

10.6 求近义词和类比词

在10.3节（word2vec的实现）中，我们在小规模数据集上训练了一个word2vec词嵌入模型，并通过词向量的余弦相似度搜索近义词。实际中，在大规模语料上预训练的词向量常常可以应用到下游自然语言处理任务中。本节将演示如何用这些预训练的词向量来求近义词和类比词。我们还将后面两节中继续应用预训练的词向量。

10.6.1 使用预训练的词向量

基于PyTorch的关于自然语言处理的常用包有官方的[torchtext](#)以及第三方的[pytorch-nlp](#)等等。你可以使用 `pip` 很方便地按照它们，例如命令行执行

```
pip install torchtext
```

详情请参见其README。

本节我们使用torchtext进行练习。下面查看它目前提供的预训练词嵌入的名称。

```
import torch
import torchtext.vocab as vocab

vocab.pretrained_aliases.keys()
```

输出：

```
dict_keys(['charngram.100d', 'fasttext.en.300d', 'fasttext.simple.300d', 'glove.42B.300d', 'glove.6B.300d', 'glove.6B.400d', 'glove.6B.50d', 'glove.6B.768d', 'glove.840B.300d', 'glove.840B.400d', 'glove.840B.50d', 'glove.840B.768d', 'glove.840B.960d', 'glove.840B.1280d', 'glove.840B.1536d', 'glove.840B.1792d', 'glove.840B.2048d', 'glove.840B.2304d', 'glove.840B.2560d', 'glove.840B.2816d', 'glove.840B.3072d', 'glove.840B.3328d', 'glove.840B.3584d', 'glove.840B.3840d', 'glove.840B.4096d', 'glove.840B.4352d', 'glove.840B.4608d', 'glove.840B.4864d', 'glove.840B.5120d', 'glove.840B.5376d', 'glove.840B.5632d', 'glove.840B.5888d', 'glove.840B.6144d', 'glove.840B.6400d', 'glove.840B.6656d', 'glove.840B.6912d', 'glove.840B.7168d', 'glove.840B.7424d', 'glove.840B.7680d', 'glove.840B.7936d', 'glove.840B.8192d', 'glove.840B.8448d', 'glove.840B.8704d', 'glove.840B.8960d', 'glove.840B.9216d', 'glove.840B.9472d', 'glove.840B.9728d', 'glove.840B.9984d', 'glove.840B.10240d', 'glove.840B.10496d', 'glove.840B.10752d', 'glove.840B.11008d', 'glove.840B.11264d', 'glove.840B.11520d', 'glove.840B.11776d', 'glove.840B.12032d', 'glove.840B.12288d', 'glove.840B.12544d', 'glove.840B.12800d', 'glove.840B.13056d', 'glove.840B.13312d', 'glove.840B.13568d', 'glove.840B.13824d', 'glove.840B.14080d', 'glove.840B.14336d', 'glove.840B.14592d', 'glove.840B.14848d', 'glove.840B.15104d', 'glove.840B.15360d', 'glove.840B.15616d', 'glove.840B.15872d', 'glove.840B.16128d', 'glove.840B.16384d', 'glove.840B.16640d', 'glove.840B.16896d', 'glove.840B.17152d', 'glove.840B.17408d', 'glove.840B.17664d', 'glove.840B.17920d', 'glove.840B.18176d', 'glove.840B.18432d', 'glove.840B.18688d', 'glove.840B.18944d', 'glove.840B.19200d', 'glove.840B.19456d', 'glove.840B.19712d', 'glove.840B.19968d', 'glove.840B.20224d', 'glove.840B.20480d', 'glove.840B.20736d', 'glove.840B.20992d', 'glove.840B.21248d', 'glove.840B.21504d', 'glove.840B.21760d', 'glove.840B.22016d', 'glove.840B.22272d', 'glove.840B.22528d', 'glove.840B.22784d', 'glove.840B.23040d', 'glove.840B.23296d', 'glove.840B.23552d', 'glove.840B.23808d', 'glove.840B.24064d', 'glove.840B.24320d', 'glove.840B.24576d', 'glove.840B.24832d', 'glove.840B.25088d', 'glove.840B.25344d', 'glove.840B.25600d', 'glove.840B.25856d', 'glove.840B.26112d', 'glove.840B.26368d', 'glove.840B.26624d', 'glove.840B.26880d', 'glove.840B.27136d', 'glove.840B.27392d', 'glove.840B.27648d', 'glove.840B.27904d', 'glove.840B.28160d', 'glove.840B.28416d', 'glove.840B.28672d', 'glove.840B.28928d', 'glove.840B.29184d', 'glove.840B.29440d', 'glove.840B.29696d', 'glove.840B.29952d', 'glove.840B.30208d', 'glove.840B.30464d', 'glove.840B.30720d', 'glove.840B.30976d', 'glove.840B.31232d', 'glove.840B.31488d', 