
微博营销数字化分析初探

—R语言

韩锷春
<http://www.DataApple.net>

议程

社交媒体营销 (**Social Media Marketing**)

1

微博情感分析 (**Sentiment Analysis**)

2

例子展示

3

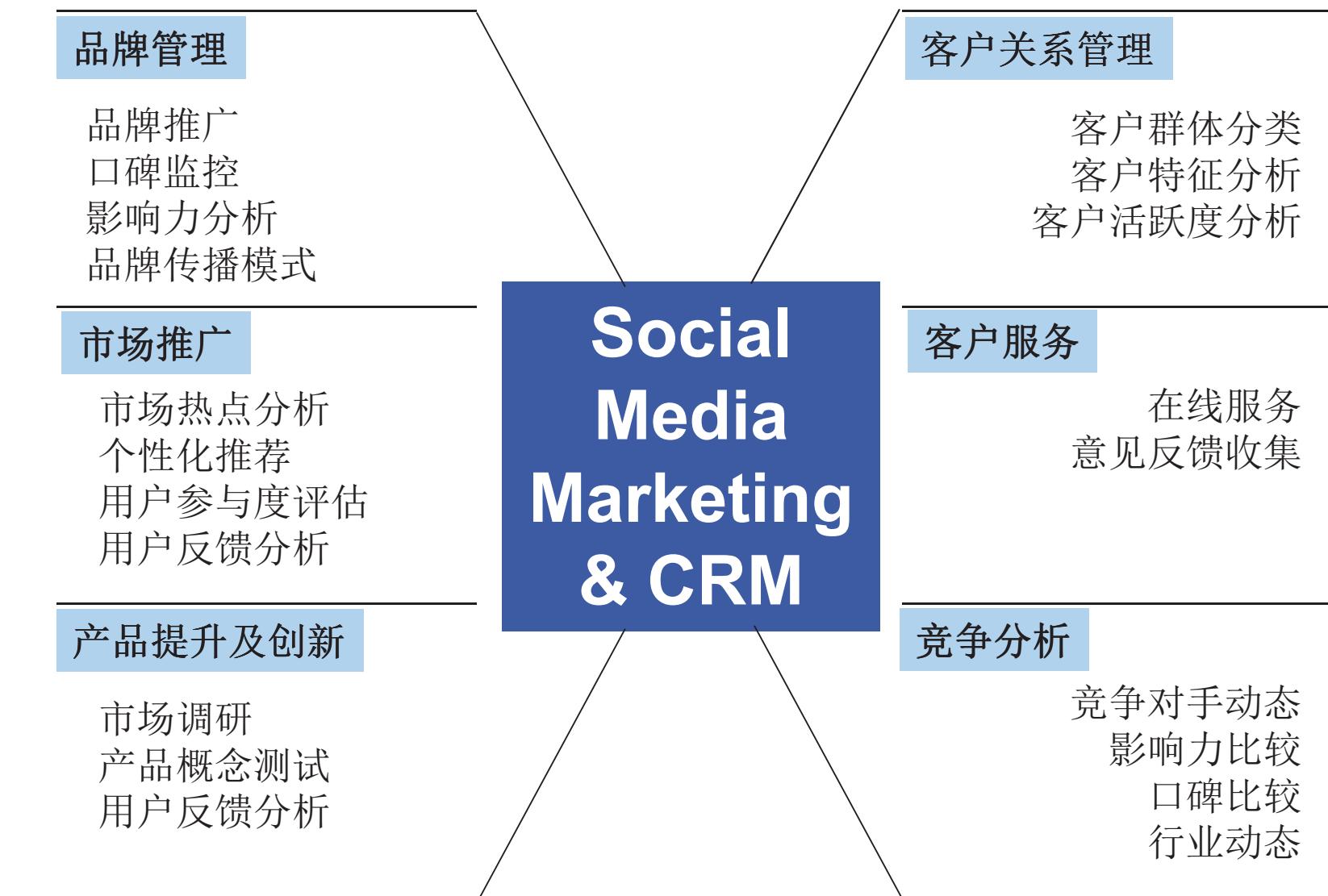
消费者行为与粉丝兴趣分析

4

总结及参考资料

5

社交媒体营销体系



微博分析的应用领域及相关工具介绍

- 官方工具 `data.weibo.com`

- ✓ 比如新浪微博的微数据、微指数、微报告
- ✓ 可进行影响力分析、粉丝分析

- 第三方应用

- ✓ 北京大学微博可视化分析系列 <http://vis.pku.edu.cn/weibova/>
- ✓ 等等

议程

社交媒体营销 (**Social Media Marketing**)

