[**文本情感分类（一）：传统模型**](http://spaces.ac.cn/archives/3360/)

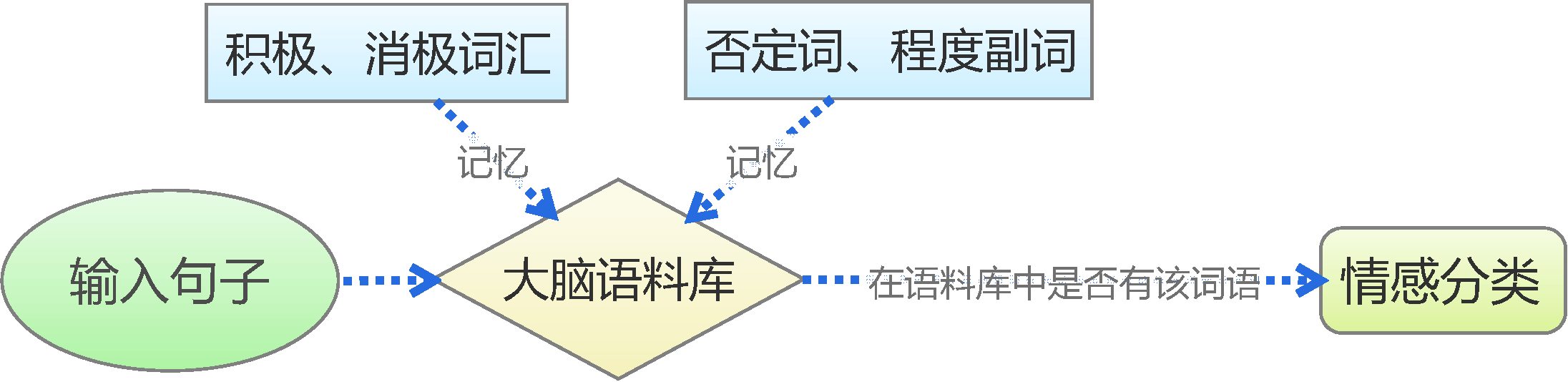
作者：苏剑林 | 发布时间：2015-06-22

**本文目录 [**[**隐藏**](javascript:content_index_toggleToc())**]**

* [基于情感词典的文本情感分类](http://spaces.ac.cn/archives/3360/#基于情感词典的文本情感分类)
* [模型结果检验](http://spaces.ac.cn/archives/3360/#模型结果检验)
* [困难所在](http://spaces.ac.cn/archives/3360/#困难所在)
* [优化思路](http://spaces.ac.cn/archives/3360/#优化思路)
* [参考文献](http://spaces.ac.cn/archives/3360/#参考文献)
* [实现平台](http://spaces.ac.cn/archives/3360/#实现平台)
* [代码列表](http://spaces.ac.cn/archives/3360/#代码列表)

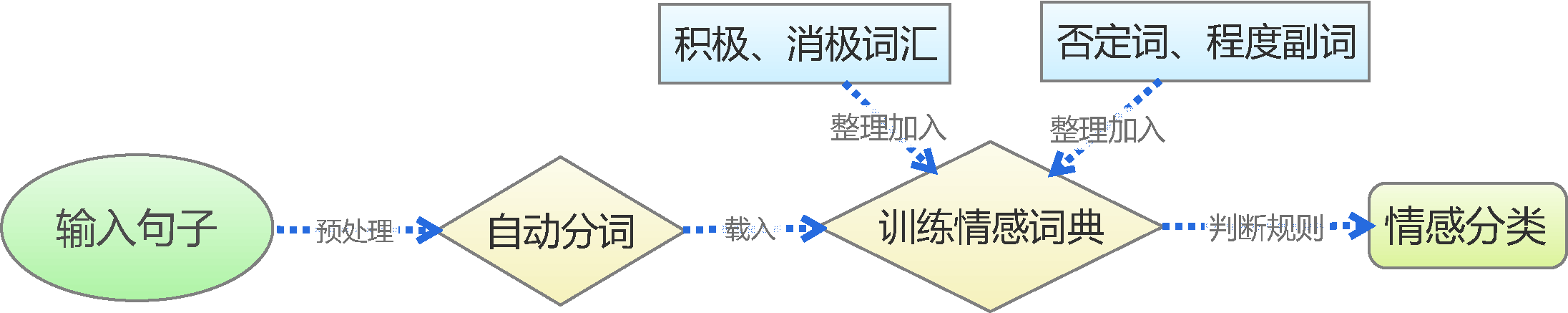
**前言：**四五月份的时候，我参加了两个数据挖掘相关的竞赛，分别是物电学院举办的“亮剑杯”，以及第三届 “泰迪杯”全国大学生数据挖掘竞赛。很碰巧的是，两个比赛中，都有一题主要涉及到中文情感分类工作。在做“亮剑杯”的时候，由于我还是初涉，水平有限，仅仅是基于传统的思路实现了一个简单的文本情感分类模型。而在后续的“泰迪杯”中，由于学习的深入，我已经基本了解深度学习的思想，并且用深度学习的算法实现了文本情感分类模型。因此，我打算将两个不同的模型都放到博客中，供读者参考。刚入门的读者，可以从中比较两者的不同，并且了解相关思路。高手请一笑置之。

