

doc2vec计算文本相似度--python实现

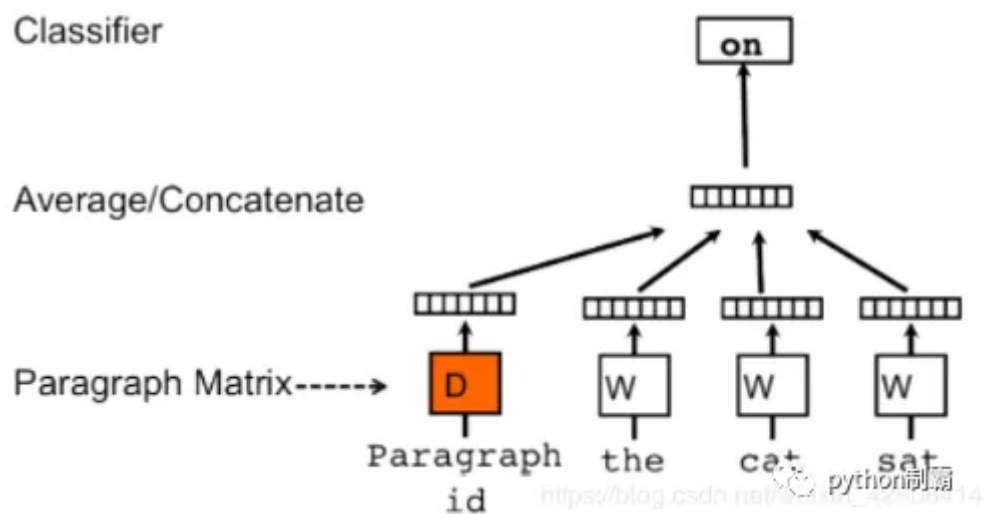
boblee python制霸 2020-05-06

Boblee人工智能硕士毕业，擅长及爱好python，基于python研究人工智能、群体智能、区块链等技术，并使用python开发前后端、爬虫等。

1.背景

doc2vec的目标是创建文档的向量化表示，而不管其长度如何。但与单词不同的是，文档并没有单词之间的逻辑结构，因此必须找到另一种方法。

Mikilov和Le使用的概念很简单但很聪明：他们使用了word2vec模型，并添加了另一个向量（下面的段落ID），如下所示：

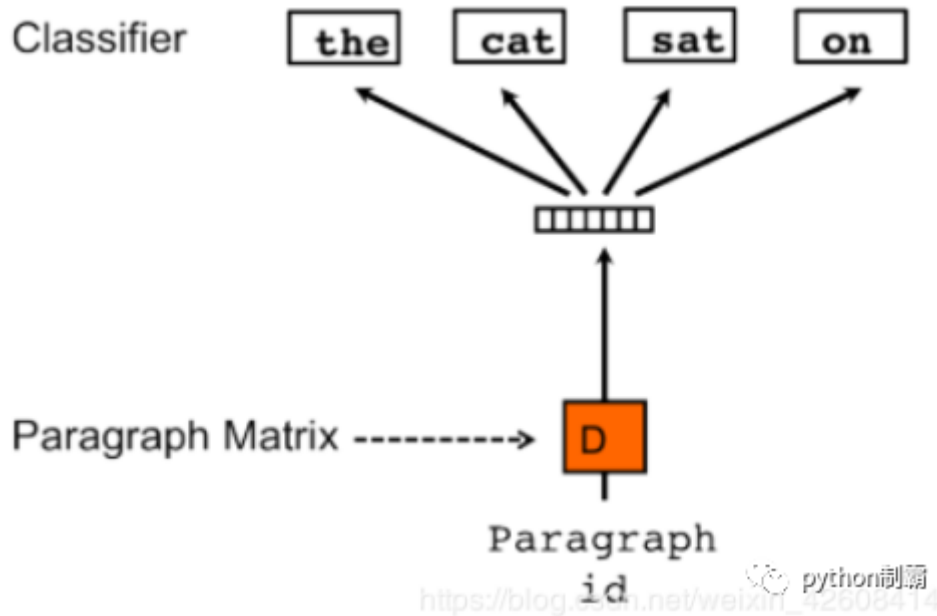


上图是word2vec中CBOW模型的一个小扩展。它不是仅是使用一些单词来预测下一个单词,我们还添加了另一个特征向量，即文档Id。

因此，当训练单词向量W时，也训练文档向量D，并且在训练结束时，它包含了文档的向量化表示。

上面的模型称为段落向量的分布式记忆的版本（PV-DM）。它充当记忆器，它能记住当前上下文中缺少的内容 - 或者段落的主题。虽然单词向量表示单词的概念，但文档向量旨在表示文档的概念。

如在doc2vec中，另一种类似于skip-gram的算法,段落向量的分布式词袋版本（PV-DBOW）。



该算法实际上更快（与word2vec相反）并且消耗更少的内存，因为不需要保存词向量。

在论文中，作者建议使用两种算法的组合，尽管PV-DM模型是优越的，并且通常会自己达到最优的结果。

doc2vec模型的使用方式：对于训练，它需要一组文档。为每个单词生成词向量W，并为每个文档生成文档向量D。该模型还训练softmax隐藏层的权重。在推理阶段，可以呈现新文档，并且固定所有权重以计算文档向量。

2.python实现

本文使用今日头条提供的文本分类数据集进行实验，<https://github.com/skdjfla/toutiao-text-classification-dataset>。

```
6552390546051039747!_102!_news_entertainment!_中国网红竟红到美国? 不多说了，连小编都心动了!_!_飞纱, 新娘, 脱口秀, 中国网, 婚礼
6552150358678831624!_102!_news_entertainment!_赵丽颖很久没有登上微博热搜了，但你们别急，她只是在憋大招而已!_!_陆贞传奇, 大红大紫, 楚乔传, 微博热搜, 赵丽颖, 花千骨, 迪丽热巴
6552408585177924099!_102!_news_entertainment!_因戴一个眼镜更改变气质的6大娱乐圈明星，你最喜欢哪一个!_!_戴上眼镜, 刘德华, 张翰, 远大前程, 杜志国, 刘亦菲
