# Python做文本情感分析之情感极性分析



一个不知死活的胖子 (/u/b04c0d6732ff) (+ 关注) 2016.08.30 05:36\* 字数 2626 阅读 27757 评论 41 喜欢 80 (/u/b04c0d6732ff)

# 导语

「NLP」最为目前及其火热的一个领域,已经逐渐渗透进越来越多产业的各项业 务中,不知死活的胖子决定对常用的应用功能挨个进行尝试,死活不论......

# 0. 介绍

「情感极性分析」是对带有感情色彩的主观性文本进行分析、处理、归纳和推理的过 程。按照处理文本的类别不同,可分为基于**新闻**评论的情感分析和基于**产品**评论的情感 分析。其中,前者多用于舆情监控和信息预测,后者可帮助用户了解某一产品在大众心 目中的口碑。

目前常见的情感极性分析方法主要是两种:基于情感词典的方法和基于机器学习的方 法。

# 1. 基于情感词典的文本情感极性分析

笔者是通过**情感打分**的方式进行文本情感极性判断, score > 0 判断为正向, score < 0 判断为负向。

#### 1.1 数据准备

# 1.1.1 情感词典及对应分数

词典来源于BosonNLP数据下载 (https://link.jianshu.com? t=http://bosonnlp.com/dev/resource)的情感词典,来源于社交媒体文本,所以词 典适用于处理社交媒体的情感分析。

词典把所有常用词都打上了唯一分数有许多不足之处。

- 之一, 不带情感色彩的停用词会影响文本情感打分。在
- 之二,由于中文的博大精深,词性的多变成为了影响模型准确度的重要原因。 一种情况是**同一个词在不同的语境下可以是代表完全相反的情感意义**,用笔者模型预 测偏差最大的句子为例(来源于朋友圈文本):

有车一族都用了这个宝贝,后果很严重哦[偷笑][偷笑][偷笑]1,交警工资估计会打5折,没有超速罚款了 [呲牙][呲牙][呲牙]2,移动联通公司大幅度裁员,电话费少了[呲牙][呲牙][呲牙]3,中石化中石油裁员2 成,路痴不再迷路,省油[悠闲][悠闲][悠闲]5,保险公司裁员2成,保费折上折2成,全国通用[憨笑][憨笑] [憨笑]买不买你自己看着办吧[调皮][调皮][调皮]

里面 严重 等词都是表达的相反意思,甚至整句话一起表示相反意思,不知死活的笔者 还没能深入研究如何用词典的方法解决这类问题,但也许可以用机器学习的方法让神 经网络进行学习能够初步解决这一问题。

æ

另外, **同一个词可作多种词性, 那么情感分数也不应相同**, 例如:

这部电影真 垃圾

垃圾 分类

很明显在第一句中  $ext{top}$  表现强烈的贬义,而在第二句中表示中性,单一评分对于这类问题的分类难免有失偏颇。

#### 1.1.2 否定词词典

否定词的出现将直接将句子情感转向相反的方向,而且通常效用是叠加的。常见的否定词: 不、没、无、非、莫、弗、勿、毋、未、否、别、無、休、难道等。

#### 1.1.3 程度副词词典

既是通过打分的方式判断文本的情感正负,那么分数绝对值的大小则通常表示情感强弱。既涉及到程度强弱的问题,那么程度副词的引入就是势在必行的。词典可从《知网》情感分析用词语集(beta版) (https://link.jianshu.com?

t=http://www.keenage.com/download/sentiment.rar)下载。词典内数据格式可参考如下格式,即共两列,第一列为程度副词,第二列是程度数值, > 1 表示强化情感, < 1 表示弱化情感。



程度副词词典

#### 1.1.4 停用词词典

中科院计算所中文自然语言处理开放平台发布了有1208个停用词的中文停用词表 (https://link.jianshu.com?t=http://www.datatang.com/data/43894),也有其他不需要积分的下载途径 (https://link.jianshu.com?t=http://www.hicode.cn/download/view-software-13784.html)。

# 1.2 数据预处理

#### 1.2.1 分词

≪

即将句子拆分为词语集合,结果如下:

e.g. 这样/的/酒店/配/这样/的/价格/还算/不错

Python常用的分词工具:

- 结巴分词 Jieba
- Pymmseg-cpp
- Loso
- smallseg

```
from collections import defaultdict
import os
import re
import jieba
import codecs
1. 文本切割
def sent2word(sentence):
   Segment a sentence to words
   Delete stopwords
   segList = jieba.cut(sentence)
   segResult = []
   for w in segList:
       segResult.append(w)
   stopwords = readLines('stop_words.txt')
   newSent = []
   for word in segResult:
       if word in stopwords:
           # print "stopword: %s" % word
           continue
           newSent.append(word)
   return newSent
```

在此笔者使用Jieba进行分词。

# 1.2.2 去除停用词

遍历所有语料中的所有词语,**删除其中的停用词** e.g. 这样/的/酒店/配/这样/的/价格/还算/不错

--> 酒店/配/价格/还算/不错

# 1.3 构建模型

# 1.3.1 将词语分类并记录其位置

将句子中各类词分别存储并标注位置。

℀

```
.....
2. 情感定位
def classifyWords(wordDict):
    # (1) 情感词
    senList = readLines('BosonNLP_sentiment_score.txt')
    senDict = defaultdict()
    for s in senList:
        senDict[s.split(' ')[0]] = s.split(' ')[1]
    # (2) 否定词
    notList = readLines('notDict.txt')
    # (3) 程度副词
    degreeList = readLines('degreeDict.txt')
    degreeDict = defaultdict()
    for d in degreeList:
        degreeDict[d.split(',')[0]] = d.split(',')[1]
    senWord = defaultdict()
    notWord = defaultdict()
    degreeWord = defaultdict()
    for word in wordDict.keys():
        if word in senDict.keys() and word not in notList and word not in degreeDict
            senWord[wordDict[word]] = senDict[word]
        elif word in notList and word not in degreeDict.keys():
            notWord[wordDict[word]] = -1
        elif word in degreeDict.keys():
           degreeWord[wordDict[word]] = degreeDict[word]
    return senWord, notWord, degreeWord
```

#### 1.3.2 计算句子得分

在此,简化的情感分数计算逻辑: **所有情感词语组的分数之和** 

#### 伪代码如下:

```
finalSentiScore = (-1) ^ (num of notWords) * degreeNum * sentiScore
finalScore = sum(finalSentiScore)
```

```
.....
3. 情感聚合
.....
def scoreSent(senWord, notWord, degreeWord, segResult):
   W = 1
   score = 0
   # 存所有情感词的位置的列表
   senLoc = senWord.keys()
   notLoc = notWord.keys()
   degreeLoc = degreeWord.keys()
   senloc = −1
   \# notloc = -1
   # degreeloc = -1
   # 遍历句中所有单词segResult, i为单词绝对位置
   for i in range(0, len(segResult)):
       # 如果该词为情感词
       if i in senLoc:
           # loc为情感词位置列表的序号
           senloc += 1
           # 直接添加该情感词分数
           score += W * float(senWord[i])
           # print "score = %f" % score
           if senloc < len(senLoc) - 1:</pre>
              # 判断该情感词与下一情感词之间是否有否定词或程度副词
              # j为绝对位置
              for j in range(senLoc[senloc], senLoc[senloc + 1]):
                  # 如果有否定词
                  if j in notLoc:
                      W ∗= -1
                  # 如果有程度副词
                  elif j in degreeLoc:
                      W *= float(degreeWord[j])
       # i定位至下一个情感词
       if senloc < len(senLoc) - 1:</pre>
           i = senLoc[senloc + 1]
   return score
```

# 1.4 模型评价

将600多条朋友圈文本的得分排序后做出散点图:

Score Distribution

其中大多数文本被判为正向文本符合实际情况,且绝大多数文本的情感得分的绝对值在 10以内,这是因为笔者在计算一个文本的情感得分时,以句号作为一句话结束的标志, 在一句话内,情感词语组的分数累加,如若一个文本中含有多句话时,则取其**所有句子** 情感得分的平均值。

℀

然而,这个模型的**缺点与局限性**也非常明显:

- 首先,段落的得分是其所有句子得分的平均值,这一方法并不符合实际情况。正如文章中先后段落有重要性大小之分,一个段落中前后句子也同样有重要性的差异。
- 其次,有一类文本使用贬义词来表示正向意义,这类情况常出现与宣传文本中,还是那个例子:

有车一族都用了这个宝贝,后果很严重哦[偷笑][偷笑][偷笑]1,交警工资估计会打5折,没有超速罚款了[呲牙][呲牙][呲牙]2,移动联通公司大幅度裁员,电话费少了[呲牙][呲牙][呲牙]3,中石化中石油裁员2成,路 瘛不再迷路,省油[悠闲][悠闲][悠闲]5,保险公司裁员2成,保费折上折2成,全国通用[憨笑][憨笑][憨笑] 买不买你自己看着办吧[调皮][调皮][调皮]2980元轩辕魔镜带回家,推广还有返利[得意] Score Distribution中得分小于 -10 的几个文本都是与这类情况相似,这也许需要深度

Score Distribution 中得分小士 - 10 的几个文本都是与这类情况相似,这也许需要深度学习的方法才能有效解决这类问题,普通机器学习方法也是很难的。

对于正负向文本的判断,该算法忽略了很多其他的否定词、程度副词和情感词搭配的情况;用于判断情感强弱也过于简单。

总之,这一模型只能用做BENCHMARK...

# 2. 基于机器学习的文本情感极性分析

# 2.1 还是数据准备

#### 2.1.1 停用词

(同1.1.4)

#### 2.1.2 正负向语料库

来源于有关中文情感挖掘的酒店评论语料 (https://link.jianshu.com? t=http://www.datatang.com/data/11936),其中正向7000条,负向3000条(笔者是不是可以认为这个世界还是充满着满满的善意呢…),当然也可以参考情感分析资源(转)(https://link.jianshu.com?t=http://www.cnblogs.com/finesite/p/3381803.html)使用其他语料作为训练集。

#### 2.1.3 验证集

Amazon上对iPhone 6s的评论,来源已不可考......

# 2.2 数据预处理

#### 2.2.1 还是要分词

(同1.2.1)

```
import numpy as np
import sys
import re
import codecs
import os
import jieba
from gensim.models import word2vec
from sklearn.cross_validation import train_test_split
from sklearn.externals import joblib
from sklearn.preprocessing import scale
from sklearn.svm import SVC
from sklearn.decomposition import PCA
from scipy import stats
from keras.models import Sequential
from keras.layers import Dense, Dropout, Activation
{\tt from\ keras.optimizers\ import\ SGD}
from sklearn.metrics import f1_score
from bayes_opt import BayesianOptimization as BO
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
def parseSent(sentence):
    seg_list = jieba.cut(sentence)
output = ''.join(list(seg_list)) # use space to join them
    return output
```

#### 2.2.2 也要去除停用词

(同1.2.2)

#### 2.2.3 训练词向量

(重点来了!)模型的输入需是数据元组,那么就需要将每条数据的词语组合转化为一个数值向量

#### 常见的转化算法有但不仅限于如下几种:

(请原谅不知死活的胖子直接用展示的ppt截图作说明,没错,我就是懒,你打我呀)

#### · Bag of Words

Bag of Words

• TF-IDF

TF-IDF

• Word2Vec

Word2Vec

在此笔者选用Word2Vec将语料转化成向量,具体步骤可参考笔者的文章问答机器人的Python分类器 (https://www.jianshu.com/p/8acb99b6671b)。

مہ

```
def getWordVecs(wordList):
    vecs = []
    for word in wordList:
        word = word.replace('\n', '')
            vecs.append(model[word])
        except KeyError:
           continue
    # vecs = np.concatenate(vecs)
    return np.array(vecs, dtype = 'float')
def buildVecs(filename):
    posInput = []
    with open(filename, "rb") as txtfile:
        # print txtfile
        for lines in txtfile:
            lines = lines.split('\n ')
            for line in lines:
               line = jieba.cut(line)
               resultList = getWordVecs(line)
               # for each sentence, the mean vector of all its vectors is used to
               if len(resultList) != 0:
                    resultArray = sum(np.array(resultList))/len(resultList)
                   posInput.append(resultArray)
    return posInput
# load word2vec model
model = word2vec.Word2Vec.load_word2vec_format("corpus.model.bin", binary = True)
# txtfile = [u'标准间太差房间还不如3星的而且设施非常陈旧.建议酒店把老的标准间从新改善.', u'在这个
posInput = buildVecs('pos.txt')
negInput = buildVecs('pos.txt')
# use 1 for positive sentiment, 0 for negative
y = np.concatenate((np.ones(len(posInput)), np.zeros(len(negInput))))
X = posInput[:]
for neg in negInput:
   X.append(neg)
X = np.array(X)
```

#### 2.2.4 标准化

虽然笔者觉得在这一问题中,标准化对模型的准确率影响不大,当然也可以尝试其他的 标准化的方法。

# standardization

X = scale(X)

