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

武学谦

[sy2303705@buaa.edu.cn](mailto:sy2303705@buaa.edu.cn)

**Abstract**

本研究基于金庸武侠小说全集17部作品构建语料库，采用Word2Vec模型训练词向量，探索武侠文本的语义表征能力。实验验证模型成功捕捉武学术语间的关联性，例如“剑术”与“掌法”的语义相似度达0.900，同时量化师徒关系词对的语义距离为0.698。结果表明，词向量技术可有效解析武侠文本中的领域特异性语义关联，为传统文化文本分析提供技术支撑。

**Introduction**

武侠文学中的武学术语和人物关系具有鲜明的领域特征，传统文本分析方法难以量化其语义关联。本研究以金庸小说为语料，通过Word2Vec模型构建词向量空间，旨在解决两个核心问题：第一，词向量能否准确表征“剑术”“掌法”等专业术语的语义相似性；第二，师徒关系等特殊人际关系在向量空间中是否呈现合理距离分布。实验结果将为武侠文本数字化处理提供方法论参考。

**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，去除常见停用词后进行分词，过滤标点符号与低频词。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.随机森林分类器

采用Skip-gram架构的Word2Vec模型，设置词向量维度为300，上下文窗口大小为3，通过负采样优化训练效率。模型经过15轮迭代训练，学习率从初始值0.025逐步衰减，确保参数收敛。

**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。实验聚焦武学术语关联性、人物关系建模与文化符号表征三个维度，揭示词向量技术在传统文化文本分析中的应用价值。得到的实验结果如下图所示：



以核心武学概念“剑术”为分析对象，模型生成的词向量空间清晰呈现武侠文本特有的语义网络。在语义相似度排名中，“掌法”以0.900的余弦相似度位居首位，这一发现与《神雕侠侣》中独孤求败“草木竹石皆可为剑”的武学境界形成文本互证，说明模型成功识别不同武学载体间的内在关联。位列第二的“武艺”达到0.896相似度，体现动作概念的上位范畴聚合效应，而“招数”0.895的相似值则反映具体技术层面的共性特征。值得注意的是，“达摩”以0.894相似度进入近邻列表，佐证模型对《天龙八部》少林武学体系的语义捕获能力，表明词向量能自发建立文化符号与武学流派的知识联结。

师徒关系词对“师父-徒弟”的语义相似度量化结果为0.698，该数值分布具有双重解释力。相较于测试集中随机词对0.183的平均相似度，师徒词对展现出显著语义关联；但与“兄弟-姐妹”0.812、“夫妻-伴侣”0.784等对称关系相比，其较低相似度暗含武侠叙事中师徒间的权力梯度。进一步分析“师父”的语义近邻，“指点”“传授”等动词以超过0.70的相似度密集分布，揭示该角色在文本中的主动教化特征；而“徒弟”则与“学艺”“服侍”等从属行为强关联，这种非对称向量分布精准映射《笑傲江湖》中岳不群传授紫霞神功与令狐冲恪守门规的叙事逻辑。

**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。模型不仅准确捕捉“剑术-掌法”等武学术语的关联性，还能量化师徒关系的语义距离。未来研究可进一步优化专业术语的表征能力，例如通过领域自适应训练提升低频武学术语的向量质量。本成果为武侠文学数字化分析与传统文化计算提供了可扩展的技术路径。‌