Report of Deep Learning for Natural Langauge Processing

Zhijin Cao
21231024@buaa.edu.cn

Abstract

本文探讨词向量在自然语言处理中的应用，利用Word2Vec生成词向量并对其性能进行验证。Word2Vec通过密集型向量表示单词，有效捕捉语义关系，采用CBOW或Skip-Gram模型计算词向量。实验以16本金庸小说为语料，预处理后训练Word2Vec模型，保存模型并验证效果。结果表明，同一小说中的人物名向量在聚类中常被归为同类，相似度分析也显示同小说人物向量更接近，验证了词向量模型的有效性。

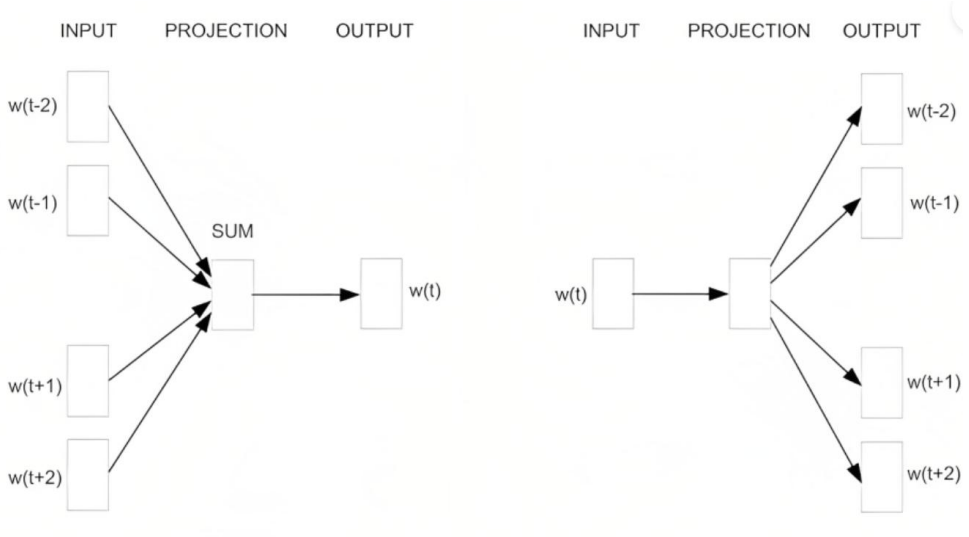
Introduction

Word2Vec

Word2vec是一种自然语言处理（NLP）技术，与传统的稀疏型向量表示（如词袋模型）相比，Word2vec通过密集型向量表示单词，其生成的向量维度通常在50到1000之间，远低于传统的TF-IDF向量的维度，能够更有效地捕捉单词之间的语义关系。

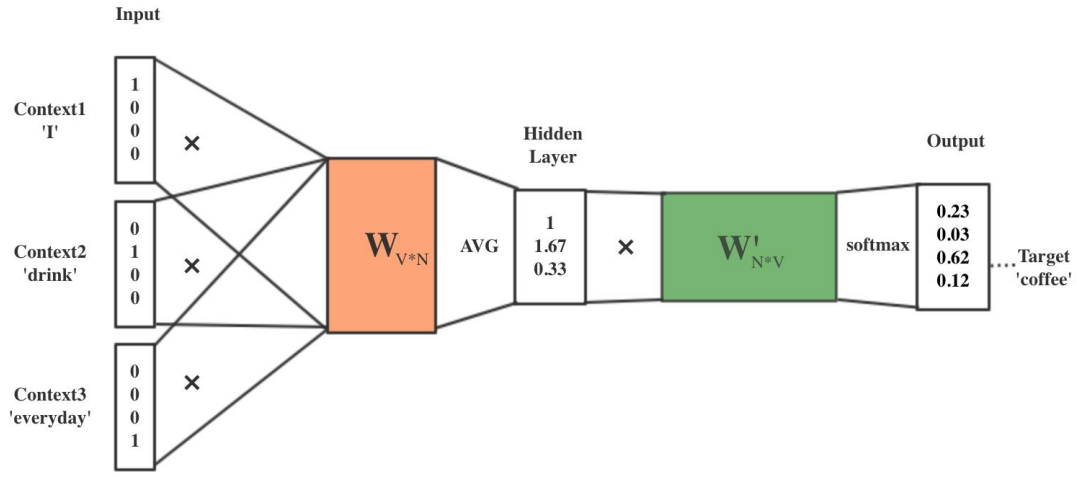
Word2vec技术的核心思想是利用单词的上下文本来推断单词的语义。它基于一个假设：在相似上下文中出现的单词往往具有相似的语义。Word2vec通过训练一个模型，学习单词的向量表示，使得在向量空间中相近的单词在语义上也相近。

