深度学习与自然语言处理第三次作业

汪婧昀
19231136@buaa.edu.cn

实验任务

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。

实验原理

1. 词向量

2. Word2Vec 模型

(1) 简介

Word2vec 是一种用于生成词向量的模型，它能够将词语映射到一个连续的向量空间中，使得语义相近的词语在向量空间中的距离也相近。词向量是自然语言处理中的一种重要技术，它能够捕捉词语之间的语义和语法关系，为文本分析、情感分析、文本分类等任务提供有力支持。

(2) 原理

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）。

(3) 训练

Word2Vec 模型通过两种主要方法来训练词向量：Skip-gram 模型和 CBOW（Continuous Bag of Words）模型。

① **Skip-gram 模型**

Skip-gram 模型的基本思想是根据当前词来预测其上下文中的词。具体来说，给定一个中心词，模型会尝试预测该词前后一定范围内的词（即上下文词）。通过这种方式，模型可以学习到词语之间的共现关系，并将这些关系编码到词向量中。

在训练过程中，模型会优化一个目标函数（如负采样或层次 softmax），以最小化预测错误。通过不断地调整词向量的参数，模型能够逐渐学习到词语之间的语义关系。

② **CBOW 模型**

与 Skip-gram 模型不同，CBOW 模型是通过上下文词来预测中心词。具体来说，给定一个词的上下文（即前后一定范围内的词），模型会尝试预测该中心词本身。

CBOW 模型的训练过程与 Skip-gram 类似，也是通过优化目标函数来最小化预测错误。不同的是，CBOW 模型更注重上下文信息对中心词的影响，因此它在某些任务中可能表现出不同的性能特点。

(4) 训练过程

Word2vec 模型的训练过程可以分为以下几个步骤： 1. 构建词汇表：从训练语料中提取所有不同的词语，构建词汇表。 2. 初始化词向量：为词汇表中的每个词语随机初始化一个词向量。 3. 构建训练样本：从训练语料中构建训练样本，每个样本包含一个中心词和其上下文词。 4. 训练神经网络：使用训练样本训练神经网络模型，优化词向量。 5. 提取词向量：训练完成后，提取每个词语对应的词向量作为最终结果。

实验结果

1.《三十三剑客图》：对该小说不是很了解，无法评价实验结果

表 1 《三十三剑客图》——赵处女

赵处女	关联度
赵国	0.6682607531547546
书籍	0.6150409579277039
却说	0.574769914150238
仙姑	0.5621033906936646
结纳	0.5392612218856812
王之意	0.4820869565010071
散文	0.46370184421539307
第三人称	0.46263349056243896
幼稚	0.45496973395347595
广选	0.44337397813796997

3. 《倚天屠龙记》:下列人物都是小说中出现的人物，并与张无忌相关，比如其中张无忌和赵敏是夫妻，和周芷若是青梅竹马，朱九真是张无忌的初恋，而朱长龄是朱九真的父亲。综上，人物关系符合小说中的人物关系，因此认为建模符合小说的实际情况。

张无忌	关联度
赵敏	0.6661372184753418
周芷若	0.6592502593994141
朱长龄	0.5535851120948792
蛛儿	0.616802990436554
朱九真	0.5621329545974731
杨不悔	0.5481631755828857

4. 《笑傲江湖》：与令狐冲相关的是他的师父（岳不群）、师妹（岳灵珊）等人，都是与令狐冲有交集的人。

表 3 《笑傲江湖》——令狐冲

令狐冲	关联度
岳不群	0.6268566846847534
盈盈	0.6258386969566345
岳灵珊	0.5815714001655579
仪琳	0.5526890754699707
岳夫人	0.5427738428115845
定静师太	0.5371097326278687

5. 《天龙八部》：表中人物都是与段誉关系密切的人。

表 4 《天龙八部》——段誉

段誉	关联度
王语嫣	0.7184401154518127
木婉清	0.6720277667045593
慕容复	0.609745442867279
钟灵	0.5889674425125122
鸠摩智	0.5772355198860168
王夫人	0.5765807032585144
段正淳	0.5681746602058411
阿朱	0.535353422164917