**基于情感词典的文本情感分类**

[](http://kexue.fm/usr/uploads/2015/06/3473700595.png)

传统的基于情感词典的文本情感分类，是对人的记忆和判断思维的最简单的模拟，如上图。我们首先通过学习来记忆一些基本词汇，如否定词语有“不”，积极词语有“喜欢”、“爱”，消极词语有“讨厌”、“恨”等，从而在大脑中形成一个基本的语料库。然后，我们再对输入的句子进行最直接的拆分，看看我们所记忆的词汇表中是否存在相应的词语，然后根据这个词语的类别来判断情感，比如“我喜欢数学”，“喜欢”这个词在我们所记忆的积极词汇表中，所以我们判断它具有积极的情感。

基于上述思路，我们可以通过以下几个步骤实现基于情感词典的文本情感分类：预处理、分词、训练情感词典、判断，整个过程可以如下图所示。而检验模型用到的原材料，包括薛云老师提供的蒙牛牛奶的评论，以及从网络购买的某款手机的评论数据（见附件）。

[](http://kexue.fm/usr/uploads/2015/06/2323544396.png)

**文本的预处理**

由网络爬虫等工具爬取到的原始语料，通常都会带有我们不需要的信息，比如额外的Html标签，所以需要对语料进行预处理。由薛云老师提供的蒙牛牛奶评论也不例外。我们队伍使用Python作为我们的预处理工具，其中的用到的库有Numpy和Pandas，而主要的文本工具为正则表达式。经过预处理，原始语料规范为如下表，其中我们用-1标注消极情感评论，1标记积极情感评论。

http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0037.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0037.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/0063.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/006F.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/006D.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/006D.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/0065.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/006E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/0074.png?V=2.7.0蒙牛又出来丢人了珍爱生命远离蒙牛http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0我一直都很爱喝蒙牛的纯牛奶一直，很爱送蒙牛http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0健康才是最好的礼物。http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/006D.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/0061.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/0072.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Math/Italic/336/006B.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0commentmark0蒙牛又出来丢人了−11珍爱生命远离蒙牛−1⋮⋮⋮1171我一直都很爱喝蒙牛的纯牛奶一直，很爱11172送蒙牛...健康才是最好的礼物。1⋮⋮⋮

**句子自动分词**

为了判断句子中是否存在情感词典中相应的词语，我们需要把句子准确切割为一个个词语，即句子的自动分词。我们对比了现有的分词工具，综合考虑了分词的准确性和在Python平台的易用性，最终选择了“结巴中文分词”作为我们的分词工具。

