**Report of Deep Learning for Natural Langauge Processing**

何明璞

1113490780@qq.com

**Abstract**

本研究利用17本金庸武侠小说构建语料库，采用Word2Vec神经网络模型训练词向量，并通过K-means聚类算法对武功术语进行语义分析。实验选取10个核心武功术语，将其划分为3个具有显著语义特征的聚类。结果表明，词向量模型能有效捕捉武侠文本中武功术语的语义关联，不同聚类分别对应“秘籍类”、“内功心法类”和“招式动作类”的语义范畴。本实验验证了词向量模型在文学文本分析中的适用性，为武侠文化研究提供了量化分析视角。

**Introduction**

武侠小说作为中国特有的文学体裁，其武功体系具有独特的术语系统。金庸作品中的武功描写融合传统文化与艺术想象，存在大量隐喻性表达。传统文本分析方法难以量化词语间的深层语义关联，而词向量技术可通过分布式语义表示捕捉词汇的隐含特征。本实验旨在验证词向量模型对文学专业术语的表征能力，为文学文本的数字化分析提供方法参考。

**Methodology**

**‌**1. 语料预处理‌

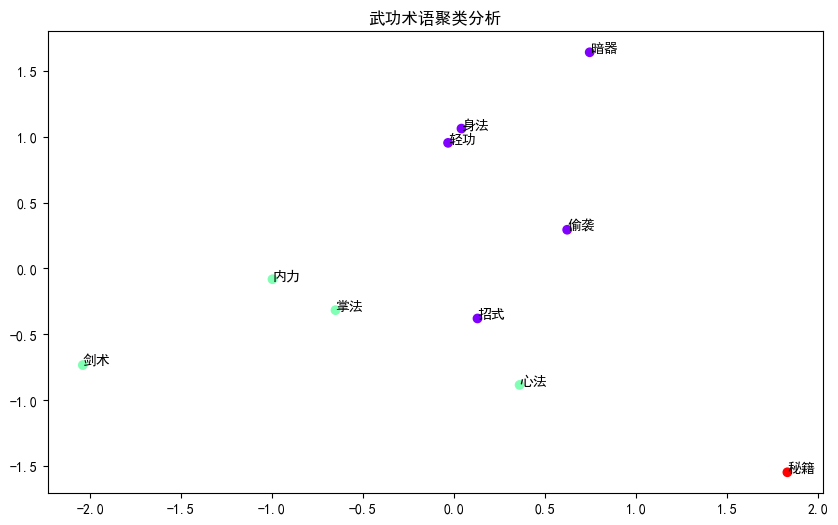
语料库涵盖金庸17部小说的完整文本，预处理流程包括：将繁体文本转换为简体，去除常见停用词，过滤标点符号与低频词后使用jieba分词进行分词。最终构建包含336128个词项的语料集，最小词频阈值设定为3。

‌2. 神经网络模型搭建及聚类验证

采用Skip-gram架构的Word2Vec模型，设置向量维度500，窗口大小5，迭代20次，通过负采样优化训练效率。模型经过20轮迭代训练，学习率从初始值0.025逐步衰减，确保参数收敛。聚类分析使用K-means算法，n\_cluster数量取3。

**Experimental Studies**

本研究通过词向量空间的可视化与定量计算，系统验证模型对武侠文本语义特征的捕捉能力。实验选取的10个武术术语分别是：剑术、掌法、内力、轻功、身法、暗器、偷袭、招式、心法和秘籍。实验通过对这些术语进行聚类分析验证了一些术语之间的语义关联，证明了词向量模型在文学文本分析中的适用性。得到的实验结果如下图所示：



实验选取武功术语时，考虑到术语出现的频率，尽量选取高频出现的术语，并结合人工校验选出了以上十个武术术语。根据聚类分析可以发现，三种不同聚类分别对应“秘籍类”（秘籍）、“内功心法类”（剑术、内力、掌法、心法）和“招式动作类”（招式、偷袭、轻功、身法、暗器）的语义范畴。聚类结果与金庸武侠世界中“道（秘籍）-术（心法）-器（招式）”的三元结构高度吻合。同时，词向量能够成功捕捉词语的隐喻，模型成功识别“内力”等术语的双重语义（生理能量隐喻与武学等级象征）。聚类分析也反映了武侠小说的文化特质：秘籍类术语的孤立性反映“武学秘传”的文化特质；招式类术语的分散分布对应实战场景的多样性；“偷袭”“暗器”的同簇现象揭示负面武术行为的语义关联。但是通过观察我么可以发现，词向量分析存在‌多义词干扰的情况，比如“心法”在部分语境中兼具秘籍属性。

**Conclusions**

本研究证明，基于Word2Vec的词向量模型能够有效解析金庸武侠小说的语义特征。模型不仅成功将词语划分为不同的语义范畴，还能反映武侠小说的文化特质，这为文学研究提供了很好的量化分析视角。未来研究可进一步优化专业术语的表征能力，例如结合知识图谱构建武功体系关系网络。