6552147830184608263!_102!_news_entertainment!_后来的我们，抢先看!_!_电影院, 前任3, 刘若英, 张一白, 田壮壮
6552472345548685837!_102!_news_entertainment!_超级英雄演员颜值身材排名，钢铁侠进不了前5，第一名很意外!_!_金刚狼3, 休·杰克曼, 神奇女侠, 绯红女巫, 超人, 金刚狼
6552385284682547716!_102!_news_entertainment!_《无限歌谣季》热播 张绍刚毛不易组合似“父子”!_!_张绍刚, 新组合, 腾讯视频, 无限歌谣季, 毛不易, 父子
6552364464715334151!_102!_news_entertainment!_张靓颖透露右耳已经问歇性失聪10年，这些年她都是怎么过来的啊!_!_中岛美嘉, 滨崎步, 张靓颖, 演唱会, 林子祥
6552310157706002702!_102!_news_entertainment!_成龙改口决定不裸捐了，20亿财产给儿子一半，你怎么看?_!_
6552286735408038403!_102!_news_entertainment!_五大“出轨”女明星，最后一个你们肯定不知道!_!_王全安, 张柏芝, 张雨绮, 吴卓林, 谢霆锋
6552269871697101315!_102!_news_entertainment!_认真搞笑的男人最帅！大张伟不张嘴就是美男子之最！蜜汁好看!_!_天天向上, 毒鸡汤, 大张伟, 美男子, 综艺, 我去上学了, 百变大咖秀
6552418723179790856!_102!_news_entertainment!_谢娜三喜临门，何炅送祝福，吴昕送祝福，只有沈梦辰不一样!_!_杜海涛, 谢娜, 何炅, 沈梦辰, 吴昕, 快本
6552283654494617859!_102!_news_entertainment!_如何评价赵丽颖?_!_
6552453686398812423!_102!_news_entertainment!_有哪些偏冷门的歌曲推荐?_!_
6552383324696871427!_102!_news_entertainment!_“整容狂人”的审美，恕欣赏不来!_!_高富帅, 阿尔维斯, 颜值, 吴彦祖, 罗德里戈
6552390851157295629!_102!_news_entertainment!_杨幂景甜徐冬冬唐嫣 不好好穿衣却美的有趣又撩人!_!_杨幂, 徐冬冬, 背带裙, 大唐荣耀, 唐嫣, 景甜
```

python中提供了doc2vec、word2vec封装好的库sklearn。sklearn使用doc2vec请见<https://radimrehurek.com/gensim/models/doc2vec.html>。

```
1 pip install sklearn
```

