In [26]:

```
import json
from sklearn.feature_extraction.text import TfidfVectorizer
import os
import re

def get_fnames():
    """Read all text files in a folder.
    """
    fnames = []
    for root,_,files in os.walk("./abstracts/awards_2002"):
        for fname in files:
            if fname[-4:] == ".txt":
                fnames.append(os.path.join(root, fname))
    return fnames

print("Number of abstracts in folder awards_2002: {}".format(len(get_fnames()
```

```
Number of abstracts in folder awards_2002: 9923
```

In [27]:

```
name_list = get_fnames()

def read_file(fname):
    with open(fname, 'r',encoding="ISO-8859-1") as f:
        # skip all lines until abstract
        for line in f:
            if "Abstract    :" in line:
                break

        # get abstract as a single string
        abstract = ' '.join([line[:-1].strip() for line in f])
        abstract = re.sub(' +', ' ', abstract)  # remove double spaces
        return abstract
```

In [28]:

```
documents = []

for i in name_list:
    documents.append(read_file(i))
```

In [29]:

```
# Fast and simple tokenization
new_vectorizer = TfidfVectorizer(stop_words = 'english', lowercase= True, ngr
word_tokenizer = new_vectorizer.build_tokenizer()
tokenized_text = [word_tokenizer(doc) for doc in documents]
```

In [30]:

```
tokenized_text[0]
```

Out[30]:

```
['This',
            'Small',
            'Business',
            'Innovation',
            'Research',
            'SBIR',
            'Phase',
            'II',
            'Project',
            'proposes',
            'to',
            'develop',
            'the',
            'database',
            'and',
            'associated',
            'software',
            'to',
            'enable',
            'analysis',
            'of',
            'protein',
            'trafficking',
            'and',
            'localization',
            'The',
            'system',
            'will',
            'be',
            'designed',
            'to',
            'enable',
            'drug',
            'discovery',
            'researchers',
            'to',
            'identify',
            'elucidate',
            'eliminate',
            'and',
            'design',
            'leads',
            'and',
            'targets',
            'while',
            'facilitating',
            'the',
            'general',
            'training',
            'of',
            'researchers',
            'During',
            'the',
            'Phase',
            'work',
            'proteins',
            'involved',
            'in',
            'trafficking',
            'and',
            'diseases',
            'related',
            'to',
            'mislocalization',
            'were',
            'identified',
            'and',
            'relational',
            'database',
            'to',
            'house',
            'information',
            'on',
            'protein',
            'trafficking',
            'was',
            'constructed',
            'Curation',
            'interface',
            'applications',
            'were',
            'created',
            'to',
            'allow',
            'remote',
            'data',
            'entry',
            'and',
            'graphical',
            'user',
            'interfaces',
            'designed',
            'to',
            'maximize',
            'the',
            'utility',
            'of',
            'the',
            'information',
            'The',
            'objective',
            'of',
            'this',
            'Phase',
            'II',
            'Project',
            'is',
            'to',
            'exhaustively',
            'populate',
            'the',
            'database',
            'from',
            'the',
            'primary',
            'journal',
            'literature',
            'Selection',
            'of',
            'proteins',
            'involved',
            'in',
            'protein',
            'trafficking',
            'will',
            'be',
            'guided',
            'by',
            'relevant',
            'human',
            'diseases',
            'and',
            'corresponding',
            'drug',
            'discovery',
            'efforts',
            'The',
            'commercial',
            'application',
            'of',
            'this',
            'project',
            'is',
            'in',
            'the',
            'area',
            'of',
            'biological',
            'informatics',
            'The',
            'potential',
            'users',
            'of',
            'the',
            'biological',
            'database',
            'to',
            'be',
            'developed',
            'in',
            'this',
            'project',
            'would',
            'include',
            'pharmaceutical',
            'and',
            'drug',
            'discovery',
            'companies']
```

In [31]:

```
### Train word vectors

import gensim # Make sure you also have cython installed to accelerate comput

import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level

# Train word2vec model
vectors = gensim.models.Word2Vec(tokenized_text, size=100, min_count=5, sg=1,
```

```
2021-02-26 21:45:03,167 : INFO : collecting all words and their counts
2021-02-26 21:45:03,168 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 21:45:03,905 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 21:45:03,906 : INFO : Loading a fresh vocabulary
2021-02-26 21:45:04,022 : INFO : effective_min_count=5 retains 20752 unique wo
rds (32% of original 63538, drops 42786)
2021-02-26 21:45:04,023 : INFO : effective_min_count=5 leaves 2585188 word cor
pus (97% of original 2656274, drops 71086)
2021-02-26 21:45:04,152 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 21:45:04,154 : INFO : sample=0.001 downsamples 26 most-common words
2021-02-26 21:45:04,155 : INFO : downsampling leaves estimated 2005620 word co
rpus (77.6% of prior 2585188)
2021-02-26 21:45:04,223 : INFO : estimated required memory for 20752 words and
100 dimensions: 26977600 bytes
2021-02-26 21:45:04,224 : INFO : resetting layer weights
2021-02-26 21:45:11,088 : INFO : training model with 4 workers on 20752 vocabu
lary and 100 features, using sg=1 hs=0 sample=0.001 negative=5 window=5
2021-02-26 21:45:12,121 : INFO : EPOCH 1 - PROGRESS: at 18.65% examples, 35364
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:13,138 : INFO : EPOCH 1 - PROGRESS: at 36.28% examples, 35218
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:14,145 : INFO : EPOCH 1 - PROGRESS: at 52.62% examples, 34535
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:15,156 : INFO : EPOCH 1 - PROGRESS: at 70.68% examples, 34554
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:16,172 : INFO : EPOCH 1 - PROGRESS: at 86.79% examples, 34232
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:16,912 : INFO : worker thread finished; awaiting finish of 3
more threads
```

```
2021-02-26 21:45:16,920 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:45:16,929 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:45:16,958 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:45:16,959 : INFO : EPOCH - 1 : training on 2656274 raw words (20
06689 effective words) took 5.9s, 341953 effective words/s
2021-02-26 21:45:17,985 : INFO : EPOCH 2 - PROGRESS: at 13.71% examples, 26892
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:18,990 : INFO : EPOCH 2 - PROGRESS: at 29.48% examples, 28972
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:20,000 : INFO : EPOCH 2 - PROGRESS: at 41.87% examples, 27874
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:21,007 : INFO : EPOCH 2 - PROGRESS: at 58.02% examples, 28652
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:22,014 : INFO : EPOCH 2 - PROGRESS: at 74.59% examples, 29411
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:23,017 : INFO : EPOCH 2 - PROGRESS: at 88.51% examples, 29337
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:23,694 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:45:23,717 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:45:23,725 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:45:23,745 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:45:23,746 : INFO : EPOCH - 2 : training on 2656274 raw words (20
06413 effective words) took 6.8s, 295817 effective words/s
2021-02-26 21:45:24,773 : INFO : EPOCH 3 - PROGRESS: at 14.12% examples, 27553
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:25,779 : INFO : EPOCH 3 - PROGRESS: at 23.78% examples, 23065
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:26,780 : INFO : EPOCH 3 - PROGRESS: at 39.77% examples, 26214
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:27,790 : INFO : EPOCH 3 - PROGRESS: at 56.02% examples, 27743
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:28,818 : INFO : EPOCH 3 - PROGRESS: at 72.95% examples, 28563
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:29,849 : INFO : EPOCH 3 - PROGRESS: at 88.51% examples, 29097
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:30,563 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:45:30,574 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:45:30,579 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:45:30,600 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:45:30,601 : INFO : EPOCH - 3 : training on 2656274 raw words (20
04774 effective words) took 6.9s, 292597 effective words/s
2021-02-26 21:45:31,671 : INFO : EPOCH 4 - PROGRESS: at 15.33% examples, 28609
```

```
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:32,671 : INFO : EPOCH 4 - PROGRESS: at 30.43% examples, 29451
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:33,681 : INFO : EPOCH 4 - PROGRESS: at 45.08% examples, 29697
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:34,689 : INFO : EPOCH 4 - PROGRESS: at 62.01% examples, 30182
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:35,715 : INFO : EPOCH 4 - PROGRESS: at 77.35% examples, 30370
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:36,728 : INFO : EPOCH 4 - PROGRESS: at 93.23% examples, 30548
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:37,112 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:45:37,113 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:45:37,124 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:45:37,172 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:45:37,174 : INFO : EPOCH - 4 : training on 2656274 raw words (20
04878 effective words) took 6.6s, 305244 effective words/s
2021-02-26 21:45:38,210 : INFO : EPOCH 5 - PROGRESS: at 15.33% examples, 29550
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:39,272 : INFO : EPOCH 5 - PROGRESS: at 31.49% examples, 30164
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:40,339 : INFO : EPOCH 5 - PROGRESS: at 47.34% examples, 30312
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:41,383 : INFO : EPOCH 5 - PROGRESS: at 63.08% examples, 29859
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:42,390 : INFO : EPOCH 5 - PROGRESS: at 76.19% examples, 29214
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:43,431 : INFO : EPOCH 5 - PROGRESS: at 90.99% examples, 29222
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:45:44,006 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:45:44,024 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:45:44,030 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:45:44,065 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:45:44,066 : INFO : EPOCH - 5 : training on 2656274 raw words (20
06161 effective words) took 6.9s, 291208 effective words/s
2021-02-26 21:45:44,067 : INFO : training on a 13281370 raw words (10028915 ef
fective words) took 33.0s, 304099 effective words/s
```

In [32]:

```
seed_word = [list(vectors.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [33]:

```
seed_word
```

Out[33]:
```
['greatly', 'every', 'includes', 'transcription', 'largest']
```

In [18]:
```python
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'greatly')
print(vectors.wv.most_similar('greatly'))
print()
```

```
2021-02-26 13:33:02,414 : INFO : precomputing L2-norms of word weight vectors
Most similar to: greatly
[('significantly', 0.8013008236885071), ('substantially', 0.7442965507507324),
('dramatically', 0.7396800518035889), ('vastly', 0.667165994644165), ('enlarge
', 0.6540445685386658), ('MPICH', 0.6524852514266968), ('streamline', 0.651308
9537620544), ('Successful', 0.6315666437149048), ('capability', 0.627498149871
8262), ('sharpen', 0.6218451261520386)]
```

In [19]:
```python
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'every')
print(vectors.wv.most_similar('every'))
print()
```

```
Most similar to: every
[('each', 0.6906091570854187), ('Every', 0.6714961528778076), ('almost', 0.660
5528593063354), ('essentially', 0.6477319002151489), ('normally', 0.6423296332
359314), ('judiciously', 0.6373945474624634), ('roughly', 0.6333438158035278),
('nearly', 0.6281598806381226), ('again', 0.6260768175125122), ('morning', 0.6
238182783126831)]
```

In [20]:
```python
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'includes')
print(vectors.wv.most_similar('includes'))
print()
```

```
Most similar to: includes
[('involves', 0.7543442249298096), ('consists', 0.7009209990501404), ('compris
es', 0.6838739514350891), ('encompasses', 0.6799794435501099), ('supports', 0.
6580832004547119), ('emphasizes', 0.6541271805763245), ('involve', 0.653008341
7892456), ('include', 0.6529262065887451), ('integrates', 0.6281484961509705),
('introduces', 0.6225360631942749)]
```

In [21]:
```python
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'transcription')
print(vectors.wv.most_similar('transcription'))
print()
```

```
Most similar to: transcription
[('chromatin', 0.8633143305778503), ('silencing', 0.8584595918655396), ('mRNA'
, 0.8561317920684814), ('meiotic', 0.8526526689529419), ('transcriptional', 0.
8488618731498718), ('repressor', 0.8470075130462646), ('homologous', 0.8436334
729194641), ('replication', 0.8433512449264526), ('mRNAs', 0.8415851593017578)
, ('virulence', 0.8390907645225525)]
```

In [22]:

```python
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'largest')
print(vectors.wv.most_similar('largest'))
print()
```

```
Most similar to: largest
[('oldest', 0.739079475402832), ('populous', 0.7280933856964111), ('southeaste
rn', 0.7225464582443237), ('towns', 0.7119206190109253), ('rest', 0.7053264379
501343), ('richest', 0.6938225626945496), ('province', 0.6938113570213318), ('
northeastern', 0.6901578307151794), ('deepest', 0.6891354322433472), ('endemic
', 0.6792290210723877)]
```

In [23]:

```python
vectors2 = gensim.models.Word2Vec(tokenized_text, size=10, min_count=1, sg=0,
```

```
2021-02-26 13:33:28,249 : WARNING : consider setting layer size to a multiple
of 4 for greater performance
2021-02-26 13:33:28,251 : INFO : collecting all words and their counts
2021-02-26 13:33:28,252 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 13:33:28,806 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 13:33:28,808 : INFO : Loading a fresh vocabulary
2021-02-26 13:33:29,051 : INFO : effective_min_count=1 retains 63538 unique wo
rds (100% of original 63538, drops 0)
2021-02-26 13:33:29,053 : INFO : effective_min_count=1 leaves 2656274 word cor
pus (100% of original 2656274, drops 0)
2021-02-26 13:33:29,346 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 13:33:29,348 : INFO : sample=0.001 downsamples 25 most-common words
2021-02-26 13:33:29,350 : INFO : downsampling leaves estimated 2081427 word co
rpus (78.4% of prior 2656274)
2021-02-26 13:33:29,524 : INFO : estimated required memory for 63538 words and
10 dimensions: 36852040 bytes
2021-02-26 13:33:29,525 : INFO : resetting layer weights
2021-02-26 13:33:46,624 : INFO : training model with 4 workers on 63538 vocabu
lary and 10 features, using sg=0 hs=0 sample=0.001 negative=5 window=5
2021-02-26 13:33:47,631 : INFO : EPOCH 1 - PROGRESS: at 84.85% examples, 17534
99 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:33:47,808 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:33:47,809 : INFO : worker thread finished; awaiting finish of 2
more threads
```

```
2021-02-26 13:33:47,812 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:33:47,817 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:33:47,818 : INFO : EPOCH - 1 : training on 2656274 raw words (20
81355 effective words) took 1.2s, 1750148 effective words/s
2021-02-26 13:33:48,821 : INFO : EPOCH 2 - PROGRESS: at 85.16% examples, 17637
94 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:33:48,994 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:33:48,995 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:33:48,999 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:33:49,003 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:33:49,004 : INFO : EPOCH - 2 : training on 2656274 raw words (20
80906 effective words) took 1.2s, 1758384 effective words/s
2021-02-26 13:33:50,007 : INFO : EPOCH 3 - PROGRESS: at 85.46% examples, 17719
85 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:33:50,176 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:33:50,177 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:33:50,180 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:33:50,184 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:33:50,185 : INFO : EPOCH - 3 : training on 2656274 raw words (20
80807 effective words) took 1.2s, 1765807 effective words/s
2021-02-26 13:33:51,189 : INFO : EPOCH 4 - PROGRESS: at 85.46% examples, 17708
08 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:33:51,360 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:33:51,361 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:33:51,365 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:33:51,369 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:33:51,370 : INFO : EPOCH - 4 : training on 2656274 raw words (20
81643 effective words) took 1.2s, 1761448 effective words/s
2021-02-26 13:33:52,374 : INFO : EPOCH 5 - PROGRESS: at 80.05% examples, 16615
58 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:33:52,686 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:33:52,687 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:33:52,692 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:33:52,697 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:33:52,698 : INFO : EPOCH - 5 : training on 2656274 raw words (20
81025 effective words) took 1.3s, 1569930 effective words/s
2021-02-26 13:33:52,699 : INFO : training on a 13281370 raw words (10405736 ef
fective words) took 6.1s, 1713386 effective words/s
```

In [24]:

```
seed_word2 = [list(vectors2.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [25]:

```
seed_word2
```

Out[25]:

```
['half', 'whether', 'neuroscience', 'Lonza', 'tightly']
```

In [26]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'half')
print(vectors2.wv.most_similar('half'))
print()
```

```
2021-02-26 13:35:45,115 : INFO : precomputing L2-norms of word weight vectors
Most similar to: half
[('thousand', 0.9656727313995361), ('80', 0.9651178121566772), ('percent', 0.9
607995748519897), ('70', 0.9601696729660034), ('torn', 0.9574507474899292), ('
parallelepipeds', 0.953171968460083), ('roughly', 0.9500912427902222), ('milli
on', 0.9493248462677002), ('catecholamines', 0.9489266872406006), ('Dickenson'
, 0.9426918029785156)]
```

In [27]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'whether')
print(vectors2.wv.most_similar('whether'))
print()
```

```
Most similar to: whether
[('what', 0.9122190475463867), ('macromodels', 0.9094517230987549), ('why', 0.
9062631130218506), ('fissions', 0.8964521288871765), ('how', 0.891663193702697
8), ('farmer', 0.8860977292060852), ('SWING', 0.8760378956794739), ('Proximate
', 0.870087206363678), ('if', 0.8693884611129761), ('manhood', 0.8572323918342
59)]
```

