

分享到

天善智能

349

文章

11万

阅读

[查看TA的文章>](#)

NLP之淘宝商品评论情感分析

2017-06-05 08:33

欢迎关注天善智能 [hellobi.com](#)，我们是专注于商业智能BI，大数据，数据分析领域的垂直社区，学习、问答、求职，一站式搞定！

对商业智能BI、大数据分析挖掘、机器学习，python，R等数据领域感兴趣的同学加微信：[tstoutiao](#)，邀请你进入数据爱好者交流群，数据爱好者们都在这儿。

前言

最近学习NLP，还在初级阶段，上次jieba分词那篇写完，还在学习哈工大的pyltp。

发现一个比较有趣的中文类库：[snownlp](#)

SnowNLP是一个Python写的类库，可以方便的处理中文文本内容。

使用环境

Python3

Features

中文分词（Character-Based Generative Model）

词性标准（TnT 3-gram 隐马）

情感分析（现在训练数据主要是买卖东西时的评价，所以对其他的一些可能效果不是很好，待解决）

文本分类（Naive Bayes）

转换成拼音

繁体转简体

提取文本关键词（TextRank算法）

提取文本摘要（TextRank算法）

tf, idf

Tokenization（分割成句子）

文本相似（BM25）

支持python3

安装snownlp

在cmd下输入：`pip3 install snownlp`

大家都在搜：WWDC

搜狐号推荐



邻章
TMT观察评论，满怀善意的批



果壳网
面向都市科技青年们的社交网...
的泛科技兴趣社区，并提供负



刘兴亮
刘兴亮的自媒体，分享以互联...
济领域的独家解读和有趣看法。



安卓每日推送
玩友汇（wyh.tv）旗下安卓每...
业发布会快媒体，手机资讯、



雷帝触网
关注互联网行业动态，大爆...
互联网行业融资，行业整合并

24小时热文

1

天猫618手机销量排
一，锤子遗憾未上

2

银联发动移动支付二
微信围剿支付宝？

3

安以轩大婚伴手礼...
定制版 被惊艳了！



性价比炸裂！联通...
量仅9元



WWDC2017大会来
MacBook Air却可能
了

0
分享到



如图，我之前已经安装过

```
snownlp分词 from snownlp import SnowNLP s = SnowNLP( u'一次满意的购物') s.words
['一', '次', '满意', '的', '购物']
```

PS:这里使用的是它自带的词典

snownlp情感分析

这里的情感分析结果是【0，1】区间上的一个值，越接近1，情感越积极，越接近0，情感越消极。

或者可以理解为positive的概率。

```
s.sentiments#positive的概率
0.8463107097139686
```

汉语转拼音 s.pinyin

```
['yi', 'ci', 'man', 'yi', 'de', 'gou', 'wu']
```

```
繁体字转简体 s = SnowNLP( u'「繁體字」「繁體中文」的叫法在臺灣亦很常見。') s.han
'「繁体字」「繁体中文」的叫法在台湾亦很常见。'
```

```
提取文章关键词 text = u'''自然语言处理(NLP)是计算机科学，人工智能，语言学关注计算机和人类
(自然)语言之间的相互作用的领域。因此，自然语言处理是与人机交互的领域有关的。在自然语言处理面
临很多挑战，包括自然语言理解，因此，自然语言处理涉及人机交互的面积。在NLP诸多挑战涉及自然
语言理解，即计算机源于人为或自然语言输入的意思，和其他涉及到自然语言生成。现代NLP算法是基
于机器学习，特别是统计机器学习。机器学习范式是不同于一般之前的尝试语言处理。语言处理任务的
实现，通常涉及直接用手的大套规则编码。许多不同类的机器学习算法已应用于自然语言处理任务。这
些算法的输入是一大组从输入数据生成的"特征"。一些最早使用的算法，如决策树，产生硬的if-then规
则类似于手写的规则，是再普通的系统体系。然而，越来越多的研究集中于统计模型，这使得基于附加
实数值的权重，每个输入要素柔软，概率的决策。此类模型具有能够表达许多不同的可能的答案，而不
是只有一个相对的确定性，产生更可靠的结果时，这种模型被包括作为较大系统的一个组成部分的优
点。自然语言处理研究逐渐从词汇语义成分的语义转移，进一步的，叙事的理解。然而人类水平的自然
语言处理，是一个人工智能完全问题。它是相当于解决中央的人工智能问题使计算机和人一样聪明，或
强大的AI。自然语言处理的未来一般也因此密切结合人工智能发展。''' s = SnowNLP(text)
s.keywords( 4) #提取关键词
['语言', '自然', '计算机', '涉及']

总结文章 s.summary(3)
['许多不同类的机器学习算法已应用于自然语言处理任务', '在NLP诸多挑战涉及自然语言理
解', '包括自然语言理解']
```