1.句子分词

```
1 import gensim
2 import numpy as np
3 import jieba
```

```
4 from gensim.models.doc2vec import Doc2Vec
5
6 def jieba_tokenize(text):
7     """
8     文本分词
9     :param text: 文本
10    :return: 分词list
11    """
12    return jieba.lcut(text)
```

2.获取训练集

```
1 def get_datasest():
2     """
3     获取doc2vec文本训练数据集
4     :return: 文本分词list, 及id
5     """
6     TaggededDocument = gensim.models.doc2vec.TaggedDocument
7     x_train = []
8     for file in open('toutiao_cat_data.txt', encoding='utf8'):
9         file = file.split('!_!_')
10        if len(file) > 3:
11            document = TaggededDocument(file[3], tags=[int(file[1])])
12            x_train.append(document)
13    return x_train
```

3.训练

```
1 def train(x_train, size=2000, epoch_num=10):
2     model_dm = Doc2Vec(x_train,min_count=1, window = 3, size = size, sample=1e
3     model_dm.train(x_train, total_examples=model_dm.corpus_count, epochs=epoch
4     model_dm.save('model')
5     return model_dm
```

4.测试

```
1 def getVecs(model, corpus, size):
2     vecs = [np.array(model.docvecs[z.tags[0]].reshape(1, size)) for z in corp
```

```

3     return np.concatenate(vecs)
4 def test():
5     model_dm = Doc2Vec.load("model")
6     test_text = ['想换个', '30', '万左右', '的', '车', ' ', ' ', '现在', '开科鲁兹',
7     inferred_vector_dm = model_dm.infer_vector(test_text)
8     sims = model_dm.docvecs.most_similar([inferred_vector_dm], topn=10)
9     return sims
10
11 if __name__ == '__main__':
12     x_train = get_datasest()
13     model_dm = train(x_train)
14
15     sims = test()
16     for count, sim in sims:
17         sentence = x_train[count]
18         words = ''
19         for word in sentence[0]:
20             words = words + word + ' '
21         print (words, sim, len(sentence[0]))

```

又一款国产中大型7座SUV将于后天正式上市，你喜欢吗？ 0.42413341999053955 27
 五菱宏光可以换发动机吗？噪音有点大，想换个大众的发动机，可行吗？ 0.34123343229293823 32
 重机云集的醉美江南机车节上谁“最美”？胡军李亚鹏告诉你答案 0.33901235461235046 29
 「新车」国产奔驰A级三厢正式亮相 豪华紧凑型轿车的新标杆 0.2854359447956085 28
 夏季暴晒很伤车！精明领导这样做，汽车寿命多延续5年，真凉爽 0.28314417600631714 29
 在华销量破10万，限量版跑车售罄，CX-8成热点，马自达信心大增 0.27753835916519165 32
 荣威MARVEL X首发如何做到“三过硬五第一”？ 0.27363264560699463 25
 越来越多夫妻自驾游不去酒店，义乌造“房车神器”，真是大开眼界 0.2627106010913849 30
 地表最强1.2T自主紧凑SUV即将上市，动力甩合资几条街 0.25406146049499
 汉腾汽车首款MPV，想做车市黑马，先问问宝骏730同意不 0.2512950301170349 28

感觉效果还是不错。对应车的句子能够提取出来。

喜欢此内容的人还喜欢

那时候....

坐怀不乱

女领导不配合防疫，耍官威找“卢书记”走后门！网友灵魂发问：你要是他该咋办？

躺倒鸭