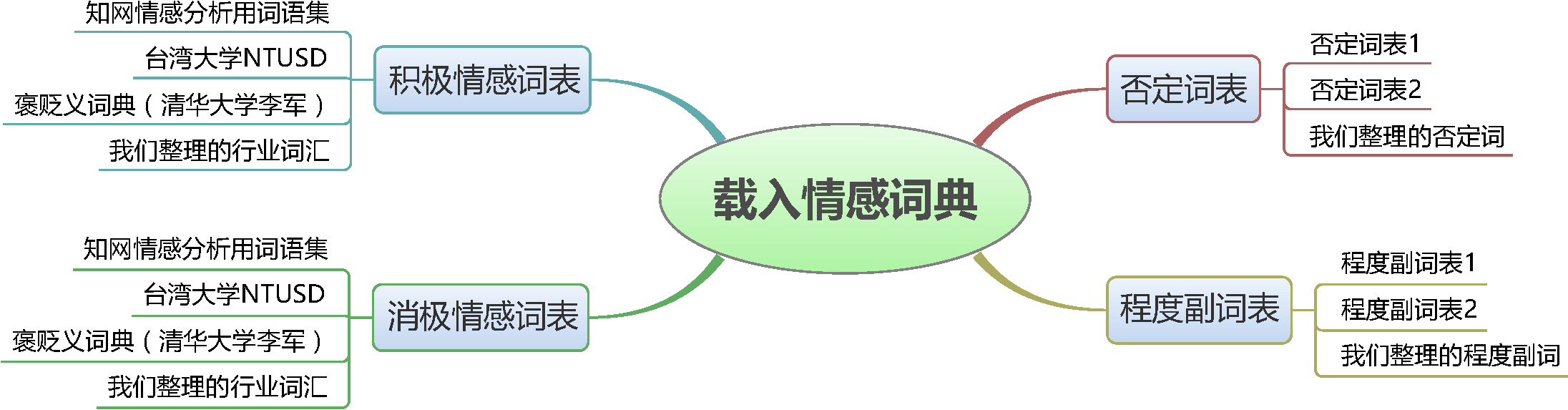
下表仅展示各常见的分词工具对其中一个典型的测试句子的分词效果：

**测试句子：**工信处女干事每月经过下属科室都要亲口交代24口交换机等技术性器件的安装工作

|  |  |
| --- | --- |
| 分词工具 | 测试结果 |
| 结巴中文分词 | 工信处/ 女干事/ 每月/ 经过/ 下属/ 科室/ 都/ 要/ 亲口/ 交代/ 24/ 口/ 交换机/ 等/ 技术性/ 器件/ 的/ 安装/ 工作 |
| 中科院分词 | 工/n 信/n 处女/n 干事/n 每月/r 经过/p 下属/v 科室/n 都/d 要/v 亲口/d 交代/v 24/m 口/q 交换机/n 等/udeng 技术性/n 器件/n 的/ude1 安装/vn 工作/vn |
| smallseg | 工信/ 信处/ 女干事/ 每月/ 经过/ 下属/ 科室/ 都要/ 亲口/ 交代/ 24/ 口/ 交换机/ 等/ 技术性/ 器件/ 的/ 安装/ 工作 |
| Yaha 分词 | 工信处 / 女 / 干事 / 每月 / 经过 / 下属 / 科室 / 都 / 要 / 亲口 / 交代 / 24 / 口 / 交换机 / 等 / 技术性 / 器件 / 的 / 安装 / 工作 |

