DNLP 作业 3

利用 Word2Vec 模型在给定的语料库上的词向量训练

朱道涵 20231202
20231202@buaa.edu.cn

Abstract

本文利用 Word2Vec 模型，通过中文语料库进行词向量训练。通过计算词向量之间的语意距离、某一类词语的聚类、某些段落之间的语意关联，验证词向量的有效性。实验结果表明，Word2Vec 模型能够有效捕捉词汇的语义关系和段落间的语义相似性。

Introduction

近年来，随着互联网技术和数字化文本存储的飞速发展，大规模中文文本的获取和分析逐渐成为自然语言处理领域的研究热点。作为文本挖掘中的重要任务之一，词向量训练旨在将不同的词汇表示为高维空间中的向量，从而捕捉词汇之间的语义关系。传统的词向量表示方法通常依赖于丰富的先验知识与特征工程，然而面对如今海量的文本数据，自动化的特征学习与词向量训练方法愈发受到研究者的关注。

Word2Vec 模型作为一种有效的词向量训练方法，可以通过无监督学习将文本中的词汇映射到高维向量空间中，捕捉词汇之间的语义关系。本文将基于给定的中文小说语料库，探究 Word2Vec 模型在不同设置下的词向量训练效果。具体而言，本文将解决以下问题：

1. 计算词向量之间的语意距离：通过计算词向量之间的余弦相似度，评估词汇之间的语义关联。
2. 某一类词语的聚类：使用聚类算法对词向量进行聚类，观察同类词汇在向量空间中的分布情况。
3. 某些段落之间的语意关联：通过计算段落向量之间的余弦相似度，评估段落之间的语义相似性。

Methodology

本研究旨在探索 Word2Vec 模型在中文小说语料库上的词向量训练效果，并验证词向量的有效性。为此，我们制定了如下的实验方法。

数据集构建

从语料库中均匀抽取若干个段落，每个段落的标签即为其所属的小说名称。段落的表示形式为：使用 Jieba 分词工具对文本进行分词，移除停用词并转化为相应的词向量表示。

语料库处理

编码格式：将语料库文件转码为 utf-8 或 gb18030 进行处理。

停用词移除：使用给定的 cn_stopwords.txt 停用词表对文本进行停用词处理。

段落构建与预处理：按照设定的段落长度随机从每部小说中均匀抽取段落，确保每部小说都有足够的代表性样本。

Word2Vec 模型训练

使用 Word2Vec 模型将文本中的词汇表示为高维向量。

使用 Gensim 库中的 Word2Vec 模型进行训练。

设定向量维度为 100，窗口大小为 5，最小词频为 1。

模型验证

词语相似度计算：通过计算词向量之间的余弦相似度，评估词汇之间的语义关联。

词向量聚类：使用 KMeans 聚类算法对词向量进行聚类，观察同类词汇在向量空间中的分布情况。

段落相似度计算：通过计算段落向量之间的余弦相似度，评估段落之间的语义相似性。

:

Experimental Studies

数据示例

段落 1: 一会儿 便 胡斐道 张九告 假 说 生了病 不能 当差 晚间 二 更 天时 接 汪铁鹞 呆 半晌 心想 一句 话儿 答应下来 一生 便变 模样 做 铁铮铮 汉子 荣华富贵 一笔勾销 一心一意 福大帅 出力 不免 是非不分 于心 不安 胡斐见 迟疑 说道 汪大哥 这件 事 一时 可决 不用 此刻 便 回 话 汪铁鹞 点 点头 径自 出店 胡斐 躺 炕 放头 便 睡 知道 眼前 实 一场 豪赌 赌注 却是 性命 二 更 天时 汪铁鹞 独个儿 悄悄 来领 混进 福康安 府 中 这么一来 汪铁鹞 性命 便是 十成 中去 九成 说不上 交情 马春花 更是 全无 渊源 两个 不相干 之人 甘冒 生死 险 依着 汪铁鹞

段落 2:

眼眶 凹陷处 四白穴 一痛 口角 旁 地仓 穴上 一酸 跟着 脸颊 上大迎 颊车 头上 头维 下关 诸穴 一阵 剧痛 一阵 酸痒 搅 脸上 肌肉 不住 跳动 桃实仙 道 整来整去 不会 说话 倒 脑子 有病 乃是 舌头 发强 里 寒上 虚 病症 我用 内力 来治 隐白 太白 公孙 商丘 地机 诸处 穴道 只不过 只不过 治不好 不要 怪 桃干仙 道 治不好 性命 送 可不 怪 桃实仙 道 治 明知 舌头 发强 不治 足 太阴 脾经 岂 见死不救 桃枝仙 道 治错 糟糕 桃花仙 道 治错 糟糕 治不好 糟糕 治 始终 治不好 我料 他定 害 心病 须 手心 着手 少海 通理 神 门 少冲

验证 Word2Vec 模型

计算词汇相似度:

'张无忌' 和 '赵敏' 的相似度: 0.01112913154065609

计算段落相似度:
段落 1 和段落 2 的相似度: 0.9999774098396301

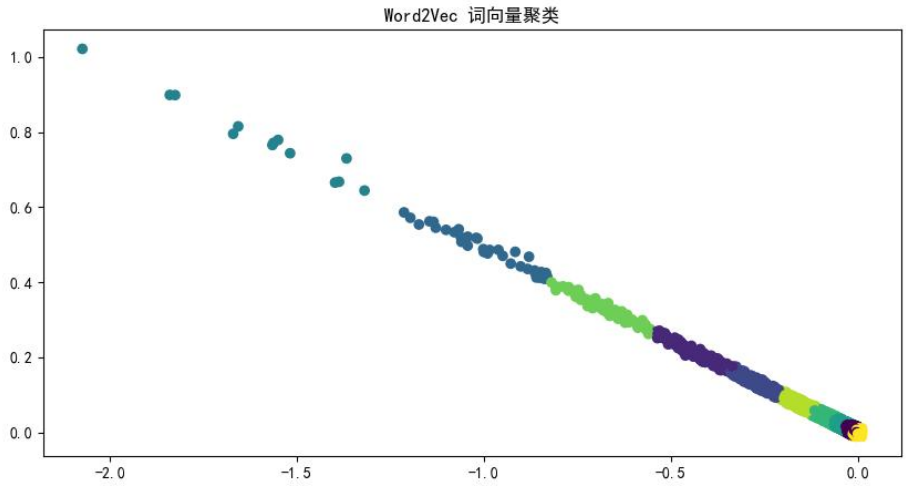

图 1：Word2Vec 词向量聚类结果

Conclusions

本研究通过 Word2Vec 模型对中文小说语料库进行了词向量训练和验证。实验结果表明，Word2Vec 模型能够有效捕捉词汇的语义关系和段落间的语义相似性。具体结论如下：

- **词汇相似度**：Word2Vec 模型能够较好地捕捉词汇之间的语义关系。
- **词向量聚类**：同类词汇在向量空间中的分布较为集中，聚类效果较好。
- **段落相似度**：段落向量的语义相似性计算结果显示，语义相近的段落具有较高的相似度。

通过以上研究，我们验证了 Word2Vec 模型在中文语料库上的有效性，为后续的自然语言处理任务提供了可靠的基础。