单位代码_____
学　　号 ZY2303803
分 类 号_____

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理（NLP）第三次课后作业

院（系）名称	自动化科学与电气工程学院
专 业 名 称	自动化
学 生 姓 名	董晨辉

2024 年 05 月

深度学习与自然语言处理（NLP）第一次课后作业

董晨辉
1127666815@qq.com

Abstract

利用给定语料库金庸语小说集，利用 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 是一种由 Google 在 2013 年提出的用于自然语言处理 (NLP) 的词嵌入技术。其主要目标是将词语表示成向量，使得具有相似语义的词在向量空间中相互靠近。通过训练神经网络模型，Word2Vec 从大量的文本数据中学习词语之间的关系和语义，从而实现这一目标。

Word2Vec 的核心概念包括词嵌入、模型架构和训练方法。词嵌入是将词语转换为实数向量的过程，这些向量捕捉了词语的语义特征，使得语义相似的词具有相近的向量表示。Word2Vec 有两种主要的模型架构：连续词袋模型（Continuous Bag of Words, CBOW）和跳字模型（Skip-Gram）。CBOW 通过上下文词预测中心词，而 Skip-Gram 通过中心词预测上下文词。训练过程中，Word2Vec 使用大量文本数据进行无监督学习，通过最小化预测误差，不断调整词向量，使得具有相似语境的词语向量越来越接近。为了提高训练效率，Word2Vec 引入了负采样（Negative Sampling）和分层 Softmax（Hierarchical Softmax）技术，这些技术可以显著减少计算量。

Word2Vec 生成的词向量可以用于多种应用。例如，在词语相似度计算中，Word2Vec 可以计算词语之间的相似度，"king" 和 "queen" 的相似度会很高，因为它们的向量表示非常接近。在文本分类和聚类任务中，词向量有助于更有效地进行情感分析、主题建模等。在信息检索和推荐系统中，Word2Vec 能够帮助改进结果的相关性和推荐的准确性。

虽然 Word2Vec 有很多优点，比如能够有效捕捉词语的语义关系和较高的计算效率，但也存在一些局限性。其静态表示意味着每个词的向量是固定的，无法处理词语在不同上下文中的多义性。此外，由于频率较低的词语训练不足，这些词的向量质量可能不高。

Methodology

具体来说，分类流程有以下步骤：

1. 准备语料库：选择一个中文文本作为实验的语料库。在本报告中，将 16 部金庸小说进行合并，得到完整的文本。
2. 文本预处理：对语料库进行预处理。删除所有的隐含符号、非中文字符和标点符号。对文本进行分词并过滤停用词。

3. 训练模型：使用 gensim 库的 Word2Vec 模型对处理后的语料库进行训练。
4. 保存和加载 Word2Vec 模型：将训练好的模型保存到文件，并在需要时加载模型。
5. 词语相似度计算：计算两个词语之间的相似度得分与语义距离。
6. KMeans 聚类与 t-SNE 降维与可视化：对词向量进行 KMeans 聚类，并计算聚类的轮廓系数。使用 t-SNE 对指定簇的词向量进行降维，并绘制散点图。

Experimental Studies

Experimental 1

实验训练(generate_model.py)条件如下：

$$model = Word2Vec(sentences = sentences, vector\_size = 200, min\_count = 10, window = 5, sg = 1, workers = 4, epoch = 20)$$

在词向最维度为 200、训练轮次为 20 的条件下，进行相似度实验(test1.py)。相似度越接近 1 说明词向量对相似度越高，实验结果如下表:

词向量对	语义距离
杨过、小龙女	0.832574725151062
峨嵋派、武当派	0.542256236076355
东方不败、韦小宝	0.20744144916534424

可以得出以下结论：

(1) 杨过和小龙女的词向量相似度高：相似度达到了 0.83 左右，说明在该训练条件下，这两个词语的向量在语义上非常接近。这与金庸小说中杨过和小龙女的关系密切相关，因此词向量捕捉到了他们之间的语义联系。

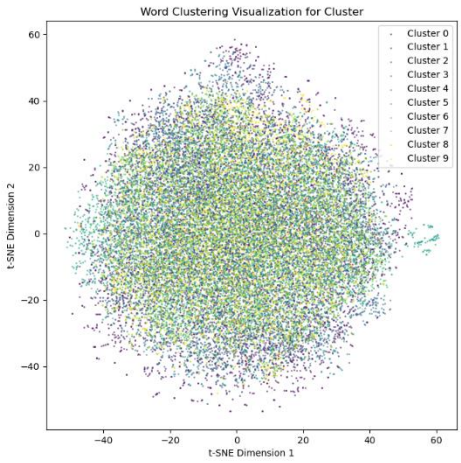
(2) 峨嵋派和武当派的词向量相似度适中：相似度约为 0.54，表示它们在一定程度上具有一些相似之处，但并不十分相近。这可能反映了这两个门派在金庸小说中的关系，既有合作又有竞争，所以它们的词向量在语义上有一定的重叠，但并不完全相同。