'glove.840B.31744d', 'glove.840B.32000d', 'glove.840B.32256d', 'glove.840B.32512d', 'glove.840B.32768d', 'glove.840B.33024d', 'glove.840B.33280d', 'glove.840B.33536d', 'glove.840B.33792d', 'glove.840B.34048d', 'glove.840B.34304d', 'glove.840B.34560d', 'glove.840B.34816d', 'glove.840B.35072d', 'glove.840B.35328d', 'glove.840B.35584d', 'glove.840B.35840d', 'glove.840B.36096d', 'glove.840B.36352d', 'glove.840B.36608d', 'glove.840B.36864d', 'glove.840B.37120d', 'glove.840B.37376d', 'glove.840B.37632d', 'glove.840B.37888d', 'glove.840B.38144d', 'glove.840B.38400d', 'glove.840B.38656d', 'glove.840B.38912d', 'glove.840B.39168d', 'glove.840B.39424d', 'glove.840B.39680d', 'glove.840B.39936d', 'glove.840B.40192d', 'glove.840B.40448d', 'glove.840B.40704d', 'glove.840B.40960d', 'glove.840B.41216d', 'glove.840B.41472d', 'glove.840B.41728d', 'glove.840B.41984d', 'glove.840B.42240d', 'glove.840B.42496d', 'glove.840B.42752d', 'glove.840B.43008d', 'glove.840B.43264d', 'glove.840B.43520d', 'glove.840B.43776d', 'glove.840B.44032d', 'glove.840B.44288d', 'glove.840B.44544d', 'glove.840B.44800d', 'glove.840B.45056d', 'glove.840B.45312d', 'glove.840B.45568d', 'glove.840B.45824d', 'glove.840B.46080d', 'glove.840B.46336d', 'glove.840B.46592d', 'glove.840B.46848d', 'glove.840B.47104d', 'glove.840B.47360d', 'glove.840B.47616d', 'glove.840B.47872d', 'glove.840B.48128d', 'glove.840B.48384d', 'glove.840B.48640d', 'glove.840B.48896d', 'glove.840B.49152d', 'glove.840B.49408d', 'glove.840B.49664d', 'glove.840B.49920d', 'glove.840B.50176d', 'glove.840B.50432d', 'glove.840B.50688d', 'glove.840B.50944d', 'glove.840B.51200d', 'glove.840B.51456d', 'glove.840B.51712d', 'glove.840B.51968d', 'glove.840B.52224d', 'glove.840B.52480d', 'glove.840B.52736d', 'glove.840B.52992d', 'glove.840B.53248d', 'glove.840B.53504d', 'glove.840B.53760d', 'glove.840B.54016d', 'glove.840B.54272d', 'glove.840B.54528d', 'glove.840B.54784d', 'glove.840B.55040d', 'glove.840B.55296d', 'glove.840B.55552d', 'glove.840B.55808d', 'glove.840B.56064d', 'glove.840B.56320d', 'glove.840B.56576d', 'glove.840B.56832d', 'glove.840B.57088d', 'glove.840B.57344d', 'glove.840B.57600d', 'glove.840B.57856d', 'glove.840B.58112d', 'glove.840B.58368d', 'glove.840B.58624d', 'glove.840B.58880d', 'glove.840B.59136d', 'glove.840B.59392d', 'glove.840B.59648d', 'glove.840B.59904d', 'glove.840B.60160d', 'glove.840B.60416d', 'glove.840B.60672d', 'glove.840B.60928d', 'glove.840B.61184d', 'glove.840B.61440d', 'glove.840B.61696d', 'glove.840B.61952d', 'glove.840B.62208d', 'glove.840B.62464d', 'glove.840B.62720d', 'glove.840B.62976d', 'glove.840B.63232d', 'glove.840B.63488d', 'glove.840B.63744d', 'glove.840B.64000d', 'glove.840B.64256d', 'glove.840B.64512d', 'glove.840B.64768d', 'glove.840B.65024d', 'glove.840B.65280d', 'glove.840B.65536d', 'glove.840B.65792d', 'glove.840B.66048d', 'glove.840B.66304d', 'glove.840B.66560d', 'glove.840B.66816d', 'glove.840B.67072d', 'glove.840B.67328d', 'glove.840B.67584d', 'glove.840B.67840d', 'glove.840B.68096d', 'glove.840B.68352d', 'glove.840B.68608d', 'glove.840B.68864d', 'glove.840B.69120d', 'glove.840B.69376d', 'glove.840B.69632d', 'glove.840B.69888d', 'glove.840B.70144d', 