成 绩 _____

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第 4 次作业

基于金庸小说用 Word2Vec 训练词向量

院（系）名称	自动化科学与电气工程学院
专 业 名 称	自动化
学 号	SY2103106
姓 名	段晓玥
指 导 教 师	秦曾昌

2022 年 5 月 19 日

一、任务描述

利用给定语料库（或者自选语料库），利用神经语言模型（如：Word2Vec，GloVe 等模型）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。

二、实验原理

计算机无法看懂人类的自然语言，也无法对其进行直接处理，因此自然语言处理第一步需要将自然文字转换为计算机能够“看懂”的数字，因此需要将自然文字进行编码，转换由数字组成的词向量。

1. 词向量的表示形式

1）One-hot 编码

One-hot 编码是一种简单的词向量表示形式。具体地，用很长的一个向量来表示一个词，其中向量的长度为词典的大小；向量的所有分量只有一个为“1”，其余全为“0”，为“1”的位置为该词在词典中的字典序。例如，当总词典中只有“我”“爱”“北航”三个词时，三个词语可以分别编码为“100”、“010”、“001”。

One-hot 编码简单易表示。然而，词典总次数很大时，one-hot 向量会很长，容易造成维度灾难；此外，one-hot 编码将每个词看成独立的个体，忽略了词与词之间的联系（如语义相近的词、语义相反的词等）。

如果将自然语言的每一个词映射成一个固定长度的短向量，将所有这些向量放在一起形成一个词向量空间，而每一向量则为该空间中的一个点，在这个空间上引入“距离”，则可以根据词之间的距离来判断它们之间的语义相似性。这便是词语的分布式表示（distributed representation）。

2）分布式表示（Distributed Representation）

上述所说，将不同的自然词语映射为词向量空间中的不同向量的做法，便是词的分布式表示。分布式表示引入了“距离”的概念从而可以衡量词与词之间的联系，这对建模自然语言的语义信息大有裨益；此外，有一个多维的词向量而非只包含 0、1 的 one-hot 向量对词语建模，使得向量可以包含更为丰富的语义信息。

Word2Vec 采用的是分布式表示的词向量。

2. Word2Vec

作为轻量级神经网络，Word2Vec 模型包括输入层、隐藏层和输出层，主要

分为 CBOW 和 Skip-gram 两种模型。其中，CBOW 模型在已知上下文$[\omega_{t-k}, \omega_{t-k+1}, \ldots, \omega_{t-1}, \omega_{t+1}, \ldots, \omega_{t+k-1}, \omega_{t+k}]$的情况下预测当前词$\omega_t$，其中滑动窗口的大小为$2k+1$；Skip-gram 模型在已知当前词$\omega_t$的情况下对上下文$[\omega_{t-k}, \omega_{t-k+1}, \ldots, \omega_{t-1}, \omega_{t+1}, \ldots, \omega_{t+k-1}, \omega_{t+k}]$进行预测。CBOW 模型和 Skip-gram 模型分别如图 1 和图 2 所示。

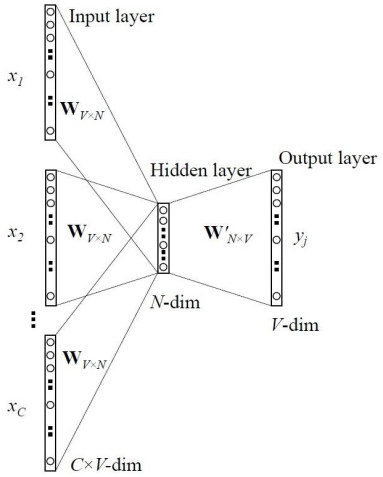

图 1：CBOW 模型示意图

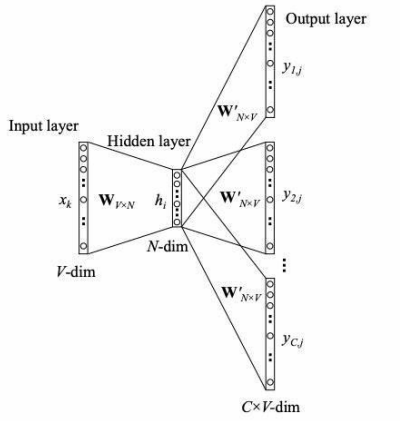

图 2：Skip-gram 模型示意图

从图 1 和图 2 可以看出，两种模型均是通过权重矩阵先将输入层的上下文或当前词的词向量映射至隐藏层向量，再用另一个权重矩阵将隐藏层向量映射至输出层词向量从而得到预测词语结果。