word2vec模式下计算词语embeddings主要有两个模型，为CBOW和Skip-Gram。



CBOW　　Skip-Gram

CBOW模型



CBOW 模型训练的基本步骤包括：

1.将上下文词进行 one-hot 表征作为模型的输入，其中词汇表的维度为 V，上下文单词数量为C ；

2.然后将所有上下文词汇的 one-hot 向量分别乘以输入层到隐层的权重矩阵W；

3.将上一步得到的各个向量相加取平均作为隐藏层向量；

4.将隐藏层向量乘以隐藏层到输出层的权重矩阵W’；

5.将计算得到的向量做 softmax 激活处理得到 V 维的概率分布，取概率最大的索引作为预测的目标词

Methodology

实验思路

本文将语料库内容进行预处理后输入Word2Vet模型进行训练，使用CBOW方法进行词向量的构建，保存模型。对几本小说中的人物进行相似向量及最不相似向量的输出、对人物姓名即name.txt文件中的全部人物名进行聚类验证模型效果。

实验数据

该实验中文数据集为16本金庸小说，具体内容见data，其中data/inf.txt中存储了16本小说的小说名，其余txt文件文件名为小说名且文件内容为小说内容。../cn_stopwords为中文停用词，../name.txt为划分聚类用到的人物名，其中全部的人名分别来自5本小说。

