Word2Vec验证词向量有效性报告

Weidong Xu
2697782204@qq.com

Abstract

实验要求：利用给定语料库，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。

Introduction

自然语言处理相关任务中要将自然语言交给机器学习中的算法来处理，通常需要将语言数学化。在将词送到神经网络训练之前需要将其编码成数值变量。常见的编码方式有两种：one-hot representation和distributed representation.

Word2Vec，是一款用于训练词向量的模型。这些模型为浅而双层的神经网络，在word2vec中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。Word2Vec主要包括CBOW模型（连续词袋模型）和Skip-gram模型（跳字模型），CBOW是给定上下文作为输入，来预测中心词，而后者则是给定中心词作为输入来预测上下文。本实验中主要使用CBOW模型，给定一个长度为T的文本序列，设时间步的词为W(t)，背景窗口大小为m，则连续词袋模型的目标函数（损失函数）是由背景词生成任一中心词的概率。表达式如下所示：

$$\sum_{t=1}^{T} P(w^{(t)} \mid w^{(t-m)},...,w^{(t-1)},w^{(t+1)},...,w^{(t+m)})$$

Experimental Studies

实验步骤：

一．语料处理

在读取语料后，首先利用jieba分词对语料进行分词，去掉txt文本中一些无意义的广告和标点符号等内容。

二．模型处理

使用开源的Gensim库提供的接口来训练word2vec模型，调用的函数如下：

vector_size:是指特征向量的维度，默认为100；

window：表示当前词与预测词在一个句子中的最大距离是多少，这里为5；

min_count：可以对字典做截断，词频少于min_count次数的单词会被丢弃掉；

epochs：迭代次数，默认为5，这里的参数不是iter，官方在新的版本中将这个关键词改为了epochs，用iter会出错

额外需要注明的是，本实验中使用的是CBOW算法，用于设置训练算法的参数sg默认为0，故在此不显示。

三．聚类分析

模型训练完毕之后，通过Gensim库中给出的接口函数（例如，most_similar()，similarity()等），输出训练之后的模型，与某个给定输入词关联度最高的词或者是给定的某两个词之间的关联性。

我选择了金庸小说集中我较为了解的五本小说《倚天屠龙记》、《天龙八部》、《射雕英雄传》、《神雕侠侣》、《笑傲江湖》作为样本，对其中的主角、门派分别进行了聚类分析。选取的五个核心词是“杨过”、“令狐冲”、“明教”、“张无忌”、“乔峰”由于原语料库是不同小说汇总成的文件，本实验有关的任务或门派均设置在单独的文本背景中进行处理，所以不会发生文本混杂的现象。

四．实验结果

```
杨过的词向量：[-1.32858511e-02  2.81959716e-02  8.70693382e-03 -2.77667008e-02
  1.40390778e-02 -3.46471630e-02 -1.47878248e-02 -7.08006416e-03
  9.12274118e-04 -1.55204197e-03 -4.06989316e-03 -6.26404351e-03
  2.01402456e-02  2.09199023e-02  2.23256499e-02  5.91121148e-03
  1.24766175e-02 -4.26766928e-03  2.88606650e-04 -7.44446320e-03
  5.40067325e-04  6.93356944e-03 -1.51123619e-02 -3.22566591e-02
 -1.14184397e-03 -2.90643121e-03 -1.48322145e-02 -1.34762367e-02
  7.81020091e-04  9.50068608e-03  1.85905900e-02  9.63654649e-03
  1.44125195e-03  1.62989064e-03 -8.30582436e-03  1.57015752e-02
 -3.05952737e-03  8.66953563e-03 -1.71496756e-02 -9.95333120e-03
  4.02984349e-03 -8.50373704e-04  1.35411071e-02 -7.60918576e-03
  1.11492733e-02 -4.89681959e-03 -7.20387045e-03 -2.40746606e-02
 -1.45101957e-02 -2.12244131e-03 -5.94531093e-03 -3.03832092e-03
 -1.37489196e-02  6.39612367e-03 -6.76329434e-03 -1.80042144e-02
  1.66328046e-02  2.30355398e-03 -1.20256422e-02  9.45151690e-03
  6.88074157e-03  4.28570202e-04 -8.66828603e-04  7.56028481e-03
  4.86016273e-03 -5.03973383e-03 -8.15231912e-03 -1.08791236e-02
 -1.16895011e-03  1.51411269e-03 -2.65503465e-03 -7.71127688e-03
 -8.46944924e-04  5.98481810e-03 -2.58110510e-03  2.02973243e-02
 -4.42091050e-03  2.43965406e-02 -8.10445490e-05  1.51772890e-03
  3.56433121e-03  1.30615886e-02 -8.93930346e-03  1.50164068e-02
  7.14924606e-03  2.07528863e-02 -1.84120815e-02  1.68601889e-02
  4.28968109e-03 -1.30887348e-02  2.84779742e-02 -1.33541916e-02
  2.19708565e-03 -9.03321616e-03  9.68291704e-03  9.89050418e-03
  3.90660297e-03 -1.94899589e-02  5.91367763e-03 -2.26848712e-03]


与杨过最相关的前20个词语
小杂种:0.7723487615585327
喜酒:0.7649268507957458
穷样:0.7582321763038635
解憾:0.7572911381721497
花枝:0.7371721267700195
年幼无知:0.7346215844154358
嘿嘿:0.7316581606864929
丘师兄:0.7287241220474243
时高时低:0.7263101935386658
檀郎:0.7243372797966003
```

