python的中文文本挖掘库snownlp进行购物评论文本情感分析实例

原创 2017年03月16日 17:11:01 7946

昨晚上发现了snownlp这个库，很开心。先说说我开心的原因。我本科毕业设计做的是文本挖掘，用R语言做的，发现R语言对文本处理特别不友好，没有很多强大的库，特别是针对中文文本的，加上那时候还没有学机器学习算法。所以很头疼，后来不得已用了一个可视化的软件RostCM，但是一般可视化软件最大的缺点是无法调参，很死板，准确率并不高。现在研一，机器学习算法学完以后，又想起来要继续学习文本挖掘了。所以前半个月开始了用python进行文本挖掘的学习，很多人都推荐我从《python自然语言处理》这本书入门，学习了半个月以后，可能本科毕业设计的时候有些基础了，再看这个感觉没太多进步，并且这里通篇将nltk库进行英文文本挖掘的，英文文本挖掘跟中文是有很大差别的，或者说学完英文文本挖掘，再做中文的，也是完全懵逼的。所以我停了下来，觉得太没效率了。然后我在网上查找关于python如何进行中文文本挖掘的文章，最后找到了snownlp这个库，这个库是国人自己开发的python类库，专门针对中文文本进行挖掘，里面已经有了算法，需要自己调用函数，根据不同的文本构建语料库就可以，真的太方便了。我只介绍一下这个库具体应用，不介绍其中的有关算法原理，因为算法原理可以自己去学习。因为我在学习这个库的时候，我查了很多资料发现很少或者基本没有写这个库的实例应用，很多都是转载官网对这个库的简介，所以我记录一下我今天的学习。

yyxyyx10 (http://blog.csd…

+ 关注

(http://blog.csdn.net/yyxyyx10)

原创	粉丝	喜欢	码云
26	14	0	未开通 (https://gite utm_sourc

他的最新文章

更多文章 (http://blog.csdn.net/yyxyyx10)

Tensorflow的应用（五） (http://blog.csdn.net/yyxyyx10/article/details/78760866)

Windows下TensorFlow的安装 (http://blog.csdn.net/yyxyyx10/article/details/78695853)

Tensorflow的应用（四） (http://blog.csdn.net/yyxyyx10/article/details/78695531)

首先简单介绍一下这个库可以进行哪些文本挖掘。snownlp主要可以进行中文分词（算法是Character-Based Generative Model）、词性标注（原理是TnT、3-gram 隐马）、情感分析（官网木有介绍原理，但是指明购物类的评论的准确率较高，其实是因为它的语料库主要是购物方面的，可以自己构建相关领域语料库，替换原来的，准确率也挺不错的）、文本分类（原理是朴素贝叶斯）、转换拼音、繁体转简体、提取文本关键词（原理是TextRank）、提取摘要（原理是TextRank）、分割句子、文本相似（原理是BM25）。官网还有更多关于该库的介绍，在看我这个文章之前，建议先看一下官网，里面有最基础的一些命令的介绍。官网链接：https://pypi.python.org/pypi/snownlp/0.11.1。

下面正式介绍实例应用。主要是中文文本的情感分析，我今天从京东网站采集了249条关于笔记本的评论文本作为练习数据，由于我只是想练习一下，没采集更多。然后人工标注每条评论的情感正负性，情感正负性就是指该条评论代表了评论者的何种态度，是褒义还是贬义。以下是样例

	A	B
1	comment	label
2	键盘很松，打字声音	-1
3	11月1日说是双十一的	-1
4	3015early的机器，首次开	1
5	2016年12月31号下单，201	1
6	5888买的 双十一 觉得很划	1
7	5988的价格买下的，不能	1
8	macbook air收到了，双十	1
9	办公用足够了，保修也没	-1
10	帮别人买的，用着感觉还	1
11	包装非常精美，非常严实，	1
12	包装很赞\(≧▽≦)/先用用	1

