

# 怎样做情感分析



不会停的蜗牛 (/u/7b67af2e61b3) [+ 关注](#)

2016.08.27 10:33\* 字数 1923 阅读 5464 评论 4 喜欢 16

(/u/7b67af2e61b3)

## 本文结构：

1. 什么是情感分析？
2. 怎么分析，技术上如何实现？

---

cs224d Day 7: 项目2-命名实体识别

2016课程地址 ([https://link.jianshu.com?  
t=https://web.archive.org/web/20160314075834/http://cs224d.stanford.edu/syllabus.ht](https://link.jianshu.com?t=https://web.archive.org/web/20160314075834/http://cs224d.stanford.edu/syllabus.html)

[ml\)](https://web.archive.org/web/20160314075834/http://cs224d.stanford.edu/syllabus.html)

项目描述地址 ([https://link.jianshu.com?  
t=https://web.archive.org/web/20160313081217/https://cs224d.stanford.edu/assignmen](https://link.jianshu.com?t=https://web.archive.org/web/20160313081217/https://cs224d.stanford.edu/assignment1/index.html)

[t1/index.html\)](https://web.archive.org/web/20160313081217/https://cs224d.stanford.edu/assignment1/index.html)

## 什么是情感分析？

就是要识别出用户对一件事一个物或一个人的看法、态度，比如一个电影的评论，一个商品的评价，一次体验的感想等等。根据对带有情感色彩的主观性文本进行分析，识别出用户的态度，是喜欢，讨厌，还是中立。在实际生活中有很多应用，例如通过对



Twitter 用户的情感分析，来预测股票走势、预测电影票房、选举结果等，还可以用来了解用户对公司、产品的喜好，分析结果可以被用来改善产品和服务，还可以发现竞争对手的优劣势等等。

## 怎么分析，技术上如何实现？

首先这是个分类问题。

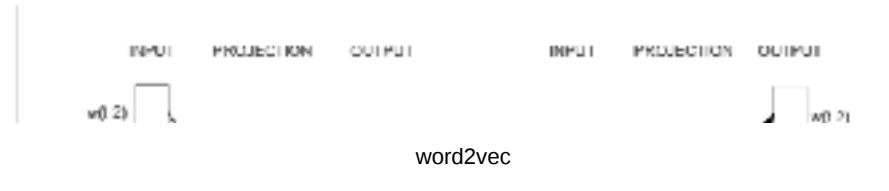
最开始的方案是在文中找到具有各种感情色彩属性的词，统计每个属性的词的个数，哪个类多，这段话就属于哪个属性。但是这存在一个问题，例如 don't like，一个属于否定，一个属于肯定，统计之后变成 0 了，而实际上应该是否定的态度。再有一种情况是，前面几句是否定，后面又是肯定，那整段到底是中立还是肯定呢，为了解决这样的问题，就需要考虑上下文的环境。

2013年谷歌发了两篇论文，介绍了 Continuous Bag of Words (CBOW) 和 Skip-gram 这两个模型，也就是 Word2Vec 方法，这两种模型都是先将每个单词转化成一个随机的 N 维向量，训练之后得到每个单词的最优表示向量，区别是，CBOW 是根据上下文来预测当前词语，Skip-gram 刚好相反，是根据当前词语来预测上下文。

Word2Vec 方法不仅可以捕捉上下文语境，同时还压缩了数据规模，让训练更快更高效。通过这个模型得到的词向量已经可以捕捉到上下文的信息。比如，可以利用基本代数公式来发现单词之间的关系（比如，“国王”-“男人”+“女人”=“王后”）。用这些自带上下文信息的词向量来预测未知数据的情感状况的话，就可以更准确。

(/apps/download?  
utm\_source=sbc)





(/apps/download?  
utm\_source=sbc)

今天的小项目，就是用 word2vec 去解决情感分析问题的。先来简单介绍一下大体思路，然后进入代码版块。

**思路分为两部分**，第一步，就是先用 word2vec 和 SGD 训练出每个单词的最优表示向量。第二步，用 Softmax Regression 对训练数据集的每个句子进行训练，得到分类器的参数，用这个参数就可以预测新的数据集的情感分类。其中训练数据集的每个句子，都对应一个 0 - 1 之间的浮点得分，将这个得分化为 0-4 整数型 5 个级别，分别属于 5 种感情类别，讨厌，有点讨厌，中立，有点喜欢，喜欢。然后将每个句子的词转化成之前训练过的词向量，这样哪些词属于哪个类就知道了，然后用分类器得到分类的边界，得到的参数就可以用来进行预测。

