



中国科学院大学
University of Chinese Academy of Sciences

自然语言处理

技术报告

班级：2023 春季自然语言处理班

组号：一

姓名：唐嘉良

学号：2020K8009907032

报告主题：基于 NNLM 和 word2vec 的中文词向量计算

2023 年 5 月 5 日

一. 报告摘要

本次实验中，我基于 n-grams 语言模型理论，利用 FNNLM、word2vec 开源模型和 tensorflow 深度学习框架等技术，对北京大学 1998 年人民日报分词语料库进行中文词向量计算，并对计算结果进行分析，给出结论与思考。

一方面，我们分别采用 C++ 和 Python 语言实现的 FNNLM（前馈神经网络语言模型）工具进行中文词向量计算，它们分别是 <http://nlg.isi.edu/software/nplm/> 和 <https://github.com/FuYanzhe2/NNLM>。选取合适的参数以达到更好的效果，最终发现前者在核心参数 embedding dimension = 100 以及 vocabulary size = 6000 时达到较好效果，但训练代价较高；而后者训练时间短，但由于自身网络复杂程度限制而在词向量相似度检查中表现欠佳。

另一方面，我们采用开源 word2vec 工具 <https://github.com/svn2github/word2vec>。选取合适的参数，在同一语料库下训练，将其与 FNN 语言模型进行全方位的对比，最终发现该 word2vec 在训练代价上远低于 FNNLM。在训练效果上，word2vec 由于作了数据平滑化处理，在低频词的近义词匹配上效果远好于 FNNLM，代价是在高频词的近义词匹配上丢失了部分准确性。

二. FNNLM

1. C++ 实现（开源模型）

采用开源模型，该模型具有两层 FNN 网络，接受输入为已分词的文本文件，输出前馈神经网络训练结果（包含词向量结果与网络参数）。

我们首先按照作业要求，选定 3-grams 语言模型，词向量维度 10，词汇表大小 1000，训练 epoch=10，采用均匀分布初始化，并将其余词汇全部用 <unk> 代替，训练得到模型。随机选取词汇计算与其余弦相似度最高的词，部分结果如下（第一行为选定词，随后是与之最相似的词汇及其相似度）：

公报	工厂/n
('释放', 0.44396305967286204)	('正在/d', 0.43871050792997535)
('结合', 0.4303669442497134)	('公布/v', 0.43348676969449834)
('工业', 0.42874544834252304)	('上百/m', 0.4136978817568357)
('寄予', 0.4125484729024546)	('电视剧/n', 0.3695310965806779)
('一九九一年', 0.39168772402986235)	('官员/n', 0.34788182206552626)
('行动', 0.3630249513607826)	('解困/v', 0.34617478068740604)
('背后', 0.35373269160320187)	('最佳/z', 0.34430927415405543)
('中共', 0.3478142831377674)	('5 0 0 0 /m', 0.3431802268289895)
('职位', 0.34451555702862025)	('贫困/a', 0.34183379141858755)

“公报”与“释放”“结合”、“工厂”与语义上相距甚远。可见模型训练效果很差。猜测是因为词向量维度过低，所能表示的差异化信息不足。另外，还有可能是因为词汇表过小，有一些词义相近但是出现频率不高的词汇没有被列入，导致匹配不到该词汇。

将词向量维度升级为 100，词汇表大小扩张成 6000，其余参数维持不变，再次训练模型。并随机选取词汇计算与其余弦相似度最高的词，部分结果如下：

香港	中国
{2562: 0.91112878897	{221: 0.8068964
2894, 2445: 0.732045	490: 0.69083648
2562成员国	221产业
3304近期	3043综合治理
3560湘西	4086执政
4076违法	5564廉洁自律
5633苏南	5994暖湿气流
2320大幅	3349攻坚
2445紧密	3490密切
5994暖湿气流	5428禁毒
4508波黑	4141活跃
5444心目	3722福利

