Report of Deep Learning for Natural Langauge Processing

冯佳铂
775193638@qq.com

Abstract

利用给定语料库金庸小说语料，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。

Introduction

使用 Word2Vec 模型训练词向量。word2vec 是 google 推出的一个 NLP 工具，它可以将所有的词向量化，可以定量的去度量词与词之间的关系，挖掘词之间的联系。

最早期的词向量编码方式为 one hot representation，这种编码简单，但却维度过高，同时词与词之间相互正交，无法较好的衡量词与词之间的关系。通过神经语言模型训练，将每个词映射到一个较短的词向量上，以此构成新的向量空间，在新的向量空间上，便可以通过普通的统计学方法研究词与词之间的关系。

word2vec 之前，一般使用 DNN 模型训练，其中又包含 CBOW(Continuous Bag-of-Words) 与 Skip-Gram 两种模型。CBOW 模型的训练输入是某一个特征词的上下文相关的词对应的词向量，输出是该特定的词的词向量；Skip-Gram 模型与 CBOW 的思路相反，即输入是一个特定词的词向量，输出的是特定词对应的上下文词向量。

当词汇达百万级别时 DNN 输出层需要进行 softmax 计算各个词的输出概率计算量过大。使用霍夫曼树来代替隐藏层和输出层的神经元，霍夫曼树的叶子节点起到输出层神经元的作用，叶子节点的个数为词汇表的大小，内部节点起隐藏层神经元的作用。

在此基础上，进一步使用 Negative Sampling（负采样）方法，进一步减少计算量。

gensim 封装了 google 的 C 语言版的 word2vec，通过它可实现词向量训练。

Methodology

.

流程如下：

1. 语料库建立：将 16 部金庸小说合并建立语料库。

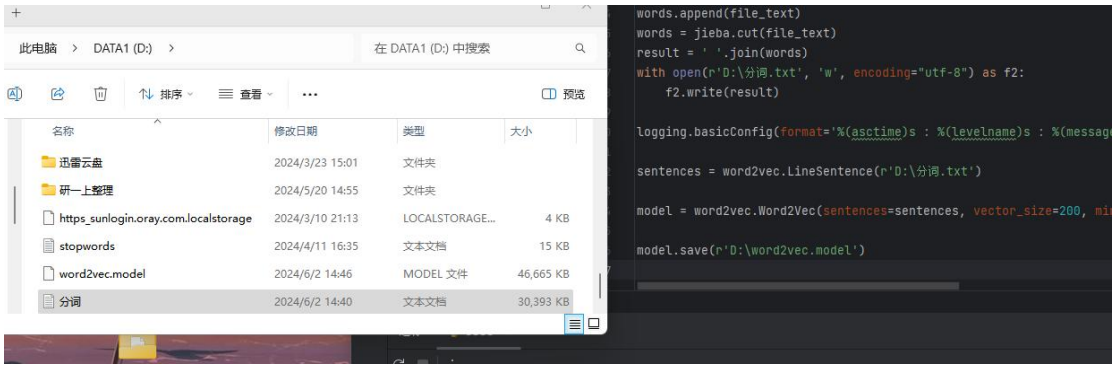
2. 文本预处理：将语料库中多余符号进行删除，如“本书来自 www.cr173.com 免费 txt 小说下载站”、“更多更新免费电子书请关注 www.cr173.com”等，一般的 NLP 处理中，需要去除停用词，但 word2vec 算法依赖于上下文，上下文可能就是停用词，可不去除停用词。使用 jieba 库函数进行分词。