1

微博情感分析 (**Sentiment Analysis**)

2

例子展示

3

消费者行为与粉丝兴趣分析

4

总结及参考资料

5

情感分析的重要性



不满意的客户*

- ✓ 只有4%的不满意的顾客会投诉
- ✓ 超过90%不满意的顾客不会再回来
- ✓ 每个不满意的顾客会告诉9个人



满意的客户*

- ✓ 维护成本是开发一个新顾客的1/5 至1/6
- ✓ 满意的客户愿意花更多的钱
- ✓ 每个高兴的顾客会告诉5个人

在社交媒体上，负面的情绪一般比正面的情绪传播得更广更快

* Lecture slides of Strategic Marketing by HKUST Prof. Joseph Salvacruz

⁵ ** 图标来自新浪微博表情图片

微博数据的抓取及预处理

● 数据获取及清洗

- ✓ 技术：调用微博平台的API、网页抓取
- ✓ 工具及资源：火车头微博数据抓取工具、Rweibo

● 文本预处理

- ✓ 技术：分词、词性标注、句法分析
- ✓ 工具及资源：ICTCLAS词法分析系统、AnsJ、Rwordseg；搜狗词库、各种停用词库

微博情感分析方法

- 主题无关的情感分析*

- ✓ 基于情感词典的方法
- ✓ 基于表情符号的方法
- ✓ 有监督的机器学习方法（SVM, Naïve Bayes等）
- ✓ 无监督的方法

- 主题相关的情感分析*

- ✓ 基于规则的方法
- ✓ 基于特征的方法

- 情感词典

- ✓ 知网 HowNet
- ✓ 台湾大学NTUSD情感词典

微博情感分析方法示例*

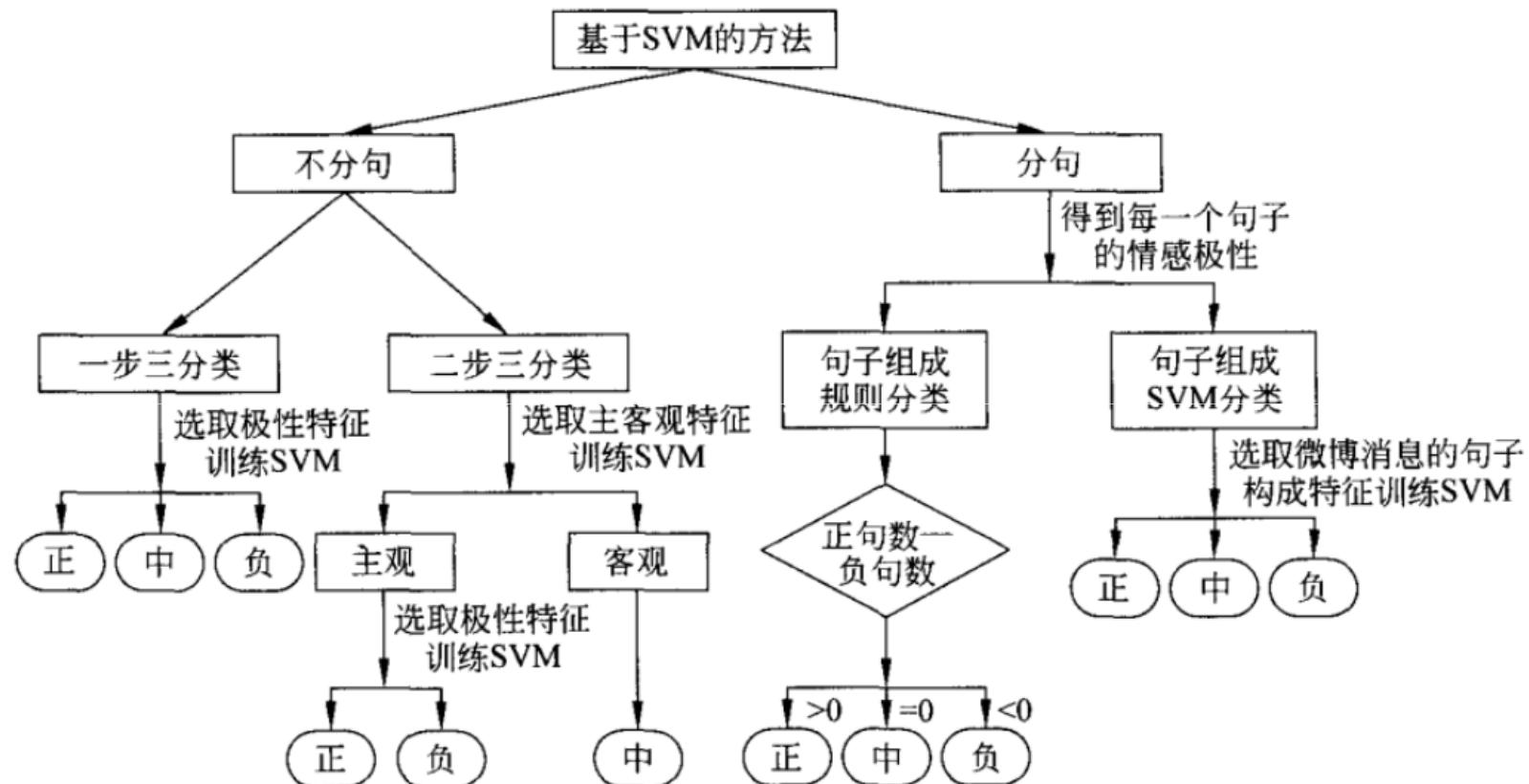


图 1 基于层次结构的多策略情感分析框架

中文微博情感分析的难点*

- 中文需要分词

- ✓ “英雄难过美人关”
- ✓ 网络新词层出不穷：“不明觉厉”“么么哒”“内什么”
- ✓ 微博语法不规范、错别字多

- 中文情感词典的构建

- ✓ 情感词的多义性

- 中文的一些特殊表达方法

- ✓ 比如：反讽 – “你行啊”

中文微博情感分析的准确率*

	基于表情符号 的规则方法	基于情感词典 的规则方法	基于 SVM 的 一步三分类方法
准确率	56.583%	55.583%	65.400%

表 1 NTCIR-08 中情感极性判别的最好结果

特征	主、客观情感判别		正、负向情感判别	
	繁体中文	简体中文	繁体中文	简体中文
精确率	56.37%	41.34%	76.48%	67.39%

表 2 COAE-09 中情感极性判别的最好结果

裁判员	P@1000	Precision	Recall	F1	Accuracy-1000
1	0.662	0.662	0.158 033	0.255 155	0.158 033
2	0.612	0.612	0.153 268	0.245 143	0.153 268