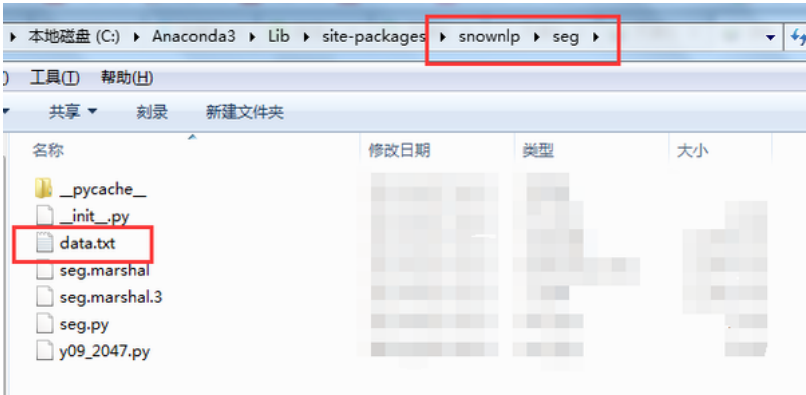
联系我们

0

分享到

机和人类(自然)语言之间的相互作用的领域', '因此', '自然语言处理是与人机交互的领域有关的', '在自然语言处理面临很多挑战', '包括自然语言理解', '因此', '自然语言处理涉及人机交互的面积', '在NLP诸多挑战涉及自然语言理解', '即计算机源于人为或自然语言输入的意思', '和其他涉及到自然语言生成', '现代NLP算法是基于机器学习', '特别是统计机器学习', '机器学习范式是不同于一般之前的尝试语言处理', '语言处理任务的实现', '通常涉及直接用手的大套规则编码', '许多不同类的机器学习算法已应用于自然语言处理任务', '这些算法的输入是一大组从输入数据生成的 "特征"', '一些最早使用的算法', '如决策树', '产生硬的if-then规则类似于手写的规则', '是再普通的系统体系', '然而', '越来越多的研究集中于统计模型', '这使得基于附加实数值的权重', '每个输入要素柔软', '概率的决策', '此类模型具有能够表达许多不同的可能的答案', '而不是只有一个相对的确定性', '产生更可靠的结果时', '这种模型被包括作为较大系统的一个组成部分的优点', '自然语言处理研究逐渐从词汇语义成分的语义转移', '进一步的', '叙事的理解', '然而人类水平的自然语言处理', '是一个人工智能完全问题', '它是相当于解决中央的人工智能问题使计算机和人一样聪明', '或强大的AI', '自然语言处理的未来一般也因此密切结合人工智能发展'] 接下来, 进入正题! from snownlp import seg#现在提供训练的包括分词, 词性标注, 情感分析, 而且都提供了我用来训练的原始文件 以分词为例 分词在snownlp/seg目录下

用data.txt可以用于训练



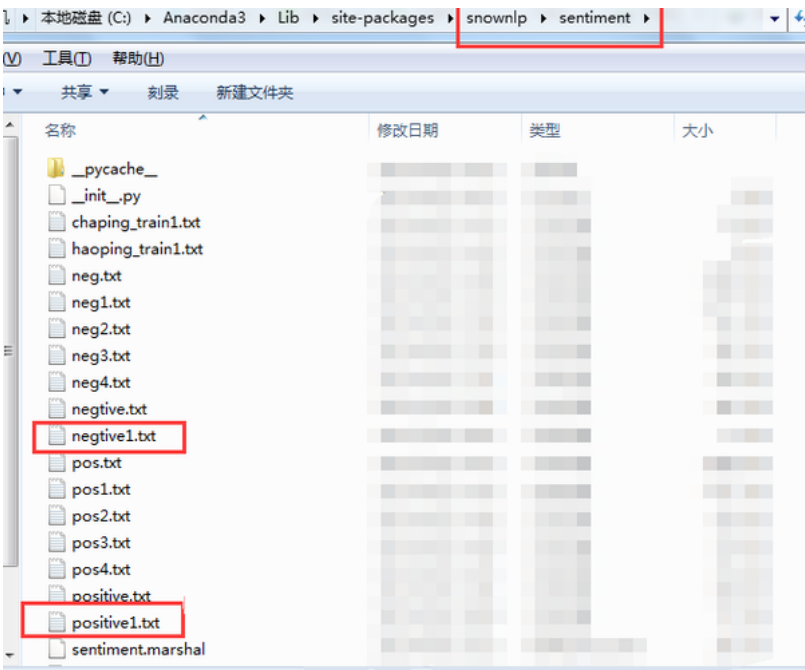
```
seg.train( 'C:\Anaconda3\Lib\site-packages\snownlp\seg\data.txt') seg.save( 'seg.marshall')
```