其中“香港”作为地区词，与“成员国”“苏南”等地区词语义接近，可见参数调整初见成效，验证了我们之前的想法。但是模型发挥不稳定，例如“中国”一词的匹配结果不如人意，甚至匹配到的同一词性的词语都很少，但观察输出我们可以发现一个有意思的事实：词语“中国”的匹配结果似乎都是经常与其一起出现的词，例如“廉洁自律”“禁毒”“攻坚”，但本身语义却不一致。

考虑到这一点，结合模型的不稳定性，猜测是由于 3-grams 语言模型窗口大小太小，训练信息比较局部与片面，导致训练出来的词向量中包含的词语关联信息有偏差，为了获取更长的词汇出现历史，我们将模型调整为 5-grams，其余参数不变。再次训练，随机选取 20 个词汇，计算匹配与其最接近的 10 个词汇，结果如下：

是/v The top10 similar words are: 而是/c similarity: 0.5709573147497415 正是/v similarity: 0.5342286343842609 还是/c similarity: 0.5319058748214434 说是/v similarity: 0.529149229476576 即/c similarity: 0.5248538190061678 那种/r similarity: 0.5233100748501344 而/c similarity: 0.5098008495188113 视为/v similarity: 0.5031900948608081 需要/v similarity: 0.48976844952188103	提高/v The top10 similar words are: 增强/v similarity: 0.7692529449649345 降低/v similarity: 0.6733302927670619 扩大/v similarity: 0.6537732259977445 规范/v similarity: 0.6380946804267499 缩小/v similarity: 0.6353469818483746 减少/v similarity: 0.6280967924095437 改善/v similarity: 0.6239109980903587 改变/v similarity: 0.623612073008025 促进/v similarity: 0.6210474451123486	会议/n The top10 similar words are: 座谈会/n similarity: 0.798794295111997 大会/n similarity: 0.7765733080552919 仪式/n similarity: 0.7759808907025094 研讨会/n similarity: 0.7444852415895983 代表大会/n similarity: 0.7403841079708696 代表/n similarity: 0.7334950814448417 茶话会/n similarity: 0.7131587320212591 决定/n similarity: 0.7089230581754951 全会/n similarity: 0.695769950576361	春节/t The top10 similar words are: 佳节/n similarity: 0.6950621321808996 新年/t similarity: 0.669039014395176 床/n similarity: 0.6661223048473353 建交/v similarity: 0.6611016889098013 新春/t similarity: 0.6163088840474411 联欢/vn similarity: 0.6041923308703334 建国/v similarity: 0.6007900885652069 鸣笛/n similarity: 0.6004642248320962 到来/v similarity: 0.5902990430365392
---	---	---	---