3	0.544	0.544	0.149 986	0.235 142	0.149 986

议程

社交媒体营销 (**Social Media Marketing**)

1

微博情感分析 (**Sentiment Analysis**)

2

例子展示

3

消费者行为模型与粉丝兴趣分析

4

总结及参考资料

5

微博上对“iOS7”的评价如何？-- 情感词典的方法分析

Step1: 新浪微博数据抓取、清洗

Step2: 装载词库、添加词库

Step3: 分词、创建 词项-文档 矩阵

Step4: 正、负情感词统计

Step5: 分类准确度比较

Step6: 问题分析

Step7: 词云展现

Step1：新浪微博数据抓取、清洗

```
library("Rweibo")

#通过网页抓取的方式获得相关微博数据
res<-web.search.content("iOS7", page = 20, sleepmean = 15,sleepsd = 5)

#去除可能存在的重复微博
res <- res[!duplicated(res[, "MID"]),]

# 将微博文本存储到一个向量中
doc=res$Weibo

# 去除微博中含有的url
doc=gsub(pattern="http:[a-zA-Z\\\.\.0-9]+","",doc)

# 去除转发微博时包含的其他微博博主ID,比如@实用小百科
doc=gsub(pattern="@(\w+)[ , : ]","",doc)

# 去除 "我在: XXX", “我在这里: ” 等位置信息
doc=gsub(pattern="我在: (\w*)","",doc)
doc=gsub(pattern="我在:(\w*)","",doc)
doc=gsub(pattern="我在这里: (\w*)","",doc)
doc=gsub(pattern="我在这里:(\w*)","",doc)
```

Step2: 装载词库、添加词库

```
library("Rwordseg")

# 安装搜狗词库 (注: 拆分为5000个词一个文件安装不会出错)
installDict("D:/sougou_1.txt","sougou_1")
.....
# 安装HowNet中文正、负情感极性词典
installDict("D:/HowNet_pos.txt","HowNet_pos")
.....
# 安装台湾大学中文正、负情感极性词典
installDict("D:/NTUSD_pos.txt","NTUSD_pos")
.....
# 添加iOS7相关词汇
textwords=c("iOS7","正式版","iPhone4s","iPhone3","九宫格","wifi")
insertWords(textwords)
```

