**R语言自然语言处理：情感分析**

**作者：黄天元，复旦大学博士在读，热爱数据科学与R，热衷推广R在工业界与学术界的应用。邮箱：huang.tian-yuan@qq.com.欢迎合作交流**

前文参考：

[HopeR：R语言自然语言处理：中文分词](https://zhuanlan.zhihu.com/p/53693842)

[HopeR：R语言自然语言处理：词性标注与命名实体识别](https://zhuanlan.zhihu.com/p/53765293)

[HopeR：R语言自然语言处理：关键词提取（TF-IDF）](https://zhuanlan.zhihu.com/p/53841879)

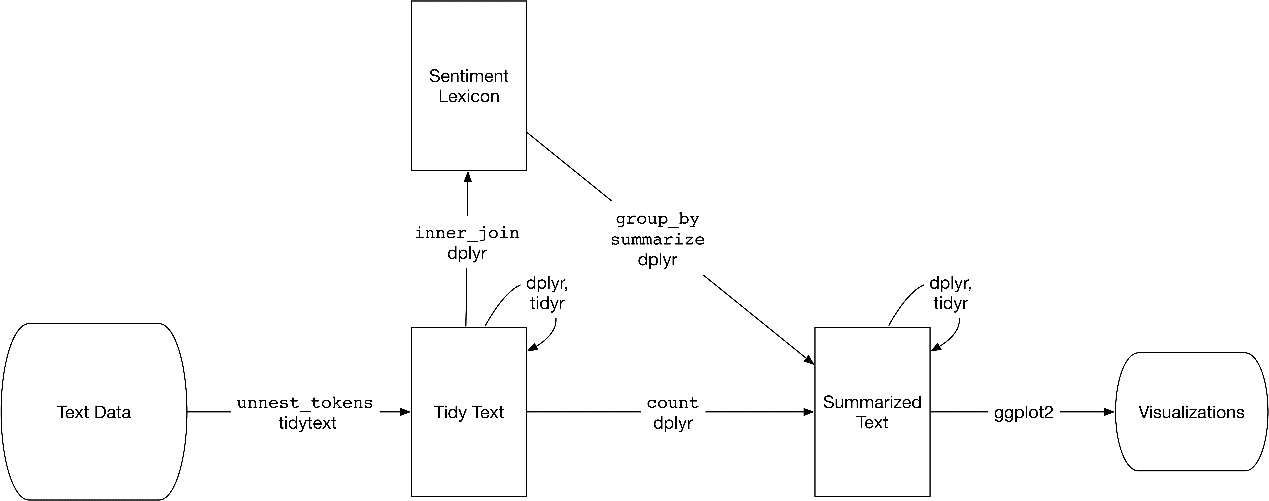
[HopeR：R语言自然语言处理：关键词提取与文本摘要（TextRank）](https://zhuanlan.zhihu.com/p/54036838)

[HopeR：R语言自然语言处理：词嵌入（Word Embedding）](https://zhuanlan.zhihu.com/p/54325648)

情感分析，就是根据一段文本，分析其表达情感的技术。比较简单的情感分析，能够辨别文本内容是积极的还是消极的（褒义/贬义）；比较复杂的情感分析，能够知道这些文字是否流露出恐惧、生气、狂喜等细致入微的情感。此外，情感的二元特性还可以表达为是否含有较大的感情波动。也就是说，狂喜和暴怒都属于感情波动，而宠辱不惊则属于稳定的情感状态。

情感分析方法主要分为两种：1.词法分析；2.机器学习。

其中，机器学习需要依赖于标注和特征提取，这里大有文章可做，但是很难提炼出共性，因此这里不进行特殊的介绍。主要就是以人工标注的情感数值作为响应变量，然后另一方面对于文本进行向量化处理（词嵌入），然后用模型进行拟合，最后得到一个好的模型对新的文本进行情感的评估。需要注意的是，对文本进行情感特征的提取也是有文章可以做的，只要能够正确认识哪些文本能够提供情感信息，就能够更好地捕捉文本的情感方向和程度。



tidytext进行词法分析的技术路线

另一个方法，就是词法分析。它的原理非常简单，事前需要定义一个情感词典。比如“喜欢”这个词我们定义为1分。那么“我喜欢你”这句话，“我”和“你”都是中性词，均为0分，“喜欢”为1分，这句话的总分就是1分。“我喜欢你，但讨厌他”，这样一句话中有“讨厌”这个词，在情感词典中分数为“-1”，那么整句话的得分就是0。这样，我们就可以对每一个文本进行分词，然后使用内连接（inner join）来提取其中的情感词语，并根据情感词语的得分，来评估这段文本的情感得分。我找了两个我认为比较好的R包来实现这个功能：

1.tidytext（[juliasilge/tidytext](http://link.zhihu.com/?target=https%3A//github.com/juliasilge/tidytext)）; 2.SentimentAnalysis（[sfeuerriegel/SentimentAnalysis](http://link.zhihu.com/?target=https%3A//github.com/sfeuerriegel/SentimentAnalysis)）。

这些包基本能够通过复制粘贴代码就能够实现对文本情感的分析，这里不进行赘述。不过这些包都是针对英文的，中文的情感词典则比较匮乏。经过搜索，比较好的有大连理工大学提供的情感词汇本体库（[大连理工大学信息检索研究室（DUTIR）-搜人搜物搜信息，重情重义重认知](http://link.zhihu.com/?target=http%3A//ir.dlut.edu.cn/EmotionOntologyDownload)）。不过我认为这些中文词库还是远远无法达到应用级的水准，真的要用，就需要自己构建情感词库。经过思考，我拟定的方法如下：1.对需要分析的文本进行分词，并进行词频的分析（TF）；2.根据出现最多的一些词，对它们可能代表的情感进行评估；3.构建自定义情感词典，导入分词器中进行分词，并通过词典表对其进行情感打分。

只有了解这些文本的特点，才能够更好地进行分词并提取出情感信息。这种分析可以广泛应用到用户评价分析、舆情分析，能够成为一个非常重要的辅助参考。针对个性化的词典，还能够精准评估一些文本是否出现了“特殊的”情感，比如极端情绪或我们特别感兴趣的情绪（不满？特别满意？威胁恐吓？）。这种分析还可以扩展到句法分析，也就是什么句型与相应的情绪具有强烈的相关性，未来这个领域大有可期。