建设/v The top10 similar words are: 培养/v similarity: 0.601517471243004 探索/v similarity: 0.5947942201545519 扶持/v similarity: 0.5264157195242369 印/v similarity: 0.5231605537685128 营造/v similarity: 0.5175784365238942 倡导/v similarity: 0.5171905555644942 该书/r similarity: 0.5125834840135394 变革/vn similarity: 0.5097493503638065 培育/v similarity: 0.4952302856737851	吃/v The top10 similar words are: 买/v similarity: 0.844492996014147 回/q similarity: 0.8366911062225713 赚/v similarity: 0.8353986414395742 还/v similarity: 0.8222576209544677 读/v similarity: 0.812349172332845 脱/v similarity: 0.801675576633962 包/v similarity: 0.8003523718887177 听/v similarity: 0.7986318707373741 退/v similarity: 0.7920468848347153	中/f The top10 similar words are: 上/f similarity: 0.7094588534748285 里/f similarity: 0.7049258170184087 那里/r similarity: 0.6438802369346298 以来/f similarity: 0.6335638509227101 近年来/l similarity: 0.6033535583268216 期间/f similarity: 0.5958660572360444 因而/c similarity: 0.595747396080676 来/f similarity: 0.5946378812719503 时刻/n similarity: 0.5841397541177612	美国/ns The top10 similar words are: 日本/ns similarity: 0.8169144509130961 伊拉克/ns similarity: 0.64343224524859 俄罗斯/ns similarity: 0.6336431472889745 英国/ns similarity: 0.6333970027890303 越南/ns similarity: 0.61508756629344 韩国/ns similarity: 0.5982756093560003 马耳他/ns similarity: 0.5957971468538014 土耳其/ns similarity: 0.5892029863357131 埃及/ns similarity: 0.5853716561402694
世界/n The top10 similar words are: 国际/n similarity: 0.7608956606299261 社会/n similarity: 0.7091367232110644 恩来/nr similarity: 0.6108340749857436 历史/n similarity: 0.5458185288857847 战线/n similarity: 0.5423959661900151 区域/n similarity: 0.5370450964519737 全球/n similarity: 0.5346560483625606 城市/n similarity: 0.5346013479389281 京剧/n similarity: 0.5298667679929256	时间/n The top10 similar words are: 头脑/n similarity: 0.7318552928106794 日子/n similarity: 0.7286515947219484 范围/n similarity: 0.7149895661766916 保护区/n similarity: 0.700423407691531 天/q similarity: 0.6938546222426089 月/n similarity: 0.6928753143546534 书/n similarity: 0.6853025422173729 年/q similarity: 0.6744089269791494 岁月/n similarity: 0.6639583661055286	条件/n The top10 similar words are: 形势/n similarity: 0.8075214277740617 态势/n similarity: 0.7562194195422769 前提/n similarity: 0.7498357053579803 另一方面/c similarity: 0.74266291025187 情况/n similarity: 0.7402492605257887 差距/n similarity: 0.7334938710378106 状态/n similarity: 0.7316319849495218 认识/n similarity: 0.7230680237725676 财力/n similarity: 0.7196639782191434	考试/vn The top10 similar words are: 选举/vn similarity: 0.7863272749422723 奖励/vn similarity: 0.7796570072688501 茶话会/n similarity: 0.768989928611868 磋商/vn similarity: 0.7678436502452238 评选/vn similarity: 0.7517982650146614 课堂/n similarity: 0.7466701461086374 兴奋剂/n similarity: 0.7459516722015037 学位/n similarity: 0.7390172511995665 租金/n similarity: 0.7362190316234845
执法/v The top10 similar words are: 积极性/n similarity: 0.7905597897754092 拼搏/v similarity: 0.7704796855545677 重要性/n similarity: 0.7540997525471702 贫穷/an similarity: 0.751036964930181 转变/vn similarity: 0.7470093410339688 解放思想/i similarity: 0.7237969396322468 要求/n similarity: 0.7098398151569437 措施/n similarity: 0.7084412309384456 宗旨/n similarity: 0.7070814891434003	牛/n The top10 similar words are: 昆/j similarity: 0.7374270183284729 高峰/n similarity: 0.7245559691290258 金/ng similarity: 0.720322643838396 肉/n similarity: 0.7197745605903756 蓝/a similarity: 0.7186201891135227 阴雨/n similarity: 0.7033381367307314 冷/a similarity: 0.7006818057810184 钵/q similarity: 0.6946362298396214 蛋/n similarity: 0.6934945745678079	年轻人/n The top10 similar words are: 路段/n similarity: 0.8148706335642181 老伴/n similarity: 0.809996113950461 那里/r similarity: 0.8018830523113739 奶奶/n similarity: 0.7958611069800644 里面/f similarity: 0.7905227515976542 双手/n similarity: 0.772835391991294 探测器/n similarity: 0.7527861152201625 恩格斯/nr similarity: 0.7486059491929702 波波夫/nr similarity: 0.7483555734169034	天空/n The top10 similar words are: 此事/r similarity: 0.8257922231902953 季节/n similarity: 0.8153852770214735 农贸市场/n similarity: 0.8144683215621378 大门/n similarity: 0.808616514656791 物品/n similarity: 0.8070450699439212 掌声/n similarity: 0.8033012783852466 航线/n similarity: 0.7963109966920585 浪/n similarity: 0.795404048505522 头/n similarity: 0.785922128535476
说法/n The top10 similar words are: 错误/n similarity: 0.8680343400788135 道理/n similarity: 0.8601702437170056 气息/n similarity: 0.8467877640262504 好处/n similarity: 0.8438510922671201 获奖/v similarity: 0.8421769315237159 那时/r similarity: 0.83587052247400678 损害/vn similarity: 0.8272342889657158 下去/v similarity: 0.8269175069719701 戏/n similarity: 0.8268017558627657	外交大臣/n The top10 similar words are: 外长/n similarity: 0.7786064669100671 议长/n similarity: 0.7632110140713916 外交部长/n similarity: 0.7542434958831225 大臣/n similarity: 0.7519437715973817 国防部长/n similarity: 0.7454683332920108 华夏/n similarity: 0.7270209200421592 总统/n similarity: 0.7194715949756397 包/nr similarity: 0.7126990328435915 发言人/n similarity: 0.7087588744760416	严肃/ad The top10 similar words are: 严格/ad similarity: 0.9169376792487335 着手/v similarity: 0.9114321791209699 广泛/ad similarity: 0.9032685863462125 统一/ad similarity: 0.9023822173855021 自行/d similarity: 0.8886347412188899 全力/d similarity: 0.8737841879743586 抓紧/v similarity: 0.8735728092657169 自主/vd similarity: 0.87051587359347 顺利/ad similarity: 0.8609627063446335	一/m The top10 similar words are: 第一/m similarity: 0.732600096641261 数/m similarity: 0.7020775112700167 双/m similarity: 0.7017474828208 五/m similarity: 0.7012037259080578 6/m similarity: 0.6961519802529197 两/m similarity: 0.6920990445462836 3/m similarity: 0.6830010905972189 9/m similarity: 0.6829813001836658 4 2/m similarity: 0.6829295611927422