'glove.840B.70400d', 'glove.840B.70656d', 'glove.840B.70912d', 'glove.840B.71168d', 'glove.840B.71424d', 'glove.840B.71680d', 'glove.840B.71936d', 'glove.840B.72192d', 'glove.840B.72448d', 'glove.840B.72704d', 'glove.840B.72960d', 'glove.840B.73216d', 'glove.840B.73472d', 'glove.840B.73728d', 'glove.840B.73984d', 'glove.840B.74240d', 'glove.840B.74496d', 'glove.840B.74752d', 'glove.840B.75008d', 'glove.840B.75264d', 'glove.840B.75520d', 'glove.840B.75776d', 'glove.840B.76032d', 'glove.840B.76288d', 'glove.840B.76544d', 'glove.840B.76800d', 'glove.840B.77056d', 'glove.840B.77312d', 'glove.840B.77568d', 'glove.840B.77824d', 'glove.840B.78080d', 'glove.840B.78336d', 'glove.840B.78592d', 'glove.840B.78848d', 'glove.840B.79104d', 'glove.840B.79360d', 'glove.840B.79616d', 'glove.840B.79872d', 'glove.840B.80128d', 'glove.840B.80384d', 'glove.840B.80640d', 'glove.840B.80896d', 'glove.840B.81152d', 'glove.840B.81408d', 'glove.840B.81664d', 'glove.840B.81920d', 'glove.840B.82176d', 'glove.840B.82432d', 'glove.840B.82688d', 'glove.840B.82944d', 'glove.840B.83200d', 'glove.840B.83456d', 'glove.840B.83712d', 'glove.840B.83968d', 'glove.840B.84224d', 'glove.840B.84480d', 'glove.840B.84736d', 'glove.840B.84992d', 'glove.840B.85248d', 'glove.840B.85504d', 'glove.840B.85760d', 'glove.840B.86016d', 'glove.840B.86272d', 'glove.840B.86528d', 'glove.840B.86784d', 'glove.840B.87040d', 'glove.840B.87296d', 'glove.840B.87552d', 'glove.840B.87808d', 'glove.840B.88064d', 'glove.840B.88320d', 'glove.840B.88576d', 'glove.840B.88832d', 'glove.840B.89088d', 'glove.840B.89344d', 'glove.840B.89600d', 'glove.840B.89856d', 'glove.840B.90112d', 'glove.840B.90368d', 'glove.840B.90624d', 'glove.840B.90880d', 'glove.840B.91136d', 'glove.840B.91392d', 'glove.840B.91648d', 'glove.840B.91904d', 'glove.840B.92160d', 'glove.840B.92416d', 'glove.840B.92672d', 'glove.840B.92928d', 'glove.840B.93184d', 'glove.840B.93440d', 'glove.840B.93696d', 'glove.840B.93952d', 'glove.840B.94208d', 'glove.840B.94464d', 'glove.840B.94720d', 'glove.840B.94976d', 'glove.840B.95232d', 'glove.840B.95488d', 'glove.840B.95744d', 'glove.840B.96000d', 'glove.840B.96256d', 'glove.840B.96512d', 'glove.840B.96768d', 'glove.840B.97024d', 'glove.840B.97280d', 'glove.840B.97536d', 'glove.840B.97792d', 'glove.840B.98048d', 'glove.840B.98304d', 'glove.840B.98560d', 'glove.840B.98816d', 'glove.840B.99072d', 'glove.840B.99328d', 'glove.840B.99584d', 'glove.840B.99840d', 'glove.840B.100096d', 'glove.840B.100352d', 'glove.840B.100608d', 'glove.840B.100864d', 'glove.840B.101120d', 'glove.840B.101376d', 'glove.840B.101632d', 'glove.840B.101888d', 'glove.840B.102144d', 'glove.840B.102400d', 'glove.840B.102656d', 'glove.840B.102912d', 'glove.840B.103168d', 'glove.840B.103424d', 'glove.840B.103680d', 'glove.840B.103936d', 'glove.840B.104192d', 'glove.840B.104448d', 'glove.840B.104704d', 'glove.840B.104960d', 'glove.840B.105216d', 'glove.840B.105472d', 'glove.840B.105728d', 'glove.840B.105984d', 'glove.840B.106240d', 'glove.840B.106496d', 'glove.840B.106752d', 'glove.840B.107008d', 'glove.840B.107264d', 'glove.840B.107520d', 'glove.840B.107776d', 'glove.840B.108032d', 'glove.840B.108288d', 