Report 4 of Deep Learning for Natural Langauge Processing

乔彪
ZY2303706

Abstract

本报告利用给定的金庸小说语料库，利用 Word2Vec 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、或者其他方法验证了词向量的有效性。

Introduction

Word2Vec 是一种计算模型，它可以依据一个词的上下文关系，将其映射到一个固定大小的向量。Word2Vec 模型的目的是捕捉给定词在其上下文中的语义和语法特征。
Word2Vec 模型有两种主要的架构：连续词袋（CBOW）和 Skip-Gram。

1) 连续词袋（CBOW）模型：CBOW 模型通过一个词的上下文（即周围的词）来预测这个词。具体来说，它取输入词的上下文中的多个词，通常是这个词周围的 n 个词，然后用这些词的向量加权求和来预测中心词的向量。
2) Skip-Gram 模型：与 CBOW 相反，Skip-Gram 模型的目的是用一个词来预测其上下文。也就是说，给定一个词作为输入，模型应该能够预测该词周围的词。Skip-Gram 在处理大型语料库时通常表现得更好，并且能够产生更具区分性的向量表示。

Word2Vec 模型的训练通常使用神经网络，并且可以通过诸如负采样的技术来加速训练过程并减少过拟合。训练完成后，这些模型可以用来获取词向量，这些向量可以用于各种任务，如文本分类、情感分析、主题建模等。

Methodology

代码流程如下所述

M1：读取所有小说并保存

该部分的作用是读取语料库中所有小说，然后进行预处理，包括去除所有的如广告、标点符号等的无用的字符，最后以字典的形式保存文本，字典的“key”为每个小说名称，“value”为每个小说中根据换行符分割后的列表。

M2:训练 Word2Vec 模型

该部分在读取上一段部分的文本字典和停顿词后，首先获得“逆”的词字典，即字典的“key”为一个小说中的词语，“value”为一个二元列表，列表中第一个元素为该词语出现过的小说名称，第二个元素为该小说的词频。然后获得将所有小说列表合并为一个列表，并将列表中的字符串分词、除去停用词，获得训练模型所需的句子列表。最后是训练 Word2Vec 模型并保存。

M3:模型数据处理

该部分执行数据处理的工作，包括对给定小说的给定词语找到其余弦相似度最大的前 10 个词、对小说中的高频词进行聚类并利用 TSNE 使聚类结果可视化以及找到指定段落余弦相似度最大的句子。

Experimental Studies

（一）余弦相似度计算列表

选择的小说有《倚天屠龙记》、《天龙八部》、《神雕侠侣》、《射雕英雄传》、《笑傲江湖》，其各自的角色为“张无忌”、“段誉”、“杨过”、“郭靖”、“令狐冲”。其各自的余弦相似度表如下列表所示。

表 1《倚天屠龙记》中“张无忌”余弦相似度表

其他关键词	所属小说	余弦相似度
周芷若	《倚天屠龙记》	0.77402
赵敏	《倚天屠龙记》	0.72874
谢逊	《倚天屠龙记》	0.66761
张翠山	《倚天屠龙记》	0.63682
蛛儿	《倚天屠龙记》	0.58456
令狐冲	《笑傲江湖》	0.57576
宋青书	《倚天屠龙记》	0.57438
金花婆婆	《倚天屠龙记》	0.55456
灭绝师太	《倚天屠龙记》	0.54857
范遥	《倚天屠龙记》	0.54599

表 2《天龙八部》中“段誉”余弦相似度表

其他关键词	所属小说	余弦相似度
虚竹	《天龙八部》	0.66897
萧峰	《天龙八部》	0.66449
木婉清	《天龙八部》	0.65606
王语嫣	《天龙八部》	0.65177
慕容复	《天龙八部》	0.65177
段正淳	《天龙八部》	0.60389
乔峰	《天龙八部》	0.58383
鸠摩智	《天龙八部》	0.58074
游坦之	《天龙八部》	0.56142
包不同	《天龙八部》	0.54266

