深度学习与自然语言处理第三次作业

ZY2314222 魏智兴

Abstract:

本作业依据Word2vec模型，通过采用金庸小说集语料库，并辅以jieba库进行分词处理，成功训练出中文单词的词向量表示。为验证所训练词向量的有效性，我们进行了系列实验。首先，通过计算特定词对之间的语意距离，我们发现相关词对的语意距离相对较小，而无关词语意距离则较大，这初步证明了模型训练的词向量在语义表示上的有效性。其次，我们进一步从文本语料库中筛选出与目标词最为相似的词汇，这一结果再次印证了词向量在语义捕捉方面的准确性。再者，我们选取文本语料库中的部分内容进行K-means聚类分析，通过聚类结果的合理性，进一步验证了词向量的有效性。最后，我们分别提取了相同小说和不同语料文本中的语段，并计算其词向量的余弦相似度。结果表明，同一小说内的语段间语意距离较小，而不同语料间的语意距离则较大，这充分证明了词向量在区分不同语义内容上的有效性。综上所述，本作业所训练的词向量在语义表示、语义捕捉以及语义区分方面均展现出了较高的有效性。

Introduction

一、Word2vec 模型

Word2Vec 是一种用于训练词向量的模型，由Google 研究团队里的Tomas Mikolov 等人于2013年提出。它的基本思想是利用上下文相似的词其词向量也应相似的原则，通过训练得到每个词的向量表示。Word2Vec模型主要包括两种训练方法：Skip-gram 和Continuous Bag of Words (CBOW)。Skip-gram模型通过当前词来预测上下文词汇，而CBOW模型则通过上下文词汇来预测当前词。这两种方法都采用了负采样和层次softmax等优化技术来加速训练过程。

Word2Vec模型的特点包括：

- 它是一种浅层神经网络模型，主要包含输入层、隐藏层和输出层。
- 输入层接收词的one-hot编码或稀疏编码作为输入。
- 隐藏层对输入进行变换，得到词的向量表示。
- 输出层根据隐藏层的输出预测下一个词或生成当前词的上下文词汇的概率分布。

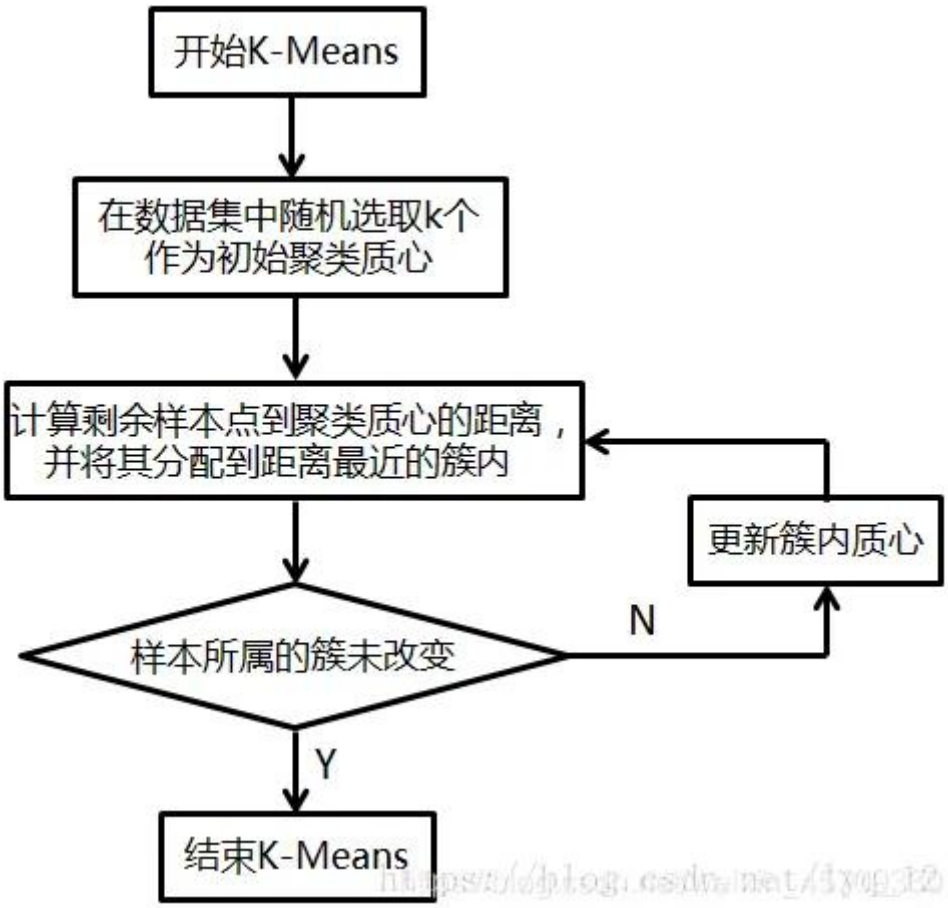
Word2Vec模型的优点包括:

- 训练速度快，能够处理大规模的语料库。
- 得到的词向量能够捕捉到词的语义信息，使得语义相似的词在向量空间中相近。
- 词向量的维度较低，便于计算和存储。

然而，Word2Vec也存在一些局限性，例如它是一种静态的词向量表示方法，无法处理一词多义的情况。此外，随着更先进的模型如BERT的出现，Word2Vec在许多NLP任务上的效果已经不是最佳选择。尽管如此，Word2Vec仍然是自然语言处理领域中非常基础和重要的模型之一。

二、 K-means 聚类

k均值聚类算法（k-means clustering algorithm）是一种迭代求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。



Methodology

利用给定语料库（金庸语小说料如下链接），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。

- 首先，通过计算特定词对之间的语意距离，我们发现相关词对的语意距离相对较小，而无关词语意距离则较大，这初步证明了模型训练的词向量在语义表示上的有效性。
- 其次，进一步从文本语料库中筛选出与目标词最为相似的词汇，这一结果再次印证了词向量在语义捕捉方面的准确性。
- 再者，选取文本语料库中的部分内容进行K-means聚类分析，通过聚类结果的合理性，进一步验证了词向量的有效性。

- 最后，分别提取了相同小说和不同语料文本中的语段，并计算其词向量的余弦相似度。

Experimental Studies

（1）计算词向量之间的语意距离

词语对	相似度
张无忌/周芷若	0.9694
大象/老虎	0.6539
静夜/清风	0.5795

与“张无忌”相似度最高的词：

张无忌	
周芷若	0.9694
张翠山	0.9682
赵敏	0.9603
殷素素	0.9525
谢逊	0.9443
'金花婆婆	0.9310

（2）K-means聚类分析，验证了词向量的有效性

选择训练好Word2vec的模型，选择如下词类进行聚类分析。

words = ['张无忌','赵敏','周芷若','张翠山','殷素素','谢逊',

'静夜','梨花','雪地','茅草','雪地','清风',

'武功','兵器','兵刃','武当山','宝刀','功夫']

得到的结果如下：

```
聚类 0: ['武功', '兵刃', '宝刀', '功夫']
聚类 1: ['张无忌', '赵敏', '周芷若', '张翠山', '殷素素', '谢逊']
聚类 2: ['静夜', '梨花', '雪地', '茅草', '雪地', '清风', '兵器', '武当山']
```

可见训练的模型很好的进行了聚类效果，有效验证了词向量的有效性。

（3）提取了相同小说和不同语料中的语段，并计算其词向量的余弦相似度

选取相同小说：

paragraph1 = "张无忌携了谢逊之手，正要并肩走开。谢逊忽道：“且慢！”指着少林僧众中的一名老僧叫道：“成昆！你站出来，当着天下众英雄之前，将诸般前因后果分说明白。群雄吃了一惊，只见这老僧弓腰曲背，形容猥琐，相貌与成昆截然不同"

paragraph2 = "张无忌见周芷若委顿在地，脸上尽是沮丧失意之情，心下大是不忍，当即上前解开她穴道，扶她起身。周芷若一挥手，推开他手臂，径自跃回峨嵋群弟子之间。只听谢逊朗声说道：今日之事，全自成昆与我二人身上所起，种种恩怨纠缠，须当由我二人了结"

不同语料：

paragraph1 = "五年来，我们坚持加强党的全面领导和党中央集中统一领导，全力推进全面建成小康社会进程，完整、准确、全面贯彻新发展理念，着力推动高质量发展，主动构建新发展格局，蹄疾步稳推进改革，扎实推进全过程人民民主，全面推进依法治国，积极发展社会主义先进文化，突出保障和改善民生，集中力量实施脱贫攻坚战，大力推进生态文明建设，坚决维护国家安全，防范化解重大风险，保持社会大局稳定，大力度推进国防和军队现代化建设"

paragraph2 = "同志们！十八大召开至今已经十年了。十年来，我们经历了对党和人民事业具有重大现实意义和深远历史意义的三件大事：一是迎来中国共产党成立一百周年，二是中国特色社会主义进入新时代，三是完成脱贫攻坚、全面建成小康社会的历史任务，实现第一个百年奋斗目标。这是中国共产党和中国人民团结奋斗赢得的历史性胜利，是彪炳中华民族发展史册的历史性胜利，也是对世界具有深远影响的历史性胜利"

相同小说余弦相似度	不同语料文本余弦相似度
0.9540	0.8540

可见对不同语料文本余弦相似度低于相同小说余弦相似度，验证了词向量的有效性。

Conclusions

通过本次作业，充分证明了词向量在区分不同语义内容上的有效性。综上所述，本作业所训练的词向量在语义表示、语义捕捉以及语义区分方面均展现出了较高的有效性。