丘处机是杨过的师祖。这一关系源于杨过的父亲杨康曾是丘处机的徒弟。

```
令狐冲的词向量：[-0.6922638   1.9787018   1.0061189  -0.6196524   0.45119393 -0.20400293
  1.7874864   1.0000676   0.05991349 -1.0373737  -0.69127643 -1.7893761
 -0.1288138  -0.79148227  1.0871345  -0.01168294 -0.87529755  0.22601633
  0.8026633  -1.7772264   1.1226429   0.83502114  0.62841624 -0.85676914
 -0.11261415  1.2088966  -1.301166   -0.72104895  0.5395437   0.762929
  1.5529774  -0.44584808 -0.28833538 -0.05235251  0.9342561  -0.02734638
  0.7975802  -0.6849228  -0.16564868 -2.074583   -0.01107912 -0.8036496
 -0.0134872  -0.50677645 -0.10880847 -0.2151708  -1.07065     0.01595299
  1.4356819  -1.2752093   0.27311373  0.28362232  1.4416894  -0.16121005
 -1.0422119  -0.9589222   1.7285646  -0.72286016  1.249918   -0.06577503
 -0.33789343 -0.12149879  0.10809949  0.479514   -2.1032796   0.87409794
 -0.8237637  -0.686744   -1.3665977   0.9430681  -1.6154265   0.28039208
  1.065824    1.1805862  -0.682848   -0.17573404 -0.13656874  0.15796876
 -0.29214984  1.3098887  -0.05913427 -0.7234849   0.33861965  1.1015621
  0.2452534   0.03898486 -0.40372565  0.11735509 -0.12248677  0.3484771
  0.56404084  1.1884488  -0.43746328 -0.6068935   2.1079156   1.1380842
  1.0461744  -0.45739046  0.7791916  -0.19885586]
```

```
与令狐冲最相关的前10个词语
黑白子:0.8351880311965942
盈盈:0.8252428770065308
林平之:0.8143759965896606
林震南:0.8097099661827087
岳夫人:0.8059633374214172
岳不群:0.7833855748176575
岳灵珊:0.7684560418128967
杨莲亭:0.7510129809379578
丹青生:0.750237762928009
王夫人:0.7379755973815918
```

令狐冲主要相关人物：林平之是其师弟，岳不群一家属其师门，其余人是与令狐冲有交集的人。

```
张无忌的词向量: [-1.3468244   1.2772491  -1.1998045  -0.2602764   1.3325298  -1.4390932
 -0.29077262  2.1700346  -0.22679313 -0.9975519   0.74084765 -0.9512809
  0.864543    1.2214223   0.85325694 -0.06900047  0.78540146 -0.42696396
  0.14530905 -2.041194   -0.73863715 -0.7073573  -0.41537547 -0.52954215
  0.1415687  -0.91572183  0.78300506  0.57008266 -1.0023104  -1.7771616
  2.2341619   1.1866965  -0.00412047 -0.05329161 -1.8132656   1.5370086
 -0.07218401 -0.29883572  0.05068571 -2.022912    0.15101732  0.17888051
 -0.35644808  0.27350634 -0.3393595  -0.2213502  -0.15527396 -0.23989207
  1.2219301   0.1704401   1.3895394   0.47458795 -0.19911963  0.56605095
 -1.2805806   0.18071665  1.2919277   0.63812643 -0.2566227   1.0028602
  0.0676926   0.06318896  1.0981171  -0.61256635 -1.1545277  -0.09271729
 -0.2566575  -0.34081653 -0.75439364  0.63489336  0.64055973 -0.34232396
  0.79677576  0.7882188   1.1420865  -0.5277288   0.98003185 -0.278677
  0.5820484  -0.46380544 -1.4713687   2.0999544  -1.1367747   0.78859615
 -0.01670607  0.39377022  0.8271907  -0.67936933  2.0805273   0.8377142
  0.5456229   0.8739085   0.08869809  0.764242    1.2899886   0.09806713
  0.6215119  -1.437655    1.4535251   0.7519684 ]
```

```
与张无忌最相关的前10个词语
张翠山:0.8899224996566772
金花婆婆:0.8835256695747375
赵敏:0.8718828558921814
周芷若:0.8711311221122742
殷素素:0.8707377910614014
谢逊:0.8494747281074524
殊儿:0.8395859599113464
朱长龄:0.8390606641769409
俞岱岩:0.8297713994979858
鹿杖客:0.8233658075332642
```

