学　号 20376310

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理

大作业 3

词向量训练与验证

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	杨佳木
学 号	**20376310**

2024 年 5 月

1.摘要

本报告的主要内容主要分为以下两部分：第一部分：使用 Word2Vec 模型训练词向量，第二部分：通过计算词向量之间的语义距离、某一类词语的聚类、某些段落直接的语意关联来验证该词向量的有效性。经过实验，取得较好的验证效果。

2.问题描述

本文基于 Word2Vec 模型在对所给语料库进行处理，并研究以下三个问题：

（1）Word2Vec 生成词向量的主要流程

（2）针对训练后的词向量应使用何种方法评价训练效果

（3）基于词向量计算段落间的语意关联方法研究

3.问题的解决方案

基于以上的三个问题，设计问题的解决方案，并将其集成于三个实验中，研究步骤如下所示：

1.基于语料库生成数据集：选择老师给定的 16 部金庸小说合并成一个完整文本，基于 jieba 分词对文本进行分词，删除隐藏符号、非中文字符、标点及停用词，得到中文词数据集。

2.使用 gensim 库的 Word2Vec 模型对语料库进行训练，得到词向量模型。

3.对模型进行保存，使用该模型计算词语之间的相似度得分和语意距离，计算词语相似度。

4.使用词向量分类模型对某一类词语进行分类，以对词向量划分进行检验。

5.使用分类模型对所得词向量对段落的语意关联进行计算。

实验 1 词向量训练

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， 给定一个长度为T的文本序列，设时间步的词为$W(t)$，背景窗口大小为m。则连续词袋模型的目标函数（损失函数）是由背景词生成任一中心词的概率。

$$\sum_{t=1}^{T} \mathrm{P}\left(\mathrm{w}^{(t)} | \mathrm{w}^{(t-m)}, \ldots, \mathrm{w}^{(t-1)}, \mathrm{w}^{(t+1)}, \ldots, \mathrm{w}^{(t+m)}\right)$$

基于以上原理，在前两侧作业的基础上，对得到的词向量组使用 gensim 库训练词向量，代码如下：

```
model1 = Word2Vec(vector_size=200, window=5, min_count=1, workers=4)
model1.build_vocab(corpus)

#训练 Word2Vec 模型
model1.train(corpus, total_examples=model.corpus_count, epochs=10)
```

其中，Word2Vec 的参数含义如下：

Sg：代表使用的模型，其中默认使用的是词袋模型

min_count：最低词频阈值，低于 min_count 的词过滤掉，这里取 1

vector_size:词向量维度，取值几十到几百，这里取 200

window：滑动窗口,当前词与上下文词最远距离，这里取 5

workers：计算使用线程数，这里取 4

实验 2 计算词向量间的语意距离

基于以上建立的词向量分类模型，对文本语料中选取的词语进行词向量划分。通过分别选取一本书和整个语料库作为训练集，对同一个词计算词向量语意距离，结果如下：

选取《天龙八部》这部书，选取人名“虚竹”作为训练词，得到语意距离前十名的词向量对如下表所示：

乌老大	0.7882791757583618
段正淳	0.7848057150840759
游坦之	0.7654115557670593
木婉清	0.756263017654419
乔峰	0.7511184811592102
那老僧	0.7416008114814758
白世镜	0.7362761497497559
钟夫人	0.7242682576179504
段誉	0.7197190523147583
苏星河	0.7190867066383362

这是因为，在天龙八部里，主人公之一虚竹和这些人关系分词密切，发生很多交集。

使用整个语料库作为数据集，得到“虚竹”语意距离前十名的词向量对如下表所示：

令狐冲	0.8014066219329834
张无忌	0.8009008765220642
杨过	0.722987711429596

张翠山	0.7213016748428345
石破天	0.7203312516212463
周芷若	0.718650758266449
俞岱岩	0.7138881683349609
洪七公	0.7134624719619751
小龙女	0.7025972604751587
段誉	0.6988698840141296