In [28]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors2.wv.most_similar('neuroscience'))
print()
```

```
Most similar to: neuroscience
[('populace', 0.9711309671401978), ('LCLUC', 0.9577138423919678), ('focusing',
0.9521087408065796), ('0231010', 0.9496864676475525), ('macroeconomics', 0.947
7670192718506), ('lifeline', 0.9440898895263672), ('codimension', 0.9428080320
358276), ('sustainability', 0.941412627696991), ('naturalization', 0.931682944
2977905), ('arena', 0.9310978651046753)]
```

In [29]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'Lonza')
print(vectors2.wv.most_similar('Lonza'))
print()
```

```
Most similar to: Lonza
[('Flory', 0.9816685318946838), ('nontransforming', 0.9795741438865662), ('Num
erous', 0.9769992828369141), ('SAR324', 0.9755681157112122), ('dictionaries',
0.9745925664901733), ('graphein', 0.9740005731582642), ('discordant', 0.973209
798336029), ('Ibn', 0.9724632501602173), ('pyridylmethyl', 0.972217321395874),
('sulfoxides', 0.9721080660820007)]
```

In [30]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'tightly')
print(vectors2.wv.most_similar('tightly'))
print()
```

```
Most similar to: tightly
[('spectrally', 0.974089503288269), ('screened', 0.9697147607803345), ('seedin
g', 0.9688395857810974), ('reversible', 0.9639790654182434), ('sum', 0.9636118
412017822), ('periodically', 0.9626333713531494), ('computa', 0.96170574426651
), ('inconvenient', 0.9616869688034058), ('prescribed', 0.9607305526733398), (
'corrections', 0.9584357142448425)]
```

In [31]:

```
vectors3 = gensim.models.Word2Vec(tokenized_text, size=100, min_count=5, sg=0
```

```
2021-02-26 13:36:05,065 : INFO : collecting all words and their counts
2021-02-26 13:36:05,067 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 13:36:05,624 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 13:36:05,625 : INFO : Loading a fresh vocabulary
2021-02-26 13:36:05,781 : INFO : effective_min_count=5 retains 20752 unique wo
rds (32% of original 63538, drops 42786)
2021-02-26 13:36:05,782 : INFO : effective_min_count=5 leaves 2585188 word cor
pus (97% of original 2656274, drops 71086)
2021-02-26 13:36:05,888 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 13:36:05,890 : INFO : sample=0.001 downsamples 26 most-common words
2021-02-26 13:36:05,891 : INFO : downsampling leaves estimated 2005620 word co
rpus (77.6% of prior 2585188)
```

```
2021-02-26 13:36:05,956 : INFO : estimated required memory for 20752 words and
100 dimensions: 26977600 bytes
2021-02-26 13:36:05,957 : INFO : resetting layer weights
2021-02-26 13:36:11,658 : INFO : training model with 4 workers on 20752 vocabu
lary and 100 features, using sg=0 hs=0 sample=0.001 negative=5 window=5
2021-02-26 13:36:12,666 : INFO : EPOCH 1 - PROGRESS: at 71.89% examples, 14199
62 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:36:13,067 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:36:13,068 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:36:13,072 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:36:13,077 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:36:13,078 : INFO : EPOCH - 1 : training on 2656274 raw words (20
05182 effective words) took 1.4s, 1415989 effective words/s
2021-02-26 13:36:14,084 : INFO : EPOCH 2 - PROGRESS: at 72.61% examples, 14368
65 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:36:14,463 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:36:14,464 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:36:14,469 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:36:14,473 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:36:14,473 : INFO : EPOCH - 2 : training on 2656274 raw words (20
05550 effective words) took 1.4s, 1440367 effective words/s
2021-02-26 13:36:15,478 : INFO : EPOCH 3 - PROGRESS: at 71.05% examples, 14088
87 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:36:15,888 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:36:15,889 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:36:15,894 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:36:15,899 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:36:15,900 : INFO : EPOCH - 3 : training on 2656274 raw words (20
05873 effective words) took 1.4s, 1409233 effective words/s
2021-02-26 13:36:16,906 : INFO : EPOCH 4 - PROGRESS: at 67.00% examples, 13244
24 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:36:17,383 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:36:17,384 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:36:17,388 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:36:17,394 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:36:17,395 : INFO : EPOCH - 4 : training on 2656274 raw words (20
05517 effective words) took 1.5s, 1343624 effective words/s
```

```
2021-02-26 13:36:18,399 : INFO : EPOCH 5 - PROGRESS: at 67.71% examples, 13429
06 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:36:18,872 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:36:18,873 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:36:18,877 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:36:18,884 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:36:18,884 : INFO : EPOCH - 5 : training on 2656274 raw words (20
06176 effective words) took 1.5s, 1349448 effective words/s
2021-02-26 13:36:18,885 : INFO : training on a 13281370 raw words (10028298 ef
fective words) took 7.2s, 1387816 effective words/s
```

In [32]:

```
seed_word3 = [list(vectors3.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [33]:

```
seed_word3
```

Out[33]:

```
['greatly', 'every', 'includes', 'transcription', 'largest']
```

In [34]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'greatly')
print(vectors3.wv.most_similar('greatly'))
print()
```

```
2021-02-26 13:36:24,757 : INFO : precomputing L2-norms of word weight vectors
Most similar to: greatly
[('significantly', 0.8782045245170593), ('substantially', 0.7808791399002075),
('dramatically', 0.6995285749435425), ('ultimately', 0.6894906163215637), ('ca
pability', 0.6457507014274597), ('improved', 0.638140082359314), ('enhanced',
0.629065215587616), ('thus', 0.6160373091697693), ('increased', 0.591135203838
3484), ('improvements', 0.5893326997756958)]
```

In [35]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'every')
print(vectors3.wv.most_similar('every'))
print()
```

```
Most similar to: every
[('ten', 0.6816279888153076), ('almost', 0.6725512742996216), ('few', 0.670748
9490509033), ('half', 0.6691499352455139), ('except', 0.6654048562049866), ('r
oughly', 0.663583517074585), ('least', 0.6603336334228516), ('per', 0.65591037
27340698), ('old', 0.6518365740776062), ('billion', 0.651665449142456)]
```

In [36]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'includes')
print(vectors3.wv.most_similar('includes'))
print()
```

```
Most similar to: includes
[('involves', 0.8537815809249878), ('combines', 0.7831259369850159), ('support
s', 0.7734456062316895), ('integrates', 0.7514086961746216), ('emphasizes', 0.
7439501881599426), ('utilizes', 0.7368146777153015), ('consists', 0.7192471623
420715), ('develops', 0.7155880331993103), ('encompasses', 0.7074522972106934)
, ('explores', 0.6937383413314819)]
```

In [37]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'transcription')
print(vectors3.wv.most_similar('transcription'))
print()
```

```
Most similar to: transcription
[('eukaryotic', 0.9008668661117554), ('transcriptional', 0.900171160697937), (
'replication', 0.8730258941650391), ('receptor', 0.8679792881011963), ('intrac
ellular', 0.859769880771637), ('signaling', 0.8548659086227417), ('cis', 0.854
0043234825134), ('putative', 0.847966194152832), ('silencing', 0.8417541384696
96), ('viral', 0.8399631977081299)]
```

In [38]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'largest')
print(vectors3.wv.most_similar('largest'))
print()
```

```
Most similar to: largest
[('oldest', 0.852716326713562), ('earliest', 0.7771295309066772), ('north', 0.
7501647472381592), ('province', 0.7465780973434448), ('Appalachian', 0.7434969
544410706), ('Great', 0.7431834936141968), ('island', 0.7421600222587585), ('r
ichest', 0.738149881362915), ('Asian', 0.7377354502677917), ('southeastern', 0
.7294509410858154)]
```

In [34]:

```
vectors4 = gensim.models.Word2Vec(tokenized_text, size=10, min_count=1, sg=1,
```

```
2021-02-26 21:46:04,472 : WARNING : consider setting layer size to a multiple
of 4 for greater performance
2021-02-26 21:46:04,474 : INFO : collecting all words and their counts
2021-02-26 21:46:04,475 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 21:46:05,118 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 21:46:05,119 : INFO : Loading a fresh vocabulary
```

```
2021-02-26 21:46:16,734 : INFO : effective_min_count=1 retains 63538 unique wo
rds (100% of original 63538, drops 0)
2021-02-26 21:46:16,736 : INFO : effective_min_count=1 leaves 2656274 word cor
pus (100% of original 2656274, drops 0)
2021-02-26 21:46:17,180 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 21:46:17,183 : INFO : sample=0.001 downsamples 25 most-common words
2021-02-26 21:46:17,184 : INFO : downsampling leaves estimated 2081427 word co
rpus (78.4% of prior 2656274)
2021-02-26 21:46:17,408 : INFO : estimated required memory for 63538 words and
10 dimensions: 36852040 bytes
2021-02-26 21:46:17,409 : INFO : resetting layer weights
2021-02-26 21:46:38,719 : INFO : training model with 4 workers on 63538 vocabu
lary and 10 features, using sg=1 hs=0 sample=0.001 negative=5 window=5
2021-02-26 21:46:39,727 : INFO : EPOCH 1 - PROGRESS: at 22.66% examples, 46071
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:40,752 : INFO : EPOCH 1 - PROGRESS: at 42.16% examples, 43670
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:41,756 : INFO : EPOCH 1 - PROGRESS: at 64.97% examples, 44252
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:42,759 : INFO : EPOCH 1 - PROGRESS: at 85.80% examples, 44130
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:43,394 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:46:43,404 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:46:43,406 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:46:43,427 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:46:43,428 : INFO : EPOCH - 1 : training on 2656274 raw words (20
82099 effective words) took 4.7s, 442510 effective words/s
2021-02-26 21:46:44,474 : INFO : EPOCH 2 - PROGRESS: at 20.11% examples, 39185
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:45,484 : INFO : EPOCH 2 - PROGRESS: at 38.47% examples, 38625
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:46,486 : INFO : EPOCH 2 - PROGRESS: at 55.54% examples, 37842
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:47,504 : INFO : EPOCH 2 - PROGRESS: at 74.29% examples, 37648
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:48,538 : INFO : EPOCH 2 - PROGRESS: at 91.71% examples, 37421
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:48,901 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:46:48,929 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:46:48,936 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:46:48,953 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:46:48,954 : INFO : EPOCH - 2 : training on 2656274 raw words (20
81026 effective words) took 5.5s, 376793 effective words/s
2021-02-26 21:46:50,000 : INFO : EPOCH 3 - PROGRESS: at 18.65% examples, 36290
```

```
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:51,029 : INFO : EPOCH 3 - PROGRESS: at 35.55% examples, 35311
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:52,050 : INFO : EPOCH 3 - PROGRESS: at 53.31% examples, 35887
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:53,072 : INFO : EPOCH 3 - PROGRESS: at 72.95% examples, 36531
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:54,084 : INFO : EPOCH 3 - PROGRESS: at 90.62% examples, 36830
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:54,555 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:46:54,561 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:46:54,574 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:46:54,590 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:46:54,592 : INFO : EPOCH - 3 : training on 2656274 raw words (20
81151 effective words) took 5.6s, 369400 effective words/s
2021-02-26 21:46:55,596 : INFO : EPOCH 4 - PROGRESS: at 19.00% examples, 38581
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:56,619 : INFO : EPOCH 4 - PROGRESS: at 39.12% examples, 39987
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:57,620 : INFO : EPOCH 4 - PROGRESS: at 59.16% examples, 40551
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:58,631 : INFO : EPOCH 4 - PROGRESS: at 78.77% examples, 40696
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:46:59,644 : INFO : EPOCH 4 - PROGRESS: at 98.11% examples, 40453
5 words/s, in_qsize 6, out_qsize 0
2021-02-26 21:46:59,688 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:46:59,703 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:46:59,719 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:46:59,742 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:46:59,743 : INFO : EPOCH - 4 : training on 2656274 raw words (20
82136 effective words) took 5.1s, 404494 effective words/s
2021-02-26 21:47:00,793 : INFO : EPOCH 5 - PROGRESS: at 18.65% examples, 36183
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:47:01,815 : INFO : EPOCH 5 - PROGRESS: at 36.97% examples, 36908
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:47:02,838 : INFO : EPOCH 5 - PROGRESS: at 56.91% examples, 38175
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:47:03,857 : INFO : EPOCH 5 - PROGRESS: at 76.19% examples, 38464
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:47:04,873 : INFO : EPOCH 5 - PROGRESS: at 93.60% examples, 38056
0 words/s, in_qsize 8, out_qsize 1
2021-02-26 21:47:05,223 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:47:05,233 : INFO : worker thread finished; awaiting finish of 2
more threads
```

```
2021-02-26 21:47:05,239 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:47:05,246 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:47:05,247 : INFO : EPOCH - 5 : training on 2656274 raw words (20
82719 effective words) took 5.5s, 378661 effective words/s
2021-02-26 21:47:05,248 : INFO : training on a 13281370 raw words (10409131 ef
fective words) took 26.5s, 392372 effective words/s
```

In [35]:

```
seed_word4 = [list(vectors4.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [36]:

```
seed_word4
```

Out[36]:

```
['half', 'whether', 'neuroscience', 'Lonza', 'tightly']
```

In [37]:

```
len(list(vectors4.wv.vocab.keys()))
```

Out[37]:

```
63538
```

In [42]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'half')
print(vectors4.wv.most_similar('half'))
print()
```

```
2021-02-26 13:37:38,820 : INFO : precomputing L2-norms of word weight vectors
Most similar to: half
[('representing', 0.9732823371887207), ('nearly', 0.972795307636261), ('black'
, 0.9664586186408997), ('least', 0.9610405564308167), ('few', 0.95456826686859
13), ('around', 0.9482489824295044), ('every', 0.9429680705070496), ('covering
', 0.9417109489440918), ('over', 0.9415010809898376), ('spanning', 0.940800130
367279)]
```

In [43]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'whether')
print(vectors4.wv.most_similar('whether'))
print()
```

```
Most similar to: whether
[('explain', 0.9642354249954224), ('skew', 0.9559688568115234), ('hypothesized
', 0.9536693096160889), ('endogenous', 0.9525479078292847), ('trait', 0.950679
6598434448), ('what', 0.9503313899040222), ('ask', 0.9473053812980652), ('if',
0.9449271559715271), ('asymmetry', 0.9439817667007446), ('unknown', 0.94338172
67417908)]
```

In [38]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors4.wv.most_similar('neuroscience'))
print()
```

```
2021-02-26 21:50:18,179 : INFO : precomputing L2-norms of word weight vectors
Most similar to: neuroscience
[('furthering', 0.9842066764831543), ('informational', 0.9749946594238281), ('
relevance', 0.9741153717041016), ('neurobiology', 0.96800696849823), ('interes
t', 0.9663718342781067), ('informatics', 0.9641363620758057), ('technological'
, 0.9639126062393188), ('advancing', 0.9617303609848022), ('playing', 0.959730
1483154297), ('sciences', 0.9593894481658936)]
```

In [45]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'Lonza')
print(vectors4.wv.most_similar('Lonza'))
print()
```

```
Most similar to: Lonza
[('domes', 0.997254490852356), ('Curation', 0.9971436262130737), ('CCC', 0.996
4454174041748), ('HPF', 0.9954668283462524), ('ImmunoPrecipitation', 0.9951314
330101013), ('Holographic', 0.9947695136070251), ('eyewear', 0.994498789310455
3), ('electrogenerated', 0.9944267272949219), ('Quantized', 0.9943544268608093
), ('HiSS', 0.9943377375602722)]
```

In [46]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'tightly')
print(vectors4.wv.most_similar('tightly'))
print()
```

```
Most similar to: tightly
[('routes', 0.9862256646156311), ('manipulating', 0.9775048494338989), ('tailo
r', 0.9773040413856506), ('inductive', 0.9732342958450317), ('reverse', 0.9725
438952445984), ('render', 0.9721719622612), ('domain', 0.9718952178955078), ('
channels', 0.9717416763305664), ('resultant', 0.9705893397331238), ('tuned', 0
.9697235226631165)]
```

In [47]:

```
vectors5 = gensim.models.Word2Vec(tokenized_text, size=100, min_count=1, sg=1
```

```
2021-02-26 13:37:57,083 : INFO : collecting all words and their counts
2021-02-26 13:37:57,084 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 13:37:57,617 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 13:37:57,618 : INFO : Loading a fresh vocabulary
2021-02-26 13:37:57,880 : INFO : effective_min_count=1 retains 63538 unique wo
```