张无忌跟周芷若是青梅竹马，父亲为张翠山，义父为谢逊，后来跟赵敏在一起，宋青书是张无忌的师哥，金花婆婆是明教的紫衫龙王，也是小昭的母亲，也是与张无忌相关。

```
明教的词向量: [ 2.1630008   2.240448    1.7479224  -1.0775363  -2.3455064  -0.89590317
  1.21734     1.3350781   0.2825392   0.60316396 -0.79488194 -0.41209048
  0.67971486 -1.491573   -1.6027249  -1.1296531   0.21864946 -1.5433532
  0.31751284 -1.3096851   0.13275604  0.20244917  0.451487    0.11442654
 -0.8715765   0.10405313 -0.60723656 -0.96444607 -0.36689612 -0.54387754
 -0.66570884  1.2797772  -0.05531019 -0.46876985  0.28239882  0.7489113
 -0.41952068 -1.5190701  -2.5771725  -2.1507983  -0.2627869   0.3363149
 -0.30059746  0.02630228  1.1066824   0.5298387  -1.3366715   0.22545971
  0.88270795  0.3066164  -0.31554905  1.2658323  -0.8915513  -0.6180083
 -1.0995188  -0.74506235  0.6704993  -1.6084002  -1.0106865  -0.9676509
  2.1125305  -0.9342753  -1.275166    1.2825323   0.56364125  0.70266944
 -1.4374524   0.04839968 -1.5261071   0.547908   -2.5169795   2.2162218
 -0.8197871  -0.15951931  1.7853405   2.6990016   0.36183432  0.2569372
  0.2616273   1.2513024   0.87933654  0.30893093 -1.6702783   1.3079062
 -1.23952    -0.46824384  0.85758096  2.0119748   0.15619136  0.7680483
  1.6547933   0.07721848 -1.8862492  -0.26317704  1.0115772   0.40558225
  0.09466172  0.3949548   0.5438792  -0.40051275]

与明教最相关的前20个词语
魔教:0.922904372215271
本教:0.9134768843650818
在世:0.8774579763412476
天鹰:0.8760349154472351
英雄:0.8697824478149414
教:0.8680177330970764
法王:0.8671013116836548
正派:0.8657382726669312
名门:0.8632087707519531
武当:0.8626278042793274
```

明教常自称本教，也常被六大门派叫做魔教。

```
乔峰的词向量: [-1.10054040e+00  1.42638242e+00  2.43640125e-01  7.39877298e-02
  3.93443644e-01 -1.41571379e+00  5.89881279e-02  1.02442479e+00
  1.22555597e-02 -4.05546755e-01  1.79792255e-01 -1.44055510e+00
 -4.83076155e-01 -3.47064406e-01 -5.32419086e-01 -8.96198213e-01
  8.02984357e-01 -2.08921766e+00 -8.38263452e-01 -1.00756824e+00
 -2.52243489e-01  1.32862506e-02  5.11466146e-01  1.48483932e-01
 -1.58053374e+00  7.29796410e-01 -6.58845782e-01  3.22459787e-01
  3.92481312e-02 -2.26552133e-02  1.66045332e+00 -1.74794376e-01
 -6.55436218e-01  6.01251781e-01 -1.19125962e-01  6.06180616e-02
  1.88423598e+00 -1.18553364e+00  3.77610445e-01 -2.60079324e-01
  6.22188866e-01 -8.47188294e-01 -4.05369073e-01 -1.10511148e+00
  5.62851191e-01 -9.21833754e-01 -5.23132920e-01 -1.51775986e-01
  5.18869698e-01  1.06284827e-01  5.82588017e-02 -5.70093513e-01
  8.22134674e-01  3.00216407e-01 -8.82756650e-01  2.57911265e-01
  1.39294331e+00 -7.86842287e-01  2.21555576e-01 -7.10275948e-01
  6.07843995e-01 -6.06885076e-01  7.24386632e-01 -8.44850540e-01
 -7.80144930e-02  1.46828592e-01  5.68108201e-01  1.68386376e+00
 -7.30943084e-01  7.52927184e-01  5.12439370e-01  1.30385458e+00
  1.00676799e+00  8.88747871e-01  1.14057446e+00 -9.62578058e-02
 -2.60877490e-01  1.72714586e-03  1.62630200e-01 -1.30194378e+00
 -1.42693162e-01 -1.05692856e-01  1.26679778e-01  1.59837854e+00
 -9.37138081e-01 -4.74761724e-01 -1.19155720e-01  8.31254601e-01
  2.36449376e-01 -3.82272422e-01  1.99537933e+00  1.12193739e+00
  1.20791160e-02 -6.77747548e-01  7.18313456e-01  7.13803247e-02
  3.88304979e-01 -4.38134700e-01 -2.95279592e-01 -3.69712204e-01]
```

```
与乔峰最相关的前10个词语
萧峰:0.8711372017860413
乌老大:0.8653621673583984
段正淳:0.8586615920066833
木婉清:0.8414422869682312
游坦之:0.8403527736663818
全冠清:0.8224583864212036
苏星河:0.820433497428894
钟灵:0.8153152465820312
钟万仇:0.8114773035049438
童姥:0.8102238178253174
```

萧峰，原名乔峰，段正淳曾是其带头大哥。