### 2.2.5 降维

根据PCA结果,发现前100维能够cover 95%以上的variance。

PCA

```
# PCA
# Plot the PCA spectrum
pca.fit(X)
plt.figure(1, figsize=(4, 3))
plt.clf()
plt.axes([.2, .2, .7, .7])
plt.plot(pca.explained_variance_, linewidth=2)
plt.axis('tight')
plt.xlabel('n_components')
plt.ylabel('explained_variance_')

X_reduced = PCA(n_components = 100).fit_transform(X)
```

#### 2.3 构建模型

# 2.3.1 SVM (RBF) + PCA

SVM (RBF)分类表现更为宽松,且使用**PCA**降维后的模型表现有明显提升,misclassified多为负向文本被分类为正向文本,其中 AUC = 0.92, KSValue = 0.7。 关于SVM的调参可以参考笔者的另一篇文章Python利用Gausian Process对Hyperparameter进行调参 (https://www.jianshu.com/p/90e6abdeb4f2)

```
.....
2.1 SVM (RBF)
   using training data with 100 dimensions
clf = SVC(C = 2, probability = True)
clf.fit(X_reduced_train, y_reduced_train)
print 'Test Accuracy: %.2f'% clf.score(X_reduced_test, y_reduced_test)
pred_probas = clf.predict_proba(X_reduced_test)[:,1]
print "KS value: %f" % KSmetric(y_reduced_test, pred_probas)[0]
# plot ROC curve
# AUC = 0.92
\# KS = 0.7
fpr,tpr,_ = roc_curve(y_reduced_test, pred_probas)
roc_auc = auc(fpr,tpr)
plt.plot(fpr, tpr, label = 'area = %.2f' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.legend(loc = 'lower right')
plt.show()
joblib.dump(clf, "SVC.pkl")
```

# 2.3.2 MLP

MLP相比于SVM (RBF),分类更为严格,PCA降维后对模型准确率影响不大,misclassified多为正向文本被分类为负向,其实是更容易overfitting,原因是语料过少,其实用神经网络未免有些小题大做,AuC = 0.91。

℀

```
.....
2.2 MLP
    using original training data with 400 dimensions
model = Sequential()
model.add(Dense(512, input_dim = 400, init = 'uniform', activation = 'tanh'))
model.add(Dropout(0.5))
model.add(Dense(256, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(128, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(64, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(32, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(1, activation = 'sigmoid'))
model.compile(loss = 'binary_crossentropy',
              optimizer = 'adam',
              metrics = ['accuracy'])
model.fit(X_train, y_train, nb_epoch = 20, batch_size = 16)
score = model.evaluate(X_test, y_test, batch_size = 16)
print ('Test accuracy: ', score[1])
pred_probas = model.predict(X_test)
# print "KS value: %f" % KSmetric(y_reduced_test, pred_probas)[0]
# plot ROC curve
# AUC = 0.91
fpr,tpr,_ = roc_curve(y_test, pred_probas)
roc_auc = auc(fpr,tpr)
plt.plot(fpr, tpr, label = 'area = %.2f' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.legend(loc = 'lower right')
plt.show()
```

# 2.4 模型评价

- 实际上,第一种方法中的第二点缺点依然存在,但相比于基于词典的情感分析方法, 基于机器学习的方法**更为客观**
- 另外由于训练集和测试集分别来自**不同领域**,所以有理由认为训练集不够充分,未来 可以考虑扩充训练集以提升准确率。

# 小礼物走一走,来简书关注我

赞赏支持



(http://cwb.assets.jianshu.io/notes/images/5234134/w



# 下载简书 App ▶

随时随地发现和创作内容



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

#### ▋被以下专题收入,发现更多相似内容

算法 (/c/f417a955c48c?utm\_source=desktop&utm\_medium=notesincluded-collection)

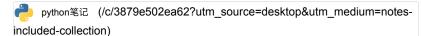


🥏 python加油站 (/c/1680257829b6?

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

文本挖掘 (/c/459400e10ca3?utm\_source=desktop&utm\_medium=notesincluded-collection)

🕠 python 学习 (/c/e80d82fbf8bf?utm\_source=desktop&utm\_medium=notesincluded-collection)



NLP (/c/41ada54bfab2?utm\_source=desktop&utm\_medium=notesincluded-collection)

python\_... (/c/52a872d9c95e?utm\_source=desktop&utm\_medium=notesincluded-collection)

(/p/233da896226a?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation) 利用Python实现中文情感极性分析 (/p/233da896226a?utm\_campaign=m...