```
rds (100% of original 63538, drops 0)
2021-02-26 13:37:57,881 : INFO : effective_min_count=1 leaves 2656274 word cor
pus (100% of original 2656274, drops 0)
2021-02-26 13:37:58,185 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 13:37:58,188 : INFO : sample=0.001 downsamples 25 most-common words
2021-02-26 13:37:58,190 : INFO : downsampling leaves estimated 2081427 word co
rpus (78.4% of prior 2656274)
2021-02-26 13:37:58,373 : INFO : estimated required memory for 63538 words and
100 dimensions: 82599400 bytes
2021-02-26 13:37:58,375 : INFO : resetting layer weights
2021-02-26 13:38:15,767 : INFO : training model with 4 workers on 63538 vocabu
lary and 100 features, using sg=1 hs=0 sample=0.001 negative=5 window=5
2021-02-26 13:38:16,799 : INFO : EPOCH 1 - PROGRESS: at 18.65% examples, 36836
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:17,814 : INFO : EPOCH 1 - PROGRESS: at 37.74% examples, 38108
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:18,825 : INFO : EPOCH 1 - PROGRESS: at 56.02% examples, 38130
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:19,857 : INFO : EPOCH 1 - PROGRESS: at 73.62% examples, 37166
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:20,859 : INFO : EPOCH 1 - PROGRESS: at 91.71% examples, 37569
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:21,264 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:38:21,276 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:38:21,314 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:38:21,327 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:38:21,328 : INFO : EPOCH - 1 : training on 2656274 raw words (20
81346 effective words) took 5.6s, 374625 effective words/s
2021-02-26 13:38:22,334 : INFO : EPOCH 2 - PROGRESS: at 18.29% examples, 36924
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:23,335 : INFO : EPOCH 2 - PROGRESS: at 35.91% examples, 36883
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:24,341 : INFO : EPOCH 2 - PROGRESS: at 53.31% examples, 36864
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:25,373 : INFO : EPOCH 2 - PROGRESS: at 70.68% examples, 36039
5 words/s, in_qsize 8, out_qsize 0
2021-02-26 13:38:26,375 : INFO : EPOCH 2 - PROGRESS: at 85.16% examples, 34990
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:27,148 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:38:27,159 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:38:27,191 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:38:27,211 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:38:27,212 : INFO : EPOCH - 2 : training on 2656274 raw words (20
81176 effective words) took 5.9s, 353883 effective words/s
```

```
2021-02-26 13:38:28,221 : INFO : EPOCH 3 - PROGRESS: at 15.75% examples, 32223
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:29,233 : INFO : EPOCH 3 - PROGRESS: at 33.15% examples, 33967
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:30,286 : INFO : EPOCH 3 - PROGRESS: at 50.62% examples, 34375
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:31,301 : INFO : EPOCH 3 - PROGRESS: at 62.35% examples, 31527
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:32,311 : INFO : EPOCH 3 - PROGRESS: at 78.77% examples, 32231
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:33,344 : INFO : EPOCH 3 - PROGRESS: at 96.52% examples, 32827
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:33,467 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:38:33,484 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:38:33,515 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:38:33,540 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:38:33,541 : INFO : EPOCH - 3 : training on 2656274 raw words (20
82315 effective words) took 6.3s, 329131 effective words/s
2021-02-26 13:38:34,547 : INFO : EPOCH 4 - PROGRESS: at 12.71% examples, 26137
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:35,558 : INFO : EPOCH 4 - PROGRESS: at 30.11% examples, 30992
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:36,562 : INFO : EPOCH 4 - PROGRESS: at 45.52% examples, 31660
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:37,582 : INFO : EPOCH 4 - PROGRESS: at 63.08% examples, 32264
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:38,595 : INFO : EPOCH 4 - PROGRESS: at 78.77% examples, 32497
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:39,621 : INFO : EPOCH 4 - PROGRESS: at 95.38% examples, 32709
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:39,832 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:38:39,835 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:38:39,853 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:38:39,889 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:38:39,890 : INFO : EPOCH - 4 : training on 2656274 raw words (20
81077 effective words) took 6.3s, 327954 effective words/s
2021-02-26 13:38:40,926 : INFO : EPOCH 5 - PROGRESS: at 15.33% examples, 30612
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:41,938 : INFO : EPOCH 5 - PROGRESS: at 32.32% examples, 32731
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:42,954 : INFO : EPOCH 5 - PROGRESS: at 48.05% examples, 32944
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:43,991 : INFO : EPOCH 5 - PROGRESS: at 66.21% examples, 33300
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:44,994 : INFO : EPOCH 5 - PROGRESS: at 82.74% examples, 33536
```

```
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:38:45,998 : INFO : EPOCH 5 - PROGRESS: at 98.88% examples, 33689
4 words/s, in_qsize 4, out_qsize 0
2021-02-26 13:38:46,011 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:38:46,020 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:38:46,048 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:38:46,058 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:38:46,059 : INFO : EPOCH - 5 : training on 2656274 raw words (20
80982 effective words) took 6.2s, 337469 effective words/s
2021-02-26 13:38:46,060 : INFO : training on a 13281370 raw words (10406896 ef
fective words) took 30.3s, 343554 effective words/s
```

In [48]:

```
seed_word5 = [list(vectors5.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [49]:

```
seed_word5
```

Out[49]:

```
['half', 'whether', 'neuroscience', 'Lonza', 'tightly']
```

In [50]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'half')
print(vectors5.wv.most_similar('half'))
print()
```

```
2021-02-26 13:39:16,488 : INFO : precomputing L2-norms of word weight vectors
Most similar to: half
[('25', 0.790177583694458), ('80', 0.7899731397628784), ('ten', 0.769554972648
6206), ('70', 0.7545516490936279), ('percent', 0.7527369856834412), ('65', 0.7
49315619468689), ('trillion', 0.7484829425811768), ('About', 0.745859086513519
3), ('150', 0.7433521747589111), ('60', 0.7410141229629517)]
```

In [51]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'whether')
print(vectors5.wv.most_similar('whether'))
print()
```

```
Most similar to: whether
[('if', 0.7773952484130859), ('how', 0.7436619400978088), ('why', 0.7058299183
84552), ('what', 0.6954272985458374), ('sex', 0.6872129440307617), ('androdioe
cy', 0.6792768239974976), ('Are', 0.6774872541427612), ('Is', 0.67665272951126
1), ('Or', 0.6747390031814575), ('ask', 0.6745706796646118)]
```

In [52]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors5.wv.most_similar('neuroscience'))
print()
```

```
Most similar to: neuroscience
[('endocrinology', 0.8928981423377991), ('neurobiology', 0.8739007115364075),
('neuroendocrinology', 0.8608860373497009), ('sociology', 0.8580520749092102),
('linguistics', 0.8549712896347046), ('epidemiology', 0.8538591861724854), ('n
europhysiology', 0.8537481427192688), ('revolutionizing', 0.8456230163574219),
('archeology', 0.8387148976325989), ('informs', 0.8276082277297974)]
```

In [53]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'Lonza')
print(vectors5.wv.most_similar('Lonza'))
print()
```

```
Most similar to: Lonza
[('04cc', 0.9778331518173218), ('Discover', 0.9769275784492493), ('lampposts',
0.9763314723968506), ('drums', 0.9755334854125977), ('COPS', 0.974900722503662
1), ('diaminopropane', 0.9746689796447754), ('lobules', 0.9742003083229065), (
'Cytoplamic', 0.9737499952316284), ('mutase', 0.9733877182006836), ('CDV', 0.9
733231067657471)]
```

In [54]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'tightly')
print(vectors5.wv.most_similar('tightly'))
print()
```

```
Most similar to: tightly
[('loosely', 0.7940118312835693), ('mimic', 0.788912296295166), ('V1', 0.77770
59078216553), ('deformable', 0.7661259174346924), ('flexibly', 0.7651008367538
452), ('seamlessly', 0.7630290389060974), ('behaviorally', 0.7627905607223511)
, ('conveniently', 0.762204647064209), ('1D', 0.7617976665496826), ('internall
y', 0.7611943483352661)]
```

In [55]:

```
vectors6 = gensim.models.Word2Vec(tokenized_text, size=100, min_count=1, sg=0
```

```
2021-02-26 13:39:46,315 : INFO : collecting all words and their counts
2021-02-26 13:39:46,317 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 13:39:46,850 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 13:39:46,852 : INFO : Loading a fresh vocabulary
2021-02-26 13:39:47,018 : INFO : effective_min_count=1 retains 63538 unique wo
rds (100% of original 63538, drops 0)
```

```
2021-02-26 13:39:47,020 : INFO : effective_min_count=1 leaves 2656274 word cor
pus (100% of original 2656274, drops 0)
2021-02-26 13:39:47,319 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 13:39:47,322 : INFO : sample=0.001 downsamples 25 most-common words
2021-02-26 13:39:47,324 : INFO : downsampling leaves estimated 2081427 word co
rpus (78.4% of prior 2656274)
2021-02-26 13:39:47,498 : INFO : estimated required memory for 63538 words and
100 dimensions: 82599400 bytes
2021-02-26 13:39:47,500 : INFO : resetting layer weights
2021-02-26 13:40:04,571 : INFO : training model with 4 workers on 63538 vocabu
lary and 100 features, using sg=0 hs=0 sample=0.001 negative=5 window=5
2021-02-26 13:40:05,579 : INFO : EPOCH 1 - PROGRESS: at 64.59% examples, 13285
88 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:06,129 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:40:06,131 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:40:06,138 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:40:06,143 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:40:06,143 : INFO : EPOCH - 1 : training on 2656274 raw words (20
80440 effective words) took 1.6s, 1326973 effective words/s
2021-02-26 13:40:07,148 : INFO : EPOCH 2 - PROGRESS: at 64.59% examples, 13324
97 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:07,686 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:40:07,687 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:40:07,692 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:40:07,699 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:40:07,699 : INFO : EPOCH - 2 : training on 2656274 raw words (20
81371 effective words) took 1.6s, 1340202 effective words/s
2021-02-26 13:40:08,703 : INFO : EPOCH 3 - PROGRESS: at 63.49% examples, 13104
48 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:09,391 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:40:09,392 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:40:09,396 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:40:09,401 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:40:09,401 : INFO : EPOCH - 3 : training on 2656274 raw words (20
81330 effective words) took 1.7s, 1224891 effective words/s
2021-02-26 13:40:10,409 : INFO : EPOCH 4 - PROGRESS: at 63.08% examples, 12972
91 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:10,987 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:40:10,988 : INFO : worker thread finished; awaiting finish of 2
```

```
more threads
2021-02-26 13:40:10,993 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:40:11,000 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:40:11,000 : INFO : EPOCH - 4 : training on 2656274 raw words (20
81534 effective words) took 1.6s, 1304021 effective words/s
2021-02-26 13:40:12,009 : INFO : EPOCH 5 - PROGRESS: at 62.35% examples, 12813
22 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:12,618 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:40:12,619 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:40:12,624 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:40:12,631 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:40:12,631 : INFO : EPOCH - 5 : training on 2656274 raw words (20
81510 effective words) took 1.6s, 1278856 effective words/s
2021-02-26 13:40:12,632 : INFO : training on a 13281370 raw words (10406185 ef
fective words) took 8.1s, 1291192 effective words/s
```

In [56]:

```
seed_word6 = [list(vectors6.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [57]:

```
seed_word6
```

Out[57]:

```
['half', 'whether', 'neuroscience', 'Lonza', 'tightly']
```

In [58]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'half')
print(vectors6.wv.most_similar('half'))
print()
```

```
2021-02-26 13:40:36,736 : INFO : precomputing L2-norms of word weight vectors
Most similar to: half
[('80', 0.8614833354949951), ('million', 0.8565804958343506), ('70', 0.8434327
840805054), ('percent', 0.8345407247543335), ('25', 0.828275203704834), ('50',
0.822907567024231), ('square', 0.8213421106338501), ('1000', 0.821179389953613
3), ('150', 0.8172222375869751), ('days', 0.8170233368873596)]
```

In [59]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'whether')
print(vectors6.wv.most_similar('whether'))
print()
```

```
Most similar to: whether
[('if', 0.8375376462936401), ('why', 0.8201342225074768), ('how', 0.8175545930
862427), ('what', 0.8073095083236694), ('How', 0.7672230005264282), ('bicoordi
nate', 0.6755235195159912), ('responses', 0.6288872957229614), ('inactivating'
, 0.6159760355949402), ('sex', 0.6069273948669434), ('Phospho', 0.597217082977
2949)]
```

In [60]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors6.wv.most_similar('neuroscience'))
print()
```

```
Most similar to: neuroscience
[('sociology', 0.8529632091522217), ('economics', 0.842451274394989), ('bioche
mistry', 0.8309021592140198), ('endocrinology', 0.8099850416183472), ('medicin
e', 0.8027428388595581), ('ecology', 0.7961971759796143), ('conservation', 0.7
931435108184814), ('microbiology', 0.7917824983596802), ('oceanography', 0.791
6997075080872), ('contemporary', 0.7900481224060059)]
```

In [61]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'Lonza')
print(vectors6.wv.most_similar('Lonza'))
print()
```

```
Most similar to: Lonza
[('Rapt', 0.946742057800293), ('Cisco', 0.9320446848869324), ('Serbia', 0.9216
082692146301), ('PBD', 0.9192733764648438), ('vegetables', 0.9182940125465393)
, ('inconsistency', 0.9156026840209961), ('vicariant', 0.9148354530334473), ('
Sternoptychidae', 0.9147069454193115), ('0209202', 0.9143208861351013), ('your
', 0.9138616919517517)]
```

In [62]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'tightly')
print(vectors6.wv.most_similar('tightly'))
print()
```

```
Most similar to: tightly
[('structurally', 0.8403840065002441), ('intrinsically', 0.8381390571594238),
('weakly', 0.8356295824050903), ('deformable', 0.8241686224937439), ('tunnels'
, 0.8216641545295715), ('gates', 0.8200643658638), ('functionally', 0.81474375
72479248), ('mechanically', 0.8121119737625122), ('intimately', 0.811967253684
9976), ('clustered', 0.8089224100112915)]
```

In [63]:

```
vectors7 = gensim.models.Word2Vec(tokenized_text, size=10, min_count=5, sg=1,
```

```
2021-02-26 13:40:50,348 : WARNING : consider setting layer size to a multiple
```

```
of 4 for greater performance
2021-02-26 13:40:50,350 : INFO : collecting all words and their counts
2021-02-26 13:40:50,351 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 13:40:50,889 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 13:40:50,891 : INFO : Loading a fresh vocabulary
2021-02-26 13:40:51,080 : INFO : effective_min_count=5 retains 20752 unique wo
rds (32% of original 63538, drops 42786)
2021-02-26 13:40:51,082 : INFO : effective_min_count=5 leaves 2585188 word cor
pus (97% of original 2656274, drops 71086)
2021-02-26 13:40:51,182 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 13:40:51,186 : INFO : sample=0.001 downsamples 26 most-common words
2021-02-26 13:40:51,187 : INFO : downsampling leaves estimated 2005620 word co
rpus (77.6% of prior 2585188)
2021-02-26 13:40:51,244 : INFO : estimated required memory for 20752 words and
10 dimensions: 12036160 bytes
2021-02-26 13:40:51,245 : INFO : resetting layer weights
2021-02-26 13:40:56,914 : INFO : training model with 4 workers on 20752 vocabu
lary and 10 features, using sg=1 hs=0 sample=0.001 negative=5 window=5
2021-02-26 13:40:57,928 : INFO : EPOCH 1 - PROGRESS: at 27.66% examples, 54589
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:58,931 : INFO : EPOCH 1 - PROGRESS: at 56.53% examples, 56123
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:40:59,936 : INFO : EPOCH 1 - PROGRESS: at 85.16% examples, 56394
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:00,444 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:41:00,448 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:41:00,455 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:41:00,473 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:41:00,474 : INFO : EPOCH - 1 : training on 2656274 raw words (20
05645 effective words) took 3.6s, 564188 effective words/s
2021-02-26 13:41:01,478 : INFO : EPOCH 2 - PROGRESS: at 27.33% examples, 54301
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:02,482 : INFO : EPOCH 2 - PROGRESS: at 54.70% examples, 54870
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:03,490 : INFO : EPOCH 2 - PROGRESS: at 82.34% examples, 54483
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:04,304 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:41:04,306 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:41:04,325 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:41:04,342 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:41:04,344 : INFO : EPOCH - 2 : training on 2656274 raw words (20
05800 effective words) took 3.9s, 518760 effective words/s
```

