机器学习 自然语言处理 谷歌 (Google) 开源项目 word2vec

关注者 23 被浏览 3955

word2vec的demo里的训练数据text8内的数据格式是什么样子的？

关注问题 写回答 1 条评论 分享 邀请回答 ...

2 个回答 默认排序

林洲汉
非典型工科男

2 人赞同了该回答

text8来源于enwiki8，而enwiki8最早是用来做文本压缩的。简单说来，enwiki8是从wikipedia上扒下来的前100，000，000个字符；而text8就是把这些字符当中各种奇怪的符号啊，非英文字符啊全都去掉，再把大写字符转化成小写字符，把数字转化成对应的英语单词之后，得到的。

所以text8中只包含27种字符：小写的从a到z，以及空格符。如果把它打出来，读起来就像是去掉了所有标点的wikipedia。楼上已经有人打出来了，我就不上图了。

Matt Mahoney有一个网页很详细地说明了这个文件是如何来的，也包含了对文本内容一些基本分析：About the Test Data

发布于 2015-10-22

▲ 2 ▼ 2 条评论 分享 收藏 感谢

匿名用户


下载知乎客户端
与世界分享知识、经验和见解

相关问题

机器学习有很多关于核函数的说法，核函数的定义和作用是什么？ 50 个回答

如何用简单易懂的例子解释隐马尔可夫模型？ 35 个回答

概率图模型（PGM）有必要系统地学习一下吗？ 27 个回答

在自然科学领域，复杂的模型(如神经网络)在逐渐淘汰掉简单的模型吗？ 29 个回答

对于一个开源 Python 量化交易平台项目的建议有哪些？ 31 个回答

私家课 · Live 推荐

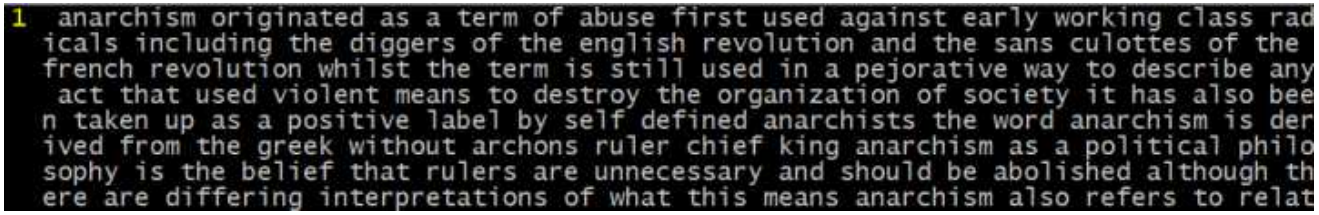
这种明文，你直接打出来看看就行了啊！！！！
不过看到有很多小伙伴们关注了，就帮他们打出来吧...

google公布的word2vec源码的输入数据要求是：分词后数据，以空格为单词的分隔符（那就意味着中文分词后数据也是可以直接跑demo的！）。

text8中的数据格式：

> **"text8" [noeol] 1L, 100000000C**

1行，10^8个字符。
如下图：



发布于 2015-03-04

▲ 5 ▼ 2 条评论 分享 收藏 感谢

写回答


少数派
共 15 节课 ▶ 试听

1024 程序员职场进阶必修课
8 场 Live, 9195 次参与


推荐算法(四):排序与建模
姚凯飞
★★★★★ 54 人参与

刘看山 · 知乎指南 · 知乎协议 · 应用 · 工作
侵权举报 · 网上有害信息举报专区
违法和不良信息举报：010-82716601
儿童色情信息举报专区
联系我们 © 2017 知乎