其中-1表示贬义，1表示褒义。由于snownlp全部是unicode编码，所以要注意数据是否为unicode编码。因为是unicode编码，所以不需要去除中文文本里面含有的英文，因为都会被转码成统一的编码（补充一下，关于编码问题，我还是不特别清楚，所以这里不多讲，还请对这方面比较熟悉的伙伴多多指教）。软件本身默认的是Ascii编码，所以第一步先设置软件的默认编码为utf-8，代码如下：

1、改变软件默认编码

```
import sys
reload(sys)
sys.setdefaultencoding('utf-8')
```

2、然后准备数据

```
import pandas as pd #加载pandas
text=pd.read_excel(u'F:/自然语言处理/评论文本.xlsx',header=0) #读取文本数据
text0=text.iloc[:,0] #提取所有数据
text1=[i.decode('utf-8') for i in text0] #上一步提取数据不是字符而是object，所以在这一步进行转码为字符
```

3、训练语料库

```
from snownlp import sentiment #加载情感分析模块
sentiment.train('E:/Anaconda2/Lib/site-packages/snownlp/sentiment/neg.txt', 'E:/Anaconda2/Lib/site-packages/snownlp/sentiment/pos.txt') #对语料库进行训练，把路径改成相应的位置。我这次练习并没有构建语料库，用了默认的，所以把路径写到了sentiment模块下。
sentiment.save('D:/pyscript/sentiment.marshal')#这一步是对上一步的训练结果进行保存，如果以后语料库没有改变，下次不用再进行训练，直接使用就可以了，所以一定要保存，保存位置可以自己决定，但是要把`snownlp/seg/__init__.py`里的`data_path`也改成你保存的位置，不然下次使用还是默认的。
```

4、进行预测

```
from snownlp import SnowNLP
senti=[SnowNLP(i).sentiments for i in text1] #遍历每条评论进行预测
```

5、进行验证准确率

预测结果为positive的概率，positive的概率大于等于0.6，我认为可以判断为积极情感，小于0.6的判断为消极情感。所以以下将概率大于等于0.6的评论标签赋为1，小于0.6的评论标签赋为-1，方便后面与实际标签进行比较。

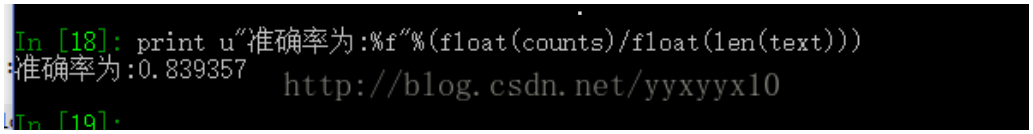
```
newsenti=[]
for i in senti:
  if (i>=0.6):
      newsenti.append(1)
  else:
      newsenti.append(-1)
text['predict']=newsenti  #将新的预测标签增加为text的某一列，所以现在text的第0列为评论文本，第1列为实际标签，第2列为预测标签
counts=0
for j in range(len(text.iloc[:,0])): #遍历所有标签，将预测标签和实际标签进行比较，相同则判断正确。
    if text.iloc[j,2]==text.iloc[j,1]:
```


他的热门文章

python的中文文本挖掘库snownlp进行购物评论文本情感分析实例 (http://blog.csdn.net/yyxyyx10/article/details/62428238)
7926

Anaconda安装出现的问题及解决心得 (http://blog.csdn.net/yyxyyx10/article/details/57083527)

counts+=1
print u"准确率为:%f"%(float(counts)/float(len(text)))#输出本次预测的准确率
运行结果为：



准确率还可以，但还不算高，原因是我考虑时间原因，并且我只是练习一下，所以没有自己构建该领域的语料库，如果构建了相关语料库，替换默认语料库，准确率会高很多。所以语料库是非常关键的，如果要正式进行文本挖掘，建议要构建自己的语料库。在没有构建新的语料库的情况下，这个83.9357%的准确率还是不错了。

