成 绩 _____

北京航空航天大學

BEIHANG UNIVERSITY

深度学习与自然语言处理

第四次大作业

院（系）名称	自动化科学与电气工程学院
专 业 名 称	模式识别
学 号	SY2103130
姓 名	殷健凯
指 导 教 师	秦曾昌

2022 年 5 月 19 日

1 问题描述

利用给定语料库（或者自选语料库），利用神经语言模型 Word2Vec 来训练词向量，通过对词向量的聚类或者其他方法来验证词向量的有效性。

2 背景知识

传统的自然语言处理将词看作是一个个孤立的符号，这样的处理方式对于系统处理不同的词语没有提供有用的信息。词映射(word embedding)实现了将一个不可量化的单词映射到一个实数向量。Word embedding 能够表示出文档中单词的语义和与其他单词的相似性等关系。它已经被广泛应用在了推荐系统和文本分类中。Word2Vec 模型则是 Word embedding 中广泛应用的模型。Word2Vec 使用一层神经网络将 one-hot（独热编码）形式的词向量映射到分布式形式的词向量。使用了 Hierarchical softmax，negative sampling 等技巧进行训练速度上的优化。

2.1 词的 one-hot 表示

使用 one-hot 表示可以使得词向量生成方式简单、生成速度快。如以下三个句子：

How are you?

Fine, thanks. And you?

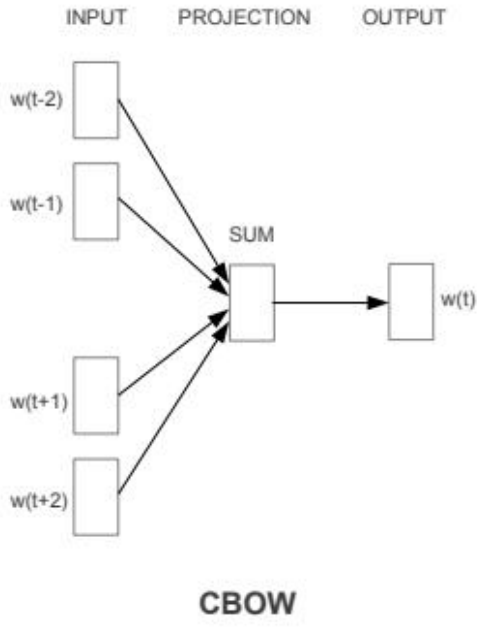
I am fine, too.

使用 one-hot 表示为：

词	编码	One-Hot表示
am	0	[1, 0, 0, 0, 0, 0, 0, 0, 0]
and	1	[0, 1, 0, 0, 0, 0, 0, 0, 0]
are	2	[0, 0, 1, 0, 0, 0, 0, 0, 0]
fine	3	[0, 0, 0, 1, 0, 0, 0, 0, 0]
how	4	[0, 0, 0, 0, 1, 0, 0, 0, 0]
i	5	[0, 0, 0, 0, 0, 1, 0, 0, 0]
thanks	6	[0, 0, 0, 0, 0, 0, 1, 0, 0]
too	7	[0, 0, 0, 0, 0, 0, 0, 1, 0]
you	8	[0, 0, 0, 0, 0, 0, 0, 0, 1]

2.2 词的分布式表示

传统的独热表示仅仅将词符号化，不包含任何语义信息。如何将语义融入到词表示中？Harris 在 1954 年提出的“分布假说”为这一设想提供了理论基础：上下文相似的词，其语义也相似。Firth 在 1957 年对分布假说进行了进一步阐述和明确：词的语义由其上下文决定。Word Embedding 正是这样的模型，而 Word2Vec 则是其中的一个典型，Word2Vec 包含两种模型，即 CBOW 模型（如图 1）和 Skip-gram（如图 2）模型。以 CBOW 模型为例，如果有一个句子“the cat sits one the mat”，在训练的时候，将“the cat sits one the”作为输入，预测出最后一个词是“mat”。