## 具体实现

接下来以一个初学者的角度来讲一下要如何利用这几个模型和算法来实现情感分析这个任务的，因为项目的代码有点多，不方便全写在文章里，可以去[这里](https://link.jianshu.com?t=https://github.com/AliceDudu/Sentiment-Analysis) (https://link.jianshu.com?t=https://github.com/AliceDudu/Sentiment-Analysis) 查看完整代码。

**第一步，用 word2vec 和 SGD 训练出每个单词的最优表示向量。**

- 执行 c7\_run\_word2vec.py
- 其中训练词向量的方法是 c5\_word2vec.py
- 同时用 c6\_sgd.py 训练参数，并且将结果保存起来，每1000次迭代保存在一个文件中 saved\_params\_1000.npy

### word2vec :

上面提到了，它有两种模型 CBOW 和 Skip-gram，每一种都可以用来训练生成最优的词向量，同时还有两种 cost function 的定义方式，一种是 Softmax cost function，一种是 Negative sampling cost function，所以在提到 word2vec 的时候，其实是可以有 4 种搭配的方法的，这个小项目里用到的是 Skip-gram 和 Negative sampling cost function 的结合方式。

### 先定义 skipgram 函数：

给一个中心词 currentWord，和它的窗口大小为 2C 的上下文 contextWords，要求出代表它们的词向量矩阵 W1 和 W2。

```
def skipgram(currentWord, C, contextWords, tokens, inputVectors, outputVectors,
             dataset, word2vecCostAndGradient = softmaxCostAndGradient):
    """ Skip-gram model in word2vec """

    currentI = tokens[currentWord]                #the order of this center word
    predicted = inputVectors[currentI, :]          #turn this word to vector representation

    cost = 0.0
    gradIn = np.zeros(inputVectors.shape)
    gradOut = np.zeros(outputVectors.shape)
    for cwd in contextWords:                      #contextWords is of 2C length
        idx = tokens[cwd]
        cc, gp, gg = word2vecCostAndGradient(predicted, idx, outputVectors, dataset)
        cost += cc                                #final cost/gradient is the sum of all
        gradOut += gg
        gradIn[currentI, :] += gp

    return cost, gradIn, gradOut
```

这里用到的成本函数是 **Negative sampling**，我们的目的就是要使这个成本函数达到最小，然后用这个极值时的参数 grad，也就是可以得到要求的 wordvectors。要增加准确度，所以可以多次生成中心词和上下文进行训练，然后取平均值，也就是函数 word2vec\_sgd\_wrapper 做的事情。

(/apps/download?  
utm\_source=sbc)



```
def negSamplingCostAndGradient(predicted, target, outputVectors, dataset, K=10):
    """ Negative sampling cost function for word2vec models """

    grad = np.zeros(outputVectors.shape)
    gradPred = np.zeros(predicted.shape)

    indices = [target]
    for k in xrange(K):
        newidx = dataset.sampleTokenIdx()
        while newidx == target:
            newidx = dataset.sampleTokenIdx()
        indices += [newidx]

    labels = np.array([1] + [-1 for k in xrange(K)])
    vecs = outputVectors[indices, :]

    t = sigmoid(vecs.dot(predicted) * labels)
    cost = -np.sum(np.log(t))

    delta = labels * (t-1)
    gradPred = delta.reshape((1, K+1)).dot(vecs).flatten()
    gradtemp = delta.reshape((K+1, 1)).dot(predicted.reshape(1, predicted.shape[0]))

    for k in xrange(K+1):
        grad[indices[k]] += gradtemp[k, :]

    return cost, gradPred, grad
```

(/apps/download?  
utm\_source=sbc)



**\*\*接着用 sgd \*\*迭代 40000 次得到训练好的 wordVectors。**

```
wordVectors0 = sgd(
    lambda vec: word2vec_sgd_wrapper(skipgram, tokens, vec, dataset, C,
        negSamplingCostAndGradient),
    wordVectors, 0.3, 40000, None, True, PRINT_EVERY=10)
```

关于 word2vec 之前有写过一篇 word2vec 模型思想和代码实现

(<https://www.jianshu.com/p/86134284fa14>)，想了解详细原理和具体怎样实现的童鞋可以去这个这里看。

**第二步，用 Softmax Regression 对训练数据集进行分类学习。**



- 执行 c10\_sentiment.py
- 其中用 c6\_sgd.py 去训练权重 weights ,
- 然后用 c8\_softmaxreg.py 根据训练好的 features , labels , weights 进行类别 label 的预测。