从上面的结果中可以看出，该模型训练效果很好，在大多数随机选取的测试用例上表现良好，例如“提高”匹配“增强”“扩大”“改善”等。这说明此前性能瓶颈确是因为3-grams语言模型的缺陷。另外，还有个别测试用例表现不佳，例如“天空”和“年轻人”，匹配到的词汇要么毫不相干，要么甚至语义相反（例如“年轻人”与“老伴”“奶奶”），在输出文件中查找得知，这一类表现不佳的词汇大部分来自词汇表尾部词汇，在语料库中出现频率太低，因此训练数据缺乏可靠性；另一方面，有些词诸如“执法”没有太多常见的近义词（即词表中该词的近义词收录较少，且出现频率均较低），所以关于其近义词词

向量的训练样本也不足，所以即便有较高的相似度，也依然存在结果的偏差。

2. Python 实现（开源模型+Tensorflow 框架）

采用开源模型，该模型是对 2003 年 Yoshua Bengio 的 FNNLM 论文的简单复现，它接受输入为已分词的文本文件，输出前馈神经网络训练结果（包含词向量结果与词汇表）。

该 FNN 仅有一层网络层，受 C++ 版本模型的结果启发，我们放弃低维词向量和小词汇表，直接将词向量维度设置为 100，词表大小为 6000，采用 5-grams 模型，训练 epoch=10，采用均匀分布初始化和 Adam 优化器，并设置梯度阈值，隐藏层神经元数目为 128。

