机器学习与自然语言处理实验三

院（系）名称	自动化科学与电气工程学院
学生姓名	赵怡然
学生学号	ZY2103208
指导老师	秦曾昌

2022 年 05 月 20 日

一、 实验背景介绍

在自然语言处理任务中，首先需要考虑词如何在计算机中表示。通常，有两种表示 方式：one-hot representation（离散表示）和 distribution representation（分布式表示）。

传统的基于规则或基于统计的自然语义处理方法将单词看作一个原子符号，这种方 法被称作 one-hot representation。one-hot representation 把每个词表示为一个长向量。这 个向量的维度是词表大小，向量中只有一个维度的值为 1，其余维度为 0，这个维度就 代表了当前的词。例如：苹果: [0,0,0,1,0,0,0 … …]。one-hot representation 相当于给每个词 分配了一个 id，这就导致这种表示方式不能展示词与词之间的关系。另外， one-hot representation 将会导致特征空间非常大，但也带来一个好处，就是在高维空间中，很多 应用任务线性可分。

分布式的方式通常称为 distribution representation，是将词转化为一种分布式的、连 续的、定长的稠密向量，其优点是可以表示词与词之间的距离关系，每一维度都有其特定的含义。

二、 实验目标

1. 利用给定语料库(或者自选语料库)，利用神经语言模型(如:Word2Vec, GloVe 等模型)来训练词向量。

2. 通过对词向量的聚类或者其他方法来验证词向量的有效性。

三、 相关原理

3.1 语言模型

语言模型生成词向量是通过训练神经网络语言模型 NNLM（neural network language model），词向量做为语言模型的附带产出。NNLM 背后的基本思想是对出现在上下文环 境里的词进行预测，这种对上下文环境的预测本质上也是一种对共现统计特征的学习。

较著名的采用神经网络语言模型生成词向量的方法有：Word2Vec、LBL、NNLM、 C&W、GloVe 等。

3.2 word2vec 模型

word2vec 是一种浅层的神经网络，由嵌入层、隐藏 层、输出层构成。其根

据输入，输出的特点可以分为两种模式：CBOW（由上下文预测 当前词），Skip Fram（由当前词预测上下文）。如下图所示。这两种方法的优化目标都 是在已知先验知识的基础上，使得预测目标值的极大似然估计值最大。为了估计极大似 然估计值，即需要计算目标词汇出现的概率，该概率的计算需要涉及词汇表中的所有词 汇。因此每次网络更新时，每一次预测都是基于全部的数据集进行的，时间开销很大。 基于此，提出了两种加快训练速度的方法，一种是负采样，一种是二叉树式的层级结构。

（1）跳字模型（skip-gram），用当前词来预测上下文。该模型定义了一个概率分布：给定一个 中心词，某个词在它上下文中出现的概率，选取词汇的向量表示，从而让概率分布值最大化。重要 的是，这个模型对于一个词汇，有且只有一个概率分布，这个概率分布就是输出，也就是出现在中 心词周围上下词的一个输出。

（2）连续词袋模型（CBOW，continuous bag of words），通过上下文来预测当前词，即基于某中心词在文本序列前后的背景词来生成该中心词。

四、 实验结果与分析

4.1 实验流程

根据上述 word2vec 原理完成给定语料库的建模，得到词语字典的词向量模型。通过欧式距离使用聚类的方式，得到某一词语最近的 10 个词语，以验证词语向量的正确性。

4.2 实验结果

训练完模型后，通过聚类的方式，完成了词语向量正确性的验证，分别得到了“令狐冲”、“屠龙刀”、“明教”、“华山”、“逃走”5 个词语最近的 10 个词语。结果如下：

令狐冲	屠龙刀	明教	华山	逃走
杨过	倚天剑	本教	恒山	溜走
岳不群	宝刀	日月神教	嵩山	逃命
胡斐	打狗棒	魔教	桃花岛	乘乱
盈盈	屠龙	首脑人物	清观	没命

天虚	短剑	光明顶	天龙	追上
凌波	至尊	五派	华山派	救兵
林平之	铁剑	首脑	论剑	追不上
小龙女	宝剑	正教	衡山	丢下
苗人凤	玄铁令	总教	五派	乘黑
岳灵珊	剑鞘	诸长老	玉女峰	冲出去

4.3 结果分析

观察上表，可以看出通过 Word2Vec 模型训练并计算出的词向量在词语的聚类上具有良好的表 现，针对人名、武器名词、教派名称、地名、动词这几个不同类型的词语都具有很好的分类效果。