北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理
第四次大作业

基于 Word2Vec 的词向量训练

院（系）名称	自动化科学与电气工程学院
学 生 学 号	ZY2103810
学 生 姓 名	游虎杰

2022 年 5 月

一、问题描述

对于每一篇金庸小说，利用 Word2Vec 模型训练词向量，将小说中的每一个词汇映射到一个高维向量，使得相近词汇对应的词向量更为接近，并利用 PCA 算法将高维词向量映射到二维空间，便于利用图像展示。

二、问题表达

词向量（Word embedding），又叫 Word 嵌入式，是自然语言处理（NLP）中的一组语言建模和特征学习技术的统称，其中来自词汇表的单词或短语被映射到实数的向量。从概念上讲，它涉及从每个单词一维的空间到具有更高维度的连续向量空间的数学嵌入。

如果将 word 看作文本的最小单元，可以将 Word Embedding 理解为一种映射，其过程是：将文本空间中的某个 word，通过一定的方法，映射或者说嵌入（embedding）到另一个数值向量空间。

Word2Vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，Word2Vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

Word2Vec 主要包括 CBOW 模型（连续词袋模型）和 Skip-gram 模型（跳字模型）本文主要使用的模型为 CBOW 模型，即给定一个长度为 T 的文本序列，设事件步的词为 W(t)，背景窗口大小为 m，则连续词袋模型的目标函数（损失函数）为由背景词生成任一中心词的概率，即：

$$\sum_{t=1}^{T} P(w^{(t)}|w^{(t-m)},\dots,w^{(t-1)},w^{(t+1)},\dots,w^{(t+m)})$$

本次实验以金庸先生的 16 本武侠小说作为语料库，利用 word2vec 模型来产生词向量，小说如下：

1、白马啸西风

2、碧血剑

3、飞狐外传

4、连城诀

5、鹿鼎记

6、三十三剑客图

7、射雕英雄传

8、神雕侠侣

9、书剑恩仇录

10、天龙八部

11、侠客行

12、笑傲江湖

13、雪山飞狐

14、倚天屠龙记

15、鸳鸯刀

16、越女剑

三、具体算法实现

1. 语料处理

在读取语料后，首先利用 jieba 分词对语料进行分词，去掉 txt 文本中一些无意义的标点符号，并将处理后的语料重新保存。

2. 模型训练

使用开源的 Gensim 库提供的接口来训练 Word2Vec 模型，调用的函数如下：

model = Word2Vec(sentences=all_terms[name], size=256, min_count=10, sg=0, window=10, iter=10)

其中 sentences 为训练 Word2Vec 模型所用语料库，size 为所训练词向量的维度；min_count 为词频下限，词频低于 min_count 的词汇会被丢除；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，本文采用 CBOW 算法，因此设置 sg=0；window 表示当前词与预测词在一个句子中的最大距离，设置为 10，iter 为迭代次数。

3. 聚类分析

模型训练完毕后，通过 most_similar 函数计算给定词汇最近的指定数目最相似的词汇，并判断计算结果的准确性。

四、运行结果

对所有语料库利用 Word2Vec 模型训练词向量，并对《射雕英雄传》、《天龙八部》、《倚天屠龙记》的训练结果进行测试，结果如下：

1. 《射雕英雄传》测试结果

图 1 为《射雕英雄传》部分词汇对应的词向量经过 PCA 降维后的结果

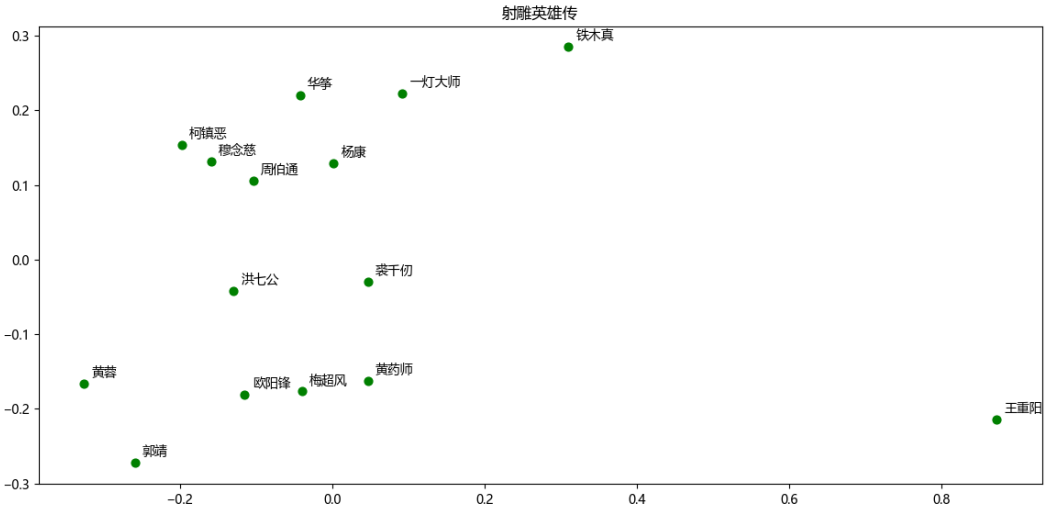

图 1 《射雕英雄传》部分词汇对应词向量经过 PCA 降维后的结果

部分书中词汇对应词向量与其他词汇对应词向量的距离如下：

郭靖:

('黄蓉', 0.9475803971290588)

('欧阳克', 0.8806752562522888)

('朱聪', 0.8687146902084351)

('洪七公', 0.8670706748962402)

('欧阳锋', 0.856984555721283)

('裘千仞', 0.8198520541191101)

('彭连虎', 0.8141355514526367)

('华筝', 0.8132104873657227)

('黄药师', 0.8066521883010864)

('拖雷', 0.8066451549530029)

('周伯通', 0.8003585338592529)

('马钰', 0.7995474338531494)

('梅超风', 0.796858549118042)

('沙通天', 0.792699933052063)

('穆念慈', 0.7877962589263916)

('柯镇恶', 0.7839921712875366)

('杨康', 0.7838218808174133)

('梁子翁', 0.7736456394195557)

('程瑶迦', 0.768312931060791)

('杨铁心', 0.7678513526916504)

黄蓉:

('郭靖', 0.9475803971290588)

('洪七公', 0.8778267502784729)

('朱聪', 0.8763206005096436)

('欧阳锋', 0.8747590780258179)

('欧阳克', 0.8739835023880005)

('周伯通', 0.8605402112007141)

('杨康', 0.8228564262390137)

('梁子翁', 0.8201038241386414)

('完颜康', 0.820087194442749)

('柯镇恶', 0.8169962763786316)

('穆念慈', 0.8161196708679199)

('彭连虎', 0.8124022483825684)

('华筝', 0.8115098476409912)

('梅超风', 0.8017178773880005)

('裘千仞', 0.7995185852050781)

('程瑶迦', 0.7985448837280273)

('马钰', 0.7962274551391602)

('灵智上人', 0.7945722341537476)

('杨铁心', 0.7903691530227661)

('沙通天', 0.7892443537712097)

杨康:

('完颜康', 0.9324432015419006)

('马钰', 0.9048839807510376)

('朱聪', 0.9035216569900513)

('王处一', 0.8961740732192993)

('丘处机', 0.8897268176078796)

('彭连虎', 0.8896654844284058)

('周伯通', 0.887956440448761)

('华筝', 0.8824740648269653)

('沙通天', 0.8810994625091553)

('洪七公', 0.8804255723953247)

('完颜洪烈', 0.8789986371994019)

('程瑶迦', 0.8768291473388672)

('梁子翁', 0.8671119213104248)

('拖雷', 0.8667126893997192)

('陆冠英', 0.861039400100708)

('裘千仞', 0.8604547381401062)

('穆念慈', 0.8589202165603638)

('哲别', 0.8568482398986816)

('杨铁心', 0.8553208112716675)

('欧阳锋', 0.8423619270324707)

