使用自己的语料训练word2vec模型

繁著 (/u/67eb7ed414d3) ＋关注
2017.08.14 14:00* 字数 751 阅读 161 评论 0 喜欢 0
(/u/67eb7ed414d3)

一、 准备环境和语料：

- 新闻20w+篇（格式： 标题。 正文 ）

> 【新闻可以自己从各大新闻网站爬取，也可以下载开源的新闻数据集，如
>
> - 互联网语料库(SogouT) (https://link.jianshu.com?t=http://www.sogou.com/labs/resource/t.php)
> - 中文文本分类数据集THUCNews (https://link.jianshu.com?t=http://thuctc.thunlp.org/)
> - 李荣陆英文文本分类语料 (https://link.jianshu.com?t=http://www.datatang.com/data/11968)
> - 谭松波中文文本分类语料 (https://link.jianshu.com?t=http://www.datatang.com/data/11970)
> - 等

- 结巴分词 (https://link.jianshu.com?t=https://github.com/fxsjy/jieba)
- word2vec (https://link.jianshu.com?t=https://radimrehurek.com/gensim/models/word2vec.html)

二、分词

先对新闻文本进行分词，使用的是结巴分词工具，将分词后的文本保存在 `seg201708.txt` ，以备后期使用。

> 安装jieba工具包：pip install jieba

```
# -*- coding: utf-8 -*-
import jieba
import io
# 加载自己的自己的金融词库
jieba.load_userdict("financialWords.txt")

def main():
    with io.open('news201708.txt','r',encoding='utf-8') as content:
        for line in content:
            seg_list = jieba.cut(line)
#           print '/'.join(seg_list)
            with io.open('seg201708.txt', 'a', encoding='utf-8') as output:
                output.write(' '.join(seg_list))

if __name__ == '__main__':
    main()
```

三、训练word2vec模型

使用python的gensim包进行训练。

> 安装gemsim包：pip install gemsim

```
from gensim.models import word2vec

def main():

    num_features = 300    # Word vector dimensionality
    min_word_count = 10   # Minimum word count
    num_workers = 16       # Number of threads to run in parallel
    context = 10          # Context window size
    downsampling = 1e-3   # Downsample setting for frequent words
    sentences = word2vec.Text8Corpus("seg201708.txt")

    model = word2vec.Word2Vec(sentences, workers=num_workers, \
            size=num_features, min_count = min_word_count, \
            window = context, sg = 1, sample = downsampling)
    model.init_sims(replace=True)
    # 保存模型，供日後使用
    model.save("model201708")

    # 可以在加载模型之后使用另外的句子来进一步训练模型
    # model = gensim.models.Word2Vec.load('/tmp/mymodel')
    # model.train(more_sentences)

if __name__ == "__main__":
    main()
```

- 参数说明

- sentences：可以是一个·ist，对于大语料集，建议使用BrownCorpus,Text8Corpus或ineSentence构建。
- sg： 用于设置训练算法，默认为0，对应CBOW算法；sg=1则采用skip-gram算法。
- size：是指特征向量的维度，默认为100。大的size需要更多的训练数据,但是效果会更好. 推荐值为几十到几百。
- window：表示当前词与预测词在一个句子中的最大距离是多少
- alpha: 是学习速率
- seed：用于随机数发生器。与初始化词向量有关。
- min_count: 可以对字典做截断. 词频少于min_count次数的单词会被丢弃掉, 默认值为5
- max_vocab_size: 设置词向量构建期间的RAM限制。如果所有独立单词个数超过这个，则就消除掉其中最不频繁的一个。每一千万个单词需要大约1GB的RAM。设置成None则没有限制。
- sample: 高频词汇的随机降采样的配置阈值，默认为1e-3，范围是(0,1e-5)
- workers参数控制训练的并行数。
- hs: 如果为1则会采用hierarchica·softmax技巧。如果设置为0（defau·t），则negative sampling会被使用。
- negative: 如果>0,则会采用negativesamp·ing，用于设置多少个noise words
- cbow_mean: 如果为0，则采用上下文词向量的和，如果为1（defau·t）则采用均值。只有使用CBOW的时候才起作用。
- hashfxn： hash函数来初始化权重。默认使用python的hash函数
- iter： 迭代次数，默认为5

- trim_rule： 用于设置词汇表的整理规则，指定那些单词要留下，哪些要被删除。可以设置为None（min_count会被使用）或者一个接受()并返回RU·E_DISCARD,uti·s.RU·E_KEEP或者uti·s.RU·E_DEFAU·T的
- sorted_vocab： 如果为1（defau·t），则在分配word index 的时候会先对单词基于频率降序排序。
- batch_words：每一批的传递给线程的单词的数量，默认为10000

四、word2vec应用

```
model = Word2Vec.load('model201708')      #模型讀取方式
model.most_similar(positive=['woman', 'king'], negative=['man']) #根据给定的条件推断相似
model.doesnt_match("breakfast cereal dinner lunch".split()) #寻找离群词
model.similarity('woman', 'man') #计算两个单词的相似度
model['computer'] #获取单词的词向量
```



