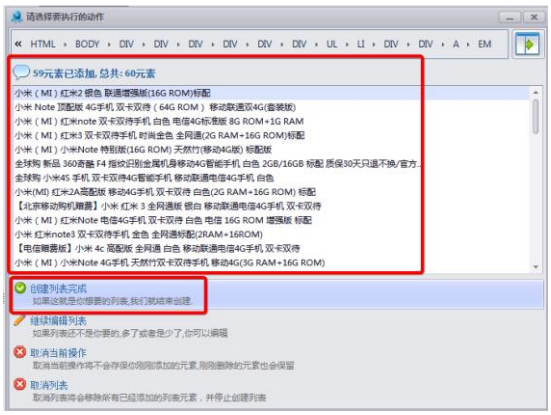


图 13 创建列表完成

在接下来的弹出窗中选择“循环”，程序将会循环点击当前页面中的第一条一直到最后一条。

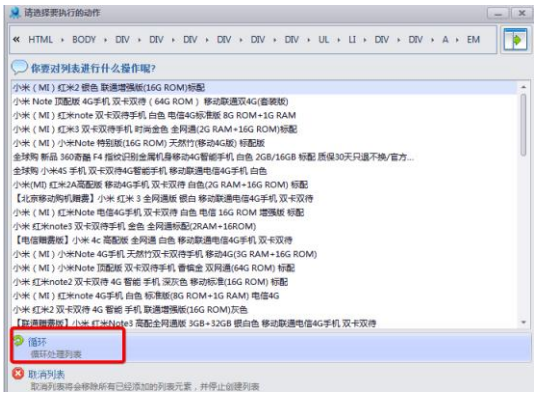


图 14 选择循环操作

完成创建循环操作之后，可以看到流程设计器中在循环翻页中增加了一个点击循环，点击“点击元素”，使其获得焦点，在右侧的高级选项中选中“滚动页面”，滚动次数为 4 次（也可以 3 次），间隔是 1 秒，滚动方式是“向下滚动一屏”。修改完毕后点击“保存”按钮。

图 15 修改点击元素属性

在网页预览页面，点击网页上的“商品评价”，在弹出动作选择窗口中选择“点击这个元素”。这样，网页中可以直接转到商品评论部分而不需要加载商品详情。

图 16 添加评论 Tab 点击动作

在左侧的流程编辑器中选中刚刚添加的“点击元素”动作，修改其右侧的高级选项，不要勾选将“在新标签中打开页面”，勾选“在页面加载完成时滚动页面”，滚动次数为 5 次，滚动方式是“向下滚动一屏”。勾选“Ajax 加载数据”，超时时间设为 2 到 5 秒之间。设置完成后点击“保存”按钮。