3. 训练模型：使用 gensim 库函数对分词后的语料库进行训练，产生对应模型。

4. 分别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，验证词向量的有效性。

Experimental Studies

将训练所得模型“word2vec.model”保存至 D 盘。



T1：计算词向量之间的语意距离

所选词：

①“郭靖” “黄蓉”： 词意距离：0.8231

②“郭靖” “杨康”： 词意距离：0.5266

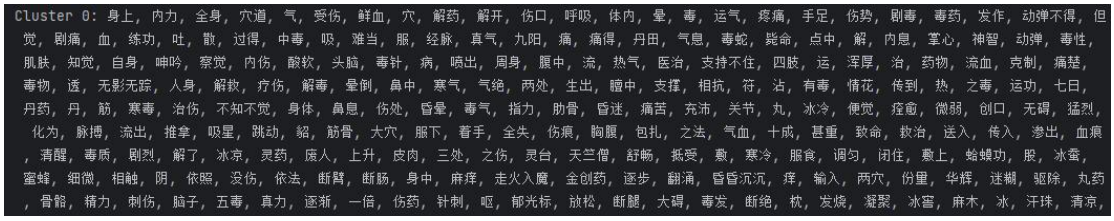
③“郭靖”“小龙女”： 词意距离：0.4966

④“郭靖”“韦小宝“： 词意距离：0.2580

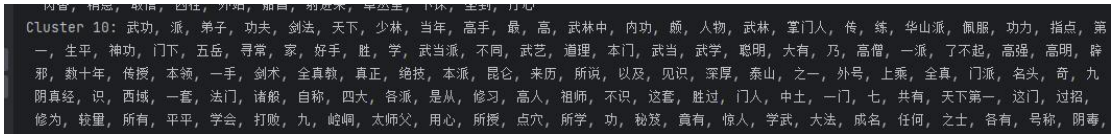
所计算词意距离满足关系，郭靖与黄蓉为夫妻，关系最为密切，词意距离为 0.8231；郭靖与杨康虽为异姓兄弟，但道不同，词意距离 0.5266；小龙女与郭靖之间由于杨过存在一定联系，但关系应略低于郭靖与杨康，词意距离为 0.4966，满足关系。郭靖与韦小宝为 2 部不同小说主角，关联程度很低，词意距离为 0.258.

T2：某一类词语的聚类

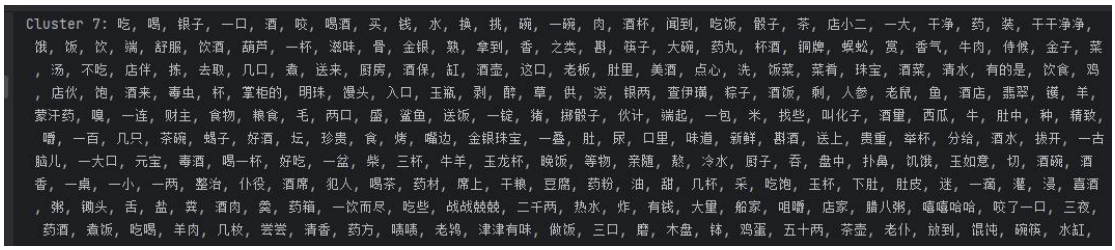
所选词：使用 K-means 聚类算法进行聚类分析
所得结果中大部分聚类比较成功，但同样存在效果较差部分。
较好的：
①与身体相关：