繁著 (/u/67eb7ed414d3) ♂
写了 30244 字，被 135 人关注，获得了 231 个喜欢
(/u/67eb7ed414d3)

+ 关注

伪文青/摄影湿/程序猿/技术宅/极客范 个人主页：weiweiblog.cn 微信公众号：goodnight_xmu

被以下专题收入，发现更多相似内容

从零开始玩转大数据 (/c/90eff33eb70f?utm_source=desktop&utm_medium=notes-included-collection)

机器学习与数据挖掘 (/c/9ca077f0fae8?utm_source=desktop&utm_medium=notes-included-collection)

首页投稿 (/c/bDHhpK?utm_source=desktop&utm_medium=notes-included-collection)

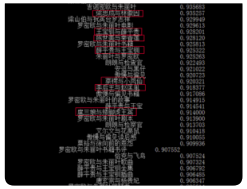
机器学习 (/c/40d81e34a7fb?utm_source=desktop&utm_medium=notes-included-collection)

NLP常用专业术语 (/p/d7ec29abbcb8?utm_campaign=maleskine&utm_c…

常用概念： 自然语言处理（NLP） 数据挖掘 推荐算法 用户画像 知识图谱 信息检索 文本分类 常用技术： 词级别：分词(Seg)，词性标注(POS)，命名实体识别（NER），未登录词识别，词向量（word2vec），词义…

御风之星 (/u/e6ae6d978f3d?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/a9445f709e8e?



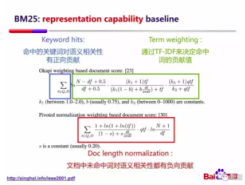
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

自然语言处理真实项目实战（20170830） (/p/a9445f709e8e?utm_campaig…

前言 本文根据实际项目撰写，由于项目保密要求，源代码将进行一定程度的删减。本文撰写的目的是进行公司培训，请勿以任何形式进行转载。由于是日语项目，用到的分词软件等，在中文任务中需要替换为相应的..

中和软件技术推进 (/u/b19707134332?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/3a9f49834c4a?



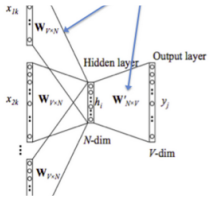
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

浅谈智能搜索和对话式OS (/p/3a9f49834c4a?utm_campaign=maleskine&…

前面的文章主要从理论的角度介绍了自然语言人机对话系统所可能涉及到的多个领域的经典模型和基础知识。这篇文章，甚至之后的文章，会从更贴近业务的角度来写，侧重于介绍一些与自然语言问答业务密切相..

我偏笑_NSNirvana (/u/2293f85dc197?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/208037b8a4f1?



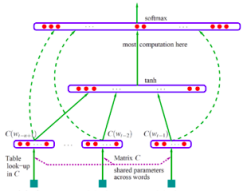
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

CS224n笔记1:自然语言处理简介 (/p/208037b8a4f1?utm_campaign=male…

关键词：自然语言处理（NLP），词向量（Word Vectors），奇异值分解（Singular Value Decompositon），Skip-gram，CBOW，负抽样（Negative Sampling），层次Softmax（Hierachical…

xiiao蜗牛 (/u/10ca3e854ec5?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/4bca99d40597?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

NLP-词嵌入学习笔记 (/p/4bca99d40597?utm_campaign=maleskine&utm…

1.NLP当前热点方向 词法/句法分析 词嵌入(word embedding) 命名实体识别(Name Entity Recognition) 机器翻译(Machine Translation) 情感分析(sentiment analysis) 文档摘要(automatic...


__Aragorn (/u/79a839788418?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

希望你成为这样的人 (/p/9600b907371d?utm_campaign=maleskine&utm…

生在一个水浒的家 有一颗红楼的心 在一个三国纷飞的年代 独自去西游


IcyMu (/u/8df367eaa377?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

马赛克做美缝麻烦吗 简单的马赛克美缝施工方法! (/p/e8a70eae83d2?utm_c…


上海好卓 (/u/d87574f05646?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

我 (/p/4c2937eb37a5?utm_campaign=maleskine&utm_content=note&u…

国歌


浅浅吟唱诗人 (/u/0244471412ab?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/fa5d3f6f851e?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

Android项目小结(二) (/p/fa5d3f6f851e?utm_campaign=maleskine&utm_…

写在前面的话 老实讲 移动开发这块人太多了。好害怕会被淘汰 厌恶安逸 又贪图安逸 以至于最后处在安逸之中失去所有年轻的想法。给自己树立过很多目标 远到一辈子 近到眼前三至五年 唯一不变的是一直停留的脚…

OneBelowZero (/u/904a91e9ca01?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/fc4dd55010b0?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

杨幂酷爱补水仪补水 水嫩软妹养成记 (/p/fc4dd55010b0?utm_campaign=m..

近日,杨幂更新了微博一则,并且附带了美美、水嫩嫩的自拍一张,手中拿着此前代言的金稻补水仪。五月的第一张自拍就这样奉献给了广告~..果然还是wuli敬业幂，这样认真工作只能给满分了。 有媒体夸杨幂保养有…

石火 (/u/7803ede8ffcb?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)