Deep Learning and Natural Langauge Processing

——第三次作业

卢田雨 ZY2303525
2589402656@qq.com

Abstract

本文利用给定语料库作为数据集，利用神经语言模型 Word2Vec 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。

Introduction

Word2vec 是一种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。对于给定一个长度为 T 的文本，假设时间步的词为 W（t），背景窗口大小为 m，则连续词袋模型目标函数是由背景词生成任一中心词概率：

$$\sum_{t=1}^{T} P(w^{(t)}|w^{(t-m)},\dots,w^{(t-1)},\dots,w^{(t+m)})$$

Experimental Studies

实验结果如下：

《倚天屠龙记》

表 1-1：倚天屠龙记词向量关联度

张无忌	关联度
周芷若	0.511336

张翠山	0.463939
谢逊	0.412231
赵敏	0.403387
宋青书	0.392989
金花婆婆	0.37242
朱长龄	0.364297

可以看出，小说中周芷若是张无忌的青梅竹马，张无忌与周芷若感情深厚，两者有很高的关联度，明显高于其他人。

表 1-2：倚天屠龙记词语聚类

明教	关联度
本教	0.415993
魔教	0.319323
峨嵋派	0.244295
贵教	0.243198

明教自称本教，被正派教称之为魔教，二者与明显关联度较高。

《天龙八部》：

表 2-1：天龙八部词向量关联度

段誉	关联度
萧峰	0.564757
段正淳	0.475494
慕容复	0.4580998
王语嫣	0.426567
鸠摩智	0.3934207
徐长老	0.3654359

可以看出，萧峰是段誉结拜兄弟，段正淳是段誉名义上的父亲，慕容复是段誉死对头，王语嫣是段誉的心上人，这些人与段誉关联度明显高于其他人。

表 2-2：天龙八部词语聚类

逍遥派	关联度
苏星河	0.338742

无崖子	0.332658
童姥	0.266579
虚竹	0.250159
丁春秋	0.247675

逍遥派掌门是无崖子，首徒是苏星河，虚竹后来成为逍遥派掌门，童姥是逍遥派大弟子。

《射雕英雄传》：

表 3-1：射雕英雄传词向量关联度

郭靖	关联度
黄蓉	0.754457
洪七公	0.689884
欧阳克	0.683156
欧阳锋	0.658934
裘千仞	0.606486
黄药师	0.597743

可以看出，郭靖的老婆是黄蓉，二人关系贯穿小说始终，因此二人关联度远高于其他人。

《神雕侠侣》：

表 4-1：神雕侠侣词向量关联度

杨过	关联度
小龙女	0.728308
李莫愁	0.685839
黄蓉	0.664348
郭靖	0.641764
陆无双	0.640486
法王	0.619647

可以看出，杨过的姑姑、师父是小龙女，二人感情线贯穿故事始终，二人关联度明显高于其他人。

表 4-2：神雕侠侣词语聚类

全真教	关联度
赵志敬	0.354839
丘道长	0.337257
重阳	0.334504
丘处机	0.287868
北斗	0.279528

赵志敬、丘处机、王重阳都是全真教的人，北斗指的是天罡北斗阵，是全真教的阵法。

《笑傲江湖》：

表 5-1：笑傲江湖词向量关联度

令狐冲	关联度
岳不群	0.778345
林平之	0.725633
岳灵珊	0.700197
田伯光	0.680995
仪琳	0.67232
岳夫人	0.66513

可以看出，令狐冲的师父君子剑岳不群二人的关系由情同父子到反目为仇是故事的经典情节，二人的关联度较高。

表 5-2：笑傲江湖词语聚类

华山派	关联度
青城派	0.560951
华山	0.504883
武当派	0.434742
嵩山	0.432196
岳不群	0.426482

岳不群是华山派掌门人，其他派系与华山派都是关系较好的派系或者同属于五岳派。

Conclusion

本次实验针对金庸五本小说《倚天屠龙记》、《天龙八部》、《射雕英雄传》、《神雕侠侣》、《笑傲江湖》作为样本，对主角与其他人关联度、门派关联度分别进行了分析，实验结果表明人物与门派的关联度均较高，符合小说的实际情况。