训练得到模型。随机选取词汇计算与其余弦相似度最高的词，部分结果如下：

金牌/n ('家庭/n', 0.4116647836539239) ('男排/n]nt', 0.3932847649493836) ('集中/v', 0.3805288043835408) ('领域/n', 0.3768357472839355) ('长期以来/l', 0.3541546308798703) ('选举/v', 0.3512167920521386) ('买/v', 0.3347685987628403) ('海洋/n', 0.3325727710577886) ('看/v', 0.3323508273490452)	特色/n ('贫困户/n', 0.4534347618460492) ('秦/nr', 0.402204307731249) ('余/nr', 0.39660469301400025) ('之中/f', 0.3718984421181844) ('支/q', 0.3532557138079492) ('房/n', 0.3527323865268041) ('过程/n', 0.3499914731138505) ('厂长/n', 0.34898496633876347) ('坚持/v', 0.3425169933089321)
华侨 ('华侨', 0.9999999926464392) ('大型', 0.8593505618851267) ('职责', 0.8517422200559747) ('/', 0.8184950098170587) ('恶', 0.811911567378585) ('电脑', 0.8023721937205178) ('违规', 0.7985725316880654) ('选举', 0.7909216884358602) ('启动', 0.7840745548390637) ('对外', 0.7835230663340944)	重申 ('重申', 1.0000000001877216) ('钻石', 0.8638899700464885) ('招生', 0.851843768564629) ('关联', 0.8445354318405096) ('会计', 0.8315493030584709) ('解开', 0.8292381391045849) ('三月', 0.8287435958443381) ('一日', 0.8251347013652339) ('指出', 0.8210182822309904) ('力量', 0.8187243238915586)

可以发现模型训练效果不如人意，在正确性上与 C++ 模型天差地别。唯一的优势是该模型训练速度极快，对北大 1998 分词语料单线程训练 1 次仅需半分钟，但对于同样的文本、模型参数和训练次数，C++ 模型在开启 10 线程的情况下训练一次需要 8 分钟。说明该模型网络复杂程度太低，需要加深网络层数以寻求性能的突破。这也提示我们需要深层的网络和高维的向量才能有效且充分地挖掘文本信息，尽管这会带来较大的计算资源开销。

最后是关于词表大小的结论，如果词表设置过小，很可能丢失掉大量的词汇信息，以至于匹配效果差，然而词汇表过大会带来大量低频词汇的词向量训练，对计算资源的要求增加，而且低频词样本过少，会导致低频词词向量训练不充分，可靠性过低，从而在相似度匹配中造成错误的判断。

三. Word2vec

利用开源模型，该 word2vec 工具主要包含两个模型：跳字模型（skip-gram）和连续词袋模型（continuous bag of words），以及两种高效训练的方法：负采样（negative sampling）和层序 softmax（hierarchical softmax）。

将该模型的参数设置为 5-grams，词向量维度 100，词汇表约为 6000，采用 CBOW 模型

和负采样方法，训练得到模型。随机选取词汇，进行词向量相似度计算，选取一部分结果如下：

```
第几个词? 71
建设/vn
The top10 similar words are:
建设/v
similarity: 0.649427793369933
物质文明/n
similarity: 0.6244636377203026
攻坚/vn
similarity: 0.6017607367928075
基本建设/l
similarity: 0.5894978655888262
党风/n
similarity: 0.5745707476217158
任务/n
similarity: 0.563391511411846
市场化/vn
similarity: 0.5617674074710098
重点/n
similarity: 0.5535417184963075
开放/vn
similarity: 0.5515051918656669
```

```
建设/vn
The top10 similar words are:
建设/v
similarity: 0.649427793369933
物质文明/n
similarity: 0.6244636377203026
攻坚/vn
similarity: 0.6017607367928075
基本建设/l
similarity: 0.5894978655888262
党风/n
similarity: 0.5745707476217158
任务/n
similarity: 0.563391511411846
市场化/vn
similarity: 0.5617674074710098
重点/n
similarity: 0.5535417184963075
开放/vn
similarity: 0.5515051918656669
```