以上是我这次的学习笔记，和大家分享一下，有不足之处请大家批评指正。我还是一个刚涉世数据挖掘、机器学习、文本挖掘领域不久的小白，有许多知识还是比较模糊，但对这数据挖掘很感兴趣。希望能多结识这方面的朋友，共同学习、共同进步。

sinat_22581761 (/sinat_22581761) 2017-07-30 20:05 2楼

(/sinat_22581761), 也在做相关的东西。目前只是用python 的爬虫，然后文本分析主要用工具了。也想全部都用python做。谢谢楼主了。

回复 2条回复

u013438638 (/u013438638) 2017-05-07 15:49 1楼

(/u013438638) 对于小白很有帮助

回复

查看 4 条热评

3

相关文章推荐

snownlp文本情感分析使用 (http://blog.csdn.net/u011961856/article/details/54573517)

snownlp为python版的文本分析工具，ubuntu安装snownlp命令为：pip install snownlp。 利用snownlp可以进行分词、词性标注、文本摘要提取、文本情感分析等，...

u011961856 (http://blog.csdn.net/u011961856) 2017年01月16日 13:57 5413

使用Python的SnowNLP模块实现情感分析 (http://blog.csdn.net/lqqlqqlqqlqq/article/details…

使用Python的SnowNLP实现情感分析SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了TextBlob的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的...

lqqlqqlqqlqq (http://blog.csdn.net/lqqlqqlqqlqq) 2016年02月27日 19:31 10339

uZR8mLPbUB48ugfEIAqspynETZ-YpAq8nWqdIAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqnHRLPjnvnfKEpyfqnHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqnW6drf) (https://passport.csdn.net/a

自然语言处理入门（2）——中文文本处理利器snownlp (http://blog.csdn.net/FlySky1991/art…

SnowNLP是一个python写的类库，可以方便的处理中文文本内容。如中文分词、词性标注、情感分析、文本分类、提取文本关键词、文本相似度计算等。snownlp示例如下所示：# -*- coding:...

FlySky1991 (http://blog.csdn.net/FlySky1991) 2017年05月31日 22:01 890

3

使用snownlp进行情感分析 (http://blog.csdn.net/Sunshine_Java_L/article/details/77586593)

使用snownlp+python进行情感分析

Sunshine_Java_L (http://blog.csdn.net/Sunshine_Java_L) 2017年08月25日 20:23 11714

SnowNLP初步使用 (http://blog.csdn.net/appleyuchi/article/details/77453166)

(python3.5) appleyuchi@Ubuntu16:~/.virtualenvs/python3.5/bin$ python Python 3.5.2 (default, Nov 17 ...

appleyuchi (http://blog.csdn.net/appleyuchi) 2017年08月21日 17:27 281

SnowNLP：处理中文文本内容 (http://blog.csdn.net/a123456ei/article/details/17205553)

这是一个比yaha更加强大的中文分词工具。yaha简单来说只是使用最短路径算法（Dijstra）实现了中文分词，而SnowNLP则实现了词性标标准、情感分析、文本分类、转换成拼音、繁体转简体、文本关键...

a123456ei (http://blog.csdn.net/a123456ei) 2013年12月08日 21:01 7742

(https://passport.csdn.net/a

snownlp情感分析源码解析 (http://blog.csdn.net/lom9357bye/article/details/78565432)

使用snownlp进行情感分析： from snownlp import SnowNLP #创建snownlp对象，设置要测试的语句 s = SnowNLP('这东西不错。。') # 调用senti...

lom9357bye (http://blog.csdn.net/lom9357bye) 2017年11月17日 21:25 139

3

python｜六款中文分词模块尝试:jieba、THULAC、SnowNLP、pynlpir、CoreNLP、pyLTP (...