分词、创建 词项-文档 矩阵

```
# 对每条微博进行分词
```

```
doc_CN=list()
```

```
for(j in 1:length(doc)){
```

```
    doc_CN[[j]]=c(segmentCN(doc[j],recognition=F))
```

```
}
```

```
# 构建语料库(Corpus对象)
```

```
library("tm")
```

```
corpus=Corpus(VectorSource(doc_CN))
```

```
# 去除停用词
```

```
stw=read.table(file.choose(),colClasses="character")
```

```
stopwords_CN<-stw[,1]
```

```
corpus=tm_map(corpus,removeWords,stopwords_CN)
```

```
# 创建 词项-文档 矩阵(tdm)
```

```
control=list(removePunctuation=T,minDocFreq=2,wordLengths = c(1,  
Inf),weighting = weightTf)
```

```
tdm=TermDocumentMatrix(corpus,control)
```

```
length(tdm$dimnames$Terms)
```

正、负情感词统计

#分别读入正、负情感词库

```
pos.words<-readLines("D:/positive.txt",encoding='UTF-8')  
neg.words<-readLines("D:/negative.txt",encoding='UTF-8')
```

#分为正向(1)、负向(0)、中性(2)情感三类

#如果正向词汇数量大于负向词汇数量，则该微博为正向情感；反之为负向；相等则为中性

```
len<-length(doc_CN)  
scores<- rep(0,times=len)  
for(i in 1:len){  
    result<-countWords(doc_CN[[i]],pos.words,neg.words)
```

.....

}

```
countWords<-function(words,pos.words,neg.words){
```

.....

```
    pos.matches<-match(words,pos.words)
```

.....

}

分类准确度比较

与手工标注的分类相比较，计算分类准确率。共500条微博

precision = 56.2%

机器分类结果：

负向(0)	正向(1)	中性(2)
108	249	143

手工分类结果：

负向(0)	正向(1)	中性(2)
188	148	164

可见，分类结果并不理想，正、负情感的对比数量与手工分类偏差较大。

存在的问题分析

■ 数据准备

- ✓ 转发的微博不要混在一起

■ 情感词典

- ✓ 基本的情感词典有不准确的地方，比如：“说”
- ✓ 基本的情感词典有不完善的地方，特别是负向情感词
- ✓ 不同词性下同一个词的情感取向不同，比如：“系统”
- ✓ 网络词汇需要及时更新，比如：“杯具”、“有点娘”，“土豪”
- ✓ 专业词汇情感意义与普通词汇不同，比如：“升级”、“蓝屏”
- ✓ 方言词汇，比如：粤语
- ✓ 错别字等等

■ 计算方法

- ✓ 加入其它特征分析，比如：表情符号、“赞”、转发次数
- ✓ 加入简单的句法分析，比如：虽然...但是...