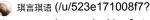
情感极性分析,即情感分类,对带有主观情感色彩的文本进行分析、归纳。情感极性分析主要有两种分类方 法:基于情感知识的方法和基于机器学习的方法。基于情感知识的方法通过一些已有的情感词典计算文本...



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

## 10款数据分析"工具",助你成为新媒体运营领域的"增长黑客" (/p/760a5917...

1、成为新媒体运营领域的"增长黑客" 1.1 一般意义上的"增长黑客" "增长黑客(Growth Hacker)",这个近几 年来风靡中国互联网界的新兴概念,滥觞于美国硅谷互联网创业圈,国内则是由范冰首先引进,他的著作--...



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

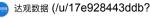
(/p/cb30cf12183f?



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

#### 达观数据情感分析架构演进 (/p/cb30cf12183f?utm\_campaign=maleskine...

在互联网日益发达的今天,许多消费者不管是通过线上电商网站或者线下门店购买商品后,包括买车、买手机等,都会到品牌官网或者一些专业网站甚至社交媒体去发表对产品的评价。对于买家来说,买前查看评...



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

(/p/08b0584a0e98?



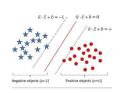
utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation) 用户研究之文本研究系列-3: 用户反馈文本挖掘基础 (/p/08b0584a0e98?ut...

导读 用户反馈文本,作为用户问题、建议、态度的载体,对产品评估和改进优化极具价值。但作为非结构化数据,用户反馈文本的处理并不完全适用结构化数据挖掘流程。本篇,将向大家阐述用户反馈文本挖掘的...

🔕 横雀 (/u/0ad2777a7750?

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

(/p/60505518d7ee?



utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation) 情感分析 (/p/60505518d7ee?utm\_campaign=maleskine&utm\_content=...

我灵愁苦,要发出言语。我心苦恼,要吐露哀情。 ——《旧·伯》7:11 —、什么是情感分析 情感分析 (SA) 又称为倾向性分析和意见挖掘,它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的...

★ 黄老邪怪 (/u/5b8f641944cd?)

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

#### 谈一谈,王老师——我们班的班主任 (/p/203b6abe6724?utm\_campaign=...

我是一个十足的淘气包,用老师的说法就是现实版的马小跳!这种情况维持在了四年级。也就是王老师来的那一年,开学的时候我们以为还是原来的班主任,但是情况不是这样,门口走进来了一个不高不矮、不胖...

🦚 听见时光飞 (/u/47806f219ae3?

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

# Python学习笔记(6) (/p/c1a26cc37e74?utm\_campaign=maleskine&utm\_...

Python 术语 "self" —— 这是一个方法参数, 总是指向当前对象实例 Points 可以在类中定义init()方法来初始化对象实例 类中定义的每个方法都必须提供self作为第一个参数 类中的每个属性前面都必须有self, 从而将数...

zhouyuhan (/u/eac34c4a6a0a?

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

### 书籍拆解50: 《沃顿商学院最受欢迎的成功课》 (/p/7329fb31bcb5?utm\_c...

关于作者: 亚当·格兰特, 沃顿商学院最年轻的终身教授, 目前在该校教授组织行为学课程。他长期担任谷歌、IBM、花旗等公司, 以及联合国和美国海陆空三军的资深顾问, 是成功研究和工作领域研究的顶级学...

玄淼 (/u/4fc62704ae9e?

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

#### 【韩喜文2017.06.03星期六】 好展馆让天下没有卖不出去的产品 好展...

ૡ૾

拥有变得不再重要,使用和共享才是。随着大量的使用,基础设施的边际成本是趋向于零的。只要是判断为基础设施的行业使用成本都会很低。用好了降低成本提升效率,看不懂一定会浪费资源。现在基础设施有...

🌍 韩喜文 (/u/147cc7a4a1d4?

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

# 第一封 (/p/9619aa26e35f?utm\_campaign=maleskine&utm\_content=not...

这是第一次在简书上写东西(虽然和她只相识还不到24个小时),但是很快就被她的美貌倾倒:简洁的界面, 丰富的内容(主要是各种干货),高效的互动真是让人相见恨晚。最近有幸读到了彭小六老师的《让未来现在...

🍃 御雪轻寒 (/u/bc8b66116fb7?

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