

peghoty
学习是一种态度!



≡ 目录视图

≡ 摘要视图

RSS 订阅

个人资料



皮果提

+ 加关注

✉ 发私信

访问：1224389次
积分：19134
等级：**BLOG > 7**
排名：第454名

原创：104篇
转载：13篇
译文：2篇
评论：750条

文章分类

数据挖掘 (35)
深度学习 (20)
语言模型 (10)
文本挖掘 (1)
强化学习 (1)
数学天地 (18)
编程知识 (5)
隐马模型 (8)
杂七杂八 (14)
机器学习 (37)
并行计算 (3)

阅读排行

word2vec 中的数学原理\ (93579)
word2vec 中的数学原理\ (55150)
受限玻尔兹曼机 (RBM) (42965)
word2vec 中的数学原理\ (39988)
word2vec 中的数学原理\ (39890)
word2vec 中的数学原理\ (39408)
牛顿法与拟牛顿法学习笔 (33615)
牛顿法与拟牛顿法学习笔 (32578)
牛顿法与拟牛顿法学习笔 (32361)
受限玻尔兹曼机 (RBM) (31081)

评论排行

word2vec 中的数学原理\ (153)
word2vec 中的数学原理\ (61)
Community Detection 算 (60)

原 word2vec 中的数学原理详解（一）目录和前言

标签：word2vec CBOw Skip-gram Hierarchical Softmax Negative Sampling

2014-07-19 22:42 93593人阅读 评论(153) 收藏 举报

≡ 分类： 语言模型 (9) ▾

版权声明：本文为博主原创文章，未经博主允许不得转载。

word2vec 是 Google 于 2013 年开源推出的一个用于获取 word vector 的工具包，它简单、高效，因此引起了很多人的关注。由于 word2vec 的作者 Tomas Mikolov 在两篇相关的论文 [3, 4] 中并没有谈及太多算法细节，因而在一定程度上增加了这个工具包的神秘感。一些按捺不住的人于是选择了通过解剖源代码的方式来一窥究竟，出于好奇，我也成为了他们中的一员。读完代码后，觉得收获颇多，整理成文，给有需要的朋友参考。

相关链接

- (一) 目录和前言
- (二) 预备知识
- (三) 背景知识
- (四) 基于 Hierarchical Softmax 的模型
- (五) 基于 Negative Sampling 的模型
- (六) 若干源码细节

受限玻尔兹曼机 (RBM)	(48)
word2vec 中的数学原理	(43)
发表在 Science 上的一种	(32)
受限玻尔兹曼机 (RBM)	(26)
利用 word2vec 训练的字	(22)
受限玻尔兹曼机 (RBM)	(21)
word2vec 中的数学原理	(20)

最新评论

word2vec 中的数学原理详解 (匹
jacksonjack001: @celia01:这个
问题貌似楼下有人解释, 说跟bp
类似! 不能求平均!

word2vec 中的数学原理详解 (匹
jacksonjack001: @neopenx:没错
吧, 我看的gensim的源码, sg算
法于cbow的主要区别就是在每个
当前词处理...

word2vec 中的数学原理详解 (匹
jacksonjack001:
@m0_37369113:应该没有问题
吧, 在更新 $v_{(w)}$ 的时候已经加上
了吧。另外我看过gensim...

受限玻尔兹曼机 (RBM) 学习笔
DouMiaoO_Oo: 想要请教一下大
家, MCMC方法是在概率分布
函数 $P(X)$ 很复杂的情况下, 我们
不好直接从分布函数中采样...

word2vec 中的数学原理详解 (三
PJ-Javis: 这的确是个坑, 我也掉
进去了

A Painless Q-learning Tutorial (-
星辰旋风: 赞

word2vec 中的数学原理详解 (一
phybrain: 求pdf 楼主
692114871@qq.com

word2vec 中的数学原理详解 (六
lreader1: 楼主, 我感觉你的亚采
样的公式写的不是很对呀

受限玻尔兹曼机 (RBM) 学习笔
zuzhangxian7307:
@qq_36010258:你好 这边我感
觉其实应该是对应状态出现的频
数。

受限玻尔兹曼机 (RBM) 学习笔
zuzhangxian7307: 楼主你好,
看了楼主的文章有豁然开朗的感
觉, 另外希望楼主能发一份pdf学
习, 谢谢!157719785...