词云展现



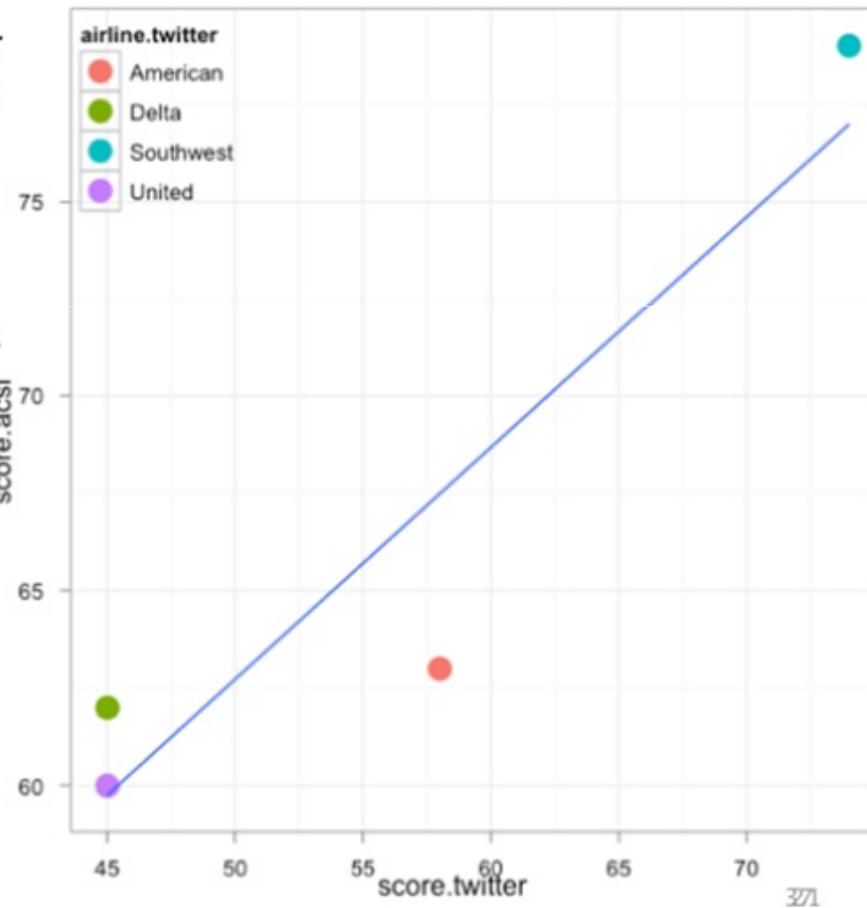
注：由于情感词典不完善等原因，上述词云分类未必准确

R by example: mining Twitter for consumer attitudes towards airlines*

an actual result!

ggplot will even run `lm()` linear (and other) regressions for you with its `geom_smooth()` layer:

```
> ggplot( compare.df ) +  
  geom_point(aes(x=score.twitter,  
                 y=score.acsi,  
                 color=airline.twitter), size=5) +  
  geom_smooth(aes(x=score.twitter,  
                 y=score.acsi, group=1), se=F,  
                 method="lm") +  
  theme_bw() +  
  opts(legend.position=c(0.2,  
                       0.85))
```



议程

社交媒体营销 (**Social Media Marketing**)

1

微博情感分析 (**Sentiment Analysis**)

2

例子展示

3

消费者行为与粉丝兴趣分析

4

总结及参考资料

5

社会阶层、年龄、性别、兴趣与消费者行为

因素	消费者行为
社会阶层 --职业、教育程度、收入	比如：在很多情况下，同一种职业的人具有相似的购买兴趣。
年龄	年龄段不同，消费兴趣往往也不同。比如年轻人更愿意尝试新的产品，而年纪大的人更具有品牌忠诚度。
性别	对产品的需要不同； 对相关市场行为（如广告内容）的反应也不同
兴趣	有些兴趣会跨越社会阶层、年龄和性别的差别，比如：旅游

粉丝的兴趣分析

- 获取粉丝基本信息及标签

- ✓ API需要授权

- 分析粉丝的所关注的微博

- ✓ API需要授权

- 分析粉丝所发布的微博内容

- ✓ 比如：利用主题模型(Topic Model)方法分析

主题模型*

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Figure 1: 几个主题的例子。

主题模型在Twitter中的应用*

- 难点

- 噪音大、篇幅短、更新快、数量大

- 主要应用模型

- ✓ LDA
 - ✓ Author-Topic Model
 - ✓ Twitter-LDA
 - ✓ 等等

我 - JackHan2008 - 的粉丝兴趣分析示例

■ 数据

- ✓ 用火车头数据采集平台采集25个粉丝9600多条微博

■ 用主题模型进行分析

- ✓ 采用R语言中的“topicmodel” package

■ 代码参考

- ✓ "topicmodels: An R Package for Fitting Topic Models" - Bettina Grun, Kurt Hornik