乍一看，从剧情上这些人没有交集，但当语料库变成整部金庸小说作品后，他们有一个共同的特征，就是都是对应作品的主人公，这一点是词向量划分的主要依据。

由上可以看出，选取人物姓名能够基于词向量进行很好地分类，再选取“刀光”一词，得到其词向量语义距离前十名的结果如下：

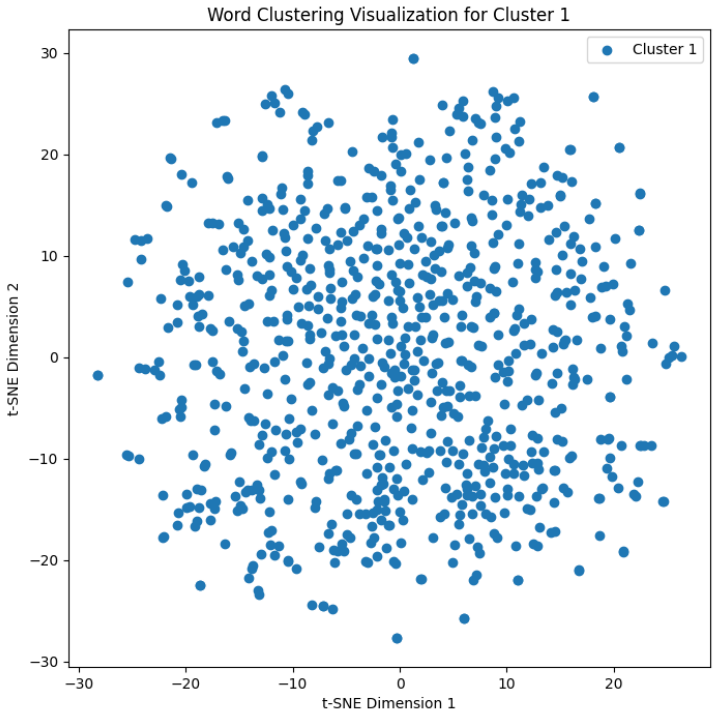
剑光	0.9076313972473145
白光	0.8949196338653564
闪动	0.8922746181488037
金光	0.8738701939582825
闪烁	0.8731573224067688
飞舞	0.869215190410614
银光	0.8691362738609314
青光	0.8588536381721497
掌影	0.841393768787384
寒光	0.8313745260238647

“刀光”一词，往往形容打斗时刀剑的舞动发出的光芒，因而，这些词与其关系都非常大。

对以上两个实验内容进行总结，可以看出通过词向量划分能够很好地对有相同特征的词进行归类。

实验 3 对某一类词向量进行聚类

使用 K-means 对整部小说集词向量进行聚类，得到聚类簇 1 结果如下所示：



可以看出词语之间聚类效果较为密集。部分聚类词如下所示：｛ 只，原来，就要，假装，想要，几次，仔细，狠狠，惊道，忽道，只想，留神，正想，第一个，偷偷，郭襄道，哭道，喝问，打断，便来，有没有，看看，怒喝，倒也不，没人，轻声道，一人道，快些，段郎，全没，别怕，黄蓉叫，啐，试，公主道，这么久，别动｝

这部分词组要是表现人物的语言以及心理活动，因而归于一类，可以看出聚类效果较好。

实验 4 对段落间的语意关联进行计算

抽取以下两个段落：

段落一： 西首锦凳上所坐的则是别派人士，其中有的是东西二宗掌门人共同出面邀请的公证人，其余则是前来观礼的嘉宾。这些人都是云南武林中的知名之士。只坐在最下首的那个青衣少年却是个无名之辈，偏是他在龚姓汉子伴作失足时嗤的一声笑。这少年乃随滇南普洱老武师马五德而来。马五德是大茶商，豪富好客，颇有孟尝之风，江湖上落魄的武师前去投奔，他必竭诚相待，因此人缘甚佳，武功却是平平。左子穆听马五德引见之时说这少年姓段，段姓是大理国的国姓，大理境内姓段的成千成万，左子穆当时听了也不以为意，心想分多半是马五德的弟子，这马老儿自身的功夫稀松平常，调教出来的弟子还高得到那里去，是以连“久仰”两字也懒得说，只拱了拱手，便肃入宾座。不料这年轻人不知天高地厚，竟当左子穆的得意弟子佯出虚招诱敌之时，失笑讥讽。

段落 2：桃实仙躺在担架之上，说道：“瞧你相貌，比我们更加人不像人、鬼不像鬼。”原来桃实仙经平一指缝好了伤口，服下灵丹妙药，又给他在顶门一拍，输入真气，立时起身行走，但毕竟失血太多，行不多时，便又晕倒，给那中年妇人提了转去。他受伤虽重，嘴头上仍是决不让人，忍不住要和那妇人顶撞几句。那妇人冷冷的道：“你们可知平大夫生平最怕的是甚么？”桃谷六仙齐道：“不知道，他怕甚么？”那妇人道：“他最怕老婆！”桃谷六仙哈哈大笑，齐声道：“他这么一个天不怕、地不怕的人，居然怕老婆，哈哈，可笑

