作业 3 报告:神经语言模型验证词向量有效性

朱隽凡
1362992167@qq.com

Abstract

本报告基于作家金庸所著书籍，使用该语料库训练 Word2Vec 模型，计算词向量之间的相似度，从而验证词向量有效性。

Introduction

word2vec 不是单一算法，而是一系列模型架构和优化，它可用于能够将单词转化为向量来表示，这样就可以定量地度量词与词之间的关系。

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分布式表示（而非 one-hot 表示）的词向量。

Word2Vec 模型主要包含了 CBOW 和 Skip-gram 两种，CBOW 模型使用目标单词的上下文作为输入，目标单词作为标签；而 Skip-gram 模型则以当前单词为输入，并预测其语境（上下文中包含的词）。二者互有优劣，通常 CBOW 适合于数据集较小的情况，而 Skip-Gram 在大型语料中表现更好。

Methodology

对给定的数据集，先提取出其中的有效部分，去除其中的广告信息和停词，使用处理后的文本训练 gensim 库中的 Word2Vec 模型，最后用训练好的模型计算两个词向量的相似度。

Experimental Studies

对我阅读过的小说《笑傲江湖》中的主要人物，计算其中主要人物与其他人物的相似度。

表 1: 令狐冲相似度

令狐冲	相似度
盈盈	0.5997354984283447
岳不群	0.5069268345832825
岳灵珊	0.5043248534202576
田伯光	0.4665871560573578
仪琳	0.45994746685028076

表 2: 盈盈相似度

盈盈	相似度
令狐冲	0.5997354388237
冲哥	0.49998822808265686
任我行	0.49508756399154663
上官云	0.46697378158569336
东方不败	0.45729267597198486

表 3: 岳不群相似度

岳不群	相似度
岳夫人	0.6588831543922424
令狐冲	0.506926953792572
岳灵珊	0.4921093285083771
冲儿	0.47856879234313965
左冷禅	0.4726393520832062

表 4: 东方不败相似度

东方不败	相似度
杨莲亭	0.6703857779502869
任我行	0.5867607593536377
童百熊	0.5232657790184021
上官云	0.47815755009651184
盈盈	0.45729270577430725

从以上相似度表中可以看出，与指定人物相似度高的人物均和该指定人物关系密切，相似度排序符合小说中的实际情况，表明 word2vec 模型的训练效果很好。

Conclusions

实验表明，word2vec 模型在中文语料库上表现良好，能够很好地衡量词向量之间的相似度。