表 3《倚天屠龙记》、《神雕侠侣》中“杨过”余弦相似度表

其他关键词	所属小说	余弦相似度
黄蓉	《倚天屠龙记》、《射雕英雄传》、《神雕侠侣》	0.73782
郭靖	《倚天屠龙记》、《射雕英雄传》、《神雕侠侣》	0.69299
小龙女	《倚天屠龙记》、《神雕侠侣》	0.68758
周伯通	《射雕英雄传》、《神雕侠侣》	0.67063
法王	《倚天屠龙记》、《神雕侠侣》	0.64193
李莫愁	《神雕侠侣》	0.63236
陆无双	《神雕侠侣》	0.62989
黄药师	《射雕英雄传》、《神雕侠侣》	0.59566
赵志敬	《神雕侠侣》	0.58910
欧阳锋	《射雕英雄传》、《神雕侠侣》	0.57316

表 4《倚天屠龙记》、《射雕英雄传》、《神雕侠侣》中“郭靖”余弦相似度表

其他关键词	所属小说	余弦相似度
黄蓉	《倚天屠龙记》、《射雕英雄传》、《神雕侠侣》	0.74727
杨过	《倚天屠龙记》、《神雕侠侣》	0.69299
欧阳锋	《射雕英雄传》、《神雕侠侣》	0.67846
洪七公	《射雕英雄传》、《神雕侠侣》	0.66360
黄药师	《射雕英雄传》、《神雕侠侣》	0.65557
柯镇恶	《射雕英雄传》、《神雕侠侣》	0.65201
欧阳克	《射雕英雄传》	0.64094
周伯通	《射雕英雄传》、《神雕侠侣》	0.62904
梅超风	《射雕英雄传》、《神雕侠侣》	0.57902
穆念慈	《射雕英雄传》、《神雕侠侣》	0.57001

表 5 《笑傲江湖》中“令狐冲”余弦相似度表

其他关键词	所属小说	余弦相似度
岳不群	《笑傲江湖》	0.73545
林平之	《笑傲江湖》	0.69134
岳夫人	《笑傲江湖》	0.62119
田伯光	《笑傲江湖》	0.61974
盈盈	《书剑恩仇录》、《侠客行》等	0.61242
左冷禅	《笑傲江湖》	0.58532
任我行	《笑傲江湖》	0.58488
张无忌	《倚天屠龙记》	0.57576
向问天	《笑傲江湖》	0.57344
仪琳	《笑傲江湖》	0.56742

由上表所示，与主角相似度较高的角色大都与主角属于同一个小说，尽管总有些例外。个人认为一个原因是这些小说同属一个作者，这些小说语言风格较为接近，如“盈盈”这一词语，不是一个人名，但在作者的好多小说里都出现了，而且频次不低，这也导致了表 5 中其排名靠前，这也许是“令狐冲”在表 1 中出现的原因。从这些表可以看出词向量总体上是有效的。

（二）聚类结果图

对 16 本小说聚类的结果如下图所述

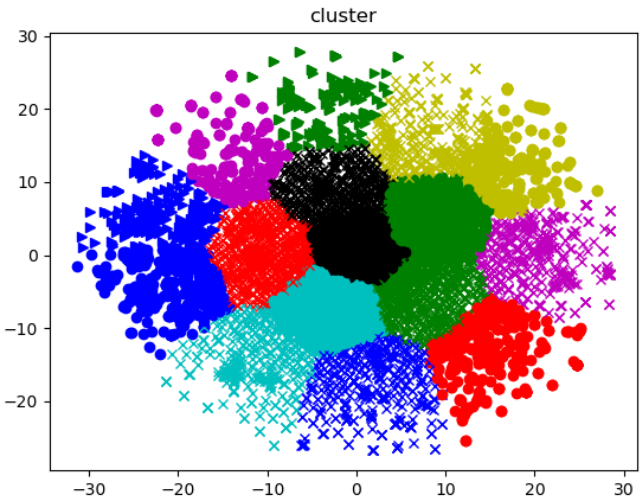

图 1 聚类结果图

这些高频词语经过聚类后分割相对明显，聚类结果总体上是成功的。

（三）指定段落余弦相似度最大的句子

找到的结果为：

['韦小宝', '一颗', '心', '怦怦', '跳', '查问', '下去', '恐怕', '师太要', '疑心', '头上']

与指定段落同属《鹿鼎记》。这说明同一个小说中的句子关联性更大。

Conclusions

从以上实验结果可以看出，Word2Vec 生成的词向量的总体是有效的。但是属于同一个作者的不同小说的相似的语言风格可能会导致词向量的准确率降低。