可以看到，该模型具备一定的效果，但是不及 FNNLM 的效果显著。针对该问题，我们尝试更大程度利用计算资源，调整词向量维度为 200，其余参数不变。总训练时长约为 20s。

为方便与 FNNLM 模型对比，我们选取了同一批词汇，挑选多个高频词与低频词代表，计算词向量。高频词结果如下：

```
是/v
The top10 similar words are:
在于/v
similarity: 0.5995491805733092
意味着/v
similarity: 0.52025530809807
对于/p
similarity: 0.5072363077917811
没有/v
similarity: 0.497731076841177
正是/v
similarity: 0.4932356199247229
作为/p
similarity: 0.4870277325345672
作为/v
similarity: 0.47401383772002625
触及/v
similarity: 0.46897393122760056
行不通/v
similarity: 0.4629459768843193
```

```
提高/v
The top10 similar words are:
增强/v
similarity: 0.7160769207494518
水平/n
similarity: 0.7002731202413132
提高/vn
similarity: 0.683202091650365
素质/n
similarity: 0.6716959286507941
更新/v
similarity: 0.6455643183149317
改进/v
similarity: 0.6428498717159682
增加/v
similarity: 0.6255291296022286
普及/v
similarity: 0.6223098373306803
扩大/v
similarity: 0.6108783141408856
```

```
会议/n
The top10 similar words are:
全会/n
similarity: 0.7354189012008978
大会/n
similarity: 0.714191210399735
讲话/v
similarity: 0.6487233636947047
理事会/n
similarity: 0.6463488314799165
座谈会/n
similarity: 0.6269235704481844
召开/v
similarity: 0.6195436680319628
[全国/n
similarity: 0.5890197638131655
选举/vn
similarity: 0.5789925190849033
审议/v
similarity: 0.5785412451176222
```

```
春节/t
The top10 similar words are:
元旦/t
similarity: 0.8118719782322485
临近/v
similarity: 0.7513286556839293
前夕/f
similarity: 0.7281587371135453
来临/v
similarity: 0.7024238008527331
节日/n
similarity: 0.6872990768982368
佳节/n
similarity: 0.6845739815711576
新春/t
similarity: 0.6803843071679885
新年/t
similarity: 0.6505797538920934
喜迎/v
similarity: 0.6460919180375192
```

在高频词的匹配效果上面，word2vec 发挥稳定且可靠，这进一步证明高频词丰富的语料可以带来其词向量训练效果的显著。但是某些高频词如“是”，其匹配效果相比 FNNLM 来说并不更加卓越，前几位的匹配结果中掺杂的非近义词更多。为更为全面地比较两种模型，我们计算低频词结果如下：

```
天空/n
The top10 similar words are:
亮/a
similarity: 0.8557922215625218
只见/v
similarity: 0.8556627405987574
灯/n
similarity: 0.8522708334823184
树/n
similarity: 0.8480841113282567
扎/v
similarity: 0.8344501018283101
不见/v
similarity: 0.8322011114834679
深处/s
similarity: 0.8282372613244473
教堂/n
similarity: 0.8260399654739513
一个个/m
similarity: 0.8250687075147259
```

```
执法/vn
The top10 similar words are:
监督/vn
similarity: 0.8441884861065164
职能/n
similarity: 0.808017900505246
检查/vn
similarity: 0.7862668136900999
党政机关/n
similarity: 0.7827668970894699
改进/v
similarity: 0.7780543231363662
行政部门/n
similarity: 0.7766423084795683
权限/n
similarity: 0.776552308881847
监察/vn
similarity: 0.7755512076088319
考核/vn
similarity: 0.7739333269905528
```

```
年轻人/n
The top10 similar words are:
身影/n
similarity: 0.7841554849811886
平时/t
similarity: 0.7571878927439967
根/n
similarity: 0.7464400787572397
个个/q
similarity: 0.7463544743203879
小伙子/n
similarity: 0.7437105130757877
孩子/n
similarity: 0.7402491606371879
想到/v
similarity: 0.7379072753130375
红薯/n
similarity: 0.7362209607388208
从小/d
similarity: 0.7354543753689528
```