(/apps/download?  
utm\_source=sbc) ×

**先将数据集分为三部分 , training set , deviation set , 和 test set.**

```
trainset = dataset.getTrainSentences()  
devset = dataset.getDevSentences()  
testset = dataset.getTestSentences()
```

在 trainset 中 , 每句话对应一个情感的得分或者说是分类 , 先将每个 word 在 token 中找到序号 , 然后在第一步训练好的 wordvectors 中找到相应的词向量。

```
trainFeatures[i, :] = getSentenceFeature(tokens, wordVectors, words)
```

然后用 sgd 和 softmax\_wrapper 迭代 10000 次去训练 weights :

```
weights = sgd(lambda weights: softmax_wrapper(trainFeatures, trainLabels, weights, r
```

**接着用 softmax regression 进行分类的预测 :**

```
_, _, pred = softmaxRegression(trainFeatures, trainLabels, weights)
```

上面用到了不同的 REGULARIZATION = [0.0, 0.00001, 0.00003, 0.0001, 0.0003, 0.001, 0.003, 0.01] , 在其中**选择 accuracy 最好的 REGULARIZATION 和相应的结果 :**



```
best_dev = 0
for result in results:
    if result["dev"] > best_dev:
        best_dev = result["dev"]
        BEST_REGULARIZATION = result["reg"]
        BEST_WEIGHTS = result["weights"]
```

(/apps/download?  
utm\_source=sbc) ×

用这个最好的参数在 test set 上进行预测：

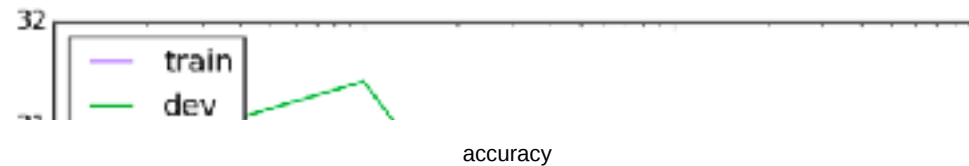
```
_, _, pred = softmaxRegression(testFeatures, testLabels, BEST_WEIGHTS)
```

并且得到 accuracy：

```
print "Test accuracy (%): %f" % accuracy(testLabels, pred)
```

下图是 accuracy 和 REGULARIZATION 在 devset 和 trainset 上的趋势：





(/apps/download?  
utm\_source=sbc) ×

以上就是 sentiment analysis 的基本实现，把它和爬虫相结合，会有很多好玩的玩儿法！

### [cs224d]

Day 1. 深度学习与自然语言处理 主要概念一览

(<https://www.jianshu.com/p/6993edef96e4>)

Day 2. TensorFlow 入门 (<https://www.jianshu.com/p/6766fbcd43b9>)

Day 3. word2vec 模型思想和代码实现 (<https://www.jianshu.com/p/86134284fa14>)

Day 4. 怎样做情感分析 (<https://www.jianshu.com/p/1909031bb1f2>)

Day 5. CS224d - Day 5: RNN快速入门 (<https://www.jianshu.com/p/bf9ddfb21b07>)

Day 6. 一文学会用 Tensorflow 搭建神经网络

(<https://www.jianshu.com/p/e112012a4b2d>)

Day 7. 用深度神经网络处理NER命名实体识别问题

(<https://www.jianshu.com/p/581832f2c458>)

Day 8. 用 RNN 训练语言模型生成文本 (<https://www.jianshu.com/p/b4c5ff7c450f>)

Day 9. RNN与机器翻译 (<https://www.jianshu.com/p/23b46605857e>)

Day 10. 用 Recursive Neural Networks 得到分析树

(<https://www.jianshu.com/p/403665b55cd4>)

Day 11. RNN的高级应用 (<https://www.jianshu.com/p/0e840f92b532>)

我是 不会停的蜗牛Alice

85后全职主妇

喜欢人工智能，行动派

创造力，思考力，学习力提升修炼进行中

欢迎您的喜欢，关注和评论！





不会停的蜗牛 (/u/7b67af2e61b3)

写了 224835 字，被 3243 人关注，获得了 1969 个喜欢

(/u/7b67af2e61b3)

+ 关注

我是 Alice 喜欢人工智能，行动派 创造力，思考力，学习力提升修炼进行中 欢迎志同道合的小伙伴们和我一...