2. 《天龙八部》测试结果

图 2 为《天龙八部》部分词汇对应的词向量经过 PCA 降维后的结果：

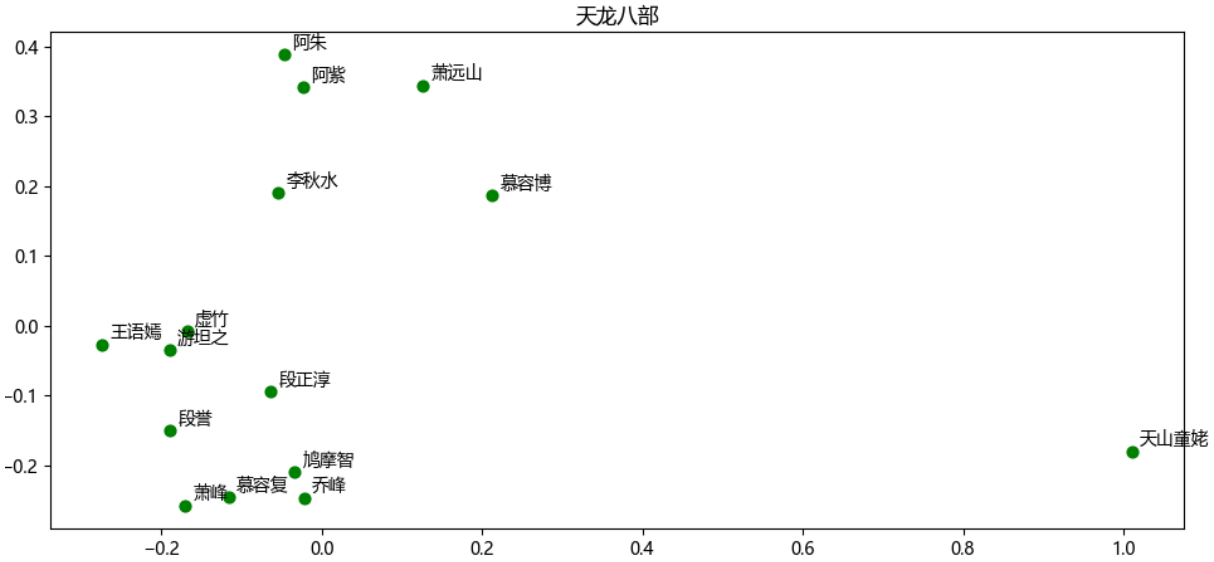

图 2 《天龙八部》部分词汇对应词向量经过 PCA 降维后的结果

部分书中词汇对应词向量与其他词汇对应词向量的距离如下：

萧峰:

('乔峰', 0.8691942691802979)

('游坦之', 0.8210697174072266)

('段誉', 0.8059802055358887)

('慕容复', 0.7688406705856323)

('鸠摩智', 0.7627589106559753)

('云中鹤', 0.755689263343811)

('木婉清', 0.7552136182785034)

('乌老大', 0.7539370059967041)

('王语嫣', 0.7528622150421143)

('丁春秋', 0.7430309057235718)

('虚竹', 0.7417379021644592)

('段正淳', 0.7385022640228271)

('包不同', 0.7383965253829956)

('李秋水', 0.730709969997406)

('段延庆', 0.7248172760009766)

('叶二娘', 0.7136490941047668)

('玄难', 0.7115432024002075)

('全冠清', 0.7105585336685181)

('司马林', 0.710083544254303)

('邓百川', 0.7062801122665405)

乔峰:

('萧峰', 0.8691942691802979)

('段正淳', 0.8471696376800537)

('慕容复', 0.8084288239479065)

('游坦之', 0.7968918085098267)

('鸠摩智', 0.7958258390426636)

('丁春秋', 0.795240044593811)

('乌老大', 0.7878570556640625)

('童姥', 0.7839413285255432)

('木婉清', 0.7769416570663452)

('段誉', 0.7670755386352539)

('包不同', 0.7661264538764954)

('王语嫣', 0.7661235928535461)

('段延庆', 0.7647024393081665)

('慕容博', 0.757604718208313)

('母亲', 0.7501741647720337)

('钟万仇', 0.7483278512954712)

('李秋水', 0.7414591908454895)

('玄难', 0.7344335317611694)

('虚竹', 0.7339366674423218)

('司马林', 0.7300753593444824)

段誉:

('慕容复', 0.8309671878814697)

('虚竹', 0.8235469460487366)

('游坦之', 0.808430016040802)

('钟灵', 0.8081443905830383)

('萧峰', 0.8059802055358887)

('王语嫣', 0.798821747303009)

('段正淳', 0.7880030870437622)

('鸠摩智', 0.7873125076293945)

('乔峰', 0.7670754790306091)

('乌老大', 0.7670257687568665)

('钟万仇', 0.766524076461792)

('段延庆', 0.7596941590309143)

('木婉清', 0.7595235109329224)

('叶二娘', 0.7484167814254761)

('云中鹤', 0.7440444827079773)

('童姥', 0.7433661222457886)

('阿紫', 0.7413511276245117)

('包不同', 0.7257860898971558)

('卓不凡', 0.7192988395690918)

('丁春秋', 0.7187897562980652)