```
2021-02-26 13:41:05,373 : INFO : EPOCH 3 - PROGRESS: at 24.44% examples, 47046
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:06,379 : INFO : EPOCH 3 - PROGRESS: at 47.34% examples, 47122
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:07,393 : INFO : EPOCH 3 - PROGRESS: at 74.29% examples, 48511
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:08,406 : INFO : EPOCH 3 - PROGRESS: at 89.26% examples, 44097
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:08,818 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:41:08,828 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:41:08,849 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:41:08,850 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:41:08,851 : INFO : EPOCH - 3 : training on 2656274 raw words (20
05248 effective words) took 4.5s, 445200 effective words/s
2021-02-26 13:41:09,856 : INFO : EPOCH 4 - PROGRESS: at 26.19% examples, 51924
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:10,885 : INFO : EPOCH 4 - PROGRESS: at 52.27% examples, 51523
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:11,892 : INFO : EPOCH 4 - PROGRESS: at 76.80% examples, 50590
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:12,900 : INFO : EPOCH 4 - PROGRESS: at 92.91% examples, 46058
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:13,138 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:41:13,145 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:41:13,163 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:41:13,170 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:41:13,171 : INFO : EPOCH - 4 : training on 2656274 raw words (20
05691 effective words) took 4.3s, 464597 effective words/s
2021-02-26 13:41:14,203 : INFO : EPOCH 5 - PROGRESS: at 25.85% examples, 49823
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:15,210 : INFO : EPOCH 5 - PROGRESS: at 51.03% examples, 50328
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:16,214 : INFO : EPOCH 5 - PROGRESS: at 67.00% examples, 43752
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:17,229 : INFO : EPOCH 5 - PROGRESS: at 92.53% examples, 45784
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 13:41:17,477 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 13:41:17,481 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 13:41:17,493 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 13:41:17,518 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 13:41:17,520 : INFO : EPOCH - 5 : training on 2656274 raw words (20
05908 effective words) took 4.3s, 461508 effective words/s
2021-02-26 13:41:17,522 : INFO : training on a 13281370 raw words (10028292 ef
fective words) took 20.6s, 486638 effective words/s
```

In [66]:

```
seed_word7 = [list(vectors7.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [67]:

```
seed_word7
```

Out[67]:

```
['greatly', 'every', 'includes', 'transcription', 'largest']
```

In [68]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'greatly')
print(vectors7.wv.most_similar('greatly'))
print()
```

```
2021-02-26 14:04:14,419 : INFO : precomputing L2-norms of word weight vectors
Most similar to: greatly
[('significantly', 0.9823163747787476), ('aid', 0.9707620143890381), ('ability
', 0.9605613350868225), ('improved', 0.9523959159851074), ('linking', 0.948238
0151748657), ('assessing', 0.9464660286903381), ('improvement', 0.944735169410
7056), ('facilitating', 0.9437584280967712), ('participatory', 0.9437243938446
045), ('Identify', 0.9425539970397949)]
```

In [69]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'every')
print(vectors7.wv.most_similar('every'))
print()
```

```
Most similar to: every
[('one', 0.9544808864593506), ('forty', 0.9348524212837219), ('fifty', 0.93360
89491844177), ('least', 0.9320022463798523), ('attracts', 0.9301199316978455),
('round', 0.9274606108665466), ('hold', 0.9169936180114746), ('given', 0.91649
7528553009), ('holding', 0.9150835871696472), ('ample', 0.9130387902259827)]
```

In [70]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'includes')
print(vectors7.wv.most_similar('includes'))
print()
```

```
Most similar to: includes
[('include', 0.9577757120132446), ('The', 0.9568184614181519), ('an', 0.952779
2930603027), ('involves', 0.9490988850593567), ('Primary', 0.9430201649665833)
, ('for', 0.9428271055221558), ('comprises', 0.9396464228630066), ('Conduct',
0.9395472407341003), ('Through', 0.9394429922103882), ('on', 0.935796439647674
6)]
```

In [71]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'transcription')
print(vectors7.wv.most_similar('transcription'))
print()
```

```
Most similar to: transcription
[('regulated', 0.98469078540802), ('replication', 0.9833722710609436), ('recep
tor', 0.9808540940284729), ('gene', 0.9795259237289429), ('signaling', 0.97838
95015716553), ('splicing', 0.9763568043708801), ('pathway', 0.9750600457191467
), ('silencing', 0.9727146625518799), ('secretion', 0.970633864402771), ('chlo
roplast', 0.9689313173294067)]
```

In [72]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'largest')
print(vectors7.wv.most_similar('largest'))
print()
```

```
Most similar to: largest
[('Saharan', 0.9502067565917969), ('mountainous', 0.9391655325889587), ('borde
r', 0.935000479221344), ('northeastern', 0.9338515996932983), ('towns', 0.9337
120056152344), ('southeastern', 0.9299823045730591), ('gatherer', 0.9286332726
478577), ('1970s', 0.9252861738204956), ('deserts', 0.924917995929718), ('biom
es', 0.9243955016136169)]
```

In [73]:

```
vectors8 = gensim.models.Word2Vec(tokenized_text, size=10, min_count=5, sg=0,
```

```
2021-02-26 14:04:22,539 : WARNING : consider setting layer size to a multiple
of 4 for greater performance
2021-02-26 14:04:22,542 : INFO : collecting all words and their counts
2021-02-26 14:04:22,542 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 14:04:23,112 : INFO : collected 63538 word types from a corpus of 2
656274 raw words and 9923 sentences
2021-02-26 14:04:23,113 : INFO : Loading a fresh vocabulary
2021-02-26 14:04:23,189 : INFO : effective_min_count=5 retains 20752 unique wo
rds (32% of original 63538, drops 42786)
2021-02-26 14:04:23,190 : INFO : effective_min_count=5 leaves 2585188 word cor
pus (97% of original 2656274, drops 71086)
2021-02-26 14:04:23,298 : INFO : deleting the raw counts dictionary of 63538 i
tems
2021-02-26 14:04:23,300 : INFO : sample=0.001 downsamples 26 most-common words
2021-02-26 14:04:23,301 : INFO : downsampling leaves estimated 2005620 word co
rpus (77.6% of prior 2585188)
2021-02-26 14:04:23,360 : INFO : estimated required memory for 20752 words and
10 dimensions: 12036160 bytes
2021-02-26 14:04:23,361 : INFO : resetting layer weights
2021-02-26 14:04:29,040 : INFO : training model with 4 workers on 20752 vocabu
lary and 10 features, using sg=0 hs=0 sample=0.001 negative=5 window=5
```

```
2021-02-26 14:04:30,047 : INFO : EPOCH 1 - PROGRESS: at 92.11% examples, 18430
59 words/s, in_qsize 7, out_qsize 0
2021-02-26 14:04:30,128 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 14:04:30,128 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 14:04:30,132 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 14:04:30,135 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 14:04:30,136 : INFO : EPOCH - 1 : training on 2656274 raw words (20
05643 effective words) took 1.1s, 1836329 effective words/s
2021-02-26 14:04:31,140 : INFO : EPOCH 2 - PROGRESS: at 85.46% examples, 17063
19 words/s, in_qsize 7, out_qsize 0
2021-02-26 14:04:31,323 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 14:04:31,324 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 14:04:31,328 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 14:04:31,332 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 14:04:31,332 : INFO : EPOCH - 2 : training on 2656274 raw words (20
05967 effective words) took 1.2s, 1680467 effective words/s
2021-02-26 14:04:32,344 : INFO : EPOCH 3 - PROGRESS: at 79.24% examples, 15765
72 words/s, in_qsize 7, out_qsize 0
2021-02-26 14:04:32,612 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 14:04:32,614 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 14:04:32,617 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 14:04:32,622 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 14:04:32,623 : INFO : EPOCH - 3 : training on 2656274 raw words (20
05277 effective words) took 1.3s, 1557944 effective words/s
2021-02-26 14:04:33,629 : INFO : EPOCH 4 - PROGRESS: at 76.53% examples, 15258
54 words/s, in_qsize 7, out_qsize 0
2021-02-26 14:04:33,939 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 14:04:33,940 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 14:04:33,944 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 14:04:33,949 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 14:04:33,950 : INFO : EPOCH - 4 : training on 2656274 raw words (20
05703 effective words) took 1.3s, 1514864 effective words/s
2021-02-26 14:04:34,956 : INFO : EPOCH 5 - PROGRESS: at 81.91% examples, 16279
41 words/s, in_qsize 7, out_qsize 0
2021-02-26 14:04:35,177 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 14:04:35,178 : INFO : worker thread finished; awaiting finish of 2
```

```
more threads
2021-02-26 14:04:35,182 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 14:04:35,185 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 14:04:35,186 : INFO : EPOCH - 5 : training on 2656274 raw words (20
05364 effective words) took 1.2s, 1626801 effective words/s
2021-02-26 14:04:35,186 : INFO : training on a 13281370 raw words (10027954 ef
fective words) took 6.1s, 1631886 effective words/s
```

In [74]:

```
seed_word8 = [list(vectors8.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [75]:

```
seed_word8
```

Out[75]:

```
['greatly', 'every', 'includes', 'transcription', 'largest']
```

In [76]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'greatly')
print(vectors8.wv.most_similar('greatly'))
print()
```

```
2021-02-26 14:04:45,154 : INFO : precomputing L2-norms of word weight vectors
Most similar to: greatly
[('significantly', 0.9625788331031799), ('ability', 0.9101622104644775), ('sub
stantially', 0.8781238794326782), ('to', 0.8764215707778931), ('necessary', 0.
8735238313674927), ('us', 0.8630886077880859), ('better', 0.8587857484817505),
('effectively', 0.8520721793174744), ('strategies', 0.850286602973938), ('shou
ld', 0.8486435413360596)]
```

In [77]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'every')
print(vectors8.wv.most_similar('every'))
print()
```

```
Most similar to: every
[('later', 0.924665093421936), ('few', 0.9168813824653625), ('classified', 0.9
053330421447754), ('chosen', 0.8952016830444336), ('thirty', 0.872394919395446
8), ('rarely', 0.8720383644104004), ('just', 0.8637912273406982), ('inaccessib
le', 0.8631974458694458), ('millions', 0.8623626232147217), ('least', 0.860414
7434234619)]
```

In [78]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'includes')
print(vectors8.wv.most_similar('includes'))
print()
```

```
Most similar to: includes
[('involves', 0.9217371344566345), ('include', 0.9062110781669617), ('combines
', 0.8593859672546387), ('capitalizes', 0.8403268456459045), ('consists', 0.83
59469771385193), ('provides', 0.8328524231910706), ('emphasize', 0.82604342699
0509), ('supports', 0.8258134126663208), ('brings', 0.8222417831420898), ('upo
n', 0.8124222755432129)]
```

In [79]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'transcription')
print(vectors8.wv.most_similar('transcription'))
print()
```

```
Most similar to: transcription
[('receptor', 0.9635428786277771), ('galaxy', 0.9552401900291443), ('eukaryoti
c', 0.9497318267822266), ('mRNA', 0.9465463161468506), ('transcriptional', 0.9
457406997680664), ('MHC', 0.9427085518836975), ('muscle', 0.9385768175125122),
('conserved', 0.9299783110618591), ('silencing', 0.9290547370910645), ('phenot
ypic', 0.9256062507629395)]
```

In [80]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'largest')
print(vectors8.wv.most_similar('largest'))
print()
```

```
Most similar to: largest
[('North', 0.9387719631195068), ('earliest', 0.923382043838501), ('America', 0
.9213746190071106), ('eighteenth', 0.9158466458320618), ('gatherers', 0.912085
9503746033), ('Asian', 0.908501148223877), ('Andean', 0.9053014516830444), ('A
sia', 0.902595043182373), ('island', 0.8988133668899536), ('19th', 0.892090439
7964478)]
```

QUESTION 2B

In [39]:

```
import json
from sklearn.feature_extraction.text import TfidfVectorizer
import os
import re

def get_fnames():
    """Read all text files in a folder.
    """
    fnames = []
    for root,_,files in os.walk("./abstracts"):
        for fname in files:
            if fname[-4:] == ".txt":
                fnames.append(os.path.join(root, fname))
    return fnames

print("Number of abstracts in folder awards: {}".format(len(get_fnames())))
```

```
Number of abstracts in folder awards: 132372
```

In [40]:

```
name_list = get_fnames()

def read_file(fname):
    with open(fname, 'r',encoding="ISO-8859-1") as f:
        # skip all lines until abstract
        for line in f:
            if "Abstract    :" in line:
                break

        # get abstract as a single string
        abstract = ' '.join([line[:-1].strip() for line in f])
        abstract = re.sub(' +', ' ', abstract)  # remove double spaces
        return abstract
```

In [41]:

```
documents_full = []

for i in name_list:
    documents_full.append(read_file(i))
```

In [42]:

```
new_vectorizer1 = TfidfVectorizer(stop_words = 'english', lowercase= True, ng
word_tokenizer1 = new_vectorizer1.build_tokenizer()
tokenized_text1 = [word_tokenizer1(doc) for doc in documents_full]
```

In [43]:

```
tokenized_text1[0]
```

Out[43]:

```
[]
```

In [44]:
```
len(tokenized_text1)
```

Out[44]:
```
132372
```

In [46]:
```
vectors9 = gensim.models.Word2Vec(tokenized_text1, size= 10, min_count=1, sg=
```

```
2021-02-26 21:56:34,045 : WARNING : consider setting layer size to a multiple
of 4 for greater performance
2021-02-26 21:56:34,048 : INFO : collecting all words and their counts
2021-02-26 21:56:34,049 : INFO : PROGRESS: at sentence #0, processed 0 words,
keeping 0 word types
2021-02-26 21:56:34,422 : INFO : PROGRESS: at sentence #10000, processed 15653
23 words, keeping 47155 word types
2021-02-26 21:56:34,922 : INFO : PROGRESS: at sentence #20000, processed 35280
75 words, keeping 79367 word types
2021-02-26 21:56:35,403 : INFO : PROGRESS: at sentence #30000, processed 54426
67 words, keeping 105577 word types
2021-02-26 21:56:35,944 : INFO : PROGRESS: at sentence #40000, processed 72055
33 words, keeping 127561 word types
2021-02-26 21:56:36,840 : INFO : PROGRESS: at sentence #50000, processed 96466
64 words, keeping 149006 word types
2021-02-26 21:56:37,624 : INFO : PROGRESS: at sentence #60000, processed 12202
268 words, keeping 167318 word types
2021-02-26 21:56:38,278 : INFO : PROGRESS: at sentence #70000, processed 14413
047 words, keeping 184849 word types
2021-02-26 21:56:38,745 : INFO : PROGRESS: at sentence #80000, processed 15940
674 words, keeping 192628 word types
2021-02-26 21:56:39,359 : INFO : PROGRESS: at sentence #90000, processed 18069
977 words, keeping 207107 word types
2021-02-26 21:56:40,049 : INFO : PROGRESS: at sentence #100000, processed 2046
5250 words, keeping 221416 word types
2021-02-26 21:56:40,529 : INFO : PROGRESS: at sentence #110000, processed 2212
4882 words, keeping 229369 word types
2021-02-26 21:56:41,135 : INFO : PROGRESS: at sentence #120000, processed 2416
0301 words, keeping 241454 word types
2021-02-26 21:56:41,904 : INFO : PROGRESS: at sentence #130000, processed 2672
9522 words, keeping 254009 word types
2021-02-26 21:56:42,108 : INFO : collected 257022 word types from a corpus of
27374377 raw words and 132372 sentences
2021-02-26 21:56:42,110 : INFO : Loading a fresh vocabulary
2021-02-26 21:56:42,996 : INFO : effective_min_count=1 retains 257022 unique w
ords (100% of original 257022, drops 0)
2021-02-26 21:56:42,997 : INFO : effective_min_count=1 leaves 27374377 word co
rpus (100% of original 27374377, drops 0)
2021-02-26 21:56:44,556 : INFO : deleting the raw counts dictionary of 257022
items
2021-02-26 21:56:44,565 : INFO : sample=0.001 downsamples 25 most-common words
2021-02-26 21:56:44,566 : INFO : downsampling leaves estimated 21283036 word c
orpus (77.7% of prior 27374377)
2021-02-26 21:56:45,478 : INFO : estimated required memory for 257022 words an
```