(3) 东方不败和韦小宝的词向量相似度较低：相似度仅为 0.21，说明它们在语义上差异较大，词向量表示的语义距离较远。这与金庸小说中东方不败和韦小宝的角色性质和所属势力有关，它们之间的联系并不密切。

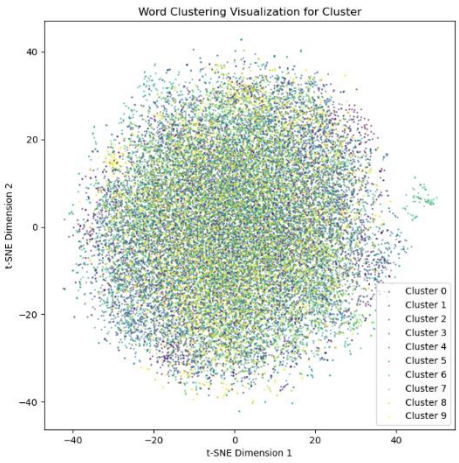
综合来看，在给定的训练条件下，Word2Vec 模型能够捕捉到金庸小说中人物和门派之间的一些语义联系，但对于某些关系较远或较复杂的词语，词向量的相似度可能不够高。

Experimental 2

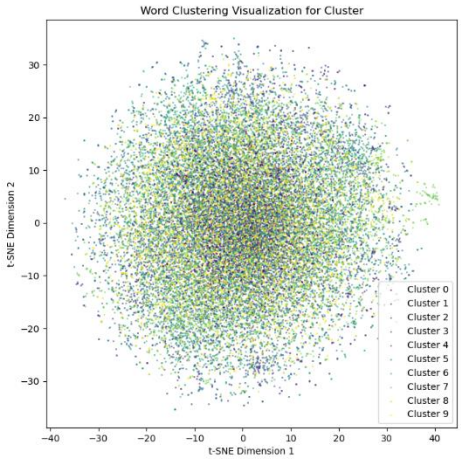
本实验使用 K-means 聚类算法随向量进行聚类分析(test2.py)，得到以下结果：



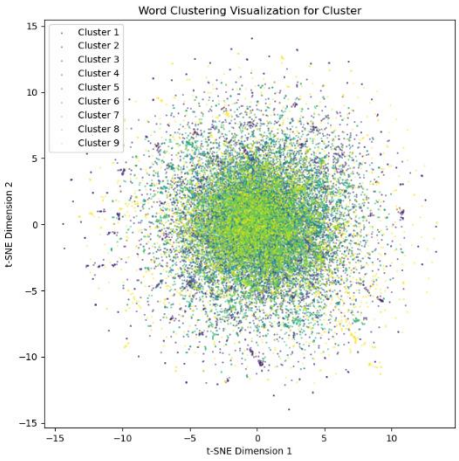
Epoch = 5



Epoch = 10



Epoch = 20



Epoch = 50

KMeans 聚类降维后的结果

从上图可以看出，Epoch = 20 时效果较好，过大会导致过拟合。

下面是部分聚类中的词汇：

武功，剑，功夫，一招，均，剑法，厉害，敌人，内力，出手，少林，欧阳锋，高手，非，适才，处，洪七公，实，极，不及，黄药师，无法，难以，竟然，高，成，未，内功，掌，之际，颇，看来，乃是，招数，练，皆，逼，华山，功力，指点，重伤，全，只须，轻功，无比，鸠摩智，尽数，生平，神功，对手，法，攻，梅超风，手法，招，左冷禅，寻常，岂知，星宿，敌，斗，生死，一身，好手，胜，学，力气，拚丁春秋，施展，不同，稍，馀，全是，刀法，不易，着实，破，占，本门，体内，双方，全然，武学，拆，石壁，非同小可，金轮法王，相斗，擒拿，从未，毕竟，抵挡，一派，毒，运气，使出，尚有，指，高强