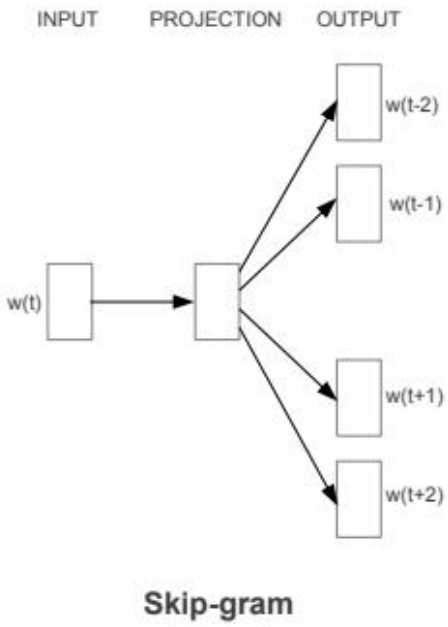


图 1　CBOW 模型示意图　　图 2　Skip-gram 模型示意图

3 实验过程

3.1 数据的预处理

为了使得数据处理更加充分，考虑了三个步骤：①语料库的读取②停词以及无用词的剔除③将金庸小说人物名、门派名称、招式名称加入 jieba 词库中进行分词。将结果保存在 data.txt 文件中。

3.2 训练过程

分别对 CBOW 模型和 Skip-gram 模型的 word2vec 进行训练，其词向量特征维度设置为 200，滑动窗口长度设置为 5，训练 10 个 epoch。

3.3 实验部分

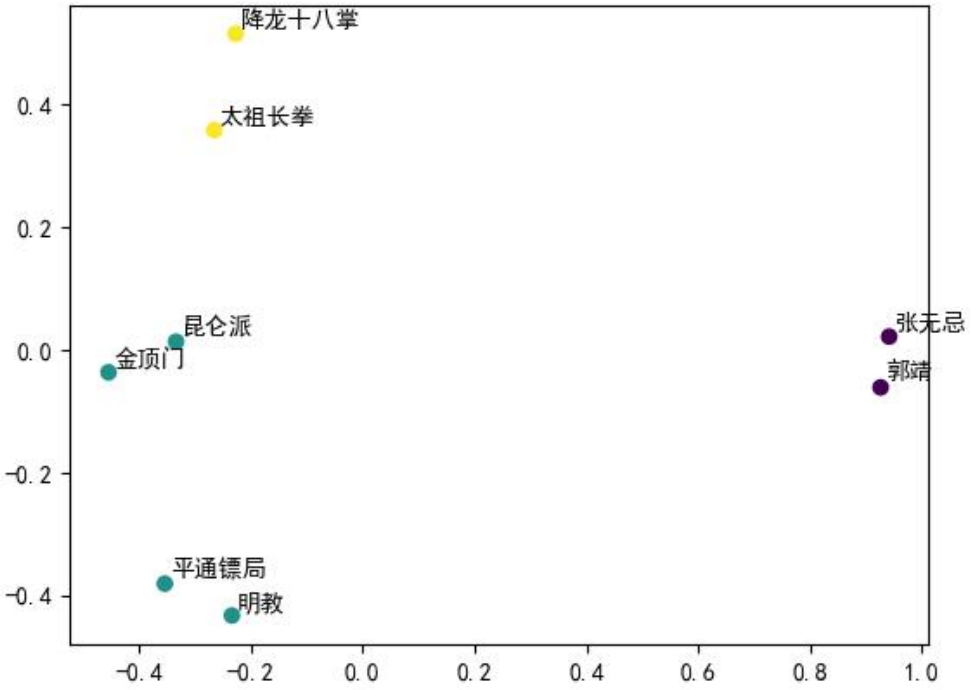
3.3.1 词语相关度展示

模型训练后，分别读取训练好的 CBOW 模型和 Skip-gram 模型，然后指定某一个词，展示与该词最相关的 5 个词。指定的词包括：明教、昆仑派、金顶门、平通镖局、张无忌、殷天正、郭靖、范百龄、降龙十八掌、太祖长拳、寒冰真气、金刚指。结果如下所示。

	1	2	3
明教	本教 0.909	五岳剑派 0.873	武当 0.869
昆仑派	崆峒派 0.890	泰山派 0.888	韦陀门 0.864
金顶门	素为 0.914	青字九打 0.914	剑侠 0.913
张无忌	张翠山 0.649	赵敏 0.647	杨逍 0.608
殷天正	韦一笑 0.759	唐文亮 0.734	杨逍 0.730
郭靖	黄蓉 0.700	杨过 0.626	欧阳锋 0.576
降龙十八掌	胡家刀法 0.852	七十二路 0.844	横扫千军 0.843
太祖长拳	新练 0.924	美女拳法 0.922	六式 0.922
寒冰真气	附有 0.910	逆冲 0.910	积储 0.905

3.3.2 Kmeans 聚类

选择["明教", "昆仑派", "金顶门", "平通镖局", "张无忌", "郭靖", "降龙十八掌", "太祖长拳"]8 个词汇，其中包含门派名、人名、招式名称，聚成 3 类，结果如下：



黄色为招式名称，绿色为门派，紫色为人物名称。