基于关键词规则的情感分析

赵海臣

背景

- 基于规则的评论评分完成雏形，由于评论评分基本上不涉及情感，因此为了企业的利益，需要控制好评、差评以及诋毁聚美相关的评论，所以需要开发相关的好评、差评、禁评的识别。

聚美优品
JUMEI.COM

基本思路

基于规则与统计的好评、差评以及禁评肯定无法覆盖100%的句子，因此预计采用统计的方式来获取词频最高的能覆盖大部分评论的好评、差评以及禁评高频词汇，采用人工审核高频关键词的方式，来制定关键词词表，通过检测关键词来获得相应的情感等级。

基本步骤

- 根据规则average_score=5.0/1.0分别筛选出极端好评/差评评论
- 对评论进行分词
- 分别统计极端好评/差评的
  - 名词词频
  - 形容词词频
  - 组合词(副词+形容词)词频
  - 组合词中副词词频
  - 组合词中形容词词频

标志性情感词

- 因为用户有可能任意输入导致噪声(正面词评论被赋予了low average score，或负面词评论给了high average score)，所以需要进行去噪处理。
- 基于朴素贝叶斯的思想，分别计算不同高频情感词的类条件概率：

$$P(sentiWord \mid emotion) = \frac{D_{sentiWord,emotion}}{D_{emotion}}$$

情感词挖掘

- 置信度
  - 为了确保某个情感词不是偶尔出现在某个情感评论中，需要确保情感词达到置信度词频阈值：

$$D_{sentiWord,emotion} > Thr_{sentiWordConf}$$

  - 在达到了置信度后，
    - 若 $D_{sentiWord,emotion1} != 0$

$$D_{sentiWord,emotion2} == 0$$

则该词的情感为emotion2

    - 若 $D_{sentiWord,emotion1} != 0$

$$D_{sentiWord,emotion2} != 0$$

则根据 $emotionRate = \frac{P(sentiWord \mid emotion1)}{P(sentiWord \mid emotion2)}$ 来判断。

副词+形容词词对

副词与形容词连续同时出现，往往副词代表形容词的程度。另外，分词器容易错误地标注单个词汇，但是副词与形容词同时错误标注的概率会小很多，因此，连续侦测副词+形容词词对的鲁棒性要高很多。

- 经过试验发现，对连续出现的副词+形容词词对会相对于单词更贴近于实际的评论语境。
  - 差评单形容词：

	A	B
1	words	count
2	差	47545
3	好	31102
4	大	15647
5	根本	14739
6	小	12348
7	坏	12127
8	一般	10240
9	便宜	8353
10	慢	8315
11	严重	6863
12	舒服	5809
13	不错	5751
14	假	5421
15	太慢	5045
16	很好	4897
17	薄	4276

差评副词+形容词组合词

	A	B
1	constructed_words	count
2	很差	5719
3	不舒服	4475
4	太大	4396
5	就坏	3408
6	没好	3017
7	挺好	1814
8	偏小	1722
9	特别差	1388
10	还不错	1371
11	非常差	1097
12	不愉快	895
13	特别大	847
14	太硬	812
15	都坏	812
16	不干净	779
17	不合适	642
18	特别小	619
19	太少	616
20	都快	566
21	特别慢	564
22	就是坏	542
23	特别薄	519

情感识别关键词识别

- 情感识别关键词采用遍历的方法，按照优先级分别侦测
  - 反义副词 + 组合副词词组
  - 组合副词词组
  - 反义副词 + 单形容词
  - 单形容词

短句的情感分

- 基于单形容词、副词+形容词来做出短句情感级别判断：
  - 基于高频形容词，人工标记形容词的基础情感分baseEmo：
    - 好 = 1
    - 漂亮 = 1.5
    - 给力 = 2
    - …
  - 副词代表形容词的程度，基于高频副词，人工标记倍数分multiRate：
    - 挺 = 1.5
    - 非常 = 2
    - 特别 = 2.5
    - …
  - 短句的情感分为

$$emoRate = baseEmo * multiRate$$

评论的情感

- 评论的情感根据情感词褒贬以及用户主观评论score来判断：
  - 若整个评论出现并只有褒义词(没有贬义词)出现，并且average_score == 5(覆盖好评90%以上)，则为强褒义评论。
  - 若整个评论出现并只有贬义词(没有褒义词)出现，并且average_score == 1(覆盖差评50%以上)，则为强贬义评论。
  - 若整个评论出现贬义词，且average_score != 5且!=1，则为适中非极端批评评论。
  - 若整个评论出现转折词
    - 可是，不过，但是，尽管，然而

    则为情感转折性句子

禁评

- 有些句子脱离了产品评论转而攻击聚美，例如侦测到诸如“假”，“伪”之类的敏感词表，需要及时侦测并屏蔽。

聚美优品
JUMEI.COM

THE END

THANK YOU!