

囡囡要努力

一日一人生，日满心安

博客园 首页 新随笔 联系 订阅 管理

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://www.cnblogs.com/baiting/p/5153877.html>

公告

昵称：囡囡要努力

园龄：5年8个月

粉丝：16

关注：3

+加关注

<	2017年12月						>
日	一	二	三	四	五	六	
26	27	28	29	30	1	2	
3	4	5	6	7	8	9	
10	11	12	13	14	15	16	
17	18	19	20	21	22	23	
24	25	26	27	28	29	30	
31	1	2	3	4	5	6	

<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：



```
1 import gensim, logging
2 import os
3
4 logging.basicConfig(format = '%(asctime)s : %(levelname)s : %(message)s', level = logging.INFO)
5 sentences = gensim.models.doc2vec.TaggedLineDocument('review_pure_text.txt')
6 model = gensim.models.Doc2Vec(sentences, size = 100, window = 5)
7 model.save('review_pure_text_model.txt')
8 print len(model.docvecs)

9 out = file('review_pure_text_vector.txt', 'w')
10 for idx, docvec in enumerate(model.docvecs):
11     for value in docvec:
12         out.write(str(value) + ' ')
13     out.write('\n')
14     print idx
15     print docvec
16 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

好文要顶

关注我

收藏该文



搜索

找找看

谷歌搜索

我的标签

机器学习(19)

论文笔记(17)

NLP工具包(6)

深度学习(6)

实用技巧(6)


推荐算法(4)

社会情感计算(4)

java 语言(4)

latex(3)

matlab 相关系数(3)



囡囡要努力

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 囡囡要努力 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](#)。

【推荐】50万行VC++源码: 大型组态工控、电力仿真CAD与GIS源码库

【促销】腾讯云技术升级10大核心产品年终让利

【推荐】高性能云服务器2折起，0.73元/日节省80%运维成本

【新闻】H3 BPM体验平台全面上线



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)

最新IT新闻:

· 郭台铭详解鸿海工业互联网战略 拟分拆在上海上市

· 扎克伯格休假照片曝光 配娃娃吃喝玩乐

· 金刚狼死侍回归漫威，迪士尼收购福克斯让好莱坞「变天」

· 乐视网发布公告：聘任刘淑青为公司总经理

· 面试软件工程师，这些准备工作你做了吗？

» 更多新闻...

 阿里云

告别高昂运维费用 云计算全面助力

40+款核心产品免费半年 再+8000津贴任意采购

立即申请

最新知识库文章:

· 以操作系统的角度述说线程与进程

· 软件测试转型之路

· 门内门外看招聘

· 大道至简，职场上做人做事做管理

· 关于编程，你的练习是不是有效的？

» 更多知识库文章...

2016年8月 (9)
2016年7月 (2)
2016年6月 (5)
2016年5月 (12)
2016年4月 (8)
2016年3月 (6)
2016年2月 (3)
2016年1月 (8)
2015年12月 (1)
2015年8月 (6)
2015年7月 (3)
2015年5月 (1)
2014年12月 (1)

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.....

--囡囡要努力

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。...

--囡囡要努力

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

1. doc2vec使用说明 (二) gensim工具包 LabeledSentence(5273)

2. Latex 中cite的使用(3186)

3. java list随机打乱(2593)

4. doc2vec使用说明 (一) gensim工具包 TaggedLineDocument(2369)

5. 如何理解 卷积 和pooling(1787)

评论排行榜

1. doc2vec使用说明 (二) gensim工具包
LabeledSentence(3)

2. BPR: Bayesian Personalized Ranking from Implicit Feedback-CoRR 2012——20160421(2)

3. Network Embedding 论文小览(2)

Copyright ©2017 囡囡要努力