这样训练好的文件就存储为seg.marshall了

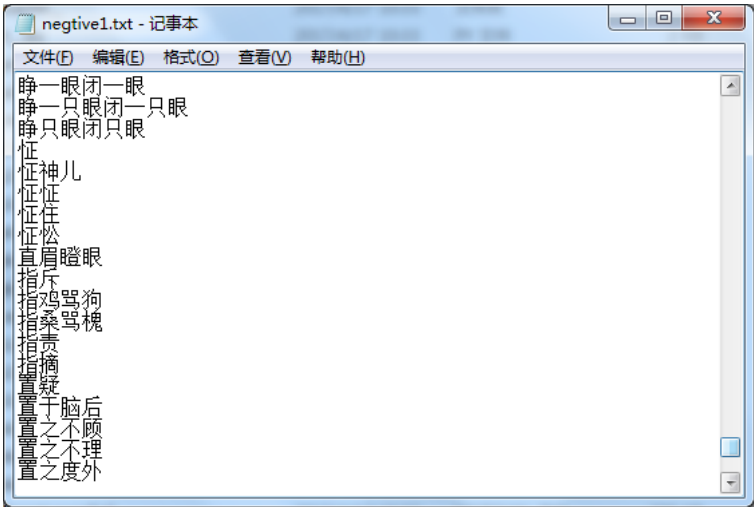
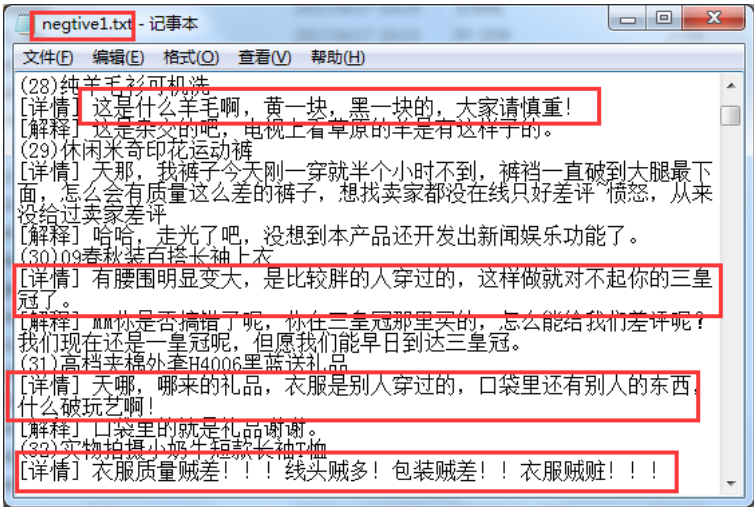
```
训练情感 from snownlp import sentiment sentiment.train( 'C:/Anaconda3/Lib/site-packages/snownlp/sentiment/negative1.txt', 'C:/Anaconda3/Lib/site-packages/snownlp/sentiment/positive1.txt') #注意路径斜线别写错 sentiment.save('C:/Anaconda3/Lib/site-packages/snownlp/sentiment/sentiment2.marshall')
```

0

分享到

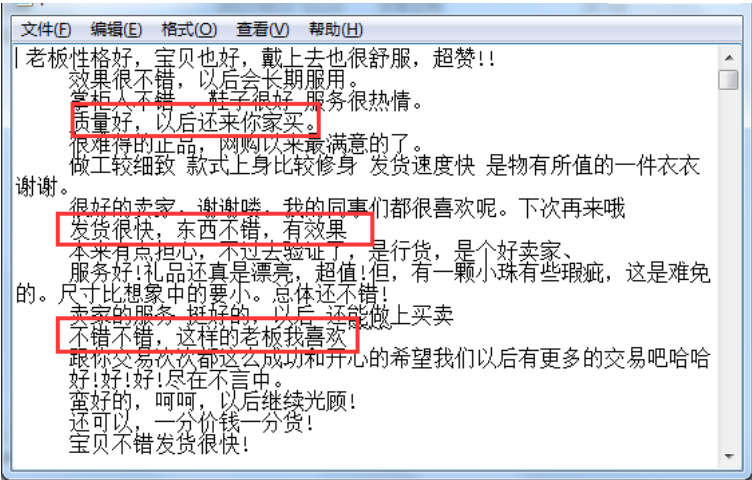


这些训练内容是自己找的，可以是情感积极消极词库，也可以是爬到的淘宝积极消极情感文档。



0

分享到



训练好了就可以计算情感啦~

```
q = SnowNLP( u'宝贝真的很不错, ,这是我第二次买了, 我的朋友都好喜欢, 穿上特别漂亮! 好性感! 质量好好! 大爱! ') q.sentiments
```

