doc2vec使用说明（一）gensim工具包TaggedLineDocument

gensim 是处理文本的很强大的工具包，基于python环境下：

1.gensim可以做什么？

它可以完成的任务，参加gensim 主页API中给出的介绍，链接如下：

http://radimrehurek.com/gensim/apiref.html

2.word2vec的使用

其中学习词向量的方法可利用，word2vec，具体使用我爱自然语言中介绍的很清楚，如下链接：

http://ju.outofmemory.cn/entry/80023

3.doc2vec/paragraph2vec的使用方法

学习文档向量，doc2vec（也就是官方网站API中的paragraph2vec)使用方法，中文资料较少,RaRe Machine Learning Blog英文博客讲解的比较详细,链接如下：

http://rare-technologies.com/doc2vec-tutorial/

因为要做文档向量的学习，我也写了个学习文档向量的例子，仅供参考，代码如下：

```
import gensim, logging
import os

logging.basicConfig(format = '%(asctime)s : %(levelname)s : %(message)s', level = logging.INFO)
sentences = gensim.models.doc2vec.TaggedLineDocument('review_pure_text.txt')
model = gensim.models.Doc2Vec(sentences, size = 100, window = 5)
model.save('review_pure_text_model.txt')
print len(model.docvecs)

out = file('review_pure_text_vector.txt', 'w')
for idx, docvec in enumerate(model.docvecs):
    for value in docvec:
      out.write(str(value) + ' ')
    out.write('\n')
    print idx
    print docvec
out.close()
```

输入文件Tweets_id_text.txt的格式就是每个doc 对应内容的分词，空格隔开，每个doc是一行

用TaggedLineDocument 实现，每个doc默认编号

博文doc2vec/paragraph2vec使用说明（二）中介绍 带多个标签的文档向量训练方法。

标签: NLP工具包


囧囧要努力
关注 - 3
粉丝 - 16
+加关注

0 0

« 上一篇：word2vec 实践

» 下一篇：机器学习中的相似性度量

posted @ 2016-01-23 19:44 囧囧要努力 阅读(2369) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

更多

随笔档案

2017年12月 (1)
2017年10月 (3)
2017年9月 (3)
2017年8月 (2)
2017年7月 (5)
2017年4月 (4)
2017年3月 (2)
2017年2月 (1)
2017年1月 (11)
2016年12月 (1)
2016年11月 (7)
2016年9月 (2)

2014年11月 (3)
2014年7月 (1)
2013年6月 (1)
2013年3月 (1)
2012年12月 (1)
2012年11月 (4)
2012年10月 (4)
2012年9月 (2)
2012年1月 (2)
2011年12月 (8)
2011年11月 (3)
2011年5月 (2)

最新评论

1. Re:doc2vec使用说明（二）gensim工具包 LabeledSentence

word2vector是将词表示为向量，doc2vector是将文本表示成向量，是这样的吗

--坚强的小红

2. Re:Network Embedding 论文小览

@tornotohi，您好，我没有看证明推导，但我认为这种embedding的方式不能完全认为是等价于矩阵分解，矩阵分解是全局的，这种deep walk 随机游走的建模方式是一种局部信息的反复抽取和利.....

--囧囧要努力

3. Re:Network Embedding 论文小览

想问一下，文中提到一些网络嵌入方法被证明等价于一些矩阵分解算法，那么这种方法和对应的矩阵分解在性能上是不是相同？

--tornoto

4. Re:doc2vec使用说明（二）gensim工具包 LabeledSentence

@free_1doc2vec 是可以用来训练句子向量的，输入的每一行是一个句子，训练出来的向量就是对应该行句子的向量。...

--囧囧要努力

5. Re:doc2vec使用说明（二）gensim工具包 LabeledSentence

可以用来训练句子向量么？

--free_1

阅读排行榜

评论排行榜