3. 《倚天屠龙记》测试结果

图 3 为《倚天屠龙记》部分词汇对应的词向量经过 PCA 降维后的结果：

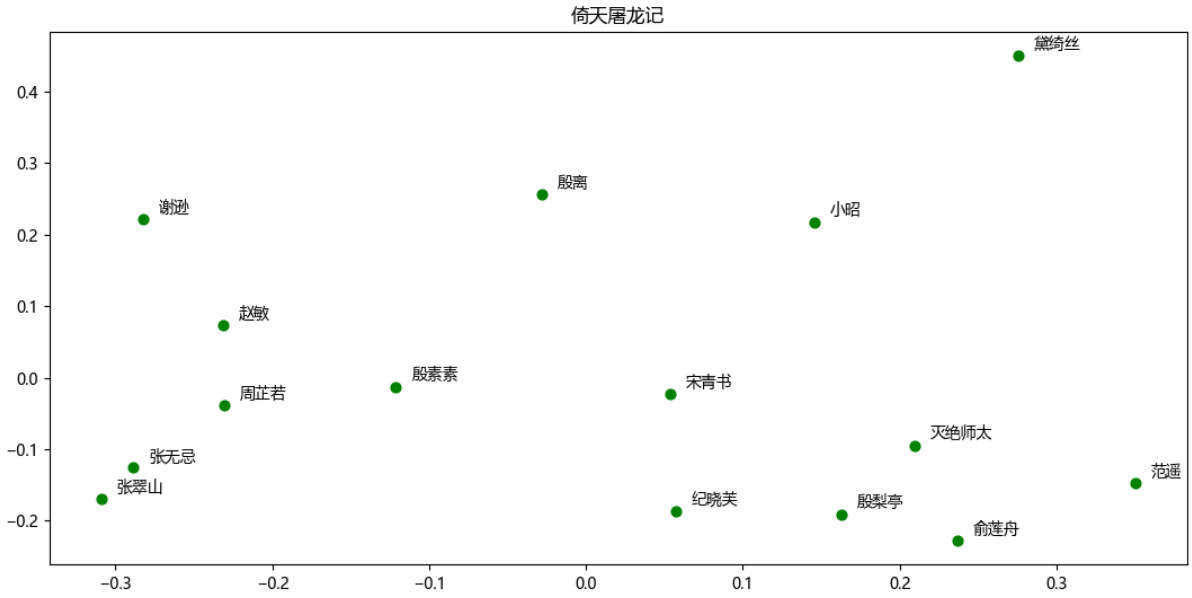

图 3 《倚天屠龙记》部分词汇对应词向量经过 PCA 降维后的结果

部分书中词汇对应词向量与其他词汇对应词向量的距离如下：

张无忌:

('张翠山', 0.9222844839096069)

('金花婆婆', 0.8542318940162659)

('周芷若', 0.8437351584434509)

('纪晓芙', 0.8395252227783203)

('殷素素', 0.8363269567489624)

('赵敏', 0.8299676179885864)

('蛛儿', 0.8265863656997681)

('朱长龄', 0.8243751525878906)

('殷梨亭', 0.8229608535766602)

('俞岱岩', 0.818655252456665)

('俞莲舟', 0.8087659478187561)

('胡青牛', 0.7918716669082642)

('丁敏君', 0.7828006148338318)

('都大锦', 0.7737926840782166)

('谢逊', 0.7713722586631775)

('宋青书', 0.7682888507843018)

('常遇春', 0.7561923265457153)

('殷离', 0.7553930282592773)

('鹿杖客', 0.7541482448577881)

('彭和尚', 0.7478709816932678)

赵敏:

('周芷若', 0.912711501121521)

('张翠山', 0.8740882277488708)

('殷素素', 0.864667534828186)

('宋青书', 0.8565881252288818)

('谢逊', 0.8383188247680664)

('张无忌', 0.8299676179885864)

('金花婆婆', 0.8291010856628418)

('殷梨亭', 0.8229625225067139)

('蛛儿', 0.8211805820465088)

('朱长龄', 0.7975161075592041)

('纪晓芙', 0.7939469814300537)

('鹿杖客', 0.7938337326049805)

('丁敏君', 0.7923796772956848)

('胡青牛', 0.7917639017105103)

('俞岱岩', 0.7896651029586792)

('郭襄', 0.7855042815208435)

('小昭', 0.7826138734817505)

('殷离', 0.7811025977134705)

('西华子', 0.7778066396713257)

('灭绝师太', 0.7767695188522339)

周芷若:

('赵敏', 0.912711501121521)

('张翠山', 0.9084430932998657)

('殷素素', 0.89710533618927)

('宋青书', 0.8586602210998535)

('殷梨亭', 0.8463558554649353)

('纪晓芙', 0.8457255363464355)

('张无忌', 0.8437351584434509)

('俞岱岩', 0.8168180584907532)

('蛛儿', 0.8160115480422974)

('胡青牛', 0.8028051853179932)

('金花婆婆', 0.7921520471572876)

('俞莲舟', 0.7897861003875732)

('小昭', 0.7892968654632568)

('灭绝师太', 0.7842851281166077)

('丁敏君', 0.7838174104690552)

('朱长龄', 0.781126081943512)

('谢逊', 0.7736884355545044)

('殷离', 0.7731884717941284)

('郭襄', 0.7707699537277222)

('西华子', 0.7668347954750061)

4. 总结

由以上实验数据与图像可以看出，Word2Vec 模型所构建词向量基本符合实际情况，能够较准确反应出各词汇之间的联系紧密程度，这说明了 Word2Vec 模型的实用性。

五、个人总结和体会

通过这次作业，我对 Word2Vec 模型和词向量的理解有了进一步的加深，对文本的 Word2Vec 建模过程有了更为熟练的掌握。同时，在编写代码的过程中，我对 python 的应用水平有了进一步提升。

六、作业代码

https://github.com/youlll/DP_NLP4.git