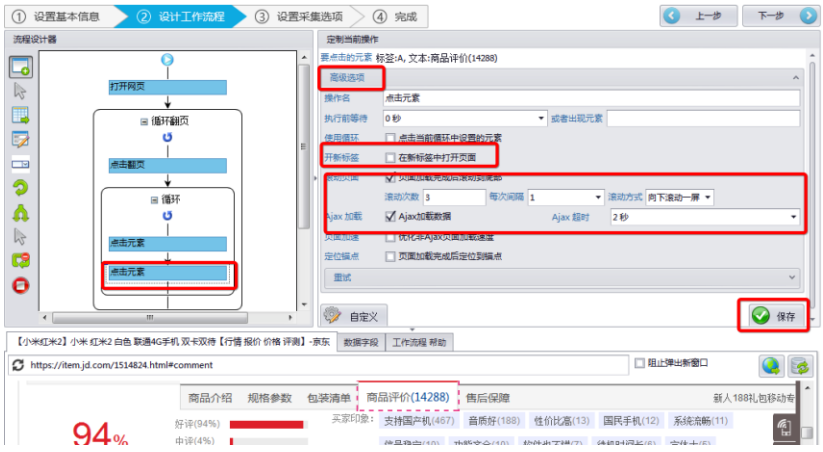


图 17 修改点击动作属性

在网页预览页面，滚到页面下方的评论分页位置，点击“下一页”，在弹出的动作选择窗口中选择“循环点击下一页”。

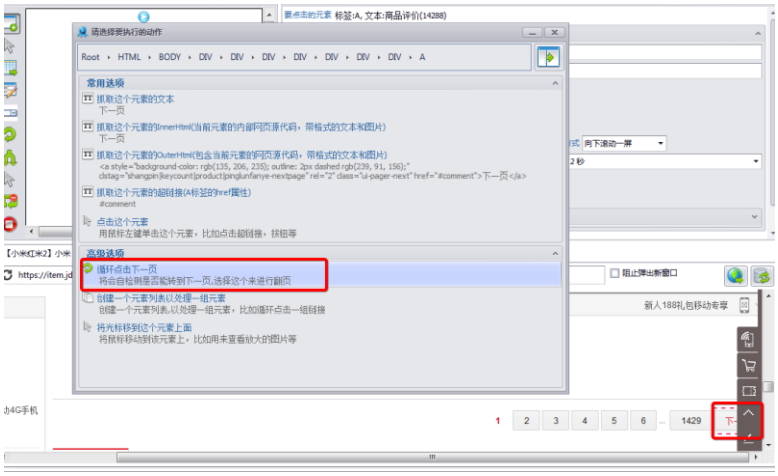


图 18 评论内容循环分页

在左侧的流程设计器中选中“点击翻页”，修改其右侧的高级选项，去掉勾选“在新标签中打开页面”，勾选“页面加载完成时滚动到底部”，点击次数为3,每隔 1 秒，滚动方式是“向下滚动一屏”，修改完毕后点击右下角的“保存按钮。