'glove.840B.108544d', 'glove.840B.108800d', 'glove.840B.109056d', 'glove.840B.109312d', 'glove.840B.109568d', 'glove.840B.109824d', 'glove.840B.110080d', 'glove.840B.110336d', 'glove.840B.110592d', 'glove.840B.110848d', 'glove.840B.111104d', 'glove.840B.111360d', 'glove.840B.111616d', 'glove.840B.111872d', 'glove.840B.112128d', 'glove.840B.112384d', 'glove.840B.112640d', 'glove.840B.112896d', 'glove.840B.113152d', 'glove.840B.113408d', 'glove.840B.113664d', 'glove.840B.113920d', 'glove.840B.114176d', 'glove.840B.114432d', 'glove.840B.114688d', 'glove.840B.114944d', 'glove.840B.115200d', 'glove.840B.115456d', 'glove.840B.115712d', 'glove.840B.115968d', 'glove.840B.116224d', 'glove.840B.116480d', 'glove.840B.116736d', 'glove.840B.116992d', 'glove.840B.117248d', 'glove.840B.117504d', 'glove.840B.117760d', 'glove.840B.118016d', 'glove.840B.118272d', 'glove.840B.118528d', 'glove.840B.118784d', 'glove.840B.119040d', 'glove.840B.119296d', 'glove.840B.119552d', 'glove.840B.119808d', 'glove.840B.120064d', 'glove.840B.120320d', 'glove.840B.120576d', 'glove.840B.120832d', 'glove.840B.121088d', 'glove.840B.121344d', 'glove.840B.121600d', 'glove.840B.121856d', 'glove.840B.122112d', 'glove.840B.122368d', 'glove.840B.122624d', 'glove.840B.122880d', 'glove.840B.123136d', 'glove.840B.123392d', 'glove.840B.123648d', 'glove.840B.123904d', 'glove.840B.124160d', 'glove.840B.124416d', 'glove.840B.124672d', 'glove.840B.124928d', 'glove.840B.125184d', 'glove.840B.125440d', 'glove.840B.125696d', 'glove.840B.125952d', 'glove.840B.126208d', 'glove.840B.126464d', 'glove.840B.126720d', 'glove.840B.126976d', 'glove.840B.127232d', 'glove.840B.127488d', 'glove.840B.127744d', 'glove.840B.128000d', 'glove.840B.128256d', 'glove.840B.128512d', 'glove.840B.128768d', 'glove.840B.129024d', 'glove.840B.129280d', 'glove.840B.129536d', 'glove.840B.129792d', 'glove.840B.130048d', 'glove.840B.130304d', 'glove.840B.130560d', 'glove.840B.130816d', 'glove.840B.131072d', 'glove.840B.131328d', 'glove.840B.131584d', 'glove.840B.131840d', 'glove.840B.132096d', 'glove.840B.132352d', 'glove.840B.132608d', 'glove.840B.132864d', 'glove.840B.133120d', 'glove.840B.133376d', 'glove.840B.133632d', 'glove.840B.133888d', 'glove.840B.134144d', 'glove.840B.134400d', 'glove.840B.134656d', 'glove.840B.134912d', 'glove.840B.135168d', 'glove.840B.135424d', 'glove.840B.135680d', 'glove.840B.135936d', 'glove.840B.136192d', 'glove.840B.136448d', 'glove.840B.136704d', 'glove.840B.136960d', 'glove.840B.137216d', 'glove.840B.137472d', 'glove.840B.137728d', 'glove.840B.137984d', 'glove.840B.138240d', 'glove.840B.138496d', 'glove.840B.138752d', 'glove.840B.139008d', 'glove.840B.139264d', 'glove.840B.139520d', 'glove.840B.139776d', 'glove.840B.140032d', 'glove.840B.140288d', 'glove.840B.140544d', 'glove.840B.140800d', 'glove.840B.141056d', 'glove.840B.141312d', 'glove.840B.141568d', 'glove.840B.141824d', 'glove.840B.142080d', 'glove.840B.142336d', 'glove.840B.142592d', 'glove.840B.142848d', 'glove.840B.143104d', 'glove.840B.143360d', 'glove.840B.143616d', 'glove.840B.143872d', 'glove.840B.144128d', 'glove.840B.144384d', 'glove.840B.144640d', 'glove.840B.144896d', 'glove.840B.145152d', 'glove.840B.145408d', 