- ✓ “[微博名人的那些事儿](#)” - 布丁Nnn

实验结果

```
> Topic <- topics(fans_TM[["VEM"]], 1)
>
> Terms <- terms(fans_TM[["VEM"]], 10)
>
> Terms[, 1:20]
   Topic 1    Topic 2    Topic 3    Topic 4    Topic 5    Topic 6    Topic 7    Topic 8    Topic 9    Topic 10   Topic 11   Topic 12   Topic 13   Topic 14   Topic 15   Topic 16   Topic 17   Topic 18   Topic 19   Topic 20
[1,] "九三学社" "价值观"  "南山区"  "投资者"  "慕尼黑"  "半拉子"  "阿拉伯"  "广东省"  "互联网"  "普通法"  "蝴蝶兰"  "工作"    "出生证"  "越来越"  "广东省"  "梅里雪山" "小孩儿"  "小分队"  "投资者"  "广州市"  "越来越"  "晒太阳"  "忍不住"  "广
[2,] "出生证"  "越来越"  "广东省"  "梅里雪山" "小孩儿"  "小分队"  "投资者"  "广州市"  "越来越"  "晒太阳"  "忍不住"  "广
[3,] "平凡人"  "欣欣向荣" "深圳市"  "统计学"  "图书馆"  "皓月当空" "服务员"  "深圳市"  "广州市"  "现实主义" "办公室"  "朝鲜
[4,] "钓鱼岛"  "仅供参考" "有感而发" "联合国"  "维也纳"  "自由泳"  "联合国"  "天河区"  "忍不住" "老爷子"  "打电话"  "乌
[5,] "俱乐部"  "足球队"  "苏格兰"  "人性化"  "一个岁"   "跳水馆"  "奥巴马"  "越秀区"  "小朋友"  "帝国主义" "动物园"  "华
[6,] "呱呱坠地" "管理者"  "一日游"  "娱乐性"  "博物馆"  "错别字"  "成功者"  "娃哈哈"  "广东省"  "打电话"  "一分钟"  "能
[7,] "谈笑风生" "意味着"  "十二生肖" "闪光灯"  "复活节"  "map"     "失败者"  "服务员"  "打电话"  "无产阶级" "小家伙"  "上
[8,] "一个半"   "必需品"  "摄影师" "眼睁睁"  "奥巴马"  "一下下"   "老爷子"  "新加坡"  "有意思"  "有没有"  "心急如焚" "主
[9,] "一分钟"   "一动不动" "新浪网"  "一点点"  "打电话"  "一下子"   "香格里拉" "办公室"  "打交道"  "老太太"  "玩意儿"  "农
[10,] "一鸣惊人" "人来疯"  "星期五"  "香格里拉" "晒太阳"  "一下数"   "梅里雪山" "南山区"  "母公司"  "一整天"  "出租车"  "办公
   Topic 13   Topic 14   Topic 15   Topic 16   Topic 17   Topic 18   Topic 19   Topic 20
[1,] "小家伙"  "不可思议" "大吉利"   "小家伙"  "红树林"   "越来越"  "小朋友"  "一不小心"
[2,] "经济学"  "表姐夫"   "一大早"  "长长的"  "停车场"   "是因为"  "有没有"  "常州市"
[3,] "一个半"   "一整天"   "吴奇隆"  "不信任感" "俱乐部"   "买不起"  "会计师"  "新加坡"
[4,] "下半场"  "一望无际" "龙马精神" "苏有朋"   "校友会"   "班主任"  "审计员"  "海阔天空"
[5,] "几分钟"  "万年历"   "从头到尾" "半夜三更" "证监会"   "管理者"  "图书馆"  "香港科技大学"
[6,] "小儿子"   "人山人海" "哈哈哈"  "一个月"   "可吸入颗粒物" "书法家"  "大部分"  "一会儿"
[7,] "小朋友"  "十多年"   "年月日"  "哥们儿"   "合伙人"   "安全带"  "三十年"  "一阵阵"
[8,] "桃源村"   "威尼斯"   "总经理"  "深棕色"  "商学院"   "看不到"  "上档次"  "三分钟"
[9,] "热闹非凡" "年月日"  "一举成名" "越来越"   "污染物"   "材料费"  "公务员"  "下定决心"
[10,] "电话会议" "更年期"  "万人迷"  "二百元"   "经济学"   "华盛顿"  "塑化剂"  "不预则废"
```

实验结果并不理想，原因是？ 文档数量太小？ 噪音词汇太多？

结论及建议

- 词典完善、数据清洗等工作是基础
- 不仅仅依赖于文本本身进行挖掘，应结合其它数据分析
- 从应用角度：立足于整个营销体系，而不仅仅是局部

参考资料

1. 谢丽星等：基于层次结构的多策略中文微博情感分析和特征抽取
2. 主题模型在文本挖掘中的应用 赵鑫，李晓明
3. [R by example: mining Twitter for consumer attitudes towards airlines](#)
4. “[微博名人的那些事儿](#)” - 布丁Nnn
5. [刘思喆 R语言环境下的文本挖掘](#)
6. Rweibo, Rwordseg包 <http://jliblog.com/app/>
7. Lecture slides of Strategic Marketing by HKUST Prof. Joseph Salvacruz

THANK YOU