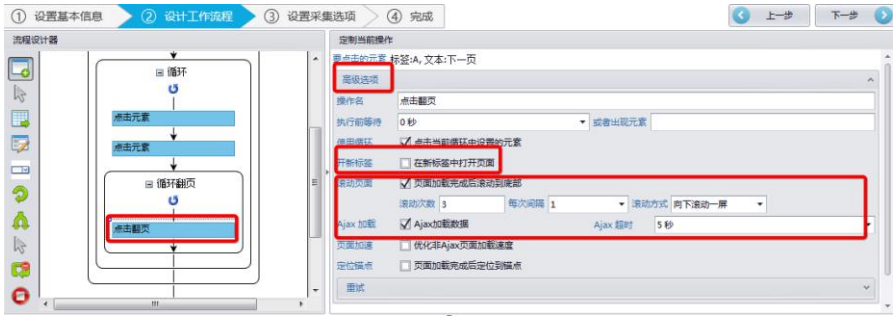


图 19 修改评论分页动作属性

在网页预览页面，点击第一条评论的内容，在弹出的动作选择窗口中选择“创建一个元素列表以处理一组元素”。

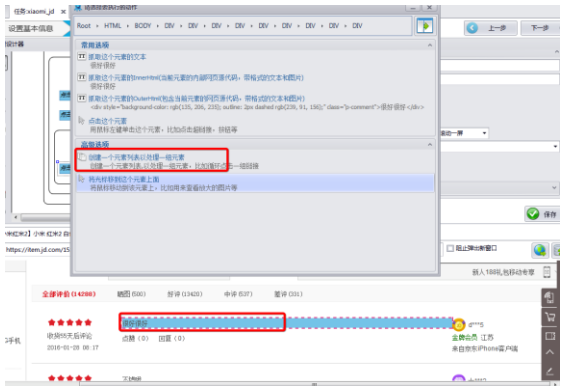


图 20 创建评论组循环处理

接下来选择“继续编辑列表”并在网页上点击第二条评论的文字内容，将其添加到列表中，添加成功之后如下图所示，点击“创建列表完成”，并选择“循环”。

图 21 创建循环评论列表

再次点击第一条评论的文字内容，在弹出的窗口中选择“抓取这个元素的文本”。如果操作成功，可以屏幕右侧看到“字段 1”中已显示网页上评论的内容。

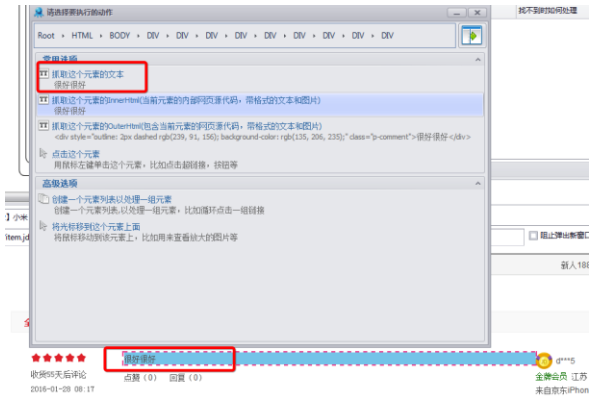


图 22 抓取评论内容

在左侧的流程设计器中点击评论的循环翻页红色圆圈位置，这样就可以查看到循环的属性信息了，如下图所示，修改循环次数为 20 次，即最多抓取 20 页评论内容。设置完成之后点击“保存”按钮。

图 23 设置评论最多抓取页数

此时，左侧流程设计器如下图所示，可以看到有 4 个循环框，分别是商品列表分页、商品循环点击进入详情、评论内容分页、循环抓取评论内容。其中点击分页按钮的操作需要放在循环之后，否则将会使用第一个页面内容忽略，需要拖动其位置来调整他们的顺序，按下图箭头所指的方向和位置调整。

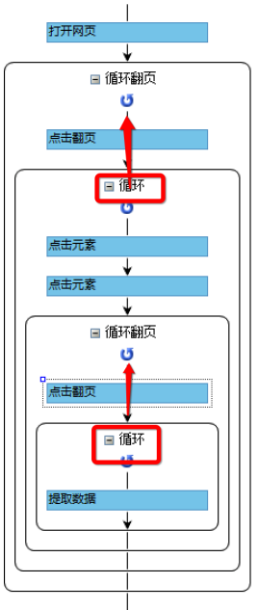


图 24 调整循环的位置

调整过程中可能会有出错提示自动修复，选择“确定”即可。调整结束之后，流程设计器中将如下图所示。

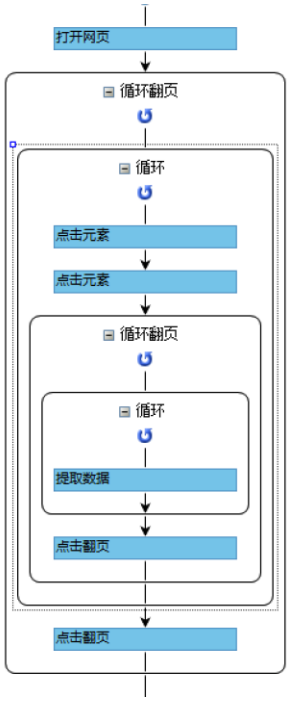


图 25 目标抓取流程

调整成功之后点击右上方的“下一步”按钮，如下图所示。不必改动红框中

的选项，直接再次点击“下一步”按钮。

图 26 设置采集选项

成功完成任务定义之后，如下图所示，点击“启动单机采集”安钮，开始从本机访问京东网站采集评论内容。

图 27 启动单机采集

采集任务正常运行的情况下，可以看到任务界面的网页内容在发生变化，当网页跳转到商品评论页面时，可以看到提取到的数据。每次分页会采集 10 条数据，总的数据量和重复数据量在窗口标题位置可以看到，当发现已采集数量增加 10 条，同时重复数量也增加 10 条时，说明网页中评论的分页出现问题了，需要设置前面提到的翻页 Ajax 的超时间或分页循环的循环次数。如果需要重新修改配置，关闭任务运行窗口即可停止运行。或者在采集到数据超过 500 条时即可以点击停止进行后续情感分析。

