单位代码
学　　号 SY2303806
分 类 号

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理（NLP）
第三次课后作业

院（系）名称　自动化科学与电气工程学院
专业名称　自动化
学生姓名　芦川川

2024 年 06 月

芦川川 自动化科学与电气工程学院 sy2303806@buaa.edu.cn

Abstract

利用给定金庸小说的语料库，利用 1～2 种神经语言模型（如：基于 Word2Vec ， LSTM， GloVe 等模型）来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。本文最终选取的是 Word2Vec 模型。

Introduction

Word2Vec 是语言模型中的一种，从大量文本预料中以无监督方式学习语义知识的模型，被广泛地应用于自然语言处理中。自然语言处理相关任务中要将自然语言交给机器学习中的算法来处理，通常需要将语言数学化，因为计算机机器只认数学符号。向量是人把自然界的东西抽象出来交给机器处理的数学性质的东西，基本上可以说向量是人对机器输入的主要方式了。词向量是对词语的向量表示，这些向量能捕获词语的语义信息，如相似意义的单词具有类似的向量。

Word2Vec 的主要作用是生成词向量，而词向量与语言模型有着密切的关系。Word2Vec 的特点是能够将单词转化为向量来表示，这样词与词之间就可以定量的去度量他们之间的关系，挖掘词之间的联系。Word2Vec 模型在自然语言处理中有着广泛的应用，包括词语相似度计算、文本分类、词性标注、命名实体识别、机器翻译、文本生成等。其主要目的是将所有词语投影到 K 维的向量空间，每个词语都可以用一个 K 维向量表示。

词向量就是用来将语言中的词进行数学化的一种方式，顾名思义，词向量就是把一个词表示成一个向量。 我们都知道词在送到神经网络训练之前需要将其编码成数值变量，常见的编码方式有两种：One-Hot Representation 和 Distributed Representation。

Word2Vec 模型的核心思想是通过词语的上下文信息来学习词语的向量表示。具体来说，Word2Vec 模型通过训练一个神经网络模型，使得给定一个词语的上下文时，能够预测该词语本身（CBOW 模型），或者给定一个词语时，能够预测其上下文（Skip-gram 模型）。Word2Vec 的训练模型本质上是只具有一个隐

含层的神经元网络。它的输入是采用 One-Hot 编码的词汇表向量，它的输出也是 One-Hot 编码的词汇表向量。使用所有的样本，训练这个神经元网络，等到收敛之后，从输入层到隐含层的那些权重，便是每一个词的采用 Distributed Representation 的词向量。

Methodology

实验步骤如下：

1、使用 jieba 库对金庸小说语料库进行分词

2、去除停用词

3、模型训练

4、设计实验计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联验证词向量的有效性

Experimental Studies

将金庸小说语料库预处理后进行分词并去除停用词，使用 Word2Vec 进行模型训练。

```
w2v_model = Word2Vec(sentences=LineSentence('./split_words.txt'), vector_size=200, window=5, min_count=5,
                     workers=20, epochs=10)
```

设置词向量的度为 200，滑动窗口大小为 5，即当前词与上下文词的最远距离为 5，词的最小出现次数为 5，低于此阈值的会被过滤掉，迭代次数为 10。

1、获取了“杨过”和“段誉”这两个词转换为词向量后相似度最高的 5 个词，结果如下表所示：

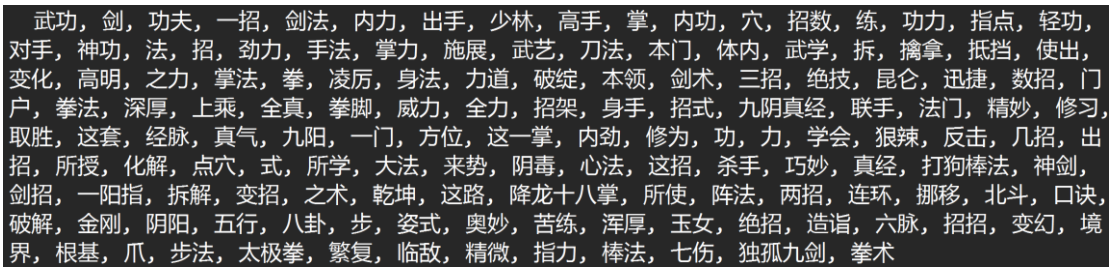
杨过	关联度
小龙女	0.710321784
郭襄	0.656961918
黄蓉	0.622366369
郭靖	0.594892025
武三通	0.592811108

段誉	关联度
王语嫣	0.745898724
木婉清	0.681169629
慕容复	0.654950976
阿朱	0.602510512
段正淳	0.592973351

可以看到，这样的结果是符合预料库内容的，这些人物关系是比较近的。

同时再次计算“韦小宝”和“鸠摩智”对应的词向量以及两个词向量的相似度，结果为 0.2433488368988037，两个词是两本书中的人物，关联程度较低。

2、使用 K-means 聚类算法对词向量进行聚类，设置聚类簇数量为 20，下图展示了一个簇的词：



这些词语都是功夫相关的词语，因此被聚到了一个簇中。

3、计算两个段落之间的相似性：

段落 1：自掌法练成以来，直至此时，方遇到周伯通这等真正的强敌。周伯通听说这是他自创的武功，兴致更高，说道：“正要见识见识！”挥手而上，仍是只用左臂。杨过抬头向天，浑若不见，呼的一掌向自己头顶空空拍出，手掌斜下，掌力化成弧形，四散落下。周伯通知道这一掌力似穹庐，圆转广被，实是无可躲闪，当下举掌相迎，“啪”的一下，双掌相交，不由得身子一晃，都只为他过于托大，殊不知他武功虽然决不弱于对方，但一掌对一掌，却无不及杨过掌力厚实雄浑。

段落 2：杨过倒持长剑，回掌相迎，砰的一声响，两股巨力相交，两人同时一晃，木梯摇了几摇，几乎折断。两人都是一惊，暗赞对手了得：“一十六年不见，他功力居然精进如斯！”杨过见情势危急，不能和他在梯上多拚掌力，长剑向上疾刺，或击小腿，或削脚掌。法王身子在上，若出金轮与之相斗，则兵刃既短，俯身弯腰实在大是不便，只得急奔上高台。杨过向他背心疾刺数剑，招招势若暴风骤雨，但法王并不回头，听风辨器，一一举轮挡开，便如背上长了眼睛一般。杨过喝采道：“贼秃！恁的了得！”法王刚刚踏上台顶回首就是一轮。杨过侧首让过，身随剑起，在半空中扑击而下。法王举金轮一挡，左手银轮便往他剑上砸去。

这两个段落皆是在《神雕侠侣》中两人相互打斗的场面，计算得出的相似性较高。

Conclusion

本次实验通过对金庸小说语料进行处理后训练了 Word2Vec 模型，通过计算词与词转换为词向量之后的相似度，使用 K-means 聚类方法对某一簇词语进行分析，以及计算两个段落之间的相似性的计算，证明了词向量的有效性。

References

[1] https://baike.baidu.com/item/Word2vec/22660840?fr=ge_ala

[2] https://zhuanlan.zhihu.com/p/371147732

[3] https://blog.csdn.net/a7303349/article/details/132385230