基于gensim的Wiki百科中文word2vec训练

xiiao蜗牛 (/u/10ca3e854ec5) + 关注
2017.07.11 00:04* 字数 946 阅读 138 评论 0 喜欢 6
(/u/10ca3e854ec5)

Word2Vec简介

Word2Vec是词（Word）的一种表示方式。不同于one-hot vector，word2vec可以通过计算各个词之间的距离，来表示词与词之间的相似度。word2vec提取了更多的特征，它使得具有相同上下文语义的词尽可能离得近一些，而不太相关的词尽可能离得较远一些。例如，【腾讯】和【网易】两个词向量将会离得很近，同理【宝马】和【保时捷】两个词向量将会离得很近。而【腾讯】和【宝马】/【保时捷】，【网易】和【宝马】/【保时捷】将会离得较远一些。因为【腾讯】和【网易】都同属于互联网类目，而【宝马】和【保时捷】都同属于汽车类目。人以类聚，物以群分嘛！互联网圈子中谈的毕竟都是互联网相关的话题，而汽车圈子中谈论的都是和汽车相关的话题。

我们怎么得到一个词的word2vec呢？下面我们将介绍如何使用python gensim得到我们想要的词向量。总的来说，包括以下几个步骤：

- wiki中文数据预处理
- 文本数据分词
- gensim word2vec训练

wiki中文数据预处理

首先，下载wiki中文数据：[zhwiki-latest-pages-articles.xml.bz2][1]。因为zhwiki数据中包含很多繁体字，所以我们想获得简体语料库，接下来需要做以下两件事：

- 使用gensim模块中的WikiCorpus从bz2中获取原始文本数据
- 使用OpenCC将繁体字转换为简体字

WikiCorpus获取原始文本数据

数据处理的python代码如下：

```
from __future__ import print_function
from gensim.corpora import WikiCorpus
import jieba
import codecs
import os
import six
from gensim.models import Word2Vec
from gensim.models.word2vec import LineSentence
import multiprocessing


class Config:
    data_path = 'xxx/zhwiki'
    zhwiki_bz2 = 'zhwiki-latest-pages-articles.xml.bz2'
    zhwiki_raw = 'zhwiki_raw.txt'
    zhwiki_raw_t2s = 'zhwiki_raw_t2s.txt'
    zhwiki_seg_t2s = 'zhwiki_seg.txt'
    embedded_model_t2s = 'embedding_model_t2s/zhwiki_embedding_t2s.model'
    embedded_vector_t2s = 'embedding_model_t2s/vector_t2s'


def dataprocess(_config):
    i = 0
    if six.PY3:
        output = open(os.path.join(_config.data_path, _config.zhwiki_raw), 'w')
    output = codecs.open(os.path.join(_config.data_path, _config.zhwiki_raw), 'w')
    wiki = WikiCorpus(os.path.join(_config.data_path, _config.zhwiki_bz2), lemmatize
    for text in wiki.get_texts():
        if six.PY3:
            output.write(b' '.join(text).decode('utf-8', 'ignore') + '\n')
        else:
            output.write(' '.join(text) + '\n')
        i += 1
        if i % 10000 == 0:
            print('Saved ' + str(i) + ' articles')
    output.close()
    print('Finished Saved ' + str(i) + ' articles')

config = Config()
dataprocess(config)
```

使用OpenCC将繁体字转换为简体字

这里，需要预先安装OpenCC，关于OpenCC在linux环境中的安装方法，请参考[这篇文章][2]。仅仅需要两行linux命令就可以完成繁体字转换为简体字的认为，而且速度很快。

```
$ cd /xxx/zhwiki/
$ opencc -i zhwiki_raw.txt -o zhwiki_t2s.txt -c t2s.json
```