THULAC四款python中中文分词的尝试。尝试的有：jieba、SnowNLP（MIT）、pynlpir（大数据搜索挖掘实验室（北京市海量语言信息处理与云计算应用工程技术研究中心））、thulac...

sinat_26917383 (http://blog.csdn.net/sinat_26917383) 2017年08月10日 17:42 2497

python 怎么安装snownlp包 (http://blog.csdn.net/qq_33160271/article/details/73469544)

https://pypi.python.org/pypi/snownlp/0.12.3linux apt-get install snownlp Windows pip install ...

qq_33160271 (http://blog.csdn.net/qq_33160271) 2017年06月19日 17:25 778

使用Python的SnowNLP模块实现情感分析 (http://blog.csdn.net/qw_xingzhe/article/details/...

SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了TextBlob的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和Te...

qw_xingzhe (http://blog.csdn.net/qw_xingzhe) 2016年10月28日 18:35 2804

Delphi7高级应用开发随书源码 (http://download.csdn.net/download/chenx...

2003年04月30日 00:00 676KB 下载

(https://passport.csdn.net/a

使用python+机器学习方法进行情感分析(详细步骤) (http://blog.csdn.net/Yan456jie/article/de…

原文地址 不是有词典匹配的方法了吗？怎么还搞多个机器学习方法。 因为词典方法和机器学习方法各有千秋。 机器学习的方法精确度更高，因为词典匹配会由于语义表达的丰富性而出现很大误差，而机器学习...

Yan456jie (http://blog.csdn.net/Yan456jie) 2016年08月18日 17:09 15935

3

基于词典的中文情感倾向分析算法设计 (http://blog.csdn.net/qw_xingzhe/article/details/5296…

基于词典的中文情感倾向分析算法设计 2014-06-05 11:23:08 By：@小和子 （计算传播学小站编辑） https://site.douban.com/146782/widg...

qw_xingzhe (http://blog.csdn.net/qw_xingzhe) 2016年10月28日 21:43 661

python - 对 '数码大冒险tri 泡泡评论' 进行简单的情感分析 (http://blog.csdn.net/PeersLee/art…

爬虫 selenium 抓取‘楚乔传’ 评论 NLP import jieba import numpy as np import pymongo from NL...

PeersLee (http://blog.csdn.net/PeersLee) 2017年07月05日 10:13 297

Python 文本挖掘：使用机器学习方法进行情感分析（一、特征提取和选择） (http://blog.csdn…

用Python 进行机器学习及情感分析，需要用到两个主要的程序包：nltk 和 scikit-learn nltk 主要负责处理特征提取（双词或多词搭配需要使用nltk 来做）和特征选择...

chenglansky (http://blog.csdn.net/chenglansky) 2014年06月16日 13:20 11914

Python贝叶斯算法进行情感分析 (http://blog.csdn.net/lsdnh521/article/details/48850551)

from __future__ import division import re from numpy import ones, array from numpy.lib.scimath impor...

lsdnh521 (http://blog.csdn.net/lsdnh521) 2015年10月01日 23:03 1279

Python爬虫和情感分析简介 (http://blog.csdn.net/u010022051/article/details/51487560)

摘要 这篇短文的目的是分享我这几天里从头开始学习Python爬虫技术的经验，并展示对爬取的文本进行情感分析(文本分类)的一些挖掘结果。 不同于其他专注爬虫技术的介绍，这里首先阐述爬取网络数据动机...

3

u010022051 (http://blog.csdn.net/u010022051) 2016年05月24日 09:29 2381

Delphi7高级应用开发随书源码 (http://download.csdn.net/download/chenx...

ZIP

(http://download.

2003年04月30日 00:00 676KB 下载

短文本情感分析 (http://blog.csdn.net/zbc1090549839/article/details/52800441)

一、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、归纳和推理的过程，其中情感分析还可以细分为情感极性（倾向）分析，情感程度分析，主客观分析等...

zbc1090549839 (http://blog.csdn.net/zbc1090549839) 2016年10月12日 21:09 8164

Python做文本挖掘的情感极性分析 (http://blog.csdn.net/starzhou/article/details/65629933)

Python做文本挖掘的情感极性分析 数据挖掘入门与实战2017-03-23 21:25:41line阅读(27)评论(0) 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...

starzhou (http://blog.csdn.net/starzhou) 2017年03月24日 13:42 2425