目 录

1 前言	3
2 预备知识	4
2.1 sigmoid 函数	4
2.2 逻辑回归	4
2.3 Bayes 公式	5
2.4 Huffman 编码	5
2.4.1 Huffman 树	6
2.4.2 Huffman 树的构造	6
2.4.3 Huffman 编码	7
3 背景知识	9
3.1 统计语言模型	9
3.2 n-gram 模型	10
3.3 神经概率语言模型	12
3.4 词向量的理解	15
4 基于 Hierarchical Softmax 的模型	18
4.1 CBOW 模型	19
4.1.1 网络结构	19
4.1.2 梯度计算	20
4.2 Skip-gram 模型	25
4.2.1 网络结构	25
4.2.2 梯度计算	25
5 基于 Negative Sampling 的模型	28
5.1 CBOW 模型	28
5.2 Skip-gram 模型	30
5.3 负采样算法	32
6 若干源码细节	34
6.1 $\sigma(x)$ 的近似计算	34
6.2 词典的存储	35
6.3 换行符	35
6.4 低频词和高频词	36
6.5 窗口及上下文	37
6.6 自适应学习率	37
6.7 参数初始化与训练	38
6.8 多线程并行	38
6.9 几点疑问和思考	38

§1 前言

word2vec 是 Google 于 2013 年开源推出的一个用于获取 word vector 的工具包, 它简单、高效, 因此引起了很多人的关注. 由于 word2vec 的作者 Tomas Mikolov 在两篇相关的论文 ([3], [4]) 中并没有谈及太多算法细节, 因而在一定程度上增加了这个工具包的神秘感. 一些按捺不住的人于是选择了通过解剖源代码的方式来一窥究竟.

第一次接触 word2vec 是 2013 年的 10 月份, 当时读了复旦大学郑晓庆老师发表的论文 [7], 其主要工作是将 SENNA 的那套算法 ([8]) 搬到中文场景. 觉得挺有意思, 于是做了一个实现 (可参见 [20]), 但苦于其中字向量的训练时间太长, 便选择使用 word2vec 来提供字向量, 没想到中文分词效果还不错, 立马对 word2vec 刮目相看了一把, 好奇心也随之增长.

后来, 陆陆续续看到了 word2vec 的一些具体应用, 而 Tomas Mikolov 团队本身也将其推广到了句子和文档 ([6]), 因此觉得确实有必要对 word2vec 里的算法原理做个了解, 以便对他们的后续研究进行追踪. 于是, 沉下心来, 仔细读了一回代码, 算是基本搞明白里面的做法了. 第一个感觉就是, “明明是个很简单的浅层结构, 为什么被那么多人沸沸扬扬地说成是 Deep Learning 呢?”

解剖 word2vec 源代码的过程中, 除了算法层面的收获, 其实编程技巧方面的收获也颇多. 既然花了功夫来读代码, 还是把理解到的东西整理成文, 给有需要的朋友提供点参考吧.

在整理本文的过程中, 和深度学习群的群友 [北流浪子](#) ([15, 16]) 进行了多次有益的讨论, 在此表示感谢. 另外, 也参考了其他人的一些资料, 都列在参考文献了, 在此对他们的工作也一并表示感谢.

