基于 Word2Vec 语言模型的词向量有效性验证

ZY2303118　Yisong Wang
2740416657@qq.com

Abstract

本次作业要求为：利用给定金庸小说语料库，选择 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。综合考虑各种因素，本次实验最终选择采用 Word2Vec 神经语言模型训练词向量，并通过对小说人物相关词向量的 K-means 聚类验证词向量的有效性。

Introduction

2.1 词向量

在自然语言处理中最细粒度的是词语，词语组成句子，句子再组成段落、篇章、文档，因此最先要处理词语，需要将自然语言交给机器学习中的算法来处理。词语是人类的抽象总结，是符号形式的（比如中文、英文、拉丁文等等），而机器只能接受数值型输入，所以需要把词语转换成数值形式。词向量就是用来将语言中的词进行数学化的一种方式，顾名思义，词向量就是把一个词表示成一个向量。从概念上讲，它涉及从每个单词一维的空间到具有更低维度的连续向量空间的数学嵌入；从目的上讲，其旨在基于语言数据的大样本中的分布属性来量化和分类语言项之间的语义相似性。

2.2 Word2Vec

Word2Vec 是 Google 研究团队里的 Tomas Mikolov 等人于 2013 年的《Distributed Representations of Words and Phrases and their Compositionality》以及后续的《Efficient Estimation of Word Representations in Vector Space》两篇文章中提出的一种高效训练词向量的模型，基本出发点是上下文相似的两个词，它们的词向量也应该相似，比如香蕉和梨在句子中可能经常出现在相同的上下文中，因此这两个词的表示向量应该就比较相似。[1]因为 Word2Vec 的最终目的不是为了得到一个语言模型，也不是要把f训练得多么完美，而是只关心模型训练完后的副产物：模型参数(这里特指神经网络的权重)，并将这些参数作为输入 x 的某种向量化的表示，这个向量便叫做——词向量。

在 Word2Vec 模型中比较重要的概念是词汇的上下文，如对于词汇 w_t 的范围为 1 的上下文就是 w_{t-1} 和 w_{t+1}，在此基础上 Word2Vex 模式下有两个模型：CBOW 和 SkipGram(图 1):

- CBOW(Continuous Bag-of-Word)：以上下文词汇预测当前词，即用 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ 去预测 w_t
- SkipGram：以当前词预测其上下文词汇，即用 w_t 去预测 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$

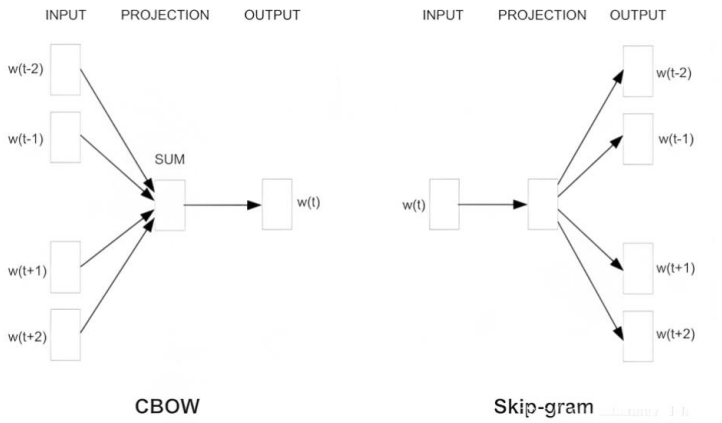


图 1 Word2Vec 下的两种模型

Methodology

依据上文理论调研，本次实验同作业一、二一致，首先对 16 本金庸所著武侠小说进行数据预处理(即分词及剔除特殊词)，采用模型 CBOW 训练词向量(即 Python gensim 库中的 Word2Vec, sg=0)，随后选择 16 本小说中的代表性人物，分析训练后与该人物相关性最强的 10 个词，最后为进一步验证模型的有效性，使用 TSNE 将训练得到的模型词向量进行降维，并使用 K-means 算法进行聚类，采用散点图进行效果展示。具体参数如下所示：

- min_count = 10，可以对字典做截断，如果词频少于 10 的单词会被丢弃；
- window = 5，表示当前词与预测词在一个句子中的最大距离是 5；
- vector_size=200，指特征向量的维度，大的 size 需要更多的训练数据，但是效果更好；
- sg=0，用于设置训练算法，其中 0 对应 CBOW 算法，1 代表 skip-gram 算法；
- workers=16，线程数；
- epochs=50，训练迭代轮数，过大会导致训练时间过长；
- n = 6，即最终 K-means 展示的聚类个数。