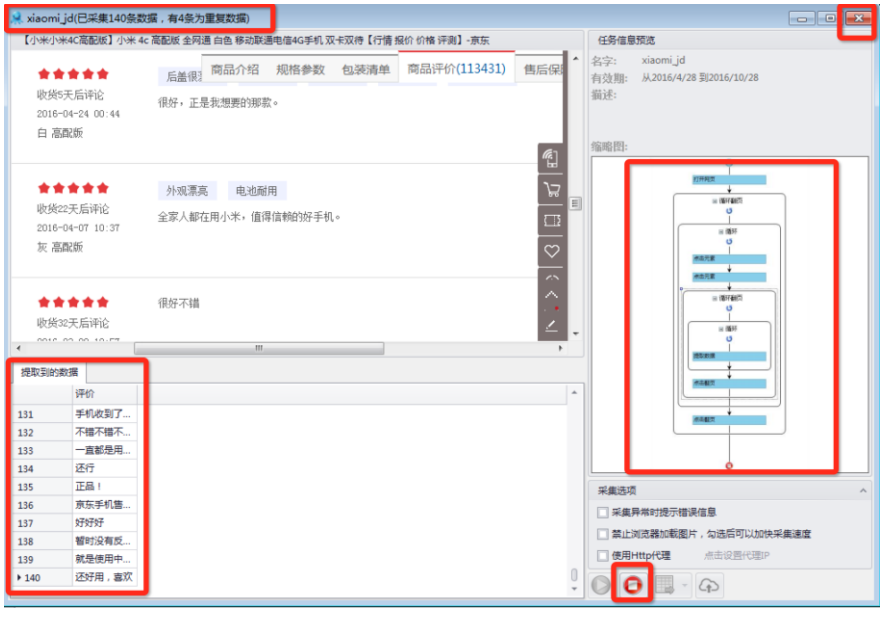


图 28 任务运行界面

任务运行结束或数据量超过 1000 条时可以点击右下角的红包停止按钮以停止单机采集，点击“导出”按钮导出为 txt 格式文本。

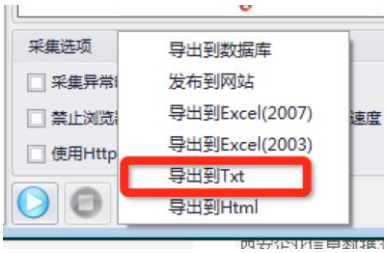


图 29 导出评论

导出成功后即可退出任务和程序。

备注：如果前述操作无法配置成功，无法抓取到数据，可以直接使用实验目录下的 xiaomi_jd_副本.otd 模板任务，导入到程序中即可抓取数据。或者直接使用抓取结果 xiaomi_jd.txt 来进行后续实验。

3 情感分析

3.1 软件介绍

ROST CM 6 是武汉大学沈阳教授研发的社会计算平台。该软件可以实现微博分析、聊天分析、全网分析、网站分析、浏览分析、分词、词频统计、英文词频统计、流量分析、聚类分析等一系列文本分析。本部分采用 ROSTCM6 来处理商品的评论去重、情感分析。

图 30 ROSTCM6 软件主界面

3.2 去除重复

前面抓取的数据文件为 xiaomi_jd.txt，由于在抓取过程中会因为用户评论内容重复或者是程序抓取的问题导致有较多重复评论，需要在情感分析之前将重复评论去除，操作方法：选择“文本处理”菜单中的“一般化处理”，在待处理文件输入框中，选择 xiaomi_jd.txt，在处理条件处选择“凡有重复的行，只保留一行”。如下图所示。

图 31 去除重复的行

执行完成后，程序自动弹出处理后的文档，可以看到重复的行已经全部去除。

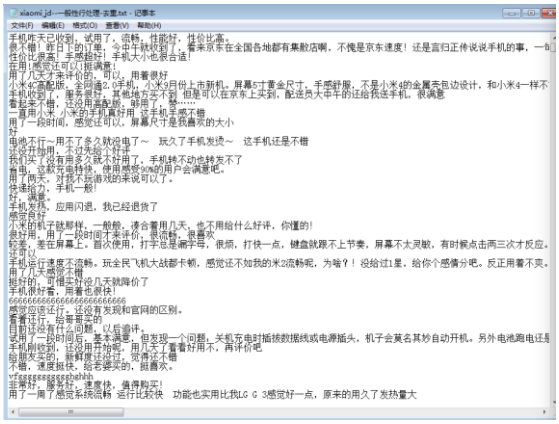


图 32 去除重复行结果

将处理后的去重结果文件关闭，然后在前面打开的一般化处理界面中选择上一步去重处理后的结果文件“xiaomi_jd—一般性行处理-去重.txt”作为待处理文件，在处理条件处选择“把所有行中包含的英文字符全部删掉”，其它选项不变，然后点击“确定”执行操作。

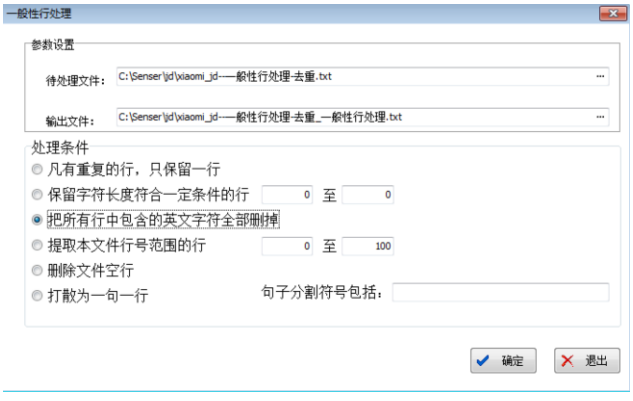


图 33 删除英文字符

处理结束后可以显示处理后的结果，可以看到所有的英文字符已经去掉。

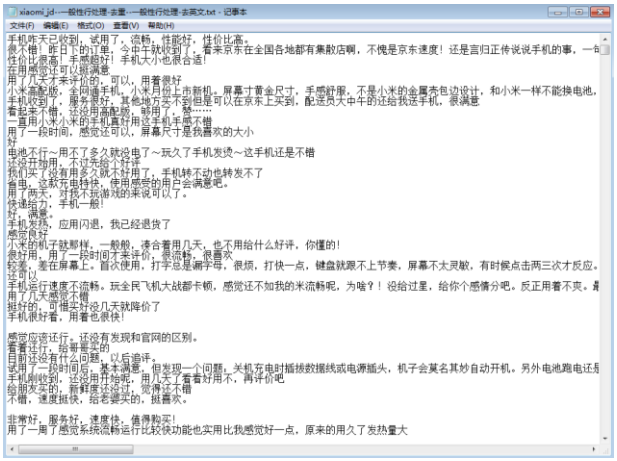


图 34 去除英文字符后的结果

3.3 分词

在主菜单“功能性分析”中选择“分词”，如下图所示。

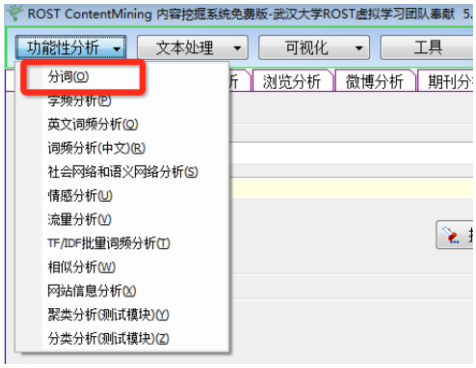


图 35 选择分词操作

将已经去除重复和去掉英文的文件作为待处理文件进行分词操作，选定之后点击“确定”按钮。

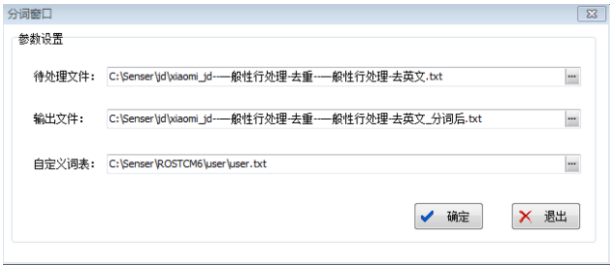


图 36 对评论内容进行分词

分词操作运行结束后自动打开分词后的结果文件，如下图所示。

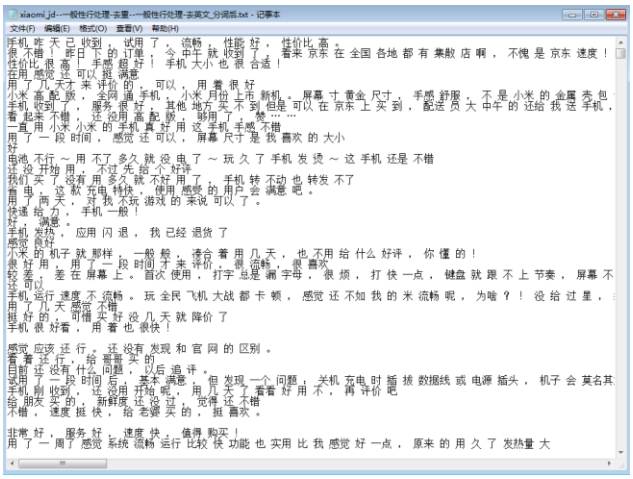


图 37 分词后结果文件

在主菜单“功能性分析”选择“词频分析”菜单，在弹出的词频统计窗口中将上一步操作中的分词结果文件作为待统计文件源，其它选项保持默认，点击“确定”按钮执行操作。