```
d 10 dimensions: 149072760 bytes
2021-02-26 21:56:45,479 : INFO : resetting layer weights
2021-02-26 21:58:09,406 : INFO : training model with 4 workers on 257022 vocab
ulary and 10 features, using sg=1 hs=0 sample=0.001 negative=5 window=5
2021-02-26 21:58:10,424 : INFO : EPOCH 1 - PROGRESS: at 2.38% examples, 340353
words/s, in_qsize 8, out_qsize 0
2021-02-26 21:58:11,427 : INFO : EPOCH 1 - PROGRESS: at 4.47% examples, 348160
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:12,438 : INFO : EPOCH 1 - PROGRESS: at 6.84% examples, 355695
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:13,445 : INFO : EPOCH 1 - PROGRESS: at 8.90% examples, 361260
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:14,449 : INFO : EPOCH 1 - PROGRESS: at 10.55% examples, 36099
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:15,449 : INFO : EPOCH 1 - PROGRESS: at 12.33% examples, 36108
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:16,460 : INFO : EPOCH 1 - PROGRESS: at 14.04% examples, 36123
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:17,472 : INFO : EPOCH 1 - PROGRESS: at 15.92% examples, 36056
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:18,486 : INFO : EPOCH 1 - PROGRESS: at 17.65% examples, 35979
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 21:58:19,507 : INFO : EPOCH 1 - PROGRESS: at 19.10% examples, 35018
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:20,521 : INFO : EPOCH 1 - PROGRESS: at 20.92% examples, 34982
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:21,552 : INFO : EPOCH 1 - PROGRESS: at 22.86% examples, 35047
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:22,569 : INFO : EPOCH 1 - PROGRESS: at 24.93% examples, 34970
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:23,590 : INFO : EPOCH 1 - PROGRESS: at 26.79% examples, 34817
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:24,608 : INFO : EPOCH 1 - PROGRESS: at 28.63% examples, 34664
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:25,636 : INFO : EPOCH 1 - PROGRESS: at 30.39% examples, 34639
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:26,637 : INFO : EPOCH 1 - PROGRESS: at 31.70% examples, 34496
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:27,642 : INFO : EPOCH 1 - PROGRESS: at 32.98% examples, 34442
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:28,649 : INFO : EPOCH 1 - PROGRESS: at 34.44% examples, 34441
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:29,659 : INFO : EPOCH 1 - PROGRESS: at 35.67% examples, 34351
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:30,676 : INFO : EPOCH 1 - PROGRESS: at 37.08% examples, 34364
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:31,679 : INFO : EPOCH 1 - PROGRESS: at 38.40% examples, 34334
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:32,682 : INFO : EPOCH 1 - PROGRESS: at 39.66% examples, 34281
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:33,690 : INFO : EPOCH 1 - PROGRESS: at 40.96% examples, 34247
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:34,695 : INFO : EPOCH 1 - PROGRESS: at 42.17% examples, 34226
```

```
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:35,697 : INFO : EPOCH 1 - PROGRESS: at 43.43% examples, 34207
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:36,700 : INFO : EPOCH 1 - PROGRESS: at 44.80% examples, 34245
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:37,721 : INFO : EPOCH 1 - PROGRESS: at 46.00% examples, 34142
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:38,755 : INFO : EPOCH 1 - PROGRESS: at 47.19% examples, 33980
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:39,760 : INFO : EPOCH 1 - PROGRESS: at 48.62% examples, 33888
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:40,796 : INFO : EPOCH 1 - PROGRESS: at 49.72% examples, 33550
6 words/s, in_qsize 5, out_qsize 2
2021-02-26 21:58:41,821 : INFO : EPOCH 1 - PROGRESS: at 51.20% examples, 33538
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:42,835 : INFO : EPOCH 1 - PROGRESS: at 52.89% examples, 33529
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:43,838 : INFO : EPOCH 1 - PROGRESS: at 54.80% examples, 33503
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:44,849 : INFO : EPOCH 1 - PROGRESS: at 57.01% examples, 33476
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:45,868 : INFO : EPOCH 1 - PROGRESS: at 59.22% examples, 33460
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:46,875 : INFO : EPOCH 1 - PROGRESS: at 61.47% examples, 33458
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:47,895 : INFO : EPOCH 1 - PROGRESS: at 63.06% examples, 33489
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:48,909 : INFO : EPOCH 1 - PROGRESS: at 64.39% examples, 33483
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:49,933 : INFO : EPOCH 1 - PROGRESS: at 65.98% examples, 33469
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:50,940 : INFO : EPOCH 1 - PROGRESS: at 67.37% examples, 33486
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:51,969 : INFO : EPOCH 1 - PROGRESS: at 68.76% examples, 33420
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:52,975 : INFO : EPOCH 1 - PROGRESS: at 70.05% examples, 33443
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:53,994 : INFO : EPOCH 1 - PROGRESS: at 71.43% examples, 33422
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:55,007 : INFO : EPOCH 1 - PROGRESS: at 72.86% examples, 33419
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:56,021 : INFO : EPOCH 1 - PROGRESS: at 74.15% examples, 33383
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:57,033 : INFO : EPOCH 1 - PROGRESS: at 75.50% examples, 33382
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 21:58:58,036 : INFO : EPOCH 1 - PROGRESS: at 77.05% examples, 33390
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:58:59,064 : INFO : EPOCH 1 - PROGRESS: at 79.08% examples, 33372
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:00,070 : INFO : EPOCH 1 - PROGRESS: at 81.21% examples, 33372
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:01,081 : INFO : EPOCH 1 - PROGRESS: at 83.51% examples, 33354
3 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 21:59:02,110 : INFO : EPOCH 1 - PROGRESS: at 85.22% examples, 33340
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:03,139 : INFO : EPOCH 1 - PROGRESS: at 86.83% examples, 33332
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:04,169 : INFO : EPOCH 1 - PROGRESS: at 88.40% examples, 33333
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:05,191 : INFO : EPOCH 1 - PROGRESS: at 89.84% examples, 33329
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:06,217 : INFO : EPOCH 1 - PROGRESS: at 91.39% examples, 33325
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:07,227 : INFO : EPOCH 1 - PROGRESS: at 92.67% examples, 33298
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 21:59:08,238 : INFO : EPOCH 1 - PROGRESS: at 93.91% examples, 33288
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:09,263 : INFO : EPOCH 1 - PROGRESS: at 95.06% examples, 33257
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:10,265 : INFO : EPOCH 1 - PROGRESS: at 96.01% examples, 33166
2 words/s, in_qsize 6, out_qsize 1
2021-02-26 21:59:11,271 : INFO : EPOCH 1 - PROGRESS: at 96.99% examples, 33049
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:12,326 : INFO : EPOCH 1 - PROGRESS: at 98.18% examples, 33010
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:13,405 : INFO : EPOCH 1 - PROGRESS: at 99.17% examples, 32896
8 words/s, in_qsize 6, out_qsize 3
2021-02-26 21:59:13,995 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 21:59:14,064 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 21:59:14,084 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 21:59:14,088 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 21:59:14,088 : INFO : EPOCH - 1 : training on 27374377 raw words (2
1281190 effective words) took 64.7s, 329039 effective words/s
2021-02-26 21:59:15,117 : INFO : EPOCH 2 - PROGRESS: at 2.22% examples, 306369
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:16,126 : INFO : EPOCH 2 - PROGRESS: at 4.13% examples, 318831
words/s, in_qsize 8, out_qsize 0
2021-02-26 21:59:17,130 : INFO : EPOCH 2 - PROGRESS: at 6.31% examples, 324027
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:18,148 : INFO : EPOCH 2 - PROGRESS: at 8.26% examples, 327328
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:19,149 : INFO : EPOCH 2 - PROGRESS: at 9.85% examples, 328841
words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:20,150 : INFO : EPOCH 2 - PROGRESS: at 11.58% examples, 32947
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:21,165 : INFO : EPOCH 2 - PROGRESS: at 13.06% examples, 32989
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 21:59:22,170 : INFO : EPOCH 2 - PROGRESS: at 14.74% examples, 32940
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:23,181 : INFO : EPOCH 2 - PROGRESS: at 16.16% examples, 32552
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:24,220 : INFO : EPOCH 2 - PROGRESS: at 17.83% examples, 32529
```

```
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:25,248 : INFO : EPOCH 2 - PROGRESS: at 19.62% examples, 32571
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:26,277 : INFO : EPOCH 2 - PROGRESS: at 21.35% examples, 32586
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:27,299 : INFO : EPOCH 2 - PROGRESS: at 23.13% examples, 32624
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:28,364 : INFO : EPOCH 2 - PROGRESS: at 25.21% examples, 32617
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 21:59:29,423 : INFO : EPOCH 2 - PROGRESS: at 26.75% examples, 32152
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:30,467 : INFO : EPOCH 2 - PROGRESS: at 27.99% examples, 31464
7 words/s, in_qsize 6, out_qsize 1
2021-02-26 21:59:31,501 : INFO : EPOCH 2 - PROGRESS: at 29.39% examples, 31141
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:32,513 : INFO : EPOCH 2 - PROGRESS: at 30.75% examples, 30924
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:33,526 : INFO : EPOCH 2 - PROGRESS: at 31.60% examples, 30457
2 words/s, in_qsize 8, out_qsize 0
2021-02-26 21:59:34,554 : INFO : EPOCH 2 - PROGRESS: at 32.65% examples, 30279
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:35,575 : INFO : EPOCH 2 - PROGRESS: at 33.70% examples, 29981
5 words/s, in_qsize 8, out_qsize 1
2021-02-26 21:59:36,578 : INFO : EPOCH 2 - PROGRESS: at 34.95% examples, 30047
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:37,587 : INFO : EPOCH 2 - PROGRESS: at 36.18% examples, 30195
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:38,623 : INFO : EPOCH 2 - PROGRESS: at 37.58% examples, 30326
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:39,639 : INFO : EPOCH 2 - PROGRESS: at 38.95% examples, 30474
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:40,666 : INFO : EPOCH 2 - PROGRESS: at 40.18% examples, 30570
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:41,671 : INFO : EPOCH 2 - PROGRESS: at 41.44% examples, 30655
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:42,727 : INFO : EPOCH 2 - PROGRESS: at 42.56% examples, 30574
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:43,747 : INFO : EPOCH 2 - PROGRESS: at 43.76% examples, 30608
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:44,759 : INFO : EPOCH 2 - PROGRESS: at 45.00% examples, 30648
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:45,768 : INFO : EPOCH 2 - PROGRESS: at 46.26% examples, 30733
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:46,771 : INFO : EPOCH 2 - PROGRESS: at 47.63% examples, 30816
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:47,784 : INFO : EPOCH 2 - PROGRESS: at 49.11% examples, 30892
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:48,785 : INFO : EPOCH 2 - PROGRESS: at 50.66% examples, 30974
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:49,790 : INFO : EPOCH 2 - PROGRESS: at 52.25% examples, 31071
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:50,824 : INFO : EPOCH 2 - PROGRESS: at 54.22% examples, 31133
2 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 21:59:51,833 : INFO : EPOCH 2 - PROGRESS: at 56.39% examples, 31206
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:52,855 : INFO : EPOCH 2 - PROGRESS: at 58.57% examples, 31271
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:53,900 : INFO : EPOCH 2 - PROGRESS: at 61.11% examples, 31333
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:54,919 : INFO : EPOCH 2 - PROGRESS: at 62.81% examples, 31416
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:55,959 : INFO : EPOCH 2 - PROGRESS: at 64.22% examples, 31476
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:56,998 : INFO : EPOCH 2 - PROGRESS: at 65.71% examples, 31500
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:58,028 : INFO : EPOCH 2 - PROGRESS: at 67.01% examples, 31457
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 21:59:59,034 : INFO : EPOCH 2 - PROGRESS: at 68.43% examples, 31472
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:00,053 : INFO : EPOCH 2 - PROGRESS: at 69.69% examples, 31479
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:01,079 : INFO : EPOCH 2 - PROGRESS: at 71.01% examples, 31513
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:02,095 : INFO : EPOCH 2 - PROGRESS: at 72.34% examples, 31486
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:03,105 : INFO : EPOCH 2 - PROGRESS: at 73.62% examples, 31495
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:04,108 : INFO : EPOCH 2 - PROGRESS: at 75.01% examples, 31554
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:05,135 : INFO : EPOCH 2 - PROGRESS: at 76.42% examples, 31569
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:06,137 : INFO : EPOCH 2 - PROGRESS: at 78.31% examples, 31604
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:07,149 : INFO : EPOCH 2 - PROGRESS: at 80.45% examples, 31634
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:00:08,149 : INFO : EPOCH 2 - PROGRESS: at 82.62% examples, 31655
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:09,169 : INFO : EPOCH 2 - PROGRESS: at 84.65% examples, 31678
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:10,178 : INFO : EPOCH 2 - PROGRESS: at 86.12% examples, 31710
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:11,181 : INFO : EPOCH 2 - PROGRESS: at 87.74% examples, 31729
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:12,186 : INFO : EPOCH 2 - PROGRESS: at 89.20% examples, 31749
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:13,188 : INFO : EPOCH 2 - PROGRESS: at 90.69% examples, 31774
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:14,194 : INFO : EPOCH 2 - PROGRESS: at 92.12% examples, 31789
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:15,199 : INFO : EPOCH 2 - PROGRESS: at 93.36% examples, 31806
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:16,214 : INFO : EPOCH 2 - PROGRESS: at 94.56% examples, 31819
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:17,233 : INFO : EPOCH 2 - PROGRESS: at 95.73% examples, 31843
4 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:00:18,253 : INFO : EPOCH 2 - PROGRESS: at 97.02% examples, 31876
```

```
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:19,257 : INFO : EPOCH 2 - PROGRESS: at 98.23% examples, 31895
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:20,292 : INFO : EPOCH 2 - PROGRESS: at 99.45% examples, 31917
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:20,659 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:00:20,662 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:00:20,676 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:00:20,734 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:00:20,736 : INFO : EPOCH - 2 : training on 27374377 raw words (2
1281943 effective words) took 66.6s, 319339 effective words/s
2021-02-26 22:00:21,765 : INFO : EPOCH 3 - PROGRESS: at 2.22% examples, 306998
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:22,783 : INFO : EPOCH 3 - PROGRESS: at 4.18% examples, 321843
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:23,784 : INFO : EPOCH 3 - PROGRESS: at 6.40% examples, 328972
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:24,788 : INFO : EPOCH 3 - PROGRESS: at 8.34% examples, 331979
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:25,824 : INFO : EPOCH 3 - PROGRESS: at 9.96% examples, 331824
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:26,842 : INFO : EPOCH 3 - PROGRESS: at 11.74% examples, 33228
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:27,855 : INFO : EPOCH 3 - PROGRESS: at 13.20% examples, 33236
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:28,873 : INFO : EPOCH 3 - PROGRESS: at 14.94% examples, 33200
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:29,886 : INFO : EPOCH 3 - PROGRESS: at 16.61% examples, 33359
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:30,893 : INFO : EPOCH 3 - PROGRESS: at 18.36% examples, 33454
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:31,929 : INFO : EPOCH 3 - PROGRESS: at 20.11% examples, 33298
4 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:00:32,945 : INFO : EPOCH 3 - PROGRESS: at 21.81% examples, 33290
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:33,958 : INFO : EPOCH 3 - PROGRESS: at 23.80% examples, 33418
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:34,971 : INFO : EPOCH 3 - PROGRESS: at 25.72% examples, 33357
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:35,996 : INFO : EPOCH 3 - PROGRESS: at 27.56% examples, 33273
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:37,033 : INFO : EPOCH 3 - PROGRESS: at 29.12% examples, 32902
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:38,043 : INFO : EPOCH 3 - PROGRESS: at 30.32% examples, 32399
1 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:00:39,048 : INFO : EPOCH 3 - PROGRESS: at 31.32% examples, 31915
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:40,055 : INFO : EPOCH 3 - PROGRESS: at 32.23% examples, 31487
5 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 22:00:41,062 : INFO : EPOCH 3 - PROGRESS: at 33.14% examples, 31090
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:42,064 : INFO : EPOCH 3 - PROGRESS: at 34.18% examples, 30790
8 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:00:43,102 : INFO : EPOCH 3 - PROGRESS: at 35.13% examples, 30424
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:44,152 : INFO : EPOCH 3 - PROGRESS: at 36.07% examples, 30175
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:45,231 : INFO : EPOCH 3 - PROGRESS: at 37.20% examples, 29970
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:00:46,233 : INFO : EPOCH 3 - PROGRESS: at 38.21% examples, 29819
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:47,244 : INFO : EPOCH 3 - PROGRESS: at 39.28% examples, 29728
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:48,259 : INFO : EPOCH 3 - PROGRESS: at 40.37% examples, 29691
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:49,277 : INFO : EPOCH 3 - PROGRESS: at 41.55% examples, 29737
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:50,283 : INFO : EPOCH 3 - PROGRESS: at 42.79% examples, 29871
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:00:51,302 : INFO : EPOCH 3 - PROGRESS: at 44.12% examples, 29979
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:52,319 : INFO : EPOCH 3 - PROGRESS: at 45.41% examples, 30105
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:53,336 : INFO : EPOCH 3 - PROGRESS: at 46.67% examples, 30195
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:54,354 : INFO : EPOCH 3 - PROGRESS: at 48.12% examples, 30281
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:55,360 : INFO : EPOCH 3 - PROGRESS: at 49.60% examples, 30354
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:56,377 : INFO : EPOCH 3 - PROGRESS: at 51.10% examples, 30440
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:57,390 : INFO : EPOCH 3 - PROGRESS: at 52.74% examples, 30498
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:58,403 : INFO : EPOCH 3 - PROGRESS: at 54.66% examples, 30570
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:00:59,424 : INFO : EPOCH 3 - PROGRESS: at 56.94% examples, 30651
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:00,465 : INFO : EPOCH 3 - PROGRESS: at 59.22% examples, 30711
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:01,484 : INFO : EPOCH 3 - PROGRESS: at 61.47% examples, 30768
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:02,485 : INFO : EPOCH 3 - PROGRESS: at 63.06% examples, 30876
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:03,508 : INFO : EPOCH 3 - PROGRESS: at 64.39% examples, 30925
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:04,518 : INFO : EPOCH 3 - PROGRESS: at 66.01% examples, 31000
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:05,525 : INFO : EPOCH 3 - PROGRESS: at 67.34% examples, 31037
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:06,544 : INFO : EPOCH 3 - PROGRESS: at 68.76% examples, 31054
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:07,577 : INFO : EPOCH 3 - PROGRESS: at 70.02% examples, 31093
```