文本数据分词

对于分词这个任务，我们直接使用了python的jieba分词模块。你也可以使用哈工大的ltp或者斯坦福的nltk python接口进行分词，准确率及权威度挺高的。不过这两个安装的时候会花费很长时间，尤其是斯坦福的。关于jieba的分词处理代码，参考如下：

```
def is_alpha(tok):
    try:
        return tok.encode('ascii').isalpha()
    except UnicodeEncodeError:
        return False


def zhwiki_segment(_config, remove_alpha=True):
    i = 0
    if six.PY3:
        output = open(os.path.join(_config.data_path, _config.zhwiki_seg_t2s), 'w',
    output = codecs.open(os.path.join(_config.data_path, _config.zhwiki_seg_t2s), 'w
    print('Start...')
    with codecs.open(os.path.join(_config.data_path, _config.zhwiki_raw_t2s), 'r', e
        for line in raw_input.readlines():
            line = line.strip()
            i += 1
            print('line ' + str(i))
            text = line.split()
            if True:
                text = [w for w in text if not is_alpha(w)]
            word_cut_seed = [jieba.cut(t) for t in text]
            tmp = ''
            for sent in word_cut_seed:
                for tok in sent:
                    tmp += tok + ' '
            tmp = tmp.strip()
            if tmp:
                output.write(tmp + '\n')
        output.close()

zhwiki_segment(config)
```

gensim word2vec训练

python的gensim模块提供了word2vec训练，为我们模型的训练提供了很大的方便。关于gensim的使用方法，可以参考[基于Gensim的Word2Vec实践][3]。
本次训练的词向量大小size为50，训练窗口为5，最小词频为5，并使用了多线程，具体代码如下：

```
def word2vec(_config, saved=False):
    print('Start...')
    model = Word2Vec(LineSentence(os.path.join(_config.data_path, _config.zhwiki_seg
                     size=50, window=5, min_count=5, workers=multiprocessing.cpu_cou
    if saved:
        model.save(os.path.join(_config.data_path, _config.embedded_model_t2s))
        model.save_word2vec_format(os.path.join(_config.data_path, _config.embedded_
    print("Finished!")
    return model


def wordsimilarity(word, model):
    semi = ''
    try:
        semi = model.most_similar(word, topn=10)
    except KeyError:
        print('The word not in vocabulary!')
    for term in semi:
        print('%s,%s' % (term[0],term[1]))

model = word2vec(config, saved=True)
```

word2vec训练已经完成，我们得到了想要的模型以及词向量，并保存到本地。下面我们分别查看同【宝马】和【腾讯】最相近的前10个词语。可以发现：和【宝马】相近的词大都属于汽车行业，而且是汽车品牌；和【腾讯】相近的词大都属于互联网行业。

```
>>> wordsimilarity(word=u'宝马', model=model)
保时捷,0.92567974329
固特异,0.888278841972
劳斯莱斯,0.884045600891
奥迪,0.881808757782
马自达,0.881799697876
亚菲特,0.880708634853
欧宝,0.877104878426
雪铁龙,0.876984715462
玛莎拉蒂,0.868475496769
桑塔纳,0.865387916565

>>> wordsimilarity(word=u'腾讯', model=model)
网易,0.880213916302
优酷,0.873666107655
腾讯网,0.87026232481
广州日报,0.859486758709
微信,0.835543811321
天涯社区,0.834927380085
李彦宏,0.832848489285
土豆网,0.831390202045
团购,0.829696238041
搜狐网,0.825544642448
```

附[相关数据及代码][4]，包含：简体字转换后文本，分词后文本，以及50维word2vec词向量。

[1]: https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2 (https://link.jianshu.com?t=https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2)

[2]: http://www.jianshu.com/p/834a02d085b6 (https://www.jianshu.com/p/834a02d085b6)

[3]: https://segmentfault.com/a/1190000008173404?from=timeline (https://link.jianshu.com?t=https://segmentfault.com/a/1190000008173404?from=timeline)

[4]: http://pan.baidu.com/s/1eRLUR9g (https://link.jianshu.com?t=http://pan.baidu.com/s/1eRLUR9g)

xiiao蜗牛 (/u/10ca3e854ec5) ♂
写了 9825 字，被 10 人关注，获得了 17 个喜欢
(/u/10ca3e854ec5)

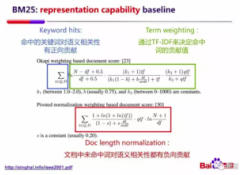
7国的天下，我要99。

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/3a9f49834c4a?