本实验采取两种模型进行训练。

三、实验内容与结果

1. 实验内容

本次实验同样采用金庸 16 本小说数据集进行 Word2Vec 模型训练。首先对

16 本小说所有语料进行读取并分词，将分词后的语料存入“./corpus.txt”中。注意分词前将“./Character_names.txt”（对应小说中出现的所有人名）、“./Kongfu_names.txt”（对应小说中出现的所有功夫名）、“./Sect_names.txt”（对应小说中出现的所有门派名）加入了 jieba 分词字典中，从而使其能够对于这些专有名词进行正确分词。

然后从“./corpus.txt”中读取所有语料，并进行 Word2Vec 的模型训练。训练过程中，CBOW 模型和 Skip-gram 模型均被训练，其词向量特征维度设置为 200，滑动窗口长度设置为 5。如下所示。

```
### training model
print("Training model...")
sentences = LineSentence('./corpus.txt')
model_cbow = models.word2vec.Word2Vec(sentences, sg=0, vector_size=200, window=5, min_count=5, workers=8)
model_cbow.save("./model_cbow.model")
model_skip_gram = models.word2vec.Word2Vec(sentences, sg=1, vector_size=200, window=5, min_count=5, workers=8)
model_skip_gram.save("./model_skip_gram.model")
```

2. 实验结果

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：黄蓉、杨过、张无忌、令狐冲、韦小宝、峨嵋派、屠龙刀、蛤蟆功、葵花宝典。其涵盖了人名、门派名、武器名、功夫名、重要物品名等。结果如下所示。

```
Results of CBOW:
Related words of 黄蓉:  [('郭靖', 0.8779825568199158), ('杨过', 0.8754432201385498), ('岳灵珊', 0.8455778360366821), ('胡斐', 0.8365615010261536), ('陆无双', 0.8222733736038208)]
Related words of 杨过:  [('黄蓉', 0.875443160533905), ('郭靖', 0.8626238703727722), ('小龙女', 0.8602283596992493), ('张无忌', 0.8487511277198792), ('胡斐', 0.8335386514663696)]
Related words of 张无忌:  [('令狐冲', 0.9240444302558899), ('张翠山', 0.8647434115409851), ('胡斐', 0.8573045134544373), ('石破天', 0.8493605256080627), ('杨过', 0.8487510681152344)]
Related words of 令狐冲:  [('张无忌', 0.9240444302558899), ('虚竹', 0.8393120169639587), ('胡斐', 0.8302900195121765), ('张翠山', 0.821201503276825), ('石破天', 0.8140281438827515)]
Related words of 韦小宝:  [('袁承志', 0.7096408605575562), ('令狐冲', 0.6980299949645996), ('康熙', 0.6850093603134155), ('张无忌', 0.6728357672691345), ('郭襄', 0.6574874520301819)]
Related words of 峨嵋派:  [('嵩山派', 0.910498321056366), ('华山派', 0.9066218733787537), ('泰山派', 0.899715904211426), ('武当派', 0.8979759216308594), ('青城派', 0.871666431427002)]
Related words of 屠龙刀:  [('宝刀', 0.8046663999557495), ('倚天剑', 0.7689284682273865), ('宝剑', 0.7675191164016724), ('打狗棒', 0.7629969120025635), ('铜牌', 0.740441620349884)]
Related words of 蛤蟆功:  [('玄冥神掌', 0.8628107905387878), ('空明拳', 0.8469604253768921), ('分筋错骨手', 0.8464518189430237), ('独孤九剑', 0.8456084132194519), ('挪移', 0.8451282382011414)]
Related words of 葵花宝典:  [('宝典', 0.8788228631019592), ('气宗', 0.8735796213150024), ('传下来', 0.8702307343482971), ('真经', 0.86048424243927), ('玉箫剑法', 0.85880446434021)]
```