'glove.840B.145664d', 'glove.840B.145920d', 'glove.840B.146176d', 'glove.840B.146432d', 'glove.840B.146688d', 'glove.840B.146944d', 'glove.840B.147200d', 'glove.840B.147456d', 'glove.840B.147712d', 'glove.840B.147968d', 'glove.840B.148224d', 'glove.840B.148480d', 'glove.840B.148736d', 'glove.840B.148992d', 'glove.840B.149248d', 'glove.840B.149504d', 'glove.840B.149760d', 'glove.840B.150016d', 'glove.840B.150272d', 'glove.840B.150528d', 'glove.840B.150784d', 'glove.840B.151040d', 'glove.840B.151296d', 'glove.840B.151552d', 'glove.840B.151808d', 'glove.840B.152064d', 'glove.840B.152320d', 'glove.840B.152576d', 'glove.840B.152832d', 'glove.840B.153088d', 'glove.840B.153344d', 'glove.840B.153600d', 'glove.840B.153856d', 'glove.840B.154112d', 'glove.840B.154368d', 'glove.840B.154624d', 'glove.840B.154880d', 'glove.840B.155136d', 'glove.840B.155392d', 'glove.840B.155648d', 'glove.840B.155904d', 'glove.840B.156160d', 'glove.840B.156416d', 'glove.840B.156672d', 'glove.840B.156928d', 'glove.840B.157184d', 'glove.840B.157440d', 'glove.840B.157696d', 'glove.840B.157952d', 'glove.840B.158208d', 'glove.840B.158464d', 'glove.840B.158720d', 'glove.840B.158976d', 'glove.840B.159232d', 'glove.840B.159488d', 'glove.840B.159744d', 'glove.840B.160000d', 'glove.840B.160256d', 'glove.840B.160512d', 'glove.840B.160768d', 'glove.840B.161024d', 'glove.840B.161280d', 'glove.840B.161536d', 'glove.840B.161792d', 'glove.840B.162048d', 'glove.840B.162304d', 'glove.840B.162560d', 'glove.840B.162816d', 'glove.840B.163072d', 'glove.840B.163328d', 'glove.840B.163584d', 'glove.840B.163840d', 'glove.840B.164096d', 'glove.840B.164352d', 'glove.840B.164608d', 'glove.840B.164864d', 'glove.840B.165120d', 'glove.840B.165376d', 'glove.840B.165632d', 'glove.840B.165888d', 'glove.840B.166144d', 'glove.840B.166400d', 'glove.840B.166656d', 'glove.840B.166912d', 'glove.840B.167168d', 'glove.840B.167424d', 'glove.840B.167680d', 'glove.840B.167936d', 'glove.840B.168192d', 'glove.840B.168448d', 'glove.840B.168704d', 'glove.840B.168960d', 'glove.840B.169216d', 'glove.840B.169472d', 'glove.840B.169728d', 'glove.840B.170000d', 'glove.840B.170256d', 'glove.840B.170512d', 'glove.840B.170768d', 'glove.840B.171024d', 'glove.840B.171280d', 'glove.840B.171536d', 'glove.840B.171792d', 'glove.840B.172048d', 'glove.840B.172304d', 'glove.840B.172560d', 'glove.840B.172816d', 'glove.840B.173072d', 'glove.840B.173328d', 'glove.840B.173584d', 'glove.840B.173840d', 'glove.840B.174096d', 'glove.840B.174352d', 'glove.840B.174608d', 'glove.840B.174864d', 'glove.840B.175120d', 'glove.840B.175376d', 'glove.840B.175632d', 'glove.840B.175888d', 'glove.840B.176144d', 'glove.840B.176400d', 'glove.840B.176656d', 'glove.840B.176912d', 'glove.840B.177168d', 'glove.840B.177424d', 'glove.840B.177680d', 'glove.840B.177936d', 'glove.840B.178192d', 'glove.840B.178448d', 'glove.840B.178704d', 'glove.840B.178960d', 'glove.840B.179216d', 'glove.840B.179472d', 'glove.840B.179728d', 'glove.840B.180000d', 'glove.840B.180256d', 'glove.840B.180512d', 'glove.840B.180768d', 'glove.840B.181024d', 'glove.840B.181280d', 'glove.840B.181536d', 'glove.840B.181792d', 'glove.840B.182048d', 'glove.840B.182
```