啊可笑！”那妇人冷冷的道：“有甚么好笑？我就是他老婆！”桃谷六仙立时不作一声。那妇人道：“我有甚么吩咐，他不敢不听。我要杀甚么人，他便会叫你们去杀。”桃谷六仙齐道：“是，是！不知平夫人要杀甚么人？”那妇人的眼光向船舱中射去，从岳不群看到岳夫人，又从岳夫人看到岳灵珊，逐一瞧向华山派群弟子，每个人都给她看得心中发毛，各人都知道，只要这个形容丑陋、全无血色的妇人向谁一指，桃谷五仙立时便会将这人撕了，纵是岳不群这样的高手，只怕也难逃毒手。

这两个段落是我随便从两篇小说中选的，直观来看并无什么联系，使用词向量计算段落间的相似度，得到结果如下所示：

段落 1 与段落 2 的语义相似度：0.5554736256599426

若选取一组明显相关的小说段落，如下所示：

段落一：林震南点头道：“老头儿怕事，这里杀伤了人命，尸体又埋在他菜园子里，他怕受到牵连，就此一走了之。”走到菜园里，指着倚在墙边的一把锄头，说道：“陈七，把死尸掘出来瞧瞧。”陈七早认定是恶鬼作祟，只锄得两下，手足俱软，直欲瘫痪在地。季镖头道：“有个屁用？亏你是吃镖行饭的！”一手接过锄头，将灯笼交在他手里，举锄扒开泥土，锄不多久，便露出死尸身上的衣服，又扒了几下，将锄头伸到尸身下，用力一挑，挑起死尸。陈七转过了头，不敢观看，却听得四人齐声惊呼，陈七一惊之下，失手抛下灯笼，蜡烛熄灭，菜园中登时一片漆黑。林平之颤声道：“咱们明明埋的是那四川人，怎地……怎地……”林震南道：“快点灯笼！”他一直镇定，此刻语音

中也有了惊惶之意。崔镖头晃火折点着灯笼，林震南弯腰察看死尸，过了半晌，道：“身上也没伤痕，一模一样的死法。”陈七鼓起勇气，向死尸瞧了一眼，尖声大叫：“史镖头，史镖头！”地下掘出来的竟是史镖头的尸身，那四川汉子的尸首却已不知去向。林震南道：“这姓萨的老头定有古怪。”抢着灯笼，奔进屋中察看，从灶下的酒坛、铁镬，直到厅房中的桌椅都细细查了一遍，不见有异。崔季二镖头和林平之也分别查看。突然听得林平之叫道：“咦！爹爹，你来看。”

段落 2：林震南循声过去，见儿子站在那少女房中，手中拿着一块绿色帕子。林平之道：“爹，一个贫家女子，怎会有这种东西？”林震南接过手来，一股淡淡幽香立时传入鼻中，那帕子甚是软滑，沉甸甸的，显是上等丝缎，再一细看，见帕子边缘以绿丝线围了三道边，一角上绣着一枝小小的红色珊瑚枝，绣工甚是精致。林震南问：“这帕子哪里找出来的？”林平之道：“掉在床底下的角落里，多半是他们匆匆离去，收拾东西时没瞧见。”林震南提着灯笼俯身又到床底照着，不见别物，沉吟道：“你说那卖酒的姑娘相貌甚丑，衣衫质料想来不会华贵，但是不是穿得十分整洁？”林平之道：“当时我没留心，但不见得污秽，倘若很脏，她来斟酒之时我定会觉得。”

这两个段落是一篇《笑傲江湖》中连续的两段，并且描述的是一件连贯的事情，使用词向量评价两段间的关系，得到：

段落 1 与段落 2 的语义相似度：0.8327980041503906

可以看到两者相似度明显变高。

总结

基于以上三个实验，可以总结出以下结论

1. 基于词向量能够很好的对具有相似特征的词语进行划分，同时能够通过对相似特征词语划分能力判断词向量的训练效果。

2. 基于词向量能够根据具有相似特征这一特点对段落进行分类。