图 3：CBOW 模型词语相关度展示

```
Results of Skip Gram:
Related words of 黄蓉:  [('郭靖', 0.6922549605369568), ('杨过', 0.6373581886291504), ('欧阳锋', 0.6218291521072388), ('洪七公', 0.611042320728302), ('郭芙', 0.5973083972930908)]
Related words of 杨过:  [('小龙女', 0.638999342918396), ('黄蓉', 0.6373582482337952), ('郭靖', 0.6055399179458618), ('郭襄', 0.5941260457038879), ('瑛姑', 0.5781541466712952)]
Related words of 张无忌:  [('张翠山', 0.6823683977127075), ('赵敏', 0.6352235078811646), ('令狐冲', 0.6224539279937744), ('杨逍', 0.6118969917297363), ('周芷若', 0.5987467765808105)]
Related words of 令狐冲:  [('张无忌', 0.6224539279937744), ('林平之', 0.6022224426269531), ('盈盈', 0.6006620526313782), ('乔峰', 0.5800546407699585), ('岳不群', 0.5766204595565796)]
Related words of 韦小宝:  [('康熙', 0.6554309725761414), ('索额图', 0.6300613284111023), ('施琅', 0.5976264476776123), ('康亲王', 0.583598792552948), ('多隆', 0.5786787271499634)]
Related words of 峨嵋派:  [('武当派', 0.7460161447525024), ('衡山派', 0.7425493597984314), ('泰山派', 0.7405927777290344), ('仙都派', 0.725755512714386), ('青海', 0.7254650592803955)]
Related words of 屠龙刀:  [('倚天剑', 0.7963321805000305), ('宝刀', 0.7563222646713257), ('屠龙', 0.7522217035293579), ('打狗棒', 0.7088170647621155), ('宝剑', 0.687846302986145)]
Related words of 蛤蟆功:  [('天山六阳掌', 0.8585114479064941), ('玄冥神掌', 0.8487555384635925), ('一阳指', 0.8467021584510803), ('天山折梅手', 0.8385077118873596), ('寒冰绵掌', 0.8366087079048157)]
Related words of 葵花宝典:  [('宝典', 0.8841336965560913), ('九阳真经', 0.8592892289161682), ('下卷', 0.8395484089851379), ('至高无上', 0.8224305510520935), ('秘要', 0.820737898349762)]
```

图 4：Skip-gram 模型词语相关度展示

从词语性质上来看，两个模型均能够准确地捕捉到词语的类型或性质。不论是 CBOW 模型还是 Skip-gram 模型，与人名最相关的词均为人名，例如 CBOW 结果与“黄蓉”最相关的“郭靖”“杨过”“岳灵珊”“胡斐”“陆无双”均为人名；与门派

名（“峨嵋派”）相关的词均为门派名称；与功夫名（“蛤蟆功”）相关的词也为功夫名称（如“玄冥神掌”“一阳指”等）；与物品名（如“葵花宝典”）相关的词均为物品（如“九阳真经”“玉箫剑法”等）。

从人物关系上来看，Skip-gram 的结果优于 CBOW 模型。例如，CBOW 模型中，与“杨过”最相关的人物为“黄蓉”，然而两个人并非出现在同一部小说中，这可能是因为“黄蓉”也为主人公，且在训练语料库中多次出现，因此会被判定为与“杨过”相关性较强的人物。而 Skip-gram 模型结果中，人名最相关的词多与该人具有紧密联系。例如，“黄蓉”最相关的词“郭靖”为其爱人；“杨过”最相关的词为“小龙女”，两人为师徒关系；“张无忌”最相关的词“张翠山”为其生父等。

可以看出，CBOW 模型，更注重文章整体的关联，很多相似度高的词跨越了多个时代、多本小说，而 skip-gram 模型更加注重局部，相似度高的词基本集中在同一本小说中。

2）tSNE 可视化聚类结果

我们从 16 本小说组成的语料库的所有词语中，选取出现次数总共超过 50 次的词作为高频词汇，并根据“stopwords”文件夹中的停用词表，过滤掉这些高频词中的停用词。

对于所有剩余的高频词，通过训练好的 skip-gram 模型得到其词向量。然后采用 K-means 聚类方法对这些词向量进行聚类，并利用 tSNE 方法将聚类结果进行可视化。聚类类别数设置为 16。可视化结果如图 5 所示。

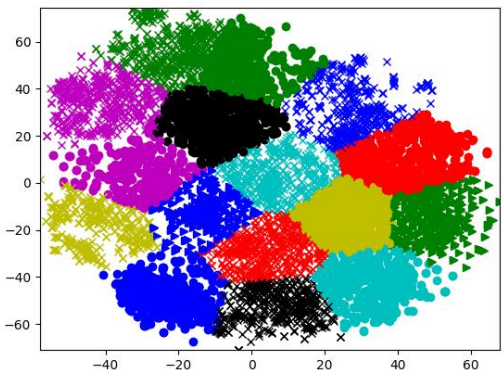

图 5：聚类结果可视化

从聚类结果可以看出，来自 16 本小说的词向量再使用 k-means 方法迭代后成功聚类，从而说明 Word2Vec 模型生成的词向量能够正确表示词语与词语之间的语义关系，生成合适的词向量。