输出：

```
['glove.42B.300d',
 'glove.840B.300d',
 'glove.twitter.27B.25d',
 'glove.twitter.27B.50d',
 'glove.twitter.27B.100d',
 'glove.twitter.27B.200d',
 'glove.6B.50d',
 'glove.6B.100d',
 'glove.6B.200d',
 'glove.6B.300d']
```

预训练的GloVe模型的命名规范大致是“模型.(数据集.)数据集词数.词向量维度”。更多信息可以参考GloVe和fastText的项目网站[1,2]。下面我们使用基于维基百科子集预训练的50维GloVe词向量。第一次创建预训练词向量实例时会自动下载相应的词向量到 `cache` 指定文件夹（默认为 `.vector_cache`），因此需要联网。

```
cache_dir = "/Users/tangshusen/Datasets/glove"
# glove = vocab.pretrained_aliases["glove.6B.50d"](cache=cache_dir)
glove = vocab.GloVe(name='6B', dim=50, cache=cache_dir) # 与上面等价
```

返回的实例主要有以下三个属性：

- `stoi` : 词到索引的字典；
- `itos` : 一个列表，索引到词的映射；
- `vectors` : 词向量。

打印词典大小。其中含有40万个词。

```
print("一共包含%d个词。" % len(glove.stoi))
```

输出：

```
一共包含 400000 个词。
```

我们可以通过词来获取它在词典中的索引，也可以通过索引获取词。

```
glove.stoi['beautiful'], glove.itos[3366] # (3366, 'beautiful')
```