```
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:08,610 : INFO : EPOCH 3 - PROGRESS: at 71.43% examples, 31130
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:09,653 : INFO : EPOCH 3 - PROGRESS: at 72.86% examples, 31156
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:10,689 : INFO : EPOCH 3 - PROGRESS: at 74.21% examples, 31186
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:11,689 : INFO : EPOCH 3 - PROGRESS: at 75.54% examples, 31221
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:12,696 : INFO : EPOCH 3 - PROGRESS: at 76.84% examples, 31181
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:13,707 : INFO : EPOCH 3 - PROGRESS: at 78.04% examples, 30972
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:14,710 : INFO : EPOCH 3 - PROGRESS: at 79.89% examples, 30947
1 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:01:15,748 : INFO : EPOCH 3 - PROGRESS: at 81.48% examples, 30820
1 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:01:16,764 : INFO : EPOCH 3 - PROGRESS: at 82.99% examples, 30655
8 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:01:17,844 : INFO : EPOCH 3 - PROGRESS: at 84.51% examples, 30504
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:18,880 : INFO : EPOCH 3 - PROGRESS: at 85.61% examples, 30381
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:19,905 : INFO : EPOCH 3 - PROGRESS: at 86.97% examples, 30323
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:20,929 : INFO : EPOCH 3 - PROGRESS: at 88.29% examples, 30290
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:21,932 : INFO : EPOCH 3 - PROGRESS: at 89.58% examples, 30284
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:01:22,948 : INFO : EPOCH 3 - PROGRESS: at 90.93% examples, 30261
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:23,967 : INFO : EPOCH 3 - PROGRESS: at 92.06% examples, 30197
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:24,988 : INFO : EPOCH 3 - PROGRESS: at 93.21% examples, 30195
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:26,014 : INFO : EPOCH 3 - PROGRESS: at 94.34% examples, 30191
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:27,028 : INFO : EPOCH 3 - PROGRESS: at 95.15% examples, 30066
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:28,086 : INFO : EPOCH 3 - PROGRESS: at 95.97% examples, 29960
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:29,093 : INFO : EPOCH 3 - PROGRESS: at 96.94% examples, 29890
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:30,107 : INFO : EPOCH 3 - PROGRESS: at 97.90% examples, 29820
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:31,142 : INFO : EPOCH 3 - PROGRESS: at 98.82% examples, 29753
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:32,144 : INFO : EPOCH 3 - PROGRESS: at 99.72% examples, 29702
2 words/s, in_qsize 4, out_qsize 3
2021-02-26 22:01:32,247 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:01:32,328 : INFO : worker thread finished; awaiting finish of 2
more threads
```

```
2021-02-26 22:01:32,332 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:01:32,334 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:01:32,335 : INFO : EPOCH - 3 : training on 27374377 raw words (2
1283281 effective words) took 71.6s, 297286 effective words/s
2021-02-26 22:01:33,401 : INFO : EPOCH 4 - PROGRESS: at 2.05% examples, 266792
words/s, in_qsize 8, out_qsize 0
2021-02-26 22:01:34,461 : INFO : EPOCH 4 - PROGRESS: at 3.79% examples, 277085
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:35,470 : INFO : EPOCH 4 - PROGRESS: at 5.61% examples, 280216
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:36,472 : INFO : EPOCH 4 - PROGRESS: at 7.21% examples, 277007
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:37,490 : INFO : EPOCH 4 - PROGRESS: at 8.75% examples, 277001
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:38,492 : INFO : EPOCH 4 - PROGRESS: at 9.85% examples, 270387
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:39,529 : INFO : EPOCH 4 - PROGRESS: at 10.98% examples, 26467
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:40,559 : INFO : EPOCH 4 - PROGRESS: at 12.20% examples, 26148
3 words/s, in_qsize 8, out_qsize 1
2021-02-26 22:01:41,601 : INFO : EPOCH 4 - PROGRESS: at 13.31% examples, 25854
5 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:01:42,627 : INFO : EPOCH 4 - PROGRESS: at 14.78% examples, 25948
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:43,643 : INFO : EPOCH 4 - PROGRESS: at 15.97% examples, 25777
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:44,678 : INFO : EPOCH 4 - PROGRESS: at 17.19% examples, 25718
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:45,737 : INFO : EPOCH 4 - PROGRESS: at 18.48% examples, 25520
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:46,777 : INFO : EPOCH 4 - PROGRESS: at 19.87% examples, 25483
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:47,785 : INFO : EPOCH 4 - PROGRESS: at 21.22% examples, 25557
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:48,793 : INFO : EPOCH 4 - PROGRESS: at 22.62% examples, 25581
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:49,803 : INFO : EPOCH 4 - PROGRESS: at 24.49% examples, 25907
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:50,806 : INFO : EPOCH 4 - PROGRESS: at 25.99% examples, 25951
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:51,854 : INFO : EPOCH 4 - PROGRESS: at 27.61% examples, 26048
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:52,876 : INFO : EPOCH 4 - PROGRESS: at 29.12% examples, 26099
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:53,879 : INFO : EPOCH 4 - PROGRESS: at 30.51% examples, 26198
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:54,903 : INFO : EPOCH 4 - PROGRESS: at 31.70% examples, 26334
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:01:55,910 : INFO : EPOCH 4 - PROGRESS: at 32.95% examples, 26608
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:56,911 : INFO : EPOCH 4 - PROGRESS: at 34.15% examples, 26685
```

```
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:57,924 : INFO : EPOCH 4 - PROGRESS: at 35.34% examples, 26827
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:58,926 : INFO : EPOCH 4 - PROGRESS: at 36.54% examples, 27027
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:01:59,933 : INFO : EPOCH 4 - PROGRESS: at 37.82% examples, 27183
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:00,939 : INFO : EPOCH 4 - PROGRESS: at 39.03% examples, 27303
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:02,000 : INFO : EPOCH 4 - PROGRESS: at 40.18% examples, 27389
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:03,004 : INFO : EPOCH 4 - PROGRESS: at 41.30% examples, 27444
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:04,006 : INFO : EPOCH 4 - PROGRESS: at 42.43% examples, 27548
3 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:02:05,010 : INFO : EPOCH 4 - PROGRESS: at 43.60% examples, 27663
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:06,016 : INFO : EPOCH 4 - PROGRESS: at 44.78% examples, 27726
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:07,048 : INFO : EPOCH 4 - PROGRESS: at 45.62% examples, 27538
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:08,053 : INFO : EPOCH 4 - PROGRESS: at 46.58% examples, 27490
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:09,061 : INFO : EPOCH 4 - PROGRESS: at 47.59% examples, 27401
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:10,076 : INFO : EPOCH 4 - PROGRESS: at 48.75% examples, 27334
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:11,145 : INFO : EPOCH 4 - PROGRESS: at 49.97% examples, 27274
1 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:02:12,169 : INFO : EPOCH 4 - PROGRESS: at 51.20% examples, 27290
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:13,205 : INFO : EPOCH 4 - PROGRESS: at 52.54% examples, 27254
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:14,236 : INFO : EPOCH 4 - PROGRESS: at 54.43% examples, 27385
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:15,270 : INFO : EPOCH 4 - PROGRESS: at 56.39% examples, 27433
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:16,276 : INFO : EPOCH 4 - PROGRESS: at 58.34% examples, 27501
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:17,292 : INFO : EPOCH 4 - PROGRESS: at 60.08% examples, 27456
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:18,299 : INFO : EPOCH 4 - PROGRESS: at 61.73% examples, 27388
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:19,318 : INFO : EPOCH 4 - PROGRESS: at 62.78% examples, 27285
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:20,348 : INFO : EPOCH 4 - PROGRESS: at 63.94% examples, 27291
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:21,364 : INFO : EPOCH 4 - PROGRESS: at 65.22% examples, 27381
6 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:22,394 : INFO : EPOCH 4 - PROGRESS: at 66.52% examples, 27384
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:23,405 : INFO : EPOCH 4 - PROGRESS: at 67.76% examples, 27383
2 words/s, in_qsize 8, out_qsize 0
```

```
2021-02-26 22:02:24,420 : INFO : EPOCH 4 - PROGRESS: at 68.82% examples, 27338
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:25,425 : INFO : EPOCH 4 - PROGRESS: at 69.94% examples, 27387
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:26,447 : INFO : EPOCH 4 - PROGRESS: at 70.86% examples, 27310
8 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:27,451 : INFO : EPOCH 4 - PROGRESS: at 71.93% examples, 27259
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:28,460 : INFO : EPOCH 4 - PROGRESS: at 72.95% examples, 27193
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:29,489 : INFO : EPOCH 4 - PROGRESS: at 73.91% examples, 27134
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:30,506 : INFO : EPOCH 4 - PROGRESS: at 74.83% examples, 27056
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:31,555 : INFO : EPOCH 4 - PROGRESS: at 75.96% examples, 27018
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:32,582 : INFO : EPOCH 4 - PROGRESS: at 77.49% examples, 27065
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:33,592 : INFO : EPOCH 4 - PROGRESS: at 79.35% examples, 27117
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:34,605 : INFO : EPOCH 4 - PROGRESS: at 81.26% examples, 27165
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:35,609 : INFO : EPOCH 4 - PROGRESS: at 82.85% examples, 27108
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:36,626 : INFO : EPOCH 4 - PROGRESS: at 84.37% examples, 27047
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:37,674 : INFO : EPOCH 4 - PROGRESS: at 85.55% examples, 27012
3 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:02:38,687 : INFO : EPOCH 4 - PROGRESS: at 86.76% examples, 26971
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:39,692 : INFO : EPOCH 4 - PROGRESS: at 88.18% examples, 27034
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:40,747 : INFO : EPOCH 4 - PROGRESS: at 89.49% examples, 27056
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:41,773 : INFO : EPOCH 4 - PROGRESS: at 90.89% examples, 27101
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:42,779 : INFO : EPOCH 4 - PROGRESS: at 92.15% examples, 27137
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:43,807 : INFO : EPOCH 4 - PROGRESS: at 93.27% examples, 27165
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:44,843 : INFO : EPOCH 4 - PROGRESS: at 94.39% examples, 27201
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:45,876 : INFO : EPOCH 4 - PROGRESS: at 95.47% examples, 27228
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:46,938 : INFO : EPOCH 4 - PROGRESS: at 96.56% examples, 27254
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:47,953 : INFO : EPOCH 4 - PROGRESS: at 97.53% examples, 27224
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:48,974 : INFO : EPOCH 4 - PROGRESS: at 98.51% examples, 27233
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:49,988 : INFO : EPOCH 4 - PROGRESS: at 99.64% examples, 27283
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:50,208 : INFO : worker thread finished; awaiting finish of 3
```

```
more threads
2021-02-26 22:02:50,241 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:02:50,256 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:02:50,263 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:02:50,264 : INFO : EPOCH - 4 : training on 27374377 raw words (2
1284152 effective words) took 77.9s, 273132 effective words/s
2021-02-26 22:02:51,304 : INFO : EPOCH 5 - PROGRESS: at 2.05% examples, 273260
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:52,304 : INFO : EPOCH 5 - PROGRESS: at 3.83% examples, 292371
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:53,308 : INFO : EPOCH 5 - PROGRESS: at 5.74% examples, 295985
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:54,324 : INFO : EPOCH 5 - PROGRESS: at 7.49% examples, 295241
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:55,378 : INFO : EPOCH 5 - PROGRESS: at 9.13% examples, 293924
words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:56,386 : INFO : EPOCH 5 - PROGRESS: at 10.47% examples, 29442
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:57,391 : INFO : EPOCH 5 - PROGRESS: at 12.02% examples, 29412
8 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:02:58,392 : INFO : EPOCH 5 - PROGRESS: at 13.26% examples, 29278
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:02:59,416 : INFO : EPOCH 5 - PROGRESS: at 14.87% examples, 29334
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:00,441 : INFO : EPOCH 5 - PROGRESS: at 16.31% examples, 29381
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:01,509 : INFO : EPOCH 5 - PROGRESS: at 17.74% examples, 29169
4 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:02,539 : INFO : EPOCH 5 - PROGRESS: at 19.39% examples, 29237
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:03,578 : INFO : EPOCH 5 - PROGRESS: at 20.77% examples, 28962
5 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:04,612 : INFO : EPOCH 5 - PROGRESS: at 22.38% examples, 29021
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:05,617 : INFO : EPOCH 5 - PROGRESS: at 23.80% examples, 28771
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:06,630 : INFO : EPOCH 5 - PROGRESS: at 25.37% examples, 28633
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:07,647 : INFO : EPOCH 5 - PROGRESS: at 26.83% examples, 28447
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:08,725 : INFO : EPOCH 5 - PROGRESS: at 28.41% examples, 28326
5 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:09,780 : INFO : EPOCH 5 - PROGRESS: at 29.83% examples, 28215
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:10,840 : INFO : EPOCH 5 - PROGRESS: at 30.96% examples, 27947
6 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:11,843 : INFO : EPOCH 5 - PROGRESS: at 31.81% examples, 27681
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:12,876 : INFO : EPOCH 5 - PROGRESS: at 32.77% examples, 27536
5 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 22:03:13,923 : INFO : EPOCH 5 - PROGRESS: at 33.86% examples, 27389
5 words/s, in_qsize 5, out_qsize 2
2021-02-26 22:03:15,088 : INFO : EPOCH 5 - PROGRESS: at 34.83% examples, 27097
2 words/s, in_qsize 8, out_qsize 1
2021-02-26 22:03:16,096 : INFO : EPOCH 5 - PROGRESS: at 35.61% examples, 26870
9 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:17,107 : INFO : EPOCH 5 - PROGRESS: at 36.68% examples, 26885
7 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:18,127 : INFO : EPOCH 5 - PROGRESS: at 37.58% examples, 26702
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:19,186 : INFO : EPOCH 5 - PROGRESS: at 38.71% examples, 26708
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:20,190 : INFO : EPOCH 5 - PROGRESS: at 39.53% examples, 26537
5 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:21,195 : INFO : EPOCH 5 - PROGRESS: at 40.30% examples, 26367
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:22,196 : INFO : EPOCH 5 - PROGRESS: at 41.41% examples, 26455
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:23,229 : INFO : EPOCH 5 - PROGRESS: at 42.17% examples, 26257
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:24,272 : INFO : EPOCH 5 - PROGRESS: at 43.11% examples, 26219
9 words/s, in_qsize 8, out_qsize 1
2021-02-26 22:03:25,293 : INFO : EPOCH 5 - PROGRESS: at 44.12% examples, 26158
4 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:26,345 : INFO : EPOCH 5 - PROGRESS: at 45.15% examples, 26160
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:27,382 : INFO : EPOCH 5 - PROGRESS: at 46.32% examples, 26273
0 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:28,383 : INFO : EPOCH 5 - PROGRESS: at 47.59% examples, 26403
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:29,385 : INFO : EPOCH 5 - PROGRESS: at 49.05% examples, 26571
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:30,425 : INFO : EPOCH 5 - PROGRESS: at 50.29% examples, 26533
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:31,469 : INFO : EPOCH 5 - PROGRESS: at 51.47% examples, 26534
4 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:32,540 : INFO : EPOCH 5 - PROGRESS: at 52.96% examples, 26551
0 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:33,638 : INFO : EPOCH 5 - PROGRESS: at 54.57% examples, 26511
7 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:34,701 : INFO : EPOCH 5 - PROGRESS: at 56.16% examples, 26424
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:35,751 : INFO : EPOCH 5 - PROGRESS: at 57.82% examples, 26370
2 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:36,804 : INFO : EPOCH 5 - PROGRESS: at 59.49% examples, 26314
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:37,825 : INFO : EPOCH 5 - PROGRESS: at 61.07% examples, 26215
8 words/s, in_qsize 5, out_qsize 2
2021-02-26 22:03:38,844 : INFO : EPOCH 5 - PROGRESS: at 62.17% examples, 26123
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:39,863 : INFO : EPOCH 5 - PROGRESS: at 63.24% examples, 26082
5 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:40,882 : INFO : EPOCH 5 - PROGRESS: at 64.27% examples, 26071
```