0.9999498509266149

```
q = SnowNLP( u'非常的好, 穿上美美哒, 喜欢的美美放心购买好了, 料子穿着很舒。') q.sentiments
```

0.9999498509266149

```
q = SnowNLP( u'布料不好, 不建议购买') q.sentiments
```

0.367781463800559

```
q = SnowNLP( u'好重的味道, 好难闻, 而且严重掉色, 请大家还是要看过再买吧') q.sentiments
```

0.2997962754770197

```
q = SnowNLP( u'太失望了') q.sentiments
```

0.1322809488637342

```
q = SnowNLP( '东西很好') q.sentiments
```

0.781183413508497

```
q = SnowNLP( '辣鸡店主, 败我钱财, 毁我青春') q.sentiments
```

0.9774554349848498

最后这个不科学。。。【捂脸】

一定是训练集太小啦~, 还需要扩充训练集

转载请保留以下内容:

本文来自天善社区ID王大伟老师的博客(公众号)。

原文链接: <https://ask.hellobi.com/blog/wangdawei/8415>

声明: 本文由入驻搜狐公众平台的作者撰写, 除搜狐官方账号外, 观点仅代表作者本人, 不代表搜狐立场。

0

分享到

广告

我来说两句 0人参与，0条评论

来说两句吧.....

登录并发表

搜狐“我来说两句”用户公约

还没有评论，快来抢沙发吧！

推荐阅读

推荐

WWDC

苹果

OPPO

乐视

VR

SpaceX

顺丰

iOS


iMac

天猫

富士康

CES


苹果WWDC大会：6大亮点 iOS11与最强iMac Pro亮相

 科技视界 · 今天 04:23

 11




苹果全面拥抱 AR/VR：iOS一夜成全球最大AR平台

 百度VR · 今天 06:16



苹果iOS11可升级名单：iPhone5/5c/iPad4被抛弃

 IT之家 · 今天 05:48



5年死了1300家创业公司，为什么却诞生五家百亿的独角兽？

 磐石之心 · 昨天 13:04


 6



全新卡罗拉，让幸福盛放！

广告 · 今天 8:40


苹果发布首款10.5英寸iPad Pro，比第一代快500倍

 科技视界 · 今天 03:30

 1

一文看尽苹果WWDC 2017：iOS 11大变革，iPad新品面世，Siri音箱正式推出



 猎云网 · 今天 03:29



nubia Z17体验：硬件封神，聊聊拍照



 数码FUN · 昨天 17:48



天猫618手机销量排名：小米第一，锤子遗憾未上榜



自媒体人冯东阳 · 昨天 21:17

这衬衣终于降价了！秒杀买一送一！天然蚕丝，夏季必备衣服！

广告 · 今天 8:40

猪队友再爆苹果秘密：iPhone8生产被推迟

中关村在线 · 今天 05:02

iOS 11发布 控制中心大变/App Store惊艳



手机中国 · 今天 03:10

安全脱毛有一招，飞利浦IPL脉冲光神器还夏日美腿



猎人IT追踪 · 昨天 22:24

诺贝尔医学奖为辟谷背书？别扯了

新京报评论 · 今天 01:51

15

夏日必备神器！10米内无蚊子！安全，无辐射！婴儿孕妇都可使用！



广告 · 今天 8:40

苹果发布 iMac Pro：售价 4999 美元起

动点科技 · 今天 02:44

阿迪彩虹鞋来袭，走路小跑打篮球全骚爆



极果 · 今天 07:00

富士康收购东芝闪存志在必得？苹果亚马逊助郭台铭

芯笔记 · 昨天 22:46

华为P10PK三台骁龙835，国产显威（内有大量动图）

这裤子值爆了！新品厂家直售199四件！全国包邮货到付款！



广告 · 今天 8:40

雷鸟I49评测：真4K+精准语音，49寸互联网电视首选

数码FUN · 昨天 17:34



短视频到下半场：快手向左求规模，美拍向右谋变现



 财经故事会 · 昨天 16:36



薪酬最高的 208 种 IT 技能



云头条 · 昨天 15:29



华为的八面曲屏手机让钢化膜厂商泪奔，你怎么看？

火火说科技 · 昨天 21:48



十年炒股未亏损，全因只看这个信号

广告 · 今天 8:40

加载更多