(/apps/download?utm\_source=sbc) ✕

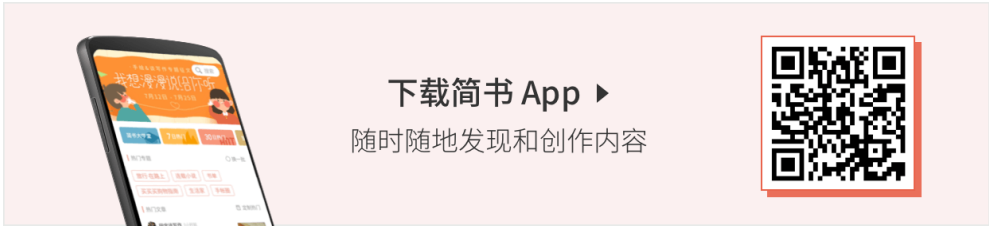
喜欢就点赞，有用就随意打赏吧 😊 Run with AI !

赞赏支持

♡ 喜欢 (/sign\_in?utm\_source=desktop&utm\_medium=not-signed-in-like-button) | 16



(http://cwb.assets.jianshu.io/notes/images/5482816/weibo/image\_c



(/apps/download?utm\_source=nbc)



登录后发表评论 (/sign\_in?utm\_source=desktop&utm\_medium=not-signed-in-comment-form)

4条评论

只看作者

按喜欢排序 按时间正序 按时间倒序





修炼手册 (/u/d994da65a00f)

2楼 · 2016.08.28 17:02

(/u/d994da65a00f)

谢谢你的分享，最近在做nlp相关的比赛~

👍 赞    💬 回复

不会停的蜗牛 (/u/7b67af2e61b3) : @ino\_jonshoo (/users/d994da65a00f) 好开心可以给你带来一点帮助

2016.08.28 21:26    💬 回复

✍ 添加新评论



匿名用户甲乙丙丁 (/u/e43f7a8e8fb6)

3楼 · 2016.09.18 08:21

(/u/e43f7a8e8fb6)

"接着用 softmax regression 进行分类的预测"，我想问一下，这里为何用softmax regression？不是每次训练出来都会有一个1到5的得分吗，然后再用这个得分和真实标签作对比？如果是softmax，训练出来的应该是一个得分的向量，表示其在每个情感的具体得分，那么这时候应该如何跟真实标签对比呢？望不吝赐教！

👍 赞    💬 回复



南唐逸少 (/u/c207ce7738de)

4楼 · 2017.10.20 15:19

(/u/c207ce7738de)

博主数据能不能提供一下，链接里没有了

👍 赞    💬 回复

被以下专题收入，发现更多相似内容



数据科学家 (/c/0adc32d3cf07?utm\_source=desktop&utm\_medium=notes-

(/apps/download?utm\_source=sbc)



included-collection)



坚持写作100天 (/c/6f43264f8299?

utm\_source=desktop&amp;utm\_medium=notes-included-collection)



Tensorflow (/c/4353fe63db7e?utm\_source=desktop&amp;utm\_medium=notes-

included-collection)



NLP (/c/173cbe9cfc2?utm\_source=desktop&amp;utm\_medium=notes-

included-collection)



数据科学 (/c/102149797c26?utm\_source=desktop&amp;utm\_medium=notes-

included-collection)

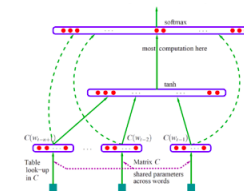


NLP-情感分析 (/c/eb12993c685a?

utm\_source=desktop&amp;utm\_medium=notes-included-collection)

(/apps/download?  
utm\_source=sbc) ×

(/p/4bca99d40597?



utm\_campaign=maleskine&amp;utm\_content=note&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**NLP-词嵌入学习笔记 (/p/4bca99d40597?utm\_campaign=maleskine&utm...**

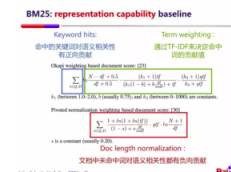
1.NLP当前热点方向 词法/句法分析 词嵌入(word embedding) 命名实体识别(Name Entity Recognition) 机器翻译(Machine Translation) 情感分析(sentiment analysis) 文档摘要(automatic...