**载入情感词典**

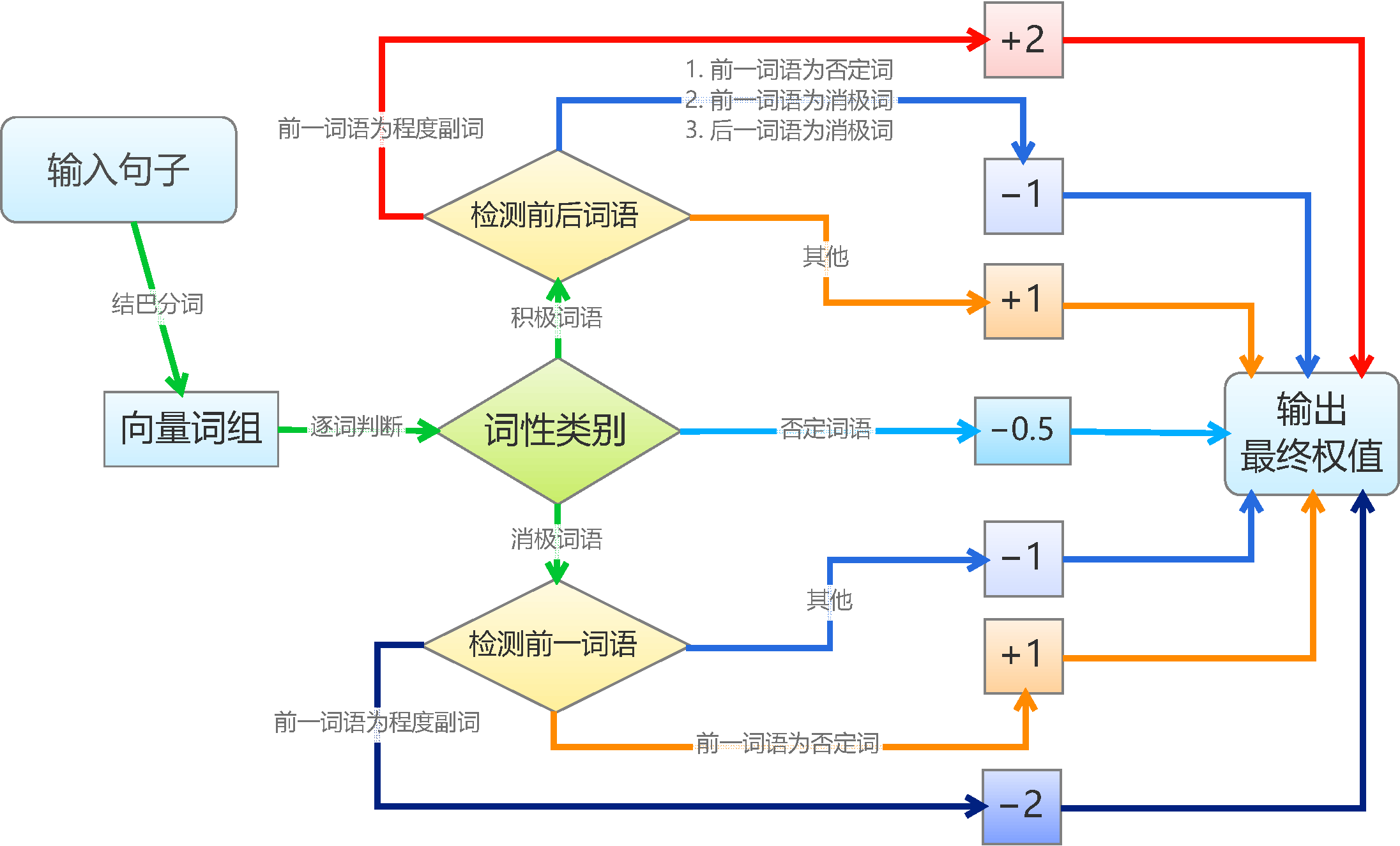
一般来说，词典是文本挖掘最核心的部分，对于文本感情分类也不例外。情感词典分为四个部分：积极情感词典、消极情感词典、否定词典以及程度副词词典。为了得到更加完整的情感词典，我们从网络上收集了若干个情感词典，并且对它们进行了整合去重，同时对部分词语进行了调整，以达到尽可能高的准确率。

[](http://kexue.fm/usr/uploads/2015/06/1152794772.png)

我们队伍并非单纯对网络收集而来的词典进行整合，而且还有针对性和目的性地对词典进行了去杂、更新。特别地，我们加入了某些行业词汇，以增加分类中的命中率。不同行业某些词语的词频会有比较大的差别，而这些词有可能是情感分类的关键词之一。比如，薛云老师提供的评论数据是有关蒙牛牛奶的，也就是饮食行业的；而在饮食行业中，“吃”和“喝”这两个词出现的频率会相当高，而且通常是对饮食的正面评价，而“不吃”或者“不喝”通常意味着对饮食的否定评价，而在其他行业或领域中，这几个词语则没有明显情感倾向。另外一个例子是手机行业的，比如“这手机很耐摔啊，还防水”，“耐摔”、“防水”就是在手机这个领域有积极情绪的词。因此，有必要将这些因素考虑进模型之中。

