
2018.4.19

作者：baby_qian

文档说明：测试不同的 Word2Vec 参数对训练词向量的性能的影响.

语料说明：

（1） 采用 25000 条原始微博评论

（2） 进行分词和词清洗之后得到 114928 个词

（3） 依据语料中词出现的频率来选取候选情感词，共有 13531 个

（4） 从（2）中选取了情感强度较明显的正负各 74 个种子情感词

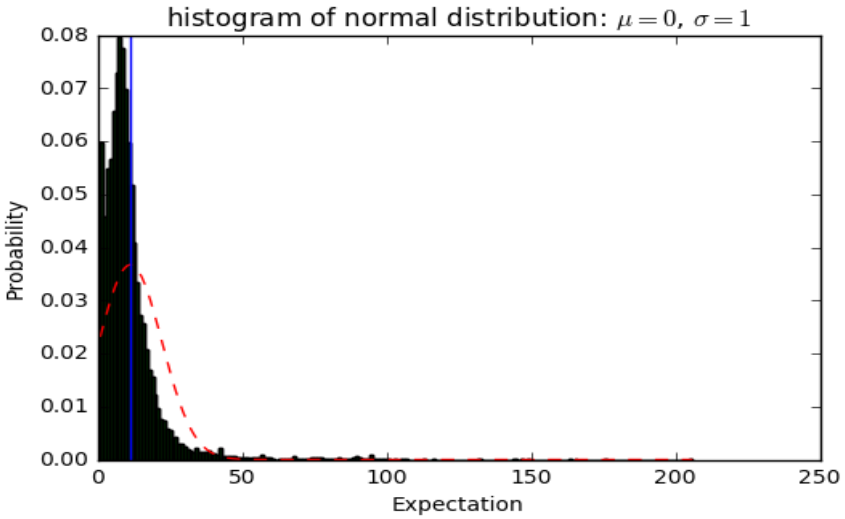
（5） 从知网的情感词典中抽取了正负各 577、422 个情感词作为测试集，这些词都包含在（2）的语料中

注意：语料的数量和分词的质量对训练结果影响最大。

一、文本预处理重点事项记录

1、句子长度分布统计

对原始微博评论进行分词、去停用词等处理后，统计所有语料中每条句子的长度分布情况，大致呈近似正态分布，其中均值为μ=10,方差σ=11。



语料库中句子长度分布统计图

二、使用 CBOW 模型训练词向量

1、 待优化的主要参数

（1） 词向量的维度 size；

（2） 滑动窗口的尺寸 windows

（3） *min_count*: 用于过滤操作，词频少于 min_count 次数的单词会被丢弃掉

（4） *hs*: 如果设置为 1，那么系统会采用 hierarchica softmax 技巧。如果设置为 0（默认情况），则系统会采用 negative samping 技巧。

（5） *cbow_mean*: 如果这个值设置为 0，那么就采用上下文词向量的总和。如果这个值设置为 1（默认情况下），那么我们就采用均值。但这个值只有在使用 CBOW 的时候才起作用

（6） *trim_rule*: 用于设置词汇表的整理规则，用来指定哪些词需要被剔除，哪些词需要保留。默认情况下，如果 word count < min_count，那么该词被剔除。这个参数也可以被设置为 None，这种情况下 min_count 会被使用。

2、 参数的实验选取情况

（1） 词向量维度 size 常用取值[100, 200, 300]

（2） 滑动窗口尺寸 Windows 常用取值(5, 10)

3、 最终选取参数及实验结果

（1） Word2Vec 模型参数

Window 的值可以依据语料中句子长度的分布情况来参考做选取

```
sentences = word2vec.Text8Corpus(r'C:\Users\大哥\Desktop\V1.0\clean_words.txt')
model = word2vec.Word2Vec(sentences, alpha=0.1, size=200, window=5, min_count=1, sg=0)
```

（2）语义距离阈值的选取（0、1、2 分别表示负向、正向、中性情感）

对于正向测试情感词，选择阈值 k_min = 0.2，k_mid = 0.3，做如下映射：

（0，0.2）--> 0; （0.2，0.3）--> 2； （0.3，1）--> 1

对于负向测试情感词，选择阈值 k_min = 0.2，k_mid = 0.3，做如下映射：

（0，0.3）--> 0; （0.3，0.6）--> 2； （0.6，1）--> 1

（2） 测试结果（精准度、召回率和 F1 值）

The hownet positive part test metrics

	precision	recall	f1-score	support
Negative words	1.00	0.73	0.84	422
Positive words	0.84	1.00	0.91	577
Neutral words	0.00	0.00	0.00	0
avg / total	0.91	0.88	0.88	999