从结果上可以看出，对于语料库中的超低频率词汇（如“天空”，位列词汇表 5000 位 -6000 位，出现频率不足 30），训练效果同样不尽如人意。但是对于亚低频词汇如“执法”“年轻人”，其匹配效果远好于 FNNLM，匹配结果语义相关性更高。

四．FNNLM v. s. Word2vec

我们计算了不同频率词汇匹配结果，并对两种模型进行了对比。word2vec 在低频词的近义词匹配上效果远好于 FNNLM，代价是在高频词的近义词匹配上丢失了部分准确性。而对于超低频词汇，二者的训练效果均不佳。出现这一差异的原因应当是由于该 word2vec 模型对高频词语料的出现频率设置了遏制机制，并对低频词语料进行频率强化，在保证高频词的训练效果下降低频词的训练精度损失，从而使得训练模型在不同频率词汇上更加平滑，属于一种“公平机制”。在该机制下，尽管我们会损失一些高频词匹配精度，但是低频词匹配精度会得到质的增强。

此外，在训练代价上，相似参数的两种模型训练代价天差地别。首先，二者的源码并不一致，FNNLM 采用 C++ 与 python(with pytorch)，而 Word2vec 采用的是 C 语言，语言自身的差异导致了训练代价的不同，Word2vec 的训练耗时远小于 FNNLM 的两个模型。

并且，在 FNNLM 的训练中，发现训练时间对词汇表大小比较敏感，这在一定程度上限制了词汇表的扩充。而且 FNNLM 在词汇表很大的时候性能急速下降，受低频词干扰严重。但是 Word2vec 的训练精度对词汇表大小并不敏感：在其他参数不变，调节词汇表大小从~6000 到~17000 的过程中，模型表现十分稳定，近义词匹配任务的准确性并未受到明显影响。

最后，我们发现模型训练结果与语料库自身有密切联系。由于我们选取的是北京大学 1998 年人民日报分词语料，其语言较为正式与官方，因此训练出来的模型在国家与地区名等书面词汇上表现极好，例如 FNNLM 模型 (C++) 对“中国”的匹配结果，和 Word2vec 对“上海”和“会议”的匹配结果：

中国/ns The top10 similar words are: 世界/n similarity: 0.5445670093676286 南非/ns similarity: 0.5387714206338987 我国/n similarity: 0.5173761029234081 法国/ns similarity: 0.5142476798366031 美国/ns similarity: 0.5103983191560669 诞生/vn similarity: 0.50939171770637 非洲/ns similarity: 0.5048203326183064 欧洲/ns similarity: 0.49323822871080014 祖国/n similarity: 0.49103895729890923	上海/ns The top10 similar words are: 广州/ns similarity: 0.779056439511092 深圳/ns similarity: 0.7497132489100412 广东/ns similarity: 0.7488853103292686 天津/ns similarity: 0.7433461095650118 南京/ns similarity: 0.719288836467624 西安/ns similarity: 0.6811918697072867 郑州/ns similarity: 0.6806294854447071 北京/ns similarity: 0.6690611895158878 杭州/ns similarity: 0.6610183881279154	会议/n The top10 similar words are: 全会/n similarity: 0.7354189012008978 大会/n similarity: 0.714191210399735 讲话/v similarity: 0.6487233636947047 理事会/n similarity: 0.6463488314799165 座谈会/n similarity: 0.6269235704481844 召开/v similarity: 0.6195436680319628 [全国/n similarity: 0.5890197638131655 选举/vn similarity: 0.5789925190849033 审议/v similarity: 0.5785412451176222
---	--	---

这说明在训练模型时应当使用全面且充足的语料库，这是模型训练精度提升的基础。