**文本情感分类**

基于情感词典的文本情感分类规则比较机械化。简单起见，我们将每个积极情感词语赋予权重1，将每个消极情感词语赋予权重-1，并且假设情感值满足线性叠加原理；然后我们将句子进行分词，如果句子分词后的词语向量包含相应的词语，就加上向前的权值，其中，否定词和程度副词会有特殊的判别规则，否定词会导致权值反号，而程度副词则让权值加倍。最后，根据总权值的正负性来判断句子的情感。基本的算法如图。

[](http://kexue.fm/usr/uploads/2015/06/2778028889.png)

要说明的是，为了编程和测试的可行性，我们作了几个假设（简化）。假设一：我们假设了所有积极词语、消极词语的权重都是相等的，这只是在简单的判断情况下成立，更精准的分类显然不成立的，比如“恨”要比“讨厌”来得严重；修正这个缺陷的方法是给每个词语赋予不同的权值，我们将在本文的第二部分探讨权值的赋予思路。假设二：我们假设了权值是线性叠加的，这在多数情况下都会成立，而在本文的第二部分中，我们会探讨非线性的引入，以增强准确性。假设三：对于否定词和程度副词的处理，我们仅仅是作了简单的取反和加倍，而事实上，各个否定词和程度副词的权值也是不一样的，比如“非常喜欢”显然比“挺喜欢”程度深，但我们对此并没有区分。

在算法的实现上，我们则选用了Python作为实现平台。可以看到，借助于Python丰富的扩展支持，我们仅用了一百行不到的代码，就实现了以上所有步骤，得到了一个有效的情感分类算法，这充分体现了Python的简洁。下面将检验我们算法的有效性。

**模型结果检验**

作为最基本的检验，我们首先将我们的模型运用于薛云老师提供的蒙牛牛奶评论中，结果是让人满意的，达到了82.02%的正确率，详细的检验报告如下表

数据内容牛奶评论正样本数http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0035.png?V=2.7.0负样本数http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0037.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0准确率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0真正率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0039.png?V=2.7.0真负率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0039.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0037.png?V=2.7.0数据内容正样本数负样本数准确率真正率真负率牛奶评论100511700.82020.82090.8197

（其中，正样本为积极情感评论，负样本为消极情感数据，

准确率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/003D.png?V=2.7.0被正确判断的样本数总样本数真正率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/003D.png?V=2.7.0被判断为积极的正样本数正样本总数真负率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/003D.png?V=2.7.0被判断为消极的负样本数负样本总数准确率=被正确判断的样本数总样本数真正率=被判断为积极的正样本数正样本总数真负率=被判断为消极的负样本数负样本总数

。）

让我们惊喜的是，将从蒙牛牛奶评论数据中调整出来的模型，直接应用到某款手机的评论数据的情感分类中，也达到了81.96%准确率！这表明我们的模型具有较好的强健性，能在不同行业的评论数据的情感分类中都有不错的表现。

数据内容手机评论正样本数http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0035.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0负样本数http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0035.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0039.png?V=2.7.0准确率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0039.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0036.png?V=2.7.0真正率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0037.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0035.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0039.png?V=2.7.0真负率http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/002E.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0038.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0035.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0数据内容正样本数负样本数准确率真正率真负率手机评论115811590.81960.75390.8852

**结论：**我们队伍初步实现了基于情感词典的文本情感分类，测试结果表明，通过简单的判断规则就能够使这一算法具有不错的准确率，同时具有较好的强健性。一般认为，正确率达80%以上的模型具有一定的生产价值，能适用于工业环境。显然，我们的模型已经初步达到了这个标准。

**困难所在**

经过两次测试，可以初步认为我们的模型正确率基本达到了80%以上。另外，一些比较成熟的商业化程序，它的正确率也只有85%到90%左右（如BosonNLP）。这说明我们这个简单的模型确实已经达到了让人满意的效果，另一方面，该事实也表明，传统的“基于情感词典的文本情感分类”模型的性能可提升幅度相当有限。这是由于文本情感分类的本质复杂性所致的。经过初步的讨论，我们认为文本情感分类的困难在以下几个方面。