10.6.2 应用预训练词向量

下面我们以GloVe模型为例，展示预训练词向量的应用。

10.6.2.1 求近义词

这里重新实现10.3节（word2vec的实现）中介绍过的使用余弦相似度来搜索近义词的算法。为了在求类比词时重用其中的求 k 近邻（ k -nearest neighbors）的逻辑，我们将这部分逻辑单独封装在 `knn` 函数中。

```
def knn(W, x, k):
    # 添加的1e-9是为了数值稳定性
    cos = torch.matmul(W, x.view((-1,))) / (
        (torch.sum(W * W, dim=1) + 1e-9).sqrt() * torch.sum(x * x).sqrt())
    _, topk = torch.topk(cos, k=k)
    topk = topk.cpu().numpy()
    return topk, [cos[i].item() for i in topk]
```

然后，我们通过预训练词向量实例 `embed` 来搜索近义词。

```
def get_similar_tokens(query_token, k, embed):
    topk, cos = knn(embed.vectors,
                    embed.vectors[embed.stoi[query_token]], k+1)
    for i, c in zip(topk[1:], cos[1:]): # 除去输入词
        print('cosine sim=%.3f: %s' % (c, (embed.itos[i])))
```

已创建的预训练词向量实例 `glove_6b50d` 的词典中含40万个词和1个特殊的未知词。除去输入词和未知词，我们从中搜索与“chip”语义最相近的3个词。

```
get_similar_tokens('chip', 3, glove)
```

输出:

```
cosine sim=0.856: chips
cosine sim=0.749: intel
cosine sim=0.749: electronics
```

接下来查找“baby”和“beautiful”的近义词。

```
get_similar_tokens('baby', 3, glove)
```

输出:

```
cosine sim=0.839: babies
cosine sim=0.800: boy
cosine sim=0.792: girl
```

```
get_similar_tokens('beautiful', 3, glove)
```

输出:

```
cosine sim=0.921: lovely
cosine sim=0.893: gorgeous
cosine sim=0.830: wonderful
```

10.6.2.2 求类比词

除了求近义词以外，我们还可以使用预训练词向量求词与词之间的类比关系。例如，“man”（男人）：“woman”（女人）:: “son”（儿子）：“daughter”（女儿）是一个类比例子：“man”之于“woman”相当于“son”之于“daughter”。求类比词问题可以定义为：对于类比关系中的4个词 $a : b :: c : d$ ，给定前3个词 a 、 b 和 c ，求 d 。设词 w 的词向量为 $\text{vec}(w)$ 。求类比词的思路是，搜索与 $\text{vec}(c) + \text{vec}(b) - \text{vec}(a)$ 的结果向量最相似的词向量。

```
def get_analogy(token_a, token_b, token_c, embed):  
    vecs = [embed.vectors[embed.stoi[t]]  
             for t in [token_a, token_b, token_c]]  
    x = vecs[1] - vecs[0] + vecs[2]  
    topk, cos = knn(embed.vectors, x, 1)  
    return embed.itos[topk[0]]
```

验证一下“男-女”类比。

```
get_analogy('man', 'woman', 'son', glove) # 'daughter'
```

“首都-国家”类比：“beijing”（北京）之于“china”（中国）相当于“tokyo”（东京）之于什么？答案应该是“japan”（日本）。

```
get_analogy('beijing', 'china', 'tokyo', glove) # 'japan'
```

“形容词-形容词最高级”类比：“bad”（坏的）之于“worst”（最坏的）相当于“big”（大的）之于什么？答案应该是“biggest”（最大的）。

```
get_analogy('bad', 'worst', 'big', glove) # 'biggest'
```

“动词一般时-动词过去时”类比：“do”（做）之于“did”（做过）相当于“go”（去）之于什么？答案应该是“went”（去过）。

```
get_analogy('do', 'did', 'go', glove) # 'went'
```

小结

- 在大规模语料上预训练的词向量常常可以应用于下游自然语言处理任务中。
- 可以应用预训练的词向量求近义词和类比词。