Experimental 3

计算段落之间的语意关联的步骤如下：

1.段落向量化:首先，将个段落表示为一个向量。可以使用类似 Word2Vec 的技术，将段落中的每个词向量化，并取平均值或加权平均值作为段落向量。

2.计算段落间相似度：使用向量表示的段落，可以通过计算它们之间的相似度来衡量它

们的语义关联性。常用的方法包括余弦相似度、欧氏距离、曼哈顿距离等。

3.相似度阀值：根据实际需求，可以设定一个相似度值，超过这个值的段落视为语义相关，否则视为不相关。

下面从语料库随机选择：

一：两人在乡间躲了三日，听得四乡饥民聚众要攻漳州、厦门。这一来，只将张朝唐吓得满腔雄心，登化乌有，眼见危邦不可居，还是急速回家的为是。其时厦门已不能再去，主仆两人一商量，决定从陆路西赴广州，再乘海船出洋。两人买了两匹坐骑，胆战心惊，沿路打听，向广东而去。幸喜一路无事，经南靖、平和，来到三河坝，已是广东省境，再过梅县、水口，向西迤逦行来。张朝唐素闻广东是富庶之地，但沿途所见，尽是饥民，心想中华地大物博，百姓人人生死系于一线，渤泥只是海外小邦，男女老幼却是安居乐业，无忧无虑，不由得大是叹息，心想中国山川雄奇，眼见者百未得一，但如此朝不保夕，还是去渤泥椰子树下唱歌睡觉安乐得多了。这一日行经鸿图嶂，山道崎岖，天色渐晚，他心中焦急起来，催马急奔。一口气奔出十多里地，到了一个小市镇上，主仆两人大喜，想找个客店借宿，哪知道市镇上静悄悄的一个人影也无。张康下马，走到一家挂着“粤东客栈”招牌的客店之外，高声叫道：“喂，店家，店家！”店房靠山，山谷响应，只听见“喂，店家，店家”的回声，店里却毫无动静。正在这时，一阵北风吹来，猎猎作响，两人都感毛骨悚然。张朝唐拔出佩剑，闯进店去，只见院子内地下倒着两具尸首，流了一大滩黑血，苍蝇绕着尸首乱飞。腐臭扑鼻，看来死者已死去多日。张康一声大叫，转身逃出店去。张朝唐四下一瞧，到处箱笼散乱，门窗残破，似经盗匪洗劫。张康见主人不出来，一步一顿的又回进店去。张朝唐道：“到别处看看。”哪知又去了三家店铺，家家都是如此。有的女尸身子赤裸，显是曾遭强暴而后被杀。一座市镇之中，到处阴风惨惨，尸臭阵阵。两人再也不敢停留，急忙上马向西。主仆两人行了十几里，天色全黑，又饿又怕，正狼狈间，张康忽道：“公子，你瞧！”张朝唐顺着他手指看去，只见远处有一点火光，喜道：“咱们借宿去。”

二：　杨过请得周伯通来和瑛姑团聚，令慈恩安心而死，又取得灵狐，一番 辛劳，连做三件好事，自是十分高兴，和郭襄、神雕一齐回到万兽山 庄。 史氏兄弟见杨过连得两头灵狐，喜感无已，当即割狐腿取血。史叔 刚服后，自行运功疗伤。 是晚万兽山庄大排筵席，公推杨过上座，席上所陈，尽是猩唇、狼腿、 熊掌、鹿胎等诸般珍异兽肉，旁人一生从未尝得一味的，这一晚筵席 中却有数十味之多。席旁放了一只大盘，盛满山珍，供神雕侠享用。 史氏兄弟和西山一窟鬼对杨过也不再说甚么感恩戴德之言，各人心中 明白，自己性命乃杨过所赐，日后不论他有甚么差遣，万死不辞。席 上各人高谈阔论，说的都是江湖上的奇闻轶事。郭襄自和杨过相见以来，一直兴高采烈，但这时却默默无言，静听各 人的说话。杨过偶尔向她望了一眼，但见她脸上微带困色，只道小姑 娘连日奔波劳碌，不免疲倦，也不以为意，那想到郭襄因和他分手在 即，良会无多，因而悄悄发愁。 喝了几巡酒，突然间外面树林中一只猿猴高声啼了起来，跟着此应彼 和，数十只猿猴齐声啼鸣。史氏兄弟微微变色。史孟捷道：“杨大哥 和西山诸兄且请安坐，小弟出去瞧瞧。”说着匆匆出厅。 各人均知林中来了强敌，但眼前有这许多好手聚集，再强的敌人也不 足惧。煞神鬼道：“最好是那霍都王子到来，大伙儿跟他斗斗，也好 让史三哥出了这口恶气……

段落间的语义相似度：0.8206448554992676

虽然两者属于不同的文章，但是其相似度依然很高。模型的词向量训练效果较好。

Conclusions

通过 Word2Vec 实验，我们成功训练了一个具有较高语义准确性的词向量模型。实验结果显示，词语之间的相似度计算得分反映了它们在语义上的联系。特别是对于那些在文本中具有密切关联的词语。

总的来说，我们的实验表明 Word2Vec 模型在捕捉词语语义关系方面具有良好的效果，为进一步的文本分析和应用提供了可靠的基础。

References

https://blog.csdn.net/shuihupo/article/details/85162237
https://blog.csdn.net/weixin_44966965/article/details/124732760