**语言系统是相当复杂的**

归根结底，这是因为我们大脑中的语言系统是相当复杂的。（1）我们现在做的是文本情感分类，文本和文本情感都是人类文化的产物，换言之，人是唯一准确的判别标准。（2）人的语言是一个相当复杂的文化产物，一个句子并不是词语的简单线性组合，它有相当复杂的非线性在里面。（3）我们在描述一个句子时，都是将句子作为一个整体而不是词语的集合看待的，词语的不同组合、不同顺序、不同数目都能够带来不同的含义和情感，这导致了文本情感分类工作的困难。

因此，文本情感分类工作实际上是对人脑思维的模拟。我们前面的模型，实际上已经对此进行了最简单的模拟。然而，我们模拟的不过是一些简单的思维定式，真正的情感判断并不是一些简单的规则，而是一个复杂的网络。

**大脑不仅仅在情感分类**

事实上，我们在判断一个句子的情感时，我们不仅仅在想这个句子是什么情感，而且还会判断这个句子的类型（祈使句、疑问句还是陈述句？）；当我们在考虑句子中的每个词语时，我们不仅仅关注其中的积极词语、消极词语、否定词或者程度副词，我们会关注每一个词语（主语、谓语、宾语等等），从而形成对整个句子整体的认识；我们甚至还会联系上下文对句子进行判断。这些判断我们可能是无意识的，但我们大脑确实做了这个事情，以形成对句子的完整认识，才能对句子的感情做了准确的判断。也就是说，我们的大脑实际上是一个非常高速而复杂的处理器，我们要做情感分类，却同时还做了很多事情。

**活水：学习预测**

人类区别于机器、甚至人类区别于其他动物的显著特征，是人类具有学习意识和学习能力。我们获得新知识的途径，除了其他人的传授外，还包括自己的学习、总结和猜测。对于文本情感分类也不例外，我们不仅仅可以记忆住大量的情感词语，同时我们还可以总结或推测出新的情感词语。比如，我们只知道“喜欢”和“爱”都具有积极情感倾向，那么我们会猜测“喜爱”也具有积极的情感色彩。这种学习能力是我们扩充我们的词语的重要方式，也是记忆模式的优化（即我们不需要专门往大脑的语料库中塞进“喜爱”这个词语，我们仅需要记得“喜欢”和“爱”，并赋予它们某种联系，以获得“喜爱”这个词语，这是一种优化的记忆模式）。

**优化思路**

经过上述分析，我们看到了文本情感分类的本质复杂性以及人脑进行分类的几个特征。而针对上述分析，我们提出如下几个改进措施。

**非线性特征的引入**

前面已经提及过，真实的人脑情感分类实际上是严重非线性的，基于简单线性组合的模型性能是有限的。所以为了提高模型的准确率，有必要在模型中引入非线性。

所谓非线性，指的是词语之间的相互组合形成新的语义。事实上，我们的初步模型中已经简单地引入了非线性——在前面的模型中，我们将积极词语和消极词语相邻的情况，视为一个组合的消极语块，赋予它负的权值。更精细的组合权值可以通过“词典矩阵”来实现，即我们将已知的积极词语和消极词语都放到同一个集合来，然后逐一编号，通过如下的“词典矩阵”，来记录词组的权值。

词语（空词）喜欢爱http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0讨厌http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0（空词）http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0喜欢http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0爱http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0034.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0讨厌http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0031.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2026.png?V=2.7.0词语（空词）喜欢爱…讨厌…（空词）012…−1…喜欢123…−2…爱234…−2…⋮⋮⋮⋮⋮⋮…讨厌−1−2−3…−2…⋮⋮⋮⋮⋮⋮…

并不是每一个词语的组合都是成立的，但我们依然可以计算它们之间的组合权值，情感权值的计算可以阅读参考文献。然而，情感词语的数目相当大，而词典矩阵的元素个数则是其平方，其数据量是相当可观的，因此，这已经初步进入大数据的范畴。为了更加高效地实现非线性，我们需要探索组合词语的优化方案，包括构造方案和储存、索引方案。