三、使用 Skip-gram 模型训练词向量

说明：在本次实验中，使用 CBOW 和 Skip-gram 两种模型分别进行词向量训练，以 F1 值作为评价指标，两种模型的效果近似，但在实现速度上，CBOW 模型会比 Skip-gram 快些，因为 Skip-gram 需要用当前词去预测预定窗口内的词的向量表达式。

四、测试结果

1、部分词测试错误情况

```
The negative predict to positive words: 简单
The negative predict to positive words: 空
The negative predict to positive words: 寂寞
The negative predict to positive words: 不当
The negative predict to positive words: 毛躁
The neural words: 旧
The negative predict to positive words: 不起眼
The negative predict to positive words: 一不小心
The negative predict to positive words: 难以
The negative predict to positive words: 不以为然
The negative predict to positive words: 迫不及待
The negative predict to positive words: 恨
The negative predict to positive words: 不屑
The negative predict to positive words: 土
The negative predict to positive words: 阴险
The negative predict to positive words: 贵
```

2、查找与某个词语义最临近的几个词（'快乐', '信心', '笑', '忧伤', '失望', '假'）

```
快乐
[('平行', 0.9988290071487427), ('氵', 0.9987861514091492), ('宄', 0.9987502694129944), ('肠粉', 0.9986573457717896),
('解决', 0.998637855052948), ('长得', 0.9986100792884827), ('拥抱', 0.9985666275024414), ('抖', 0.9985376596450806),
('厚重', 0.9984683990478516), ('馋嘴', 0.9983986616134644)]
信心
[('空降', 0.25666534900665283), ('钅', 0.25609147548675537), ('老鬼', 0.24912865459918976), ('在建', 0.2451672255992889
4), ('暴躁', 0.23470807075500488), ('逃不出', 0.23140671849250793), ('母亲', 0.22615084052085876), ('霞', 0.224498018622
39838), ('锁水', 0.21997234225273132), ('妮子', 0.21652469038963318)]
笑
[('美女', 0.9998432993888855), ('衰', 0.9998259544372559), ('牙病', 0.9998137354850769), ('搞', 0.9998121857643127),
('分享', 0.9998101592063904), ('手段', 0.9998065233230591), ('楚河', 0.999779224395752), ('谢谢', 0.9997727274894714),
('最美', 0.9997707605361938), ('美丽', 0.9997648000717163)]
忧伤
[('长见识', 0.27405285835266113), ('动听', 0.272847443819046), ('悔不当初', 0.2617332637310028), ('水说', 0.2447218000888
8245), ('墓', 0.240851491689682), ('不信来', 0.23218680918216705), ('尊敬', 0.22545559704303741), ('淘小', 0.22407123446
464539), ('探案', 0.21836566925048828), ('情侣', 0.21738380193710327)]
失望
[('转发', 0.9998719096183777), ('广告', 0.9998480677604675), ('女', 0.9998106956481934), ('号', 0.9997910261154175),
('不到', 0.9997879266738892), ('默默', 0.9997866153717041), ('点', 0.999783992767334), ('美丽', 0.9997826218605042),
('世界', 0.9997813701629639), ('哈哈哈哈', 0.9997798204421997)]
假
[('节奏', 0.9998799562454224), ('感动', 0.9998668432235718), ('片子', 0.9998475313186646), ('广告', 0.9998247027397156),
('点', 0.9998184442520142), ('转发', 0.9998167753219604), ('每次', 0.9998089671134949), ('女', 0.999808132648468), ('电
影', 0.999805212020874), ('美女', 0.9998011589050293)]
```

3、部分情感词情感强度量化结果

```
'夸': -0.07713738108381026,
'夸奖': 0.07174791148910928,
'奇妙': 0.011086158670960883,
'好受': -0.02575970537066963,
'好吃': 0.8921780016806569,
'好听': 0.07941235049401184,
'好好': 0.8940230216107306,
'好开心': 0.01803297656462402,
'好心': -0.0014878167814852015,
'好玩': 0.8382694438301228,
'好玩儿': -0.00531067282119349,
'好用': 0.8938900840679511,
'好看': 0.8936799925129005,
'好笑': 0.8914377105279159,
'好评': 0.8920565006021108,
'好过': 0.8928628902555116,
'妖艳': -0.10180324015862031,
'妙': 0.025156521433460956,
'妙不可言': 0.8905917635137165,
'妥妥': -0.05666031956207007,
'威严': -0.00981265728085025,
'威武': 0.8934816657203702,
'孜孜不倦': -0.0358967977132321,
'孝顺': 0.03041581601860958,
'安好': 0.0522497340752051,
'安康': -0.0030491740199067863,
'安稳': 0.023854148221781123,
```