```
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:41,903 : INFO : EPOCH 5 - PROGRESS: at 65.39% examples, 26075
0 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:42,915 : INFO : EPOCH 5 - PROGRESS: at 66.69% examples, 26126
5 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:43,960 : INFO : EPOCH 5 - PROGRESS: at 68.02% examples, 26163
8 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:44,973 : INFO : EPOCH 5 - PROGRESS: at 69.06% examples, 26157
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:46,006 : INFO : EPOCH 5 - PROGRESS: at 70.15% examples, 26183
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:47,046 : INFO : EPOCH 5 - PROGRESS: at 71.08% examples, 26111
0 words/s, in_qsize 7, out_qsize 1
2021-02-26 22:03:48,054 : INFO : EPOCH 5 - PROGRESS: at 72.27% examples, 26133
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:49,083 : INFO : EPOCH 5 - PROGRESS: at 73.27% examples, 26094
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:50,089 : INFO : EPOCH 5 - PROGRESS: at 74.15% examples, 26015
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:51,119 : INFO : EPOCH 5 - PROGRESS: at 75.08% examples, 25966
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:52,128 : INFO : EPOCH 5 - PROGRESS: at 76.36% examples, 26029
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:53,133 : INFO : EPOCH 5 - PROGRESS: at 77.99% examples, 26085
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:03:54,193 : INFO : EPOCH 5 - PROGRESS: at 79.80% examples, 26105
7 words/s, in_qsize 8, out_qsize 2
2021-02-26 22:03:55,231 : INFO : EPOCH 5 - PROGRESS: at 81.70% examples, 26156
6 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:56,248 : INFO : EPOCH 5 - PROGRESS: at 83.93% examples, 26228
6 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:03:57,259 : INFO : EPOCH 5 - PROGRESS: at 85.35% examples, 26277
8 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:03:58,321 : INFO : EPOCH 5 - PROGRESS: at 86.70% examples, 26275
4 words/s, in_qsize 5, out_qsize 2
2021-02-26 22:03:59,384 : INFO : EPOCH 5 - PROGRESS: at 88.14% examples, 26336
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:00,463 : INFO : EPOCH 5 - PROGRESS: at 89.49% examples, 26370
1 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:04:01,479 : INFO : EPOCH 5 - PROGRESS: at 90.89% examples, 26427
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:02,527 : INFO : EPOCH 5 - PROGRESS: at 92.12% examples, 26446
2 words/s, in_qsize 8, out_qsize 3
2021-02-26 22:04:03,566 : INFO : EPOCH 5 - PROGRESS: at 93.28% examples, 26490
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:04,569 : INFO : EPOCH 5 - PROGRESS: at 94.32% examples, 26514
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:05,605 : INFO : EPOCH 5 - PROGRESS: at 95.39% examples, 26549
3 words/s, in_qsize 8, out_qsize 1
2021-02-26 22:04:06,614 : INFO : EPOCH 5 - PROGRESS: at 96.36% examples, 26562
3 words/s, in_qsize 5, out_qsize 2
2021-02-26 22:04:07,638 : INFO : EPOCH 5 - PROGRESS: at 97.47% examples, 26588
6 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 22:04:08,646 : INFO : EPOCH 5 - PROGRESS: at 98.36% examples, 26580
3 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:04:09,703 : INFO : EPOCH 5 - PROGRESS: at 99.42% examples, 26593
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:10,185 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:04:10,188 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:04:10,224 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:04:10,226 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:04:10,226 : INFO : EPOCH - 5 : training on 27374377 raw words (2
1285476 effective words) took 80.0s, 266204 effective words/s
2021-02-26 22:04:10,228 : INFO : training on a 136871885 raw words (106416042
effective words) took 360.8s, 294921 effective words/s
```

In [47]:

```
vectors9.train(tokenized_text1, total_examples=vectors9.corpus_count, epochs=
```

```
2021-02-26 22:04:39,225 : WARNING : Effective 'alpha' higher than previous tra
ining cycles
2021-02-26 22:04:39,229 : INFO : training model with 4 workers on 257022 vocab
ulary and 10 features, using sg=1 hs=0 sample=0.001 negative=5 window=5
2021-02-26 22:04:40,279 : INFO : EPOCH 1 - PROGRESS: at 2.55% examples, 358649
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:41,286 : INFO : EPOCH 1 - PROGRESS: at 4.38% examples, 334630
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:42,296 : INFO : EPOCH 1 - PROGRESS: at 6.53% examples, 334164
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:43,362 : INFO : EPOCH 1 - PROGRESS: at 8.34% examples, 325247
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:44,393 : INFO : EPOCH 1 - PROGRESS: at 9.82% examples, 320936
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:45,427 : INFO : EPOCH 1 - PROGRESS: at 11.49% examples, 31845
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:46,453 : INFO : EPOCH 1 - PROGRESS: at 12.82% examples, 31687
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:47,555 : INFO : EPOCH 1 - PROGRESS: at 14.35% examples, 31152
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:48,597 : INFO : EPOCH 1 - PROGRESS: at 15.75% examples, 30710
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:49,632 : INFO : EPOCH 1 - PROGRESS: at 17.07% examples, 30297
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:50,647 : INFO : EPOCH 1 - PROGRESS: at 18.36% examples, 29755
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:51,658 : INFO : EPOCH 1 - PROGRESS: at 20.00% examples, 29799
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:52,670 : INFO : EPOCH 1 - PROGRESS: at 21.39% examples, 29607
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:53,679 : INFO : EPOCH 1 - PROGRESS: at 22.94% examples, 29561
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:54,682 : INFO : EPOCH 1 - PROGRESS: at 24.02% examples, 28886
```

```
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:55,691 : INFO : EPOCH 1 - PROGRESS: at 25.95% examples, 29069
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:56,735 : INFO : EPOCH 1 - PROGRESS: at 27.84% examples, 29262
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:57,795 : INFO : EPOCH 1 - PROGRESS: at 29.28% examples, 29080
2 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:04:58,805 : INFO : EPOCH 1 - PROGRESS: at 30.64% examples, 28985
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:04:59,818 : INFO : EPOCH 1 - PROGRESS: at 31.84% examples, 29050
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:00,903 : INFO : EPOCH 1 - PROGRESS: at 33.01% examples, 29010
2 words/s, in_qsize 8, out_qsize 2
2021-02-26 22:05:01,970 : INFO : EPOCH 1 - PROGRESS: at 34.18% examples, 28869
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:02,996 : INFO : EPOCH 1 - PROGRESS: at 35.25% examples, 28783
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:04,047 : INFO : EPOCH 1 - PROGRESS: at 36.38% examples, 28804
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:05,087 : INFO : EPOCH 1 - PROGRESS: at 37.52% examples, 28711
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:06,152 : INFO : EPOCH 1 - PROGRESS: at 38.52% examples, 28514
9 words/s, in_qsize 8, out_qsize 2
2021-02-26 22:05:07,245 : INFO : EPOCH 1 - PROGRESS: at 39.46% examples, 28288
2 words/s, in_qsize 4, out_qsize 3
2021-02-26 22:05:08,289 : INFO : EPOCH 1 - PROGRESS: at 40.49% examples, 28220
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:09,312 : INFO : EPOCH 1 - PROGRESS: at 41.57% examples, 28232
5 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:10,334 : INFO : EPOCH 1 - PROGRESS: at 42.68% examples, 28270
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:11,359 : INFO : EPOCH 1 - PROGRESS: at 43.87% examples, 28322
8 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:12,368 : INFO : EPOCH 1 - PROGRESS: at 45.17% examples, 28501
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:13,391 : INFO : EPOCH 1 - PROGRESS: at 46.47% examples, 28652
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:14,393 : INFO : EPOCH 1 - PROGRESS: at 47.82% examples, 28770
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:15,436 : INFO : EPOCH 1 - PROGRESS: at 49.18% examples, 28789
6 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:05:16,470 : INFO : EPOCH 1 - PROGRESS: at 50.66% examples, 28857
2 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:05:17,480 : INFO : EPOCH 1 - PROGRESS: at 52.12% examples, 28920
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:18,480 : INFO : EPOCH 1 - PROGRESS: at 53.99% examples, 29040
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:19,500 : INFO : EPOCH 1 - PROGRESS: at 55.93% examples, 29077
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:20,502 : INFO : EPOCH 1 - PROGRESS: at 58.03% examples, 29150
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:21,510 : INFO : EPOCH 1 - PROGRESS: at 59.98% examples, 29159
8 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 22:05:22,526 : INFO : EPOCH 1 - PROGRESS: at 61.77% examples, 29094
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:23,551 : INFO : EPOCH 1 - PROGRESS: at 63.18% examples, 29150
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:24,579 : INFO : EPOCH 1 - PROGRESS: at 64.51% examples, 29232
6 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:05:25,598 : INFO : EPOCH 1 - PROGRESS: at 66.17% examples, 29349
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:26,604 : INFO : EPOCH 1 - PROGRESS: at 67.68% examples, 29485
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:27,697 : INFO : EPOCH 1 - PROGRESS: at 69.09% examples, 29537
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:28,738 : INFO : EPOCH 1 - PROGRESS: at 70.21% examples, 29507
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:29,788 : INFO : EPOCH 1 - PROGRESS: at 71.61% examples, 29565
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:30,858 : INFO : EPOCH 1 - PROGRESS: at 72.92% examples, 29546
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:31,882 : INFO : EPOCH 1 - PROGRESS: at 74.26% examples, 29613
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:32,912 : INFO : EPOCH 1 - PROGRESS: at 75.54% examples, 29631
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:33,915 : INFO : EPOCH 1 - PROGRESS: at 77.15% examples, 29720
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:34,924 : INFO : EPOCH 1 - PROGRESS: at 79.19% examples, 29783
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:35,937 : INFO : EPOCH 1 - PROGRESS: at 81.30% examples, 29842
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:36,937 : INFO : EPOCH 1 - PROGRESS: at 83.70% examples, 29920
0 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:37,989 : INFO : EPOCH 1 - PROGRESS: at 85.35% examples, 29956
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:38,989 : INFO : EPOCH 1 - PROGRESS: at 86.87% examples, 29983
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:40,001 : INFO : EPOCH 1 - PROGRESS: at 88.22% examples, 29975
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:41,053 : INFO : EPOCH 1 - PROGRESS: at 89.62% examples, 29988
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:42,075 : INFO : EPOCH 1 - PROGRESS: at 91.06% examples, 30003
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:43,083 : INFO : EPOCH 1 - PROGRESS: at 92.37% examples, 30021
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:44,108 : INFO : EPOCH 1 - PROGRESS: at 93.44% examples, 29997
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:45,114 : INFO : EPOCH 1 - PROGRESS: at 94.58% examples, 30017
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:46,156 : INFO : EPOCH 1 - PROGRESS: at 95.63% examples, 29998
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:47,159 : INFO : EPOCH 1 - PROGRESS: at 96.86% examples, 30043
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:48,162 : INFO : EPOCH 1 - PROGRESS: at 97.90% examples, 30009
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:49,172 : INFO : EPOCH 1 - PROGRESS: at 98.91% examples, 29992
```

```
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:50,176 : INFO : EPOCH 1 - PROGRESS: at 99.87% examples, 29948
0 words/s, in_qsize 5, out_qsize 0
2021-02-26 22:05:50,232 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:05:50,254 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:05:50,263 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:05:50,290 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:05:50,291 : INFO : EPOCH - 1 : training on 27374377 raw words (2
1282478 effective words) took 71.1s, 299516 effective words/s
2021-02-26 22:05:51,307 : INFO : EPOCH 2 - PROGRESS: at 2.10% examples, 287341
words/s, in_qsize 6, out_qsize 1
2021-02-26 22:05:52,371 : INFO : EPOCH 2 - PROGRESS: at 3.79% examples, 283194
words/s, in_qsize 8, out_qsize 0
2021-02-26 22:05:53,386 : INFO : EPOCH 2 - PROGRESS: at 5.57% examples, 281405
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:54,392 : INFO : EPOCH 2 - PROGRESS: at 7.58% examples, 296127
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:55,400 : INFO : EPOCH 2 - PROGRESS: at 9.13% examples, 294371
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:56,406 : INFO : EPOCH 2 - PROGRESS: at 10.63% examples, 30004
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:57,409 : INFO : EPOCH 2 - PROGRESS: at 12.17% examples, 30093
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:58,440 : INFO : EPOCH 2 - PROGRESS: at 13.63% examples, 30236
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:05:59,447 : INFO : EPOCH 2 - PROGRESS: at 15.26% examples, 30168
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:00,451 : INFO : EPOCH 2 - PROGRESS: at 16.80% examples, 30488
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:01,472 : INFO : EPOCH 2 - PROGRESS: at 18.54% examples, 30721
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:02,516 : INFO : EPOCH 2 - PROGRESS: at 20.36% examples, 30916
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:03,545 : INFO : EPOCH 2 - PROGRESS: at 22.05% examples, 31063
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:04,580 : INFO : EPOCH 2 - PROGRESS: at 23.98% examples, 31182
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:05,599 : INFO : EPOCH 2 - PROGRESS: at 25.81% examples, 31110
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:06,611 : INFO : EPOCH 2 - PROGRESS: at 27.56% examples, 31104
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:07,626 : INFO : EPOCH 2 - PROGRESS: at 29.28% examples, 31143
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:08,635 : INFO : EPOCH 2 - PROGRESS: at 30.64% examples, 30930
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:09,670 : INFO : EPOCH 2 - PROGRESS: at 31.94% examples, 30980
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:10,683 : INFO : EPOCH 2 - PROGRESS: at 33.27% examples, 31133
0 words/s, in_qsize 8, out_qsize 1
```

```
2021-02-26 22:06:11,731 : INFO : EPOCH 2 - PROGRESS: at 34.77% examples, 31299
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:12,772 : INFO : EPOCH 2 - PROGRESS: at 36.02% examples, 31354
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:13,775 : INFO : EPOCH 2 - PROGRESS: at 37.23% examples, 31284
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:14,784 : INFO : EPOCH 2 - PROGRESS: at 38.30% examples, 31124
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:15,829 : INFO : EPOCH 2 - PROGRESS: at 39.53% examples, 31093
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:16,893 : INFO : EPOCH 2 - PROGRESS: at 40.68% examples, 30999
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:17,898 : INFO : EPOCH 2 - PROGRESS: at 41.74% examples, 30929
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:18,899 : INFO : EPOCH 2 - PROGRESS: at 42.84% examples, 30897
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:19,912 : INFO : EPOCH 2 - PROGRESS: at 44.15% examples, 30955
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:20,952 : INFO : EPOCH 2 - PROGRESS: at 45.20% examples, 30827
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:21,960 : INFO : EPOCH 2 - PROGRESS: at 46.23% examples, 30715
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:22,979 : INFO : EPOCH 2 - PROGRESS: at 47.32% examples, 30595
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:23,995 : INFO : EPOCH 2 - PROGRESS: at 48.72% examples, 30583
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:25,004 : INFO : EPOCH 2 - PROGRESS: at 49.87% examples, 30421
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:26,042 : INFO : EPOCH 2 - PROGRESS: at 51.10% examples, 30336
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:27,050 : INFO : EPOCH 2 - PROGRESS: at 52.78% examples, 30423
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:28,060 : INFO : EPOCH 2 - PROGRESS: at 54.80% examples, 30538
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:29,061 : INFO : EPOCH 2 - PROGRESS: at 57.07% examples, 30616
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:30,064 : INFO : EPOCH 2 - PROGRESS: at 59.01% examples, 30591
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:31,075 : INFO : EPOCH 2 - PROGRESS: at 61.11% examples, 30582
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:32,081 : INFO : EPOCH 2 - PROGRESS: at 62.45% examples, 30507
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:33,113 : INFO : EPOCH 2 - PROGRESS: at 63.61% examples, 30418
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:34,129 : INFO : EPOCH 2 - PROGRESS: at 64.76% examples, 30377
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:35,146 : INFO : EPOCH 2 - PROGRESS: at 66.27% examples, 30405
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:06:36,165 : INFO : EPOCH 2 - PROGRESS: at 67.58% examples, 30413
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:37,171 : INFO : EPOCH 2 - PROGRESS: at 68.85% examples, 30387
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:38,202 : INFO : EPOCH 2 - PROGRESS: at 69.83% examples, 30280
```