**情感词典的自动扩充**

在如今的网络信息时代，新词的出现如雨后春笋，其中包括“新构造网络词语”以及“将已有词语赋予新的含义”；另一方面，我们整理的情感词典中，也不可能完全包含已有的情感词语。因此，自动扩充情感词典是保证情感分类模型时效性的必要条件。目前，通过网络爬虫等手段，我们可以从微博、社区中收集到大量的评论数据，为了从这大批量的数据中找到新的具有情感倾向的词语，我们的思路是无监督学习式的词频统计。

我们的目标是“自动扩充”，因此我们要达到的目的是基于现有的初步模型来进行无监督学习，完成词典扩充，从而增强模型自身的性能，然后再以同样的方式进行迭代，这是一个正反馈的调节过程。虽然我们可以从网络中大量抓取评论数据，但是这些数据是无标注的，我们要通过已有的模型对评论数据进行情感分类，然后在同一类情感（积极或消极）的评论集合中统计各个词语的出现频率，最后将积极、消极评论集的各个词语的词频进行对比。某个词语在积极评论集中的词频相当高，在消极评论集中的词频相当低，那么我们就有把握将该词语添加到消极情感词典中，或者说，赋予该词语负的权值。

举例来说，假设我们的消极情感词典中并没有“黑心”这个词语，但是“可恶”、“讨厌”、“反感”、“喜欢”等基本的情感词语在情感词典中已经存在，那么我们就会能够将下述句子正确地进行情感分类：

句子这个黑心老板太可恶了我很反感这黑心企业的做法很讨厌这家黑心店铺这家店铺真黑心！http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0权值http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0030.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0句子权值这个黑心老板太可恶了−2我很反感这黑心企业的做法−2很讨厌这家黑心店铺−2这家店铺真黑心！0⋮⋮

其中，由于消极情感词典中没有“黑心”这个词语，所以“这家店铺真黑心！”就只会被判断为中性（即权值为0）。分类完成后，对所有词频为正和为负的分别统计各个词频，我们发现，新词语“黑心”在负面评论中出现很多次，但是在正面评论中几乎没有出现，那么我们就将黑心这个词语添加到我们的消极情感词典中，然后更新我们的分类结果：

句子这个黑心老板太可恶了我很反感这黑心企业的做法很讨厌这家黑心店铺这家店铺真黑心！http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0权值http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0033.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/2212.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/0032.png?V=2.7.0http://kexue.fm/sci/MathJax/fonts/HTML-CSS/TeX/png/Main/Regular/336/22EE.png?V=2.7.0句子权值这个黑心老板太可恶了−3我很反感这黑心企业的做法−3很讨厌这家黑心店铺−3这家店铺真黑心！−2⋮⋮

于是我们就通过无监督式的学习扩充了词典，同时提高了准确率，增强了模型的性能。这是一个反复迭代的过程，前一步的结果可以帮助后一步的进行。

**本文结论**

综合上述研究，我们得出如下结论：

基于情感词典的文本情感分类是容易实现的，其核心之处在于情感词典的训练。

语言系统是相当复杂的，基于情感词典的文本情感分类只是一个线性的模型，其性能是有限的。

在文本情感分类中适当地引入非线性特征，能够有效地提高模型的准确率。

引入扩充词典的无监督学习机制，可以有效地发现新的情感词，保证模型的强健性和时效性。

**参考文献**

Deep Learning（深度学习）学习笔记整理：<http://blog.csdn.net/zouxy09/article/details/8775360>  
Yoshua Bengio, Réjean Ducharme Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model, 2003  
一种新的语言模型：<http://blog.sciencenet.cn/blog-795431-647334.html>  
评论数据的情感分析数据集：<http://www.datatang.com/data/11857>  
“结巴”中文分词：<https://github.com/fxsjy/jieba>  
NLPIR汉语分词系统：<http://ictclas.nlpir.org/>  
smallseg：<https://code.google.com/p/smallseg/>  
yaha分词：<https://github.com/jannson/yaha>  
情感分析用词语集（beta版）：<http://www.keenage.com/html/c_bulletin_2007.htm>  
NTUSD-简体中文情感极性词典：<http://www.datatang.com/data/11837>  
程度副词及强度和否定词表：<http://www.datatang.com/data/44198>  
现有情感词典汇总：<http://www.datatang.com/data/46922>  
BosonNLP：<http://bosonnlp.com/product>