Experimental Studies

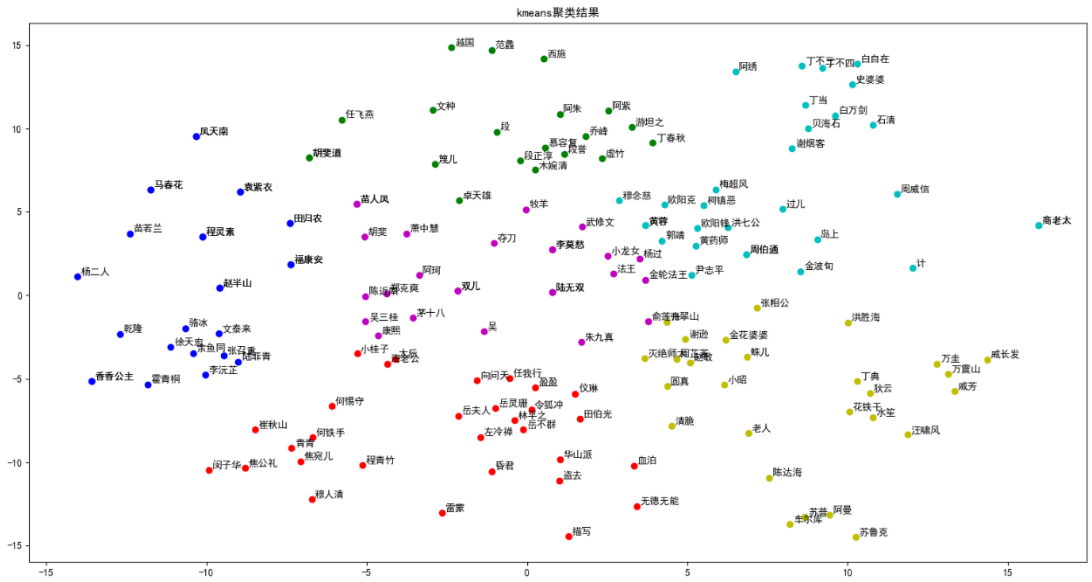
根据上述理论和方法，若采用如上参数设置，首先可以得到在每本小说主角选择的情况下，10 个其余相关词汇的分析如下(不完全展示)：

袁承志	胡斐	狄云	韦小宝	郭靖	杨过
青青 0.692357	袁紫衣 0.622910	丁典 0.588524	康熙 0.686654	黄蓉 0.715503	小龙女 0.670385
何铁手 0.606784	程灵素 0.606463	花铁干 0.569835	双儿 0.580549	欧阳锋 0.704734	郭靖 0.645077
崔秋山 0.547045	胡斐道 0.605740	水笙 0.541206	太后 0.571762	黄药师 0.670342	黄蓉 0.635589
洪胜海 0.535076	马春花 0.603988	万圭 0.526477	郑克爽 0.555381	杨过 0.645077	法王 0.613668
焦宛儿 0.526296	苗人凤 0.570122	汪啸风 0.510427	海老公 0.554056	柯镇恶 0.644685	金轮法王 0.606174
闵子华 0.499976	田归农 0.535895	万震山 0.480082	茅十八 0.542978	洪七公 0.636582	陆无双 0.595133
焦公礼 0.494829	凤天南 0.515066	令狐冲 0.477047	陈近南 0.533464	穆念慈 0.630195	李莫愁 0.594198
何惕守 0.474766	福康安 0.511578	戚芳 0.466382	小桂子 0.528651	欧阳克 0.620796	周伯通 0.582963
穆人清 0.463557	商老太 0.478238	胡斐 0.455563	阿珂 0.527906	梅超风 0.590421	过儿 0.569850
程青竹 0.462624	赵半山 0.461918	戚长发 0.451516	吴三桂 0.516149	周伯通 0.589474	尹志平 0.559828

图 2 部分小说相关词分析

观察分析可知，词向量相似度较高的词在小说中也有一定的关系，如韦小宝是康熙身边的大红人，黄蓉是郭靖的妻子，而黄药师是黄蓉的父亲，柯镇恶是郭靖的师傅，其他等人也和郭靖或多或少有着不可忽视的联系，再如神雕侠侣中，小龙女是杨过的师傅和妻子，金轮法王作为大反派在剧中也和杨过多次有过交集，由此可见，Word2Vec 模型的训练效果良好，且词向量的相关性分析是合理的。

通过对词向量的训练，采用 K-means 聚类得到的散点图如图 3 所示：



考虑到显示效果因素，图中设定的 n 为 6。从中可以清晰地看到同一本小说的人物基本上被分在了一类，极少数有被分类错误的情况，由此可见，整体效果良好，且充分验证了词向量的有效性。

References

[1] v_JULY_v. (2024-04-05). 如何通俗理解 Word2Vec (23 年修订版).https://blog.csdn.net/v_JULY_v/article/details/102708459.