参考文献

- [1] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. **Learning representations by backpropagating errors**. *Nature*, 323(6088):533-536, 1986.
- [2] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. **A neural probabilistic language model**. *Journal of Machine Learning Research (JMLR)*, 3:1137-1155, 2003.
- [3] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. **Efficient Estimation of Word Representations in Vector Space**. arXiv:1301.3781, 2013.
- [4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. **Distributed Representations of Words and Phrases and their Compositionality**. arXiv:1310.4546, 2013.
- [5] Tomas Mikolov, Quoc V. Le, Ilya Sutskever. **Exploiting Similarities among Languages for Machine Translation**. arXiv:1309.4168v1, 2013.
- [6] Quoc V. Le, Tomas Mikolov. **Distributed Representations of Sentences and Documents**. arXiv:1405.4053, 2014.
- [7] Xiaoqing Zheng, Hanyang Chen, Tianyu Xu. **Deep Learning for Chinese Word Segmentation and POS tagging**. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 647-657.
- [8] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuksa. **Natural Language Processing (Almost) from Scratch**. *Journal of Machine Learning Research (JMLR)*, 12:2493-2537, 2011.
- [9] Michael U Gutmann and Aapo Hyvärinen. **Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics**. *The Journal of Machine Learning Research*, 13:307-361, 2012.
- [10] 百度百科中的“哈夫曼树”词条.
- [11] 吴军. **《数学之美》**. 人民邮电出版社, 2012.
- [12] <http://ml.nec-labs.com/senna/>
- [13] <http://www.loooker.com/archives/5621>
- [14] licstar. **Deep Learning in NLP (一) 词向量和语言模型**.
<http://licstar.net/archives/328>
- [15] 深度学习 word2vec 笔记之基础篇.
<http://blog.csdn.net/mytestmy/article/details/26961315>

- [16] 深度学习 word2vec 笔记之算法篇.
<http://blog.csdn.net/mytestmy/article/details/26969149>
- [17] 邓澍军, 陆光明, 夏龙. Deep Learning 实战之 word2vec, 2014.
- [18] 杨超. Word2Vec 的一些理解.
<http://www.zhihu.com/question/21661274/answer/19331979>
- [19] 基于权值的微博用户采样算法研究.
<http://blog.csdn.net/itplus/article/details/9079297>
- [20] 利用 word2vec 训练的字向量进行中文分词.
<http://blog.csdn.net/itplus/article/details/17122431>
- [21] Yoav Goldberg, Omer Levy. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. arXiv: 1402.3722v1, 2014. (<http://arxiv.org/pdf/1402.3722v1.pdf>)

作者: peghoty

出处: <http://blog.csdn.net/itplus/article/details/37969519>

欢迎转载/分享, 但请务必声明文章出处.



- ▲ 上一篇 一种并行随机梯度下降法
- ▼ 下一篇 word2vec 中的数学原理详解 (二) 预备知识

相关文章推荐

- word2vec 中的数学原理详解 (一) 目录和前言
- 【直播】70天软考冲刺计划--任铄
- word2vec原理解析
- 【直播】打通Linux脉络 进程、线程、调度--宋宝华
- word2vec中的数学原理详解
- 【直播】机器学习之凸优化--马博士
- word2vec 中的数学原理详解 (二) 预备知识
- 【套餐】MATLAB基础+MATLAB数据分析与统计--...
- word2vec数学原理
- 【课程】3小时掌握Docker最佳实战--徐西宁
- word2vec 中的数学原理详解 (五) 基于 Negative...
- 【课程】深度学习基础与TensorFlow实践--AI100
- Word2Vec数据集
- word2vec 中的数学原理详解 (四) 基于 Hierarchi...
- word2vec工具下载
- word2vec 中的数学原理详解

查看评论

100楼 phybrain 2017-08-14 16:27发表



求pdf 楼主 692114871@qq.com

99楼 scar_tom 2017-08-09 11:31发表



跪求PDF, 非常非常感谢, 经典总结。919453887@qq.com

98楼 springXu 2017-07-07 08:24发表



MARK下

97楼 张小彬的代码人生 2017-07-06 10:12发表



楼下这么多的伸手党也是醉了, 直接看博客不就行了, 干嘛非要PDF。。。

96楼 qq_20571985 2017-06-20 12:44发表

关闭