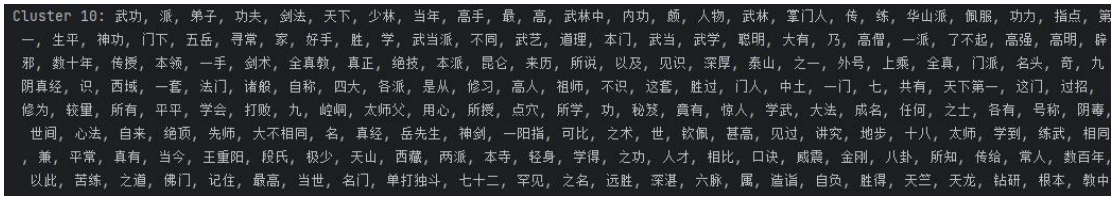
②与武功相关：



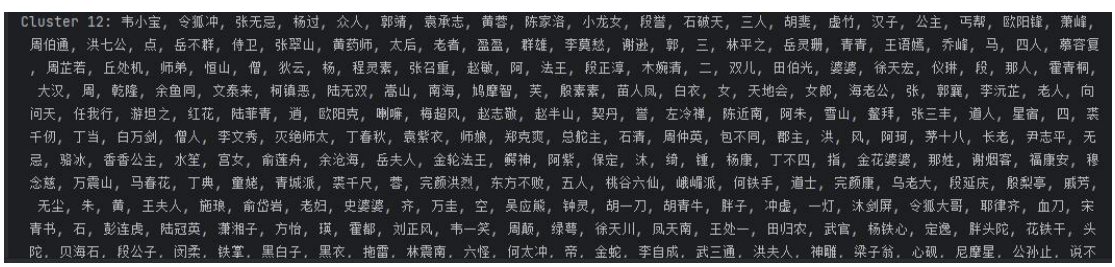
③与吃喝相关：



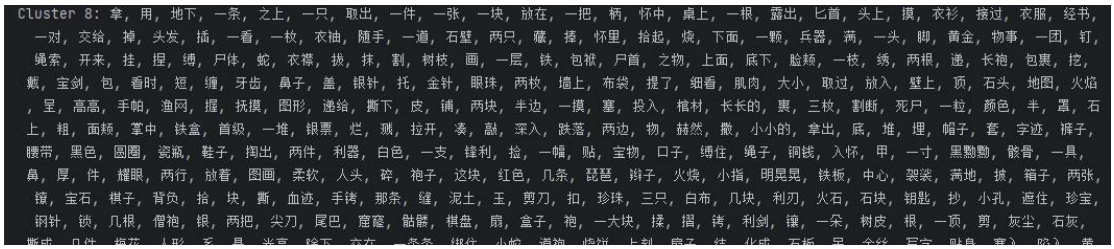
④与武功门派相关：

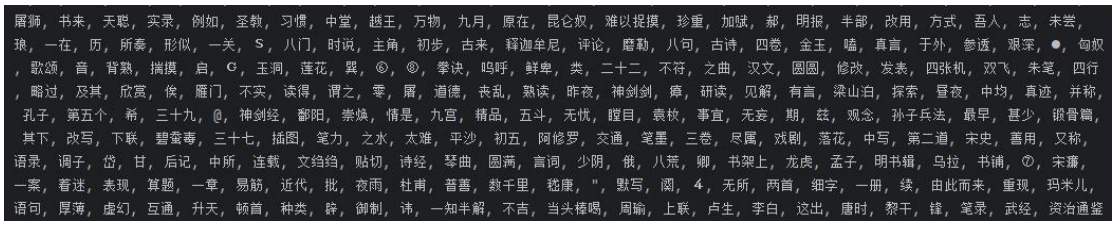


⑤人物相关：



较差的:





T3：计算某些段落直接的语意关联

基于 word2vec 的句向量,选取最简单的相加求均值:
计算公式:

$$\text{sen\_vec} = \frac{\sum_{i=1}^{m} \text{vec}_i}{m}$$

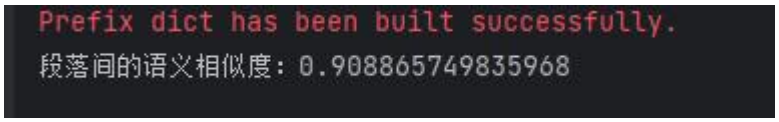
sen_vec 为句向量，m 为句中词数量，vec_i为词的词向量。

①选取两端语义相近的段落进行计算:（选取倚天屠龙记中围攻光明顶中有关七伤拳的两段讨论）

宗维侠强道:“七伤拳是我崆峒绝技，怎能说有害无益？当年我掌门师祖木灵子以七伤拳威震天下，名扬四海，寿至九十一岁，怎么说会损害自身？你这不是胡说八道麽？”张无忌道:“木灵子前辈想必内功深湛，自然能练，不但无害，反而强壮肝腑。依晚辈之见，宗前辈的内功如不到那个境界，若要强练，只怕终归无用。”

宗维侠是崆峒名宿，虽知他所说的不无有理，但在各派高手之前，被这少年指摘本派的镇山绝技无用，如何不恼？大声喝道:“凭你也配说我崆峒绝技有用无用。既说无用，那就来试试。”张无忌淡淡一笑，说道:“七伤拳自是神妙精奥的绝技，拳力刚中有柔，柔中有刚，七般拳劲各不相同，吞吐闪烁，变幻百端，敌手委实难防难挡……”宗维侠听他赞誉七伤拳的神妙，说来语语中肯，不禁脸露微笑，不住点头，却听他继续说道:“……晚辈只是说内功修为倘若不到，那便练之有害无益。”

计算结果:段落间语义相似度为:0.9089

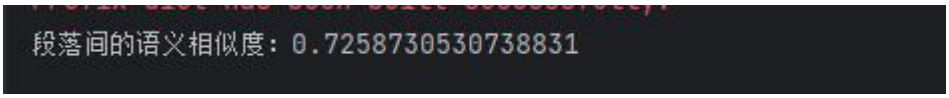


②选取两篇小说中不相关的两端进行计算：

原来黄药师对妻子情深意重，兼之爱妻为他而死，当时一意便要以死相殉。他自知武功深湛，上吊服毒，一时都不得便死，死了之后，尸身又不免受岛上哑仆糟蹋，于是去大陆捕拿造船巧匠，打造了这艘花船。这船的龙骨和寻常船只无异，但船底木材却并非用铁钉钉结，而是以生胶绳索胶缠在一起，泊在港中之时固是一艘极为华丽的花船，但如驶入大海，给浪涛一打，必致沉没。他本拟将妻子遗体放入船中，驾船出海，当波涌舟碎之际，按玉箫吹起《碧海潮生曲》，与妻子一齐葬身万丈洪涛之中，如此潇洒倜傥以终此一生，方不辱没了当世武学大宗匠的身分，但每次临到出海，总是既不忍携女同行，又不忍将她抛下不顾，终于造了墓室，先将妻子的棺木厝下。这艘船却是每年油漆，历时常新。要待女儿长大，有了妥善归宿，再行此事。