**实现平台**

我们队所做的编程工具，在以下环境中测试完成：

**Windows 8.1** 微软操作系统。

**Python** 3.4 开发平台/编程语言。选择3.x而不是2.x版本的主要原因是3.x版本对中文字符的支持更好。

**Numpy** Python的一个数值计算库，为Python提供了快速的多维数组处理的能力。

**Pandas** Python的一个数据分析包。

**结巴分词** Python平台的一个中文分词工具，也有Java、C++、Node.js等版本。

**代码列表**

资源：[情感极性词典.zip](http://kexue.fm/usr/uploads/2017/09/1922797046.zip)

**预处理**

#-\*- coding: utf-8 -\*-

import numpy as np #导入numpy

import pandas as pd

import jieba

def yuchuli(s,m): #导入文本，文本预处理

wenjian = pd.read\_csv(s, delimiter=' xxx ', encoding='utf-8', \

header= None, names=['comment']) #导入文本

wenjian = wenjian['comment'].str.replace('(<.\*?>.\*?<.\*?>)','').str.replace('(<.\*?>)','')\

.str.replace('(@.\*?[ :])',' ') #替换无用字符

wenjian = pd.DataFrame({'comment':wenjian[wenjian != '' ]})

wenjian.to\_csv('out\_'+s, header=False, index=False)

wenjian['mark'] = m #样本标记

return wenjian.reset\_index()

neg = yuchuli('data\_neg.txt',-1)

pos = yuchuli('data\_pos.txt',1)

mydata = pd.concat([neg,pos],ignore\_index=True)[['comment','mark']] #结果文件

#预处理基本结束

**加载情感词典**

#开始加载情感词典

negdict = [] #消极情感词典

posdict = [] #积极情感词典

nodict = [] #否定词词典

plusdict = [] #程度副词词典

sl = pd.read\_csv('dict/neg.txt', header=None, encoding='utf-8')

for i in range(len(sl[0])):

negdict.append(sl[0][i])

sl = pd.read\_csv('dict/pos.txt', header=None, encoding='utf-8')

for i in range(len(sl[0])):

posdict.append(sl[0][i])

sl = pd.read\_csv('dict/no.txt', header=None, encoding='utf-8')

for i in range(len(sl[0])):

nodict.append(sl[0][i])

sl = pd.read\_csv('dict/plus.txt', header=None, encoding='utf-8')

for i in range(len(sl[0])):

plusdict.append(sl[0][i])

#加载情感词典结束

**预测函数**

#预测函数

def predict(s, negdict, posdict, nodict, plusdict):

p = 0

sd = list(jieba.cut(s))

for i in range(len(sd)):

if sd[i] in negdict:

if i>0 and sd[i-1] in nodict:

p = p + 1

elif i>0 and sd[i-1] in plusdict:

p = p - 2

else: p = p - 1

elif sd[i] in posdict:

if i>0 and sd[i-1] in nodict:

p = p - 1

elif i>0 and sd[i-1] in plusdict:

p = p + 2

elif i>0 and sd[i-1] in negdict:

p = p - 1

elif i<len(sd)-1 and sd[i+1] in negdict:

p = p - 1

else: p = p + 1

elif sd[i] in nodict:

p = p - 0.5

return p

#预测函数结束

**简单的测试**

#简单的测试

tol = 0

yes = 0

mydata['result'] = 0

for i in range(len(mydata)):

print(i)

tol = tol + 1

if predict(mydata.loc[i,'comment'], negdict, posdict, nodict, plusdict)\*mydata.loc[i,'mark'] > 0:

yes = yes + 1

mydata.loc[i,'result'] = 1

print(yes/tol)

***转载到请包括本文地址：***[*http://spaces.ac.cn/archives/3360/*](http://spaces.ac.cn/archives/3360/)