```
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:39,213 : INFO : EPOCH 2 - PROGRESS: at 71.05% examples, 30284
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:40,220 : INFO : EPOCH 2 - PROGRESS: at 72.23% examples, 30226
8 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:06:41,259 : INFO : EPOCH 2 - PROGRESS: at 73.30% examples, 30122
3 words/s, in_qsize 8, out_qsize 1
2021-02-26 22:06:42,268 : INFO : EPOCH 2 - PROGRESS: at 74.40% examples, 30068
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:43,299 : INFO : EPOCH 2 - PROGRESS: at 75.83% examples, 30122
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:44,368 : INFO : EPOCH 2 - PROGRESS: at 77.16% examples, 30051
6 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:06:45,389 : INFO : EPOCH 2 - PROGRESS: at 78.98% examples, 30046
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:46,426 : INFO : EPOCH 2 - PROGRESS: at 81.21% examples, 30116
7 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:06:47,455 : INFO : EPOCH 2 - PROGRESS: at 83.56% examples, 30161
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:48,477 : INFO : EPOCH 2 - PROGRESS: at 84.90% examples, 30078
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:49,484 : INFO : EPOCH 2 - PROGRESS: at 86.24% examples, 30072
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:50,485 : INFO : EPOCH 2 - PROGRESS: at 87.45% examples, 29992
5 words/s, in_qsize 8, out_qsize 3
2021-02-26 22:06:51,558 : INFO : EPOCH 2 - PROGRESS: at 88.87% examples, 29967
2 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:06:52,608 : INFO : EPOCH 2 - PROGRESS: at 90.30% examples, 30009
1 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:06:53,629 : INFO : EPOCH 2 - PROGRESS: at 91.88% examples, 30057
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:54,634 : INFO : EPOCH 2 - PROGRESS: at 93.16% examples, 30125
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:55,636 : INFO : EPOCH 2 - PROGRESS: at 94.45% examples, 30203
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:56,643 : INFO : EPOCH 2 - PROGRESS: at 95.63% examples, 30255
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:57,650 : INFO : EPOCH 2 - PROGRESS: at 96.86% examples, 30295
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:58,654 : INFO : EPOCH 2 - PROGRESS: at 98.13% examples, 30358
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:06:59,680 : INFO : EPOCH 2 - PROGRESS: at 99.25% examples, 30373
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:00,292 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:07:00,328 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:07:00,336 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:07:00,352 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:07:00,353 : INFO : EPOCH - 2 : training on 27374377 raw words (2
1281193 effective words) took 70.1s, 303760 effective words/s
```

```
2021-02-26 22:07:01,391 : INFO : EPOCH 3 - PROGRESS: at 2.26% examples, 310603
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:02,393 : INFO : EPOCH 3 - PROGRESS: at 4.05% examples, 310919
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:03,395 : INFO : EPOCH 3 - PROGRESS: at 6.26% examples, 321531
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:04,406 : INFO : EPOCH 3 - PROGRESS: at 7.99% examples, 316419
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:05,408 : INFO : EPOCH 3 - PROGRESS: at 9.61% examples, 318558
words/s, in_qsize 8, out_qsize 0
2021-02-26 22:07:06,422 : INFO : EPOCH 3 - PROGRESS: at 10.98% examples, 31364
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:07,446 : INFO : EPOCH 3 - PROGRESS: at 12.41% examples, 31063
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:08,448 : INFO : EPOCH 3 - PROGRESS: at 13.93% examples, 31268
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:09,448 : INFO : EPOCH 3 - PROGRESS: at 15.80% examples, 31706
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:10,454 : INFO : EPOCH 3 - PROGRESS: at 17.53% examples, 32101
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:11,470 : INFO : EPOCH 3 - PROGRESS: at 19.23% examples, 32011
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:12,520 : INFO : EPOCH 3 - PROGRESS: at 20.95% examples, 32012
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:13,596 : INFO : EPOCH 3 - PROGRESS: at 22.42% examples, 31504
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:14,596 : INFO : EPOCH 3 - PROGRESS: at 23.98% examples, 31285
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:15,599 : INFO : EPOCH 3 - PROGRESS: at 25.86% examples, 31290
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:16,600 : INFO : EPOCH 3 - PROGRESS: at 27.51% examples, 31196
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:17,602 : INFO : EPOCH 3 - PROGRESS: at 29.25% examples, 31256
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:18,629 : INFO : EPOCH 3 - PROGRESS: at 30.67% examples, 31088
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:19,634 : INFO : EPOCH 3 - PROGRESS: at 31.84% examples, 31020
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:20,639 : INFO : EPOCH 3 - PROGRESS: at 33.16% examples, 31181
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:21,647 : INFO : EPOCH 3 - PROGRESS: at 34.62% examples, 31334
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:22,654 : INFO : EPOCH 3 - PROGRESS: at 35.93% examples, 31504
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:23,665 : INFO : EPOCH 3 - PROGRESS: at 37.34% examples, 31646
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:24,690 : INFO : EPOCH 3 - PROGRESS: at 38.40% examples, 31419
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:25,727 : INFO : EPOCH 3 - PROGRESS: at 39.46% examples, 31232
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:26,746 : INFO : EPOCH 3 - PROGRESS: at 40.56% examples, 31128
7 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:07:27,749 : INFO : EPOCH 3 - PROGRESS: at 41.63% examples, 31055
```

```
7 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:07:28,764 : INFO : EPOCH 3 - PROGRESS: at 42.70% examples, 30976
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:29,768 : INFO : EPOCH 3 - PROGRESS: at 43.93% examples, 30988
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:30,796 : INFO : EPOCH 3 - PROGRESS: at 45.00% examples, 30874
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:31,813 : INFO : EPOCH 3 - PROGRESS: at 46.35% examples, 31017
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:32,821 : INFO : EPOCH 3 - PROGRESS: at 47.86% examples, 31181
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:33,847 : INFO : EPOCH 3 - PROGRESS: at 49.34% examples, 31234
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:34,895 : INFO : EPOCH 3 - PROGRESS: at 50.62% examples, 31089
3 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:07:35,923 : INFO : EPOCH 3 - PROGRESS: at 51.89% examples, 30971
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:36,969 : INFO : EPOCH 3 - PROGRESS: at 53.51% examples, 30922
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:38,010 : INFO : EPOCH 3 - PROGRESS: at 55.17% examples, 30794
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:39,013 : INFO : EPOCH 3 - PROGRESS: at 57.53% examples, 30904
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:40,035 : INFO : EPOCH 3 - PROGRESS: at 59.45% examples, 30838
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:41,051 : INFO : EPOCH 3 - PROGRESS: at 61.47% examples, 30801
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:42,065 : INFO : EPOCH 3 - PROGRESS: at 62.81% examples, 30751
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:43,074 : INFO : EPOCH 3 - PROGRESS: at 64.10% examples, 30761
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:44,079 : INFO : EPOCH 3 - PROGRESS: at 65.42% examples, 30806
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:45,086 : INFO : EPOCH 3 - PROGRESS: at 66.67% examples, 30729
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:46,116 : INFO : EPOCH 3 - PROGRESS: at 68.19% examples, 30778
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:47,143 : INFO : EPOCH 3 - PROGRESS: at 69.25% examples, 30695
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:48,147 : INFO : EPOCH 3 - PROGRESS: at 70.61% examples, 30790
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:49,173 : INFO : EPOCH 3 - PROGRESS: at 71.80% examples, 30712
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:50,212 : INFO : EPOCH 3 - PROGRESS: at 73.03% examples, 30656
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:51,228 : INFO : EPOCH 3 - PROGRESS: at 74.21% examples, 30618
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:52,241 : INFO : EPOCH 3 - PROGRESS: at 75.58% examples, 30671
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:53,248 : INFO : EPOCH 3 - PROGRESS: at 76.61% examples, 30556
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:54,274 : INFO : EPOCH 3 - PROGRESS: at 78.60% examples, 30593
7 words/s, in_qsize 8, out_qsize 0
```

```
2021-02-26 22:07:55,280 : INFO : EPOCH 3 - PROGRESS: at 80.40% examples, 30547
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:56,284 : INFO : EPOCH 3 - PROGRESS: at 82.18% examples, 30501
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:57,298 : INFO : EPOCH 3 - PROGRESS: at 84.33% examples, 30522
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:58,306 : INFO : EPOCH 3 - PROGRESS: at 85.76% examples, 30548
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:07:59,315 : INFO : EPOCH 3 - PROGRESS: at 87.19% examples, 30532
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:00,325 : INFO : EPOCH 3 - PROGRESS: at 88.56% examples, 30490
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:01,383 : INFO : EPOCH 3 - PROGRESS: at 89.75% examples, 30428
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:02,416 : INFO : EPOCH 3 - PROGRESS: at 91.03% examples, 30370
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:03,426 : INFO : EPOCH 3 - PROGRESS: at 92.21% examples, 30332
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:04,428 : INFO : EPOCH 3 - PROGRESS: at 93.24% examples, 30290
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:05,442 : INFO : EPOCH 3 - PROGRESS: at 94.37% examples, 30290
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:06,465 : INFO : EPOCH 3 - PROGRESS: at 95.58% examples, 30346
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:07,485 : INFO : EPOCH 3 - PROGRESS: at 96.80% examples, 30378
4 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:08:08,499 : INFO : EPOCH 3 - PROGRESS: at 97.98% examples, 30390
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:09,513 : INFO : EPOCH 3 - PROGRESS: at 99.01% examples, 30377
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:10,320 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:08:10,351 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:08:10,372 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:08:10,391 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:08:10,393 : INFO : EPOCH - 3 : training on 27374377 raw words (2
1283801 effective words) took 70.0s, 303896 effective words/s
2021-02-26 22:08:11,409 : INFO : EPOCH 4 - PROGRESS: at 2.05% examples, 279551
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:12,415 : INFO : EPOCH 4 - PROGRESS: at 3.83% examples, 294681
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:13,419 : INFO : EPOCH 4 - PROGRESS: at 5.88% examples, 305061
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:14,441 : INFO : EPOCH 4 - PROGRESS: at 7.94% examples, 314716
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:15,447 : INFO : EPOCH 4 - PROGRESS: at 9.68% examples, 321438
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:16,485 : INFO : EPOCH 4 - PROGRESS: at 11.49% examples, 32374
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:17,491 : INFO : EPOCH 4 - PROGRESS: at 12.96% examples, 32658
```

```
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:18,498 : INFO : EPOCH 4 - PROGRESS: at 14.69% examples, 32739
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:19,502 : INFO : EPOCH 4 - PROGRESS: at 16.34% examples, 32906
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:20,530 : INFO : EPOCH 4 - PROGRESS: at 17.99% examples, 32803
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:21,536 : INFO : EPOCH 4 - PROGRESS: at 19.79% examples, 32878
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:22,541 : INFO : EPOCH 4 - PROGRESS: at 21.63% examples, 33124
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:23,580 : INFO : EPOCH 4 - PROGRESS: at 23.57% examples, 33140
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:24,633 : INFO : EPOCH 4 - PROGRESS: at 25.68% examples, 33275
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:25,653 : INFO : EPOCH 4 - PROGRESS: at 27.71% examples, 33409
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:26,663 : INFO : EPOCH 4 - PROGRESS: at 29.35% examples, 33269
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:27,666 : INFO : EPOCH 4 - PROGRESS: at 30.98% examples, 33328
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:28,667 : INFO : EPOCH 4 - PROGRESS: at 32.29% examples, 33351
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:29,692 : INFO : EPOCH 4 - PROGRESS: at 33.63% examples, 33287
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:30,708 : INFO : EPOCH 4 - PROGRESS: at 35.05% examples, 33364
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:31,720 : INFO : EPOCH 4 - PROGRESS: at 36.35% examples, 33473
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:32,724 : INFO : EPOCH 4 - PROGRESS: at 37.66% examples, 33410
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:33,741 : INFO : EPOCH 4 - PROGRESS: at 39.09% examples, 33504
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:34,752 : INFO : EPOCH 4 - PROGRESS: at 40.36% examples, 33530
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:35,792 : INFO : EPOCH 4 - PROGRESS: at 41.71% examples, 33580
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:36,836 : INFO : EPOCH 4 - PROGRESS: at 42.73% examples, 33303
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:37,839 : INFO : EPOCH 4 - PROGRESS: at 43.87% examples, 33146
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:08:38,855 : INFO : EPOCH 4 - PROGRESS: at 44.86% examples, 32881
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:39,869 : INFO : EPOCH 4 - PROGRESS: at 45.90% examples, 32682
9 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:08:40,920 : INFO : EPOCH 4 - PROGRESS: at 46.87% examples, 32380
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:41,952 : INFO : EPOCH 4 - PROGRESS: at 48.03% examples, 32191
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:42,983 : INFO : EPOCH 4 - PROGRESS: at 49.55% examples, 32210
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:44,022 : INFO : EPOCH 4 - PROGRESS: at 50.92% examples, 32131
7 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 22:08:45,063 : INFO : EPOCH 4 - PROGRESS: at 52.21% examples, 31966
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:46,109 : INFO : EPOCH 4 - PROGRESS: at 53.81% examples, 31823
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:47,157 : INFO : EPOCH 4 - PROGRESS: at 55.88% examples, 31824
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:48,162 : INFO : EPOCH 4 - PROGRESS: at 58.07% examples, 31869
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:49,170 : INFO : EPOCH 4 - PROGRESS: at 60.33% examples, 31908
2 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:08:50,184 : INFO : EPOCH 4 - PROGRESS: at 62.26% examples, 31942
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:51,211 : INFO : EPOCH 4 - PROGRESS: at 63.58% examples, 31891
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:52,241 : INFO : EPOCH 4 - PROGRESS: at 64.61% examples, 31728
2 words/s, in_qsize 8, out_qsize 1
2021-02-26 22:08:53,333 : INFO : EPOCH 4 - PROGRESS: at 66.08% examples, 31636
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:54,357 : INFO : EPOCH 4 - PROGRESS: at 67.34% examples, 31611
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:55,398 : INFO : EPOCH 4 - PROGRESS: at 68.57% examples, 31496
6 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:08:56,431 : INFO : EPOCH 4 - PROGRESS: at 69.64% examples, 31393
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:57,450 : INFO : EPOCH 4 - PROGRESS: at 70.70% examples, 31317
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:58,458 : INFO : EPOCH 4 - PROGRESS: at 71.80% examples, 31191
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:08:59,459 : INFO : EPOCH 4 - PROGRESS: at 73.00% examples, 31132
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:00,587 : INFO : EPOCH 4 - PROGRESS: at 74.23% examples, 31045
0 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:09:01,596 : INFO : EPOCH 4 - PROGRESS: at 75.47% examples, 31033
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:02,597 : INFO : EPOCH 4 - PROGRESS: at 76.84% examples, 31029
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:03,598 : INFO : EPOCH 4 - PROGRESS: at 78.59% examples, 31001
5 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:09:04,632 : INFO : EPOCH 4 - PROGRESS: at 80.54% examples, 30974
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:05,643 : INFO : EPOCH 4 - PROGRESS: at 82.67% examples, 30986
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:06,660 : INFO : EPOCH 4 - PROGRESS: at 84.69% examples, 31023
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:07,666 : INFO : EPOCH 4 - PROGRESS: at 86.18% examples, 31068
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:08,671 : INFO : EPOCH 4 - PROGRESS: at 87.81% examples, 31110
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:09,674 : INFO : EPOCH 4 - PROGRESS: at 89.35% examples, 31167
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:10,694 : INFO : EPOCH 4 - PROGRESS: at 90.85% examples, 31191
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:11,736 : INFO : EPOCH 4 - PROGRESS: at 92.31% examples, 31222
```

```
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:12,743 : INFO : EPOCH 4 - PROGRESS: at 93.57% examples, 31273
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:13,753 : INFO : EPOCH 4 - PROGRESS: at 94.81% examples, 31319
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:14,756 : INFO : EPOCH 4 - PROGRESS: at 95.97% examples, 31346
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:15,767 : INFO : EPOCH 4 - PROGRESS: at 97.27% examples, 31381
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:16,779 : INFO : EPOCH 4 - PROGRESS: at 98.51% examples, 31435
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:17,781 : INFO : EPOCH 4 - PROGRESS: at 99.75% examples, 31481
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:17,919 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:09:17,949 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:09:17,969 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:09:17,983 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:09:17,985 : INFO : EPOCH - 4 : training on 27374377 raw words (2
1281976 effective words) took 67.6s, 314871 effective words/s
2021-02-26 22:09:19,047 : INFO : EPOCH 5 - PROGRESS: at 2.38% examples, 325480
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:20,052 : INFO : EPOCH 5 - PROGRESS: at 4.47% examples, 340064
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:21,067 : INFO : EPOCH 5 - PROGRESS: at 6.67% examples, 339634
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:22,099 : INFO : EPOCH 5 - PROGRESS: at 8.61% examples, 339490
words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:23,124 : INFO : EPOCH 5 - PROGRESS: at 10.23% examples, 34005
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:24,134 : INFO : EPOCH 5 - PROGRESS: at 12.02% examples, 34083
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:25,137 : INFO : EPOCH 5 - PROGRESS: at 13.51% examples, 34113
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:26,192 : INFO : EPOCH 5 - PROGRESS: at 15.39% examples, 34018
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:27,249 : INFO : EPOCH 5 - PROGRESS: at 17.11% examples, 34088
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:28,271 : INFO : EPOCH 5 - PROGRESS: at 19.06% examples, 34293
0 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:09:29,284 : INFO : EPOCH 5 - PROGRESS: at 20.89% