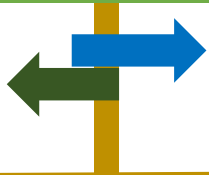


基于微博爬虫的輿情分析

作者 李玉珍

学校 东北大学

电话 18640376585



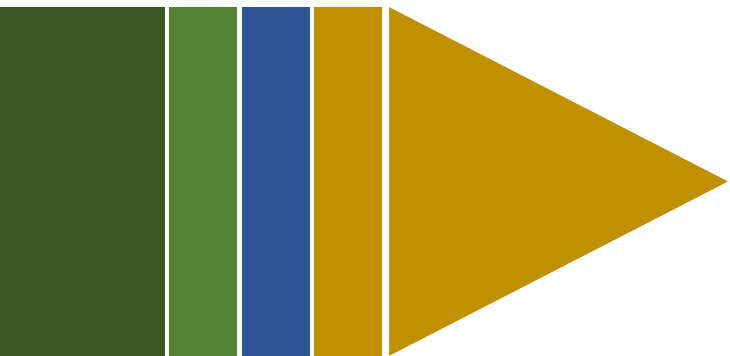
主要内容

1.爬虫与舆情分析的主要流程

2. 爬虫效率与性能分析

3.项目中的主要工具、问题及解决方式

4.爬虫及舆情分析结果展示与代码简介



1.爬虫与舆情分析的主要流程

1.1爬虫切入点

本课题主要对新浪微博进行爬虫，但是遗憾的是新浪微博并没有提供以“关键字+时间+区域”方式获取官方API。但是庆幸的是，新浪提供了高级搜索功能。

A screenshot of the Weibo Advanced Search interface. At the top, there's a title bar "微博高级搜索" with a close button. Below it, the "关键词" (Keywords) field contains "招商银行". The "类型" (Type) section has radio buttons for "全部" (selected), "热门", "原创", "关注人", "认证用户", and "媒体". The "包含" (Include) section has radio buttons for "全部" (selected), "含图片", "含视频", "含音乐", and "含短链". The "时间" (Time) section shows a date range from "2017-05-02" to "2017-05-02", with dropdown menus for "选择时间". The "地点" (Location) section has dropdown menus for "省/直辖市" and "城市/地区". At the bottom, there are two buttons: "搜索微博" (Search Weibo) in green and "取消" (Cancel) in blue.

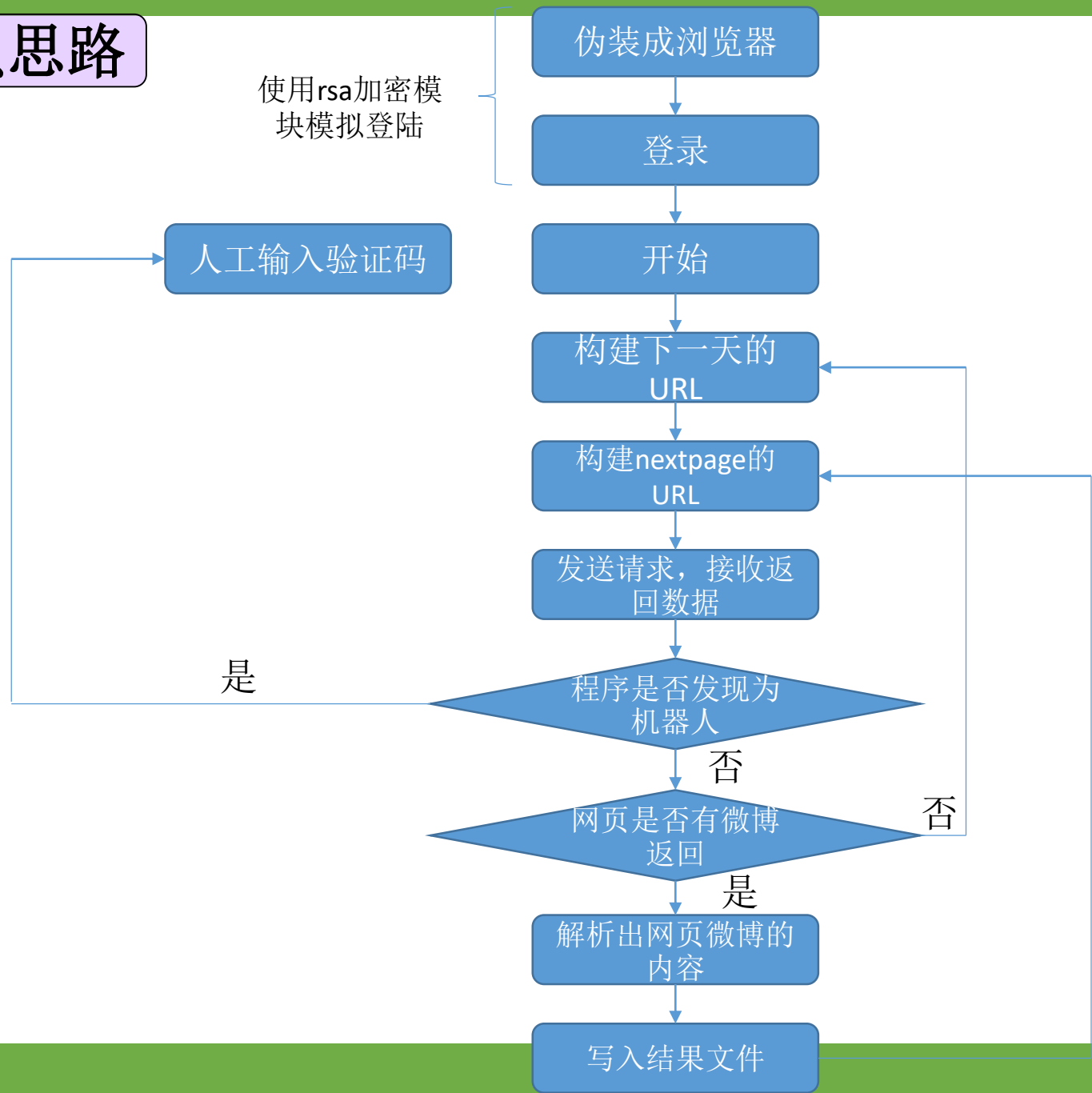
点击搜索微博后，我们看地址栏：

<http://s.weibo.com/weibo/%25E6%258B%259B%25E5%2595%2586%25E9%2593%25B6%25E8%25A1%258C&typeall=1&suball=1×cope=custom:2017-05-02:2017-05-02&Refer=g>

| | |
|---------------------|--|
| 解析如下固定地址部分： | http://s.weibo.com/wb/ |
| 关键字（2次URLEncode编码）： | %25E6%258B%259B%25E5%2595%2586%25E9%2593%25B6%25E8%25A1%258C |
| 搜索时间范围： | timescope=custom:2017-05-02:2017-05-02 |
| 可忽略项： | Refer=g |
| 某次请求的页数（未出现）： | page=1（某页请求页数） |

既然是这么回事，我们接下来就可以使用网页爬虫的方式获取“关键字+时间”的微博了.....

1.2爬虫思路



思路:

本课题所采用的爬虫语言是Python。在对新浪微博进行爬虫之前，首先需要模拟登陆，这里所采用的办法是：使用rsa加密模块进行模拟登陆。

接下来要构造URL，爬取网页，然后解析网页中的微博信息，如图所示。

另外，高级搜索最多返回50页微博。时间范围（timescope）可设置为1天，如2017-05-02:201-05-02。

1.3 舆情分析

信息抽取

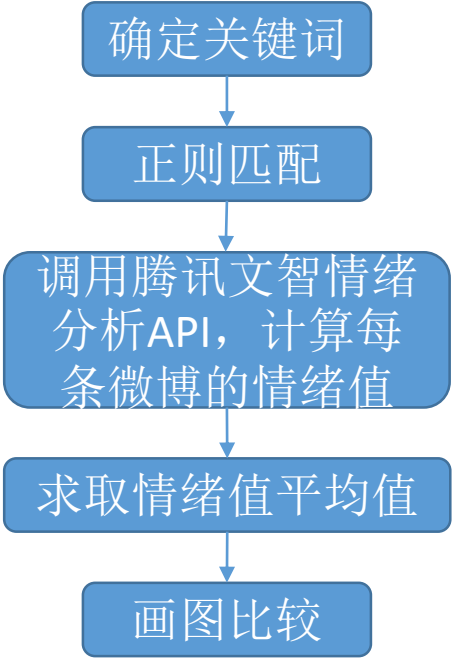
爬虫得到微博信息存储在weiboData.xls这个EXCEL文件中，我抽取的是5017-05-02开始的最近10天的信息，一共691条微博信息。要想进行舆情分析，就必须对爬虫信息进行抽取。我通过关键词正则匹配的方式，从爬虫得到的信息中抽取了和招行相关相关的服务，黑金卡、信用卡、手机银行、一季度盈利额超过交行、支持银联二维码支付等重点信息。

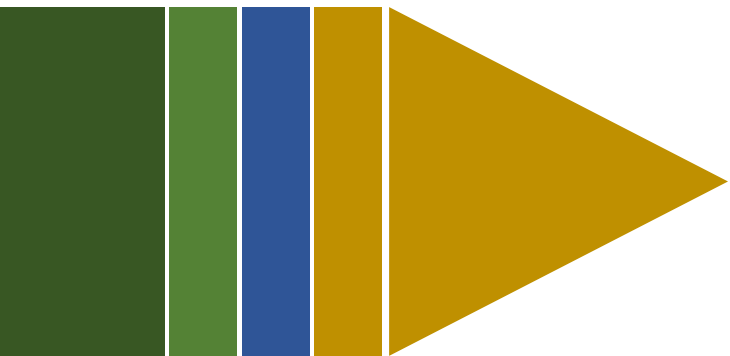
情绪分析

情绪分析算法过于复杂，短时间内，做不来。这个关键在于词典，是找现成的：

- 1. 中文情感极性词典 NTUSD
- 2. 大连理工情感词汇本体库

后来选择了大连理工的词汇库。然后把要分析的词在词库中匹配，计算，但是发现效果很差，后来废弃了所有东西，直接使用腾讯文智的情感分析服务（收费）。





2.爬虫效率与性能分析

2.1爬虫性能

1、由于第一次接触网络爬虫这方面的知识，在图书馆借了很多书，查了很多资料，最后终于短时间内实现了微博爬虫。但是遗憾的是，我仅仅做到了实现，并没有进行优化，所以爬虫的效率性能都比较差。计算了一下，大概每分钟几百条（为了防止被新浪网站发现爬虫，每次网页请求之间设置了休眠，所以用时比较长。）

2、接下来有时间准备研究一下分布式爬虫、多线程爬虫等，这样优化后一定能提高爬虫性能和效率。

A decorative graphic on the left side of the slide. It consists of four vertical bars of different colors: dark green, light green, blue, and yellow. To the right of these bars is a large yellow triangle pointing to the right.

3.项目中的主要工具、问题及解决方式

3.1 项目主要用到的工具

Python

Python是我爬虫用的主要编程语言。

RSA加密算法模块

RSA加密模块主要用来模拟登陆新浪微博。

xlwt , xlwd 模块

这两个模块主要用于向EXCEL中读写。

Matplotlib 模块

这个模块主要用于对舆情分析的结果进行可视化分析。

腾讯文智情感分析 API

在对微博进行情绪判定的时候，调用这个API。

3.1 项目主要用到的工具

腾讯文智情感分析API

在对微博进行情绪判定的时候，调用这个API。调用这个API后，我们可以得到所分析内容的正面情绪和负面情绪分别是多少。然后对相关微博的情绪值求平均，就可以得到对应微博信息的**正面和负面**的情绪值。

腾讯文智情感分析API，初次使用，可以申请免费试用。具体用法，请参考如下官网：

- 1.<https://www.qqcloud.com/document/product/271/2072>
- 2.<https://www.qqcloud.com/document/developer-resource/494/7244>

2. 输入参数

| 参数名称 | 必选 | 类型 | 描述 |
|---------|----|--------|-------------------------------------|
| content | 是 | String | 待分析的文本（只能为utf8编码） |
| type | 是 | Int | （可选参数，默认为4）1：电商；2：APP；3：美食；4：酒店和其他。 |

3. 输出参数

| 参数名称 | 类型 | 描述 |
|----------|--------|-----------------|
| code | Int32 | 错误码。0：成功，其他值：失败 |
| message | String | 错误信息 |
| positive | Double | 正面情感概率 |
| negative | Double | 负面情感概率 |

输入：

```
https://wenzhi.api.qqcloud.com/v2/index.php?
    Action=TextSentiment
    &Nonce=345122
    &Region=sz
    &SecretId=AKIDz8krbsJ5yKBZQpn74WFkmLPx3gnPhESA
    &Timestamp=1408704141
    &Signature=HgIYOPcx5lN6gz8JsCFBNAWp2oQ
    &content=双万兆服务器就是好，只是内存小点
```

输出：

```
{
  "code": 0,
  "message": "",
  "negative": 0.138263002038002,
  "positive": 0.8617370128631592
}
```

3.2主要问题及解决方式

爬虫失败，无法获得更多微博内容

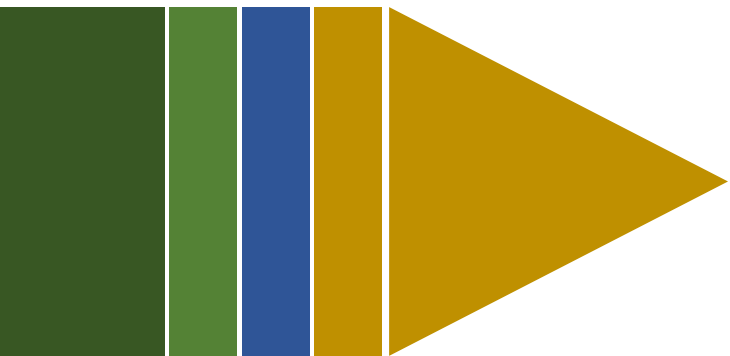
查资料后发现，新浪官方网站有反爬虫机制，所以在登录的时候，一定要伪装成浏览器。伪装成浏览器后，问题解决。

从抓到的微博中抽取重点内容失败

在进行内容抽取的时候，发现正则表达式对中文汉字并不适用。查资料后，发现可以对汉字进行Unicode编码，经过编码后就可以进行正则匹配了。以关键词“服务”为例，其Unicode编码为670d52a1，正则表达式为：
`pattern = re.compile(u"(\u670d\u52a1)+")`

情绪判定失败

在情绪判定的时候，我选择了在大连理工情感词汇本体库，但是由于词库，词不够全，以及我自己算法的一些问题，获得的结果很差。后来查资料后，发现，腾讯有腾讯文智情感分析API，新手可以获得免费调用机会。按照官方文档，调用后，成功就算出每条微博的正面情绪和负面情绪。



4.爬虫及舆情分析结果展示与代码简介

4.1 爬虫结果

运行mymain.py文件后，按照提示依次输入：关键词、爬虫起始时间、最近N天信息的N，如下：

```
E:\scrapy>mymain.py
Login Succeed!
Enter the keyword(type 'quit' to exit ):招商银行
Enter the start time(Format:YYYY-mm-dd):2017-05-02
Enter the recent N days messages:10
```

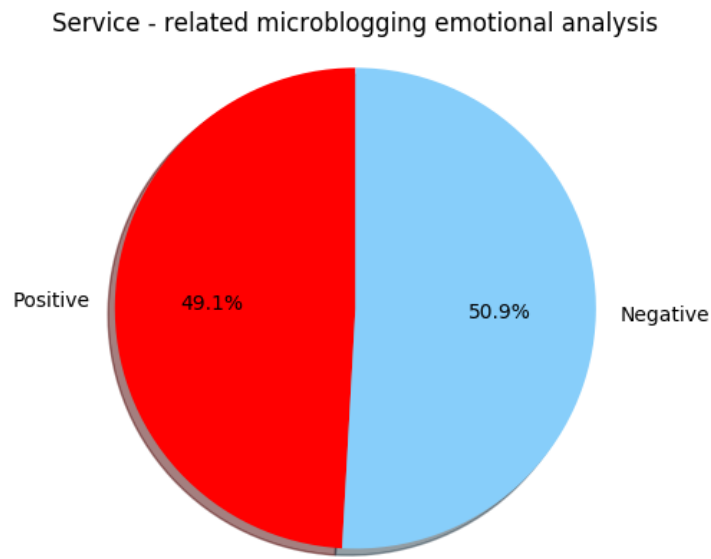
爬虫获得的微博信息写入weiboData.xls文件中，一共获得691条微博：

[illegible]

4.2 舆情分析结果

招行服务相关微博舆情分析

| 服务相关微博舆情 | 条数 | 总正面情绪 | 总负面情绪 | 平均正面情绪 | 平均负面情绪 |
|----------|----|--------|--------|--------|--------|
| | 36 | 17.681 | 18.319 | 0.4911 | 0.5089 |



结论

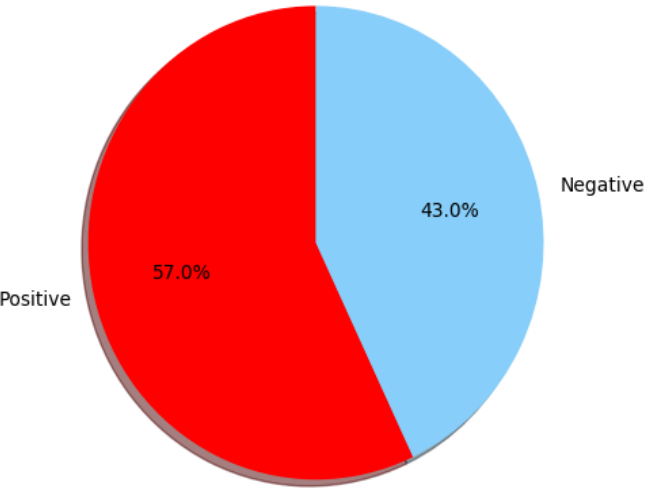
从分析结果来看，顾客对招行服务的正面情绪为49.1%，负面情绪为50.9%，负面情绪比正面情绪多了1.8个百分点，所以，我们招行的服务态度还是有待改善。

4.2 舆情分析结果

招行产品黑金卡相关微博舆情分析

| 黑金卡相关微博舆情 | 条数 | 总正面情绪 | 总负面情绪 | 平均正面情绪 | 平均负面情绪 |
|-----------|----|--------|-------|--------|--------|
| | 21 | 11.973 | 9.027 | 0.570 | 0.430 |

Black_gold_card - related microblogging emotional analysis



结论

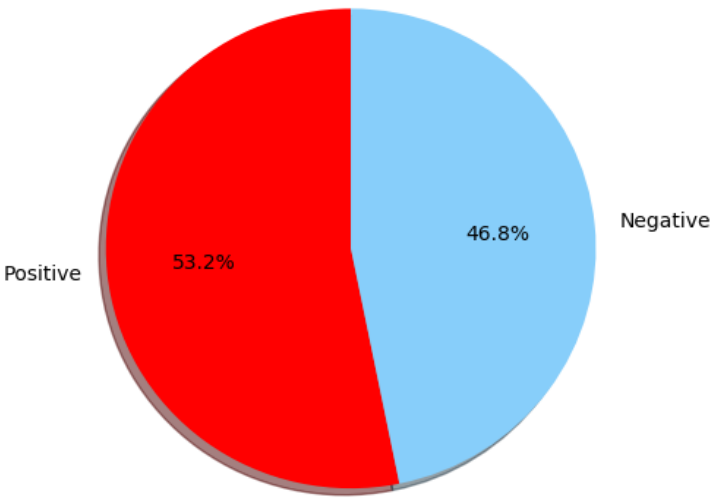
从分析结果来看，顾客对招行产品黑金卡的正面态度57.0%，负面情绪为43.0%，正面情绪比负面情绪多了14个百分点，所以，说明顾客总体对我们招行的黑金卡还是比较满意的，可以长期广泛推广销售。

4.2 舆情分析结果

招行产品信用卡相关微博舆情分析

| 信用卡相关微博舆情 | 条数 | 总正面情绪 | 总负面情绪 | 平均正面情绪 | 平均负面情绪 |
|-----------|----|--------|--------|--------|--------|
| | 85 | 45.197 | 39.803 | 0.532 | 0.468 |

Creditcard - related microblogging emotional analysis



结论

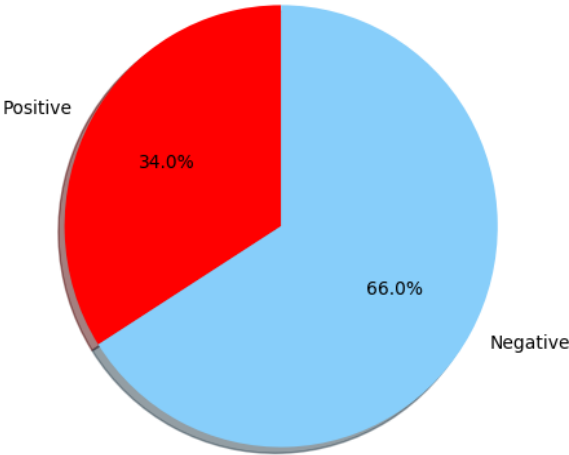
从分析结果来看，顾客对招行产品信用卡的正面情绪53.2%，负面情绪为46.8%，正面情绪比负面情绪多了6.4个百分点，所以，说明顾客总体对我们招行的信用卡还是比较满意的，可以长期广泛推广销售。

4.2 舆情分析结果

招行产品手机银行相关微博舆情分析

| 手机银行相关微博舆情 | 条数 | 总正面情绪 | 总负面情绪 | 平均正面情绪 | 平均负面情绪 |
|------------|----|-------|-------|--------|--------|
| | 4 | 1.360 | 2.640 | 0.340 | 0.660 |

Phone_bank - related microblogging emotional analysis



结论

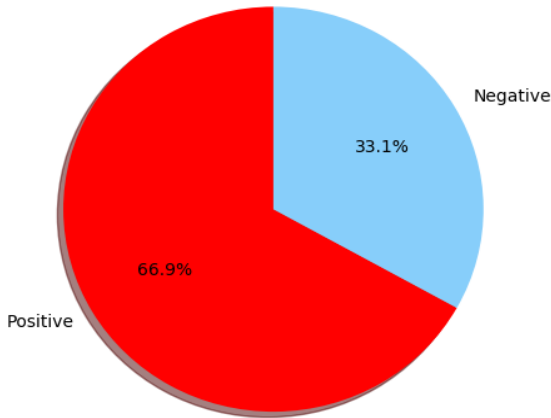
从分析结果来看，顾客对招行产品手机银行的正面情绪为34.0%，负面情绪为66.0%，负面情绪比正面情绪多了32个百分点，所以，说明顾客总体对我们招行的手机银行是非常不满意，手机银行产品有待改善。

4.2 舆情分析结果

招行事件一季度盈利额超过交行相关微博舆情分析

| 一季度盈利额 相关微博舆情 | 条数 | 总正面情绪 | 总负面情绪 | 平均正面情绪 | 平均负面情绪 |
|------------------|----|-------|-------|--------|--------|
| | 6 | 4.014 | 1.986 | 0.669 | 0.331 |

Profitability - related microblogging emotional analysis



结论

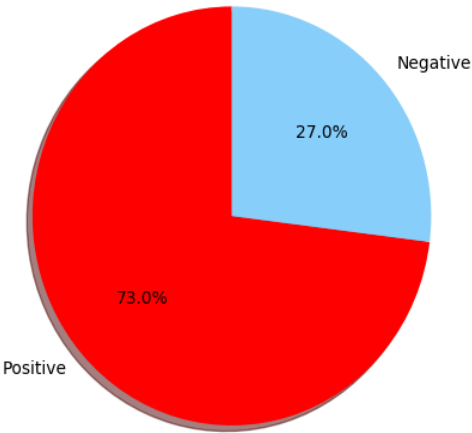
从分析结果来看，顾客对招行一季度盈利额超过交行的正面情绪为66.9%，负面情绪为33.1%，正面情绪比负面情绪多了33.8个百分点，所以，说明顾客总体对我们招行的一季度盈利额超过交行这件事是很称赞的，祝愿我行越来越好，早日超过四大行。

4.2 舆情分析结果

招行事件支持银联二维码支付相关微博舆情分析

| 支持二维码支付 相关微博舆情 | 条数 | 总正面情绪 | 总负面情绪 | 平均正面情绪 | 平均负面情绪 |
|-------------------|----|-------|-------|--------|--------|
| | 1 | 0.730 | 0.270 | 0.730 | 0.270 |

Code_payment - related microblogging emotional analysis



结论

从分析结果来看，顾客对招行事件支持银联二维码支付的正面情绪为73.0%，负面情绪为27.0%，正面情绪比负面情绪多了46个百分点，所以，说明顾客总体对我们招行支持银联二维码支付这件事是很称赞的，祝愿我行能推出更多优秀便捷的产品。





4.3代码简介

所有的源代码以及结果放在所提交的analysis_result这个文件夹下，这个文件夹下存放三个文件夹如下：

| 名称 | 修改日期 | 类型 | 大小 |
|--|----------------|-----|----|
|  Crawler_source_code_result | 2017/5/5 8:30 | 文件夹 | |
|  Data_analysis_results | 2017/5/5 17:09 | 文件夹 | |
|  Data_analysis_source_code | 2017/5/4 13:13 | 文件夹 | |

从上到下依次存放着爬虫源代码以及结果、数据分析结果、数据分析源代码。












Crawler_source_code_result 文件下：

| | | | |
|--|----------------|---------------------|--------|
|  collectWeiboDataByKeyword.py | 2017/5/3 9:01 | Python File | 15 KB |
|  login.py | 2016/2/17 4:16 | Python File | 4 KB |
|  mymain.py | 2017/5/2 17:46 | Python File | 1 KB |
|  weiboData.xls | 2017/5/3 18:11 | Microsoft Excel ... | 230 KB |

从上到下依次存放着爬虫模块、模拟登陆模块、主程序模块、爬虫结果。准备爬虫的时候运行mymain.py文件即可。

4.3代码简介

Data_analysis_source_code 文件下：

| | | | |
|---|----------------|--------------|------|
|  qcloudapi-sdk-python-master | 2017/3/1 1:47 | 文件夹 | |
|  src | 2017/5/4 10:18 | 文件夹 | |
|  .gitignore | 2017/3/1 1:47 | GITIGNORE 文件 | 1 KB |
|  analysisplot.py | 2017/5/5 15:57 | Python File | 4 KB |
|  black_gold_card.py | 2017/5/4 10:32 | Python File | 4 KB |
|  Code_Payment.py | 2017/5/4 11:06 | Python File | 3 KB |
|  creditcard.py | 2017/5/4 10:41 | Python File | 4 KB |
|  demo.py | 2017/3/1 1:47 | Python File | 1 KB |
|  phonebank.py | 2017/5/4 10:50 | Python File | 3 KB |
|  README.md | 2017/3/1 1:47 | MD 文件 | 4 KB |
|  serveranalysis.py | 2017/5/4 10:06 | Python File | 4 KB |

第一个文件夹存放着腾讯文智情绪分析API模块，analysisplot.py 是画图模块，其余的文件是和招行相关的产品、服务、重点事件的抽取和情绪值分析。

Data_analysis_results 文件下：

| 名称 | 修改日期 | 类型 | 大小 |
|--|----------------|---------------------|-------|
|  Black_gold_card.png | 2017/5/4 13:12 | PNG 文件 | 29 KB |
|  black_gold_card.xls | 2017/5/4 10:32 | Microsoft Excel ... | 33 KB |
|  Code_payment.png | 2017/5/4 13:12 | PNG 文件 | 27 KB |
|  Code_Payment.xls | 2017/5/4 11:09 | Microsoft Excel ... | 26 KB |
|  Creditcard.png | 2017/5/4 13:12 | PNG 文件 | 28 KB |
|  creditcard.xls | 2017/5/4 10:47 | Microsoft Excel ... | 51 KB |
|  Phone_bank.png | 2017/5/4 13:12 | PNG 文件 | 28 KB |
|  phonebank.xls | 2017/5/4 10:55 | Microsoft Excel ... | 27 KB |
|  Profitability.png | 2017/5/4 13:12 | PNG 文件 | 28 KB |
|  server.png | 2017/5/4 13:12 | PNG 文件 | 27 KB |
|  server.xls | 2017/5/4 10:16 | Microsoft Excel ... | 39 KB |

其中EXCEL文件是和招行相关的产品、服务、重点事件的抽取结果的存放；图片是和招行相关的产品、服务、重点事件情绪分析的结果展示。

参考文献

1. <https://sanwen8.cn/p/415Cgz9.html>
2. <http://dataunion.org/24057.html>
3. <http://blog.csdn.net/amyque/article/details/50933143>
4. <http://blog.csdn.net/gatieme/article/details/43235791>
5. <https://www.qcloud.com/document/product/271/2072>
6. <https://www.qcloud.com/document/developer-resource/494/7244>
7. <https://www.anotherhome.net/2920>
8. <http://m.blog.csdn.net/article/details?id=38149451>

**谢谢，欢迎老师批
评指正！**