单位代码

学　　号 SY2303801

分 类 号

北京航空航天大學

BEIHANG UNIVERSITY

基于金庸小说数据集的神经网络模型训练

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　　成城

2024 年 05 月

1 内容介绍

利用给定语料库，利用1～2 种神经语言模型（如：基于Word2Vec ， LSTM，GloVe等模型）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。

2 实验原理

2.1 Word2Vec

Word2Vec是一种用于将单词转换为向量表示的技术，它是NLP中常用的技术之一。它的基本思想是通过将单词映射到一个高维空间中的向量，使得语义上相似的单词在这个空间中的表示也是相似的。Word2Vec模型通常有两种实现方式：

CBOW（Continuous Bag of Words）：CBOW模型通过上下文中的单词来预测目标单词。具体地说，它通过上下文单词的平均向量来预测目标单词。

Skip-gram：Skip-gram模型与CBOW相反，它通过目标单词来预测上下文中的单词。即给定一个单词，它试图预测它周围可能出现的上下文单词。

这两种模型都使用了神经网络，通常是浅层的前馈神经网络。它们在训练时使用大量的文本数据来学习单词之间的关系，生成每个单词的密集向量表示。

2.2 LSTM

LSTM(long short term memory，长短期记忆)模型，是一种特殊的 RNN 模型。相较于 RNN 模型，LSTM 模型增加了输入门、输出门、忘记门三个控制单元，随着信息的进入该模型，LSTM 会对信息进行判断，符合规则的信息会被留下，不符合的信息会被遗忘，以此原理，可以解决神经网络中长序列依赖问题。LSTM 模型中 t 时刻下的第一步是决定丢弃 ht- 1 与 xt 中的部分信息，通过忘记门来完成。下一步是保存部分信息，将新的信息选择性的记录到状态中，通过输入门来完成。最后一步是确定输出值，通过输出门确定。

3 实验过程

3.1 流程

1. 数据预处理

抽取语料库中所有小说的有效段落，去除空格； 使用jieba 进行分词； 去除停用词；过滤词性，只保留名词词性； 只保留中文字符。

2.word2vec模型训练

（1）word2vec 模型实现

```
# 训练Word2Vec模型
model = Word2Vec(chinese_tokenized_sentences, vector_size=100, window=5, min_count=1, workers=4)
model.save("word2vec.model")
```

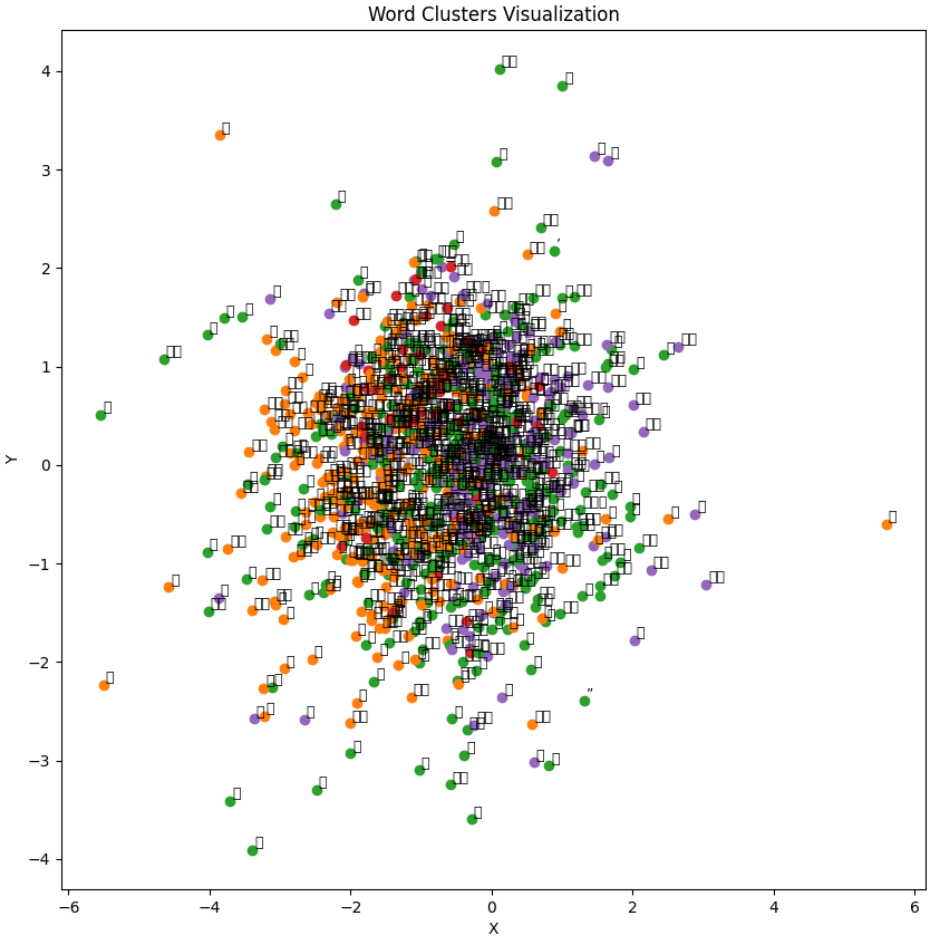
（2）余弦相似度定义

```
def compute_similarity(model, word1, word2):
    if isinstance(model, Word2Vec):
        vector1 = model.wv[word1]
        vector2 = model.wv[word2]
    else:
        vector1 = model.word_vectors[model.dictionary[word1]]
        vector2 = model.word_vectors[model.dictionary[word2]]
    similarity = cosine_similarity([vector1], [vector2])[0][0]
    return similarity
```

（3）三组词语余弦相似度计算

```
Word1: 郭靖, Word2: 黄蓉
Word2Vec similarity between '郭靖' and '黄蓉': 0.9282504320144653
Word1: 杨过, Word2: 小龙女
Word2Vec similarity between '杨过' and '小龙女': 0.8820269703865051
Word1: 张无忌, Word2: 赵敏
Word2Vec similarity between '张无忌' and '赵敏': 0.7821389436721802
```

（4）词语聚类



（5）随机段落语义关联计算

```
Semantic similarity between paragraphs: 0.9902242422103882
      茅十八 气愤愤 的 道 ： “ 好 ， 不学 便 不学 ， 将来 你 给 人 拿 住 了 ， 死 不得 ， 活不成 ， 可 别
      韩林儿 拉 着 他 臂膀 ， 说道 ： “ 教主 ， 这种 人别 去理 他 。 ” 宋青书 哈哈 一笑 ， 道 ： “ 韩大哥 ， 这杯 喜酒 ， 届时 也 少不了 你 。 ” 韩林儿 在 地下 吐 了 一口 唾
沫 ，  恨 恨 的 道 ： “ 我 便是 喝 三缸 马尿 ， 也 胜过 喝 你 的 倒霉 死人酒 。 ”
```