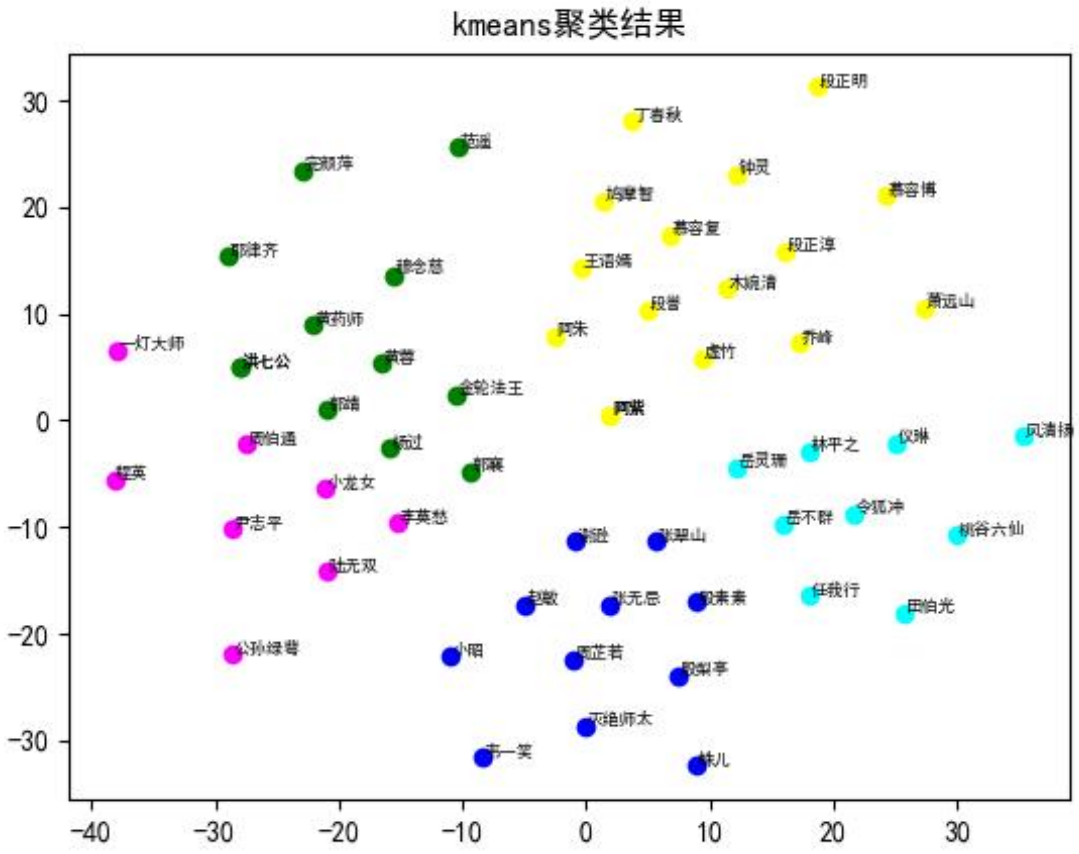
代码解析

该实验的主函数在get_model.py中，其中get_file_data获取file_path下文件内容并进行预处理，get_all_data获取数据库中的全部内容，主函数训练Word2Vet模型并保存在model1.model中。加载文档数据get_file_data部分，其内容包括将文本数据按路径加载，去除无用内容，按模式对数据进行拆分，将拆分后的数据进行筛选并去除停词。

对于该实验的验证，model_similarity_word.py对相似向量及最不相似向量进行验证，而model_cluster.py对聚类的结果进行验证。

Experimental Studies and Conclusions

运行代码，在运行model_cluster.py后，可以查看人名的聚类分布图，具体如下图所示



由上图结果中，可以看到，聚类基本上按照出现的小说所分布，即同一本书中出现的人物大都聚集在同一处。也有分错类的情况，不过基本上都是次要人物及在多本书中出现的人物才有这一问题。这也有可能与为了可视化，将多维词向量转换为二维造成的信息损失有关。总体来讲聚类效果较好，可说明词向量的有效性。

在运行model_similarity_word.py后，可以得到五个人物名字向量最相似的向量，如下表所示

Table 1: 与“郭靖”一词排名前十的相似词语及其相似度

郭靖	相似词语	相似度
1	黄蓉	0.733151
2	杨过	0.720621
3	欧阳锋	0.718854
4	欧阳克	0.695841
5	黄药师	0.672324
6	洪七公	0.669443
7	周伯通	0.654899
8	裘千仞	0.642117
9	穆念慈	0.639153
10	小龙女	0.630885

Table 2：与“杨过”一词排名前十的相似词语及其相似度

杨过	相似词语	相似度
1	小龙女	0.760509
2	郭靖	0.720621
3	黄蓉	0.712926
4	赵志敬	0.693857
5	法王	0.681658
6	李莫愁	0.669833
7	陆无双	0.669290
8	周伯通	0.645465
9	金轮王	0.625326
10	郭襄	0.623921

Table 3：与“段誉”一词排名前十的相似词语及其相似度

段誉	相似词语	相似度
1	王语嫣	0.700067

2	虚竹	0.693405
3	木婉清	0.688108
4	萧峰	0.678785
5	慕容复	0.668726
6	阿朱	0.636431
7	游坦之	0.629142
8	段正淳	0.626
9	鸠摩智	0.596280
10	阿紫	0.594929

Table 4：与“令狐冲”一词排名前十的相似词语及其相似度

令狐冲	相似词语	相似度
1	岳不群	0.763781
2	林平之	0.728944
3	岳夫人	0.699991
4	仪琳	0.657496
5	岳灵珊	0.653711
6	田伯光	0.653065
7	盈盈	0.650812
8	任我行	0.627017
9	张无忌	0.609888
10	石破天	0.596379

Table 5：与“张无忌”一词排名前十的相似词语及其相似度

张无忌	相似词语	相似度
1	周芷若	0.786402
2	赵敏	0.726930
3	谢逊	0.701524

4	张翠山	0.700218
5	灭绝师太	0.641186
6	殷素素	0.636601
7	金花婆婆	0.628115
8	令狐冲	0.609888
9	宋青书	0.577496
10	蛛儿	0.574701

从上面的结果可以看出，出现在同一部小说中、关系紧密、属于同一派别的人物名字向量往往具有更高的相似性，这说明了名字在词向量中的独特性，也说明了词向量模型的有效性。

在运行model_similarity_word.py后，还可以得到"令狐冲 林平之 岳灵珊 岳不群 宁中则 陆大有 左冷禅"这7个人物名字向量中与其他最不相似的向量，为“左冷禅”。这是合理的，除“左冷禅”外，其他人名均属于华山派，说明同一门派的名字词语见具有更高的相似性，侧面说明了词向量模型的有效性。

References

[1] https://blog.csdn.net/weixin_42663984/article/details/116739799