宗维侠强道："七伤拳是我崆峒绝技，怎能说有害无益？当年我掌门师祖木灵子以七伤拳威震天下，名扬四海，寿至九十一岁，怎么说会损害自身？你这不是胡说八道麽？"张无忌道："木灵子前辈想必内功深湛，自然能练，不但无害，反而强壮肝腑。依晚辈之见，宗前辈的内功如不到那个境界，若要强练，只怕终归无用。"

计算结果：段落间语义相似度为：0.7259



有所下降，但感觉下降的不够多，排除是因为同一个人写的小说缘故

选用一篇其他小说段落

这时，李达康的专车也驶临了收费站。侯亮平下车，当道一站，举起手掌做出停车的手势。李达康的轿车缓缓停下。后面跟踪而来的张华华的警车也在李达康专车的左侧停了下来。李达康的司机下了车，走到侯亮平和陆亦可面前：你们想干啥？知道这是谁的车吗？侯亮平说：我只知道被传唤人欧阳菁在这台车上。司机满脸的不屑：我说同志，欧阳菁是谁的夫人，你不会不知道吧？侯亮平说：欧阳菁是谁的夫人与我们检察院办案没关系！陆亦可解释道：有人举报了欧阳菁副行长，我们要请她去谈一谈！司机不无傲慢地说：知道吗？欧阳副行长是市委李达康书记的夫人，这台车也是李达康书记的专车！

计算结果为：0.7963

可见计算段落之间语义关联，具备一定参考价值，具体关系值得商榷。

Conclusions

通过 word2vec 模型训练了一个具备一定意义的词向量模型，可在一定程度上验证词向量的有效性。

References

[1] 用 gensim 学习 word2vec - 刘建平 Pinard - 博客园 (cnblogs.com)
[2] 基于 word2vec 的中文词向量训练_word2vec 能用汉字吗-CSDN 博客
[3] 基于 Word2vec 词聚类的关键词实现_word2vec 词聚类的关键词提取算法-CSDN 博客
[4] 基于 Word2vec 文本聚类_词向量模型 word2vec 与 聚类-CSDN 博客
[5] NLP 词向量和句向量方法总结及实现-CSDN 博客

Part 2 ： 计算中文平均信息熵

Abstract

对金庸先生 16 篇小说建立中文语料库，并计算其中文平均信息熵。

Introduction

信息是个很抽象的概念。人们常常说信息很多，或者信息较少，但却很难说清楚信息到底有多少。直至香农提出了“信息熵”的概念，才解决了对信息的量化度量问题。分别利用基于词的一元模型、二元模型和三元模型计算所给中文语料库的信息熵。

Methodology

.

M1：一元模型

$H(x) = -\sum_{x \epsilon X} P(x) \log P(x)$，其中$P(x)$近似等于每个词在语料库中出现的频率。

M2：二元模型

$H(X|Y) = -\sum_{x \epsilon X, y \epsilon Y} P(x,y) \log P(x|y)$，其中$P(x,y)$近似等于每个二元词在语料库中出现的频率，条件概率$P(x|y)$近似等于每个二元词组在语料库中出现的频数与该二元词组的第一个词为词首的二元词组的频数的比值。

M3：三元模型

$H(X|Y,Z) = -\sum_{x \epsilon X, y \epsilon Y, z \epsilon Z} P(x,y,z) \log P(x|y,z)$，其中$P(x,y,z)$近似等于每个三元词组在语料库中出现的频率，条件概率$P(x|y,z)$近似等于每个三元词组在语料库中出现的频数与以该三元词组的前两个词为词首的三元词组的频数的比值。

Experimental Studies

T1：验证齐夫定律

模型	词信息熵
一元模型	8.8132
二元模型	5.4837
三元模型	3.0216

Conclusions

随着模型复杂度的增加，信息熵下降，信息熵的值越大，文本中不同字符或词组的出现频率相差越大，文本的特征值越分散；模型变复杂后，考虑词与词之间的搭配，上下文之间的联系，能够更准确的扑捉到某些语言现象，从而使得这些现象的出现变得更加可预测，所以信息熵降低。

References

[1] 刘鸿博.《Python 自然语言处理》.[Z].undefined.电子工业出版社.2022.

[2]中文信息熵的计算_汉语信息熵-CSDN 博客[3]

[3]Brown, P. F. , Pietra, S. D. A. , Pietra, V. D. J. , Lai, J. C. , & Mercer, R. L. . (1992). An estimate of an upper bound for the. Computational Lingus, 18(1), 31-40.