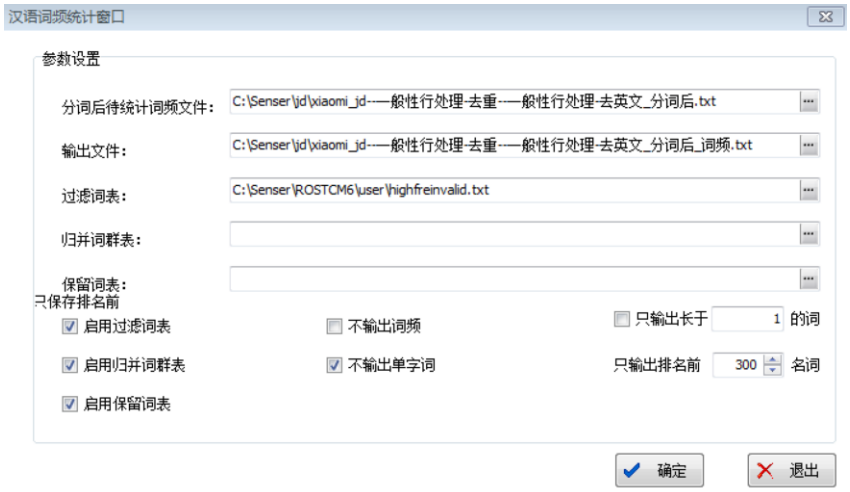
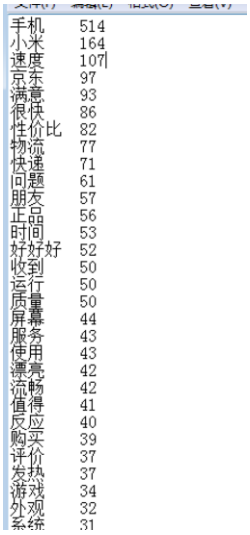


图 38 词频统计

词频统计结束后将弹出统计结果文件，如下图所示，可以发现其中除了“京东”，“小米”、“手机”之外，包括了更多的情感词汇，如“满意”、“很快”等。这些分词的词频代表了手机用户的关注点。这些统计后的结果可以在后续的数据可视化中使用标签云的方式直观地展示出来。



手机	514
小米	164
速度	107
京东	97
满意	93
很快	86
性价比	82
物流	77
快递	71
问题	61
朋友	57
正品	56
时间	53
好好好	52
收到	50
运行	50
质量	50
屏幕	44
服务	43
使用	43
漂亮	42
流畅	42
值得	41
反应	40
购买	39
评价	37
发热	37
游戏	34
外观	32
系统	31

图 39 词频结果统计

3.4 情感分析

在主菜单“功能性分析”中，选择“情感分析”菜单，将前面分词的结果文件作为待分析文件。（注意：是分词后的文件，而不是词频统计文件），其它选项不变，点击“分析”按钮执行操作。