\_\_Aragorn (/u/79a839788418?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)


(/p/3a9f49834c4a?



utm\_campaign=maleskine&amp;utm\_content=note&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

## 浅谈智能搜索和对话式OS (/p/3a9f49834c4a?utm\_campaign=maleskine&...

前面的文章主要从理论的角度介绍了自然语言人机对话系统所可能涉及到的多个领域的经典模型和基础知识。这篇文章，甚至之后的文章，会从更贴近业务的角度来写，侧重于介绍一些与自然语言问答业务密切相..

 我偏笑\_NS Nirvana (/u/2293f85dc197?

utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation) (/apps/download?utm\_source=sbc) ✕


(/p/d443aab9bcb1?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

## 在 TensorFlow 上使用 LSTM 进行情感分析 (/p/d443aab9bcb1?utm\_camp...

你可以从 Github 上面下载到所有的源代码。在这篇教程中，我们将介绍如何将深度学习技术应用到情感分析中。该任务可以被认为是一个句子，一段话，或者是一个文档中，将作者的情感分为积极的，消极的..

 chen\_h (/u/b20d6310182a?

utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)


(/p/15411de409f1?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

## 使用 TensorFlow 做文本情感分析 (/p/15411de409f1?utm\_campaign=mal...

使用 TensorFlow 做文本情感分析 本文将通过使用TensorFlow中的LSTM神经网络方法探索高效的深度学习方法。作者：Adit Deshpande July 13, 2017 翻译来源：https://www.oreilly.com/learning/perf...

 Datartisan数据工匠 (/u/ad75474d9e73?

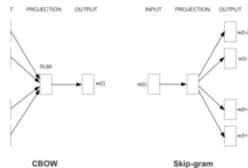
utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/7864843880e5?

utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

词向量生成模型---word2vec (/p/7864843880e5?utm\_...

在各种大举深度学习大旗的公司中，Google公司无疑是旗举得最高的，口号喊得最响亮的那一个。2013年末，Google发布的word2vec工具引起了一帮人的...



chaaff (/u/3fa714982f82?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation) (/apps/download?utm\_source=sbc) X

(/p/f2794c1b0a90?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

D女郎丰胸真的有用吗？会不会有激素？多久有效果？多少钱一套？ (/p/f279...

D女郎丰胸有用吗？会不会有激素？多久有效果？多少钱一套？ 爱美是女人的天性，没有哪个女人不希望自己能够拥有水嫩光滑的肌肤，凹凸有致的身材，能够轻松让自己心仪的异性为自己神魂颠倒。所以现在的女..

ranyue999 (/u/2de5897a9cc5?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/7339c4d02492?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

孩子爸照顾俩儿子第一天 (/p/7339c4d02492?utm\_campaign=maleskine&...

2017年9月12日，我踏上了开往南京的列车，开始了南京十日学习之旅！我家老公也开始了照顾两个儿子生活的艰巨任务..... 7点34分列车准时驶出郑州站，我开始睡觉，但是睡不着，脑子里都是：儿子这会儿到学...

蓝莓提拉米苏two (/u/dcfb56582d65?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/bef58f024c0d?

utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

## 郭天王换女票了，咱们也换个新闻女 (/p/bef58f024c0d...

►这两天想必大家微博、盆友圈已经被网红这个词刷爆了 至于原因嘛，就是天王郭富城正式宣布脱单。目测新的天王嫂，的确还是美美哒的 但是一路看下...



小小星球 (/u/884a0da63886?)

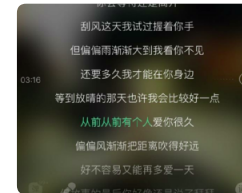
utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)



(/apps/download?  
utm\_source=sbc)



(/p/e6290609351a?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

## 很喜欢的人结婚了 (/p/e6290609351a?utm\_campaign=maleskine&utm\_c...

“从前从前 有个人爱你很久 但偏偏 风渐渐 把距离吹得好远” “三八节是跟媳妇结婚一周年的日子啊，大小情人爱你们”，你朋友圈发布说。还看到你可爱的女儿，跟你 长得好像。一家其乐融融。而我终究是和你没有一...



喵小姐说 (/u/c30db65105fc?)

utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

## 在一起 (/p/0391dadb4891?utm\_campaign=maleskine&utm\_content=not...

如果分开只会让我们变得陌生，那就永远在一起。如果再见我们还是陌生人，那就不要再见了



铃兰君子 (/u/6d6597b08448?)

utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)