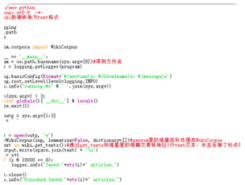
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

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/ec27062bd453?



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

利用Python实现wiki中文语料的word2vec模型构建 (/p/ec27062bd453?ut…

本实例主要介绍的是选取wiki中文语料，并使用python完成Word2vec模型构建的实践过程，不包含原理部分，旨在一步一步的了解自然语言处理的基本方法和步骤。文章主要包含了开发环境准备、数据的获取、数..


atLee (/u/5ac8d2e3d059?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/05800a28c5e4?

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

使用 word2vec 训练wiki中英文语料库 (/p/05800a28c5e4?utm_campaign…

上学期读了有关word2vec的两篇paper之后，不是很明白，这学期重新花时间再读，并且根据这两篇paper进行一个词向量相关的实验，选来选去，发现网上有大神就wiki中英文语料库进行训练，鉴于渣渣水平，于是…

howe_howe (/u/b2d143a2d95f?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

机器学习(Machine Learning)&深度学习(Deep Learning)资料(Chapter 1) (…

机器学习(Machine Learning)&深度学习(Deep Learning)资料(Chapter 1) 注:机器学习资料篇目一共500条,篇目二开始更新 希望转载的朋友，你可以不用联系我．但是一定要保留原文链接，因为这个项目还在继续也…

Albert陈凯 (/u/185a3c553fc6?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/6e67647f1c1a?

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

刘奇葆要求 (/p/6e67647f1c1a?utm_campaign=maleskine&utm_content=…

中共中央政治局委员、书记处书记、中宣部部长刘奇葆，今天在中国摄影家协会第九次全国代表大会讲话中要求全国摄影人： 一、记录时代，用光用影用情怀 二、聚焦人民，见物见人见精神 三、苦练内功，有技有…

通讯员增祥 (/u/aa8fdf2c5d72?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/0fda3c9d8f7a?

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

2017-08-16 (/p/0fda3c9d8f7a?utm_campaign=maleskine&utm_content=…

我的生活里没有诗 长期以来，人们都会因为我的职业问我“你哪里人？”我说我是云南人，第二句“云南很好噢！”我说是的，是很好，第三个问题来了“你是什么族”，我说白族，但我没说完，我父亲是白族母亲是彝…


李敏行云lemon (/u/ddc68b36ba33?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/13e8ed4f9c45?

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

《不想把你遗忘》 (/p/13e8ed4f9c45?utm_campaign=maleskine&utm_co…

昨夜梦中仿佛又来到你孤单的身旁 穿行在都市的大街小巷 我却一个人、黯然神伤 越是思念越是彷徨 今夜难眠数着黑灰色天上的星光 沉浸在无际的爱情荷塘 有许多哀愁、涌出心房 越是想念越是难忘 我不想让记忆深..


康建华 (/u/d756922cc38e?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/74334862e47e?

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

财报扭亏为盈，京东能否成为下一个亚马逊？ (/p/74334862e47e?utm_cam…

5月8日，京东发布了2017财年第一季度业绩报告。净收入为762亿元(约合111亿美元)，同比增长41.2%。基于非美国通用会计准则，净利润为人民币14亿元(约合2亿美元)，去年同期亏损2亿元。受一季报业绩刺激，…


太保乱谈 (/u/c1756dc6ab65?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

卫生巾使用防过敏大攻略 (/p/17d42dd85b5d?utm_campaign=maleskine&…

竹纤维本色生活 生活中，因有些女性体质的问题，如选用的卫生巾不适合就引起过敏。为避免和防止过敏源，首要办法是停用引起过敏反应的卫生巾，同时找到原因，比如这个品牌卫生巾中是不是含让你过敏的香..


本色竹纤维生活馆 (/u/6c06aca8064b?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)