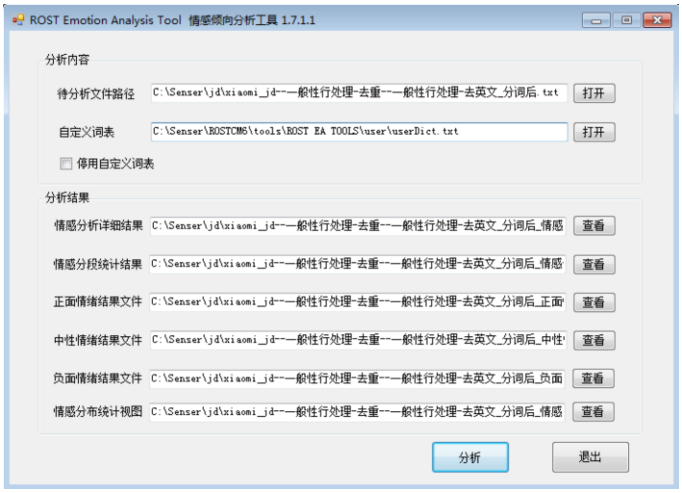


图 40 情感分析

分析结束后，可以点击分析结果文件路径右侧的“查看”按钮查看结果。下图分别是情感分析分断统计结果和情感分布视图。其中前者表示积极、中极、消极情绪所占的条数和比例，以及各情绪分段统计的情况，可以看出某一情绪的强烈程度。后者仅显示三种情绪所占的比例。

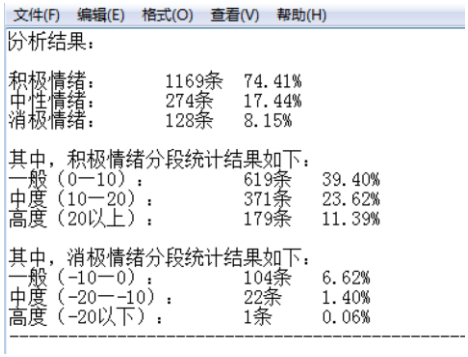


图 41 情感分析分段统计结果

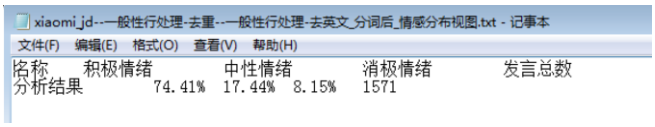


图 42 情感分布视图

3.5 云标签视图

在程序主菜单中，选择“可视化”菜单下的“标签云”菜单，在弹出窗口中，点击“打开”按钮，打开前面处理好的词频统计文件(xiaomi_jd--一般性行处理-去重--一般性行处理-去英文_分词后_词频.txt)，去掉其中的“手机”、“小米”、“京东”这几个非情感词，然后点击“显示”按钮，并且调整“最大字体”的大小，可以看到以下标签云显示。可以清楚看到“满意”、“流畅”等正向词汇，也可以看到“反应发热”、“电池”等负面词汇，说明了用户的关注焦点。

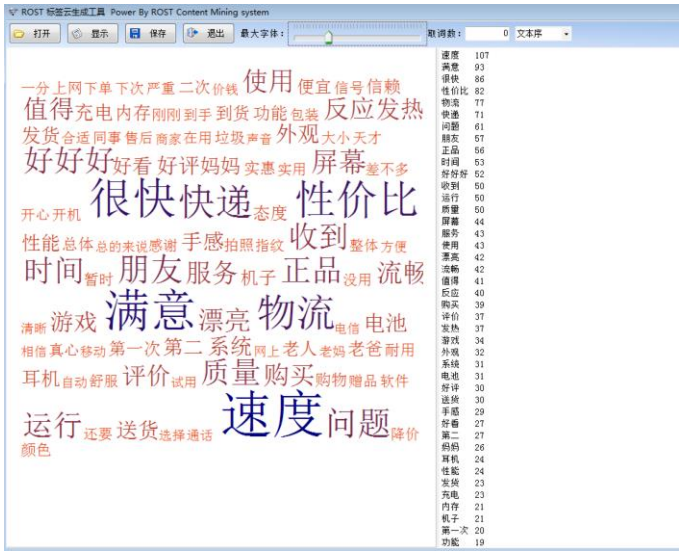


图 43 标签云视图

另外一些词汇如“速度”、“评价”等需要结合上下文进行查看用户的意思。

3.6 实验结果总结

我们从情感分析和词频统计的结果可以看到大部分用户对于小米手机还是积极的情绪，说明小米手机对于用户来说还是被认可的，从云标签中可以看出手机用户喜欢的是“性价比高”，主要是用于购买给父母、朋友等。但因为积极情绪的比例未能超过 80%，说明小米手机的用户认可度还需要加强，特别是云标签图中所显示出来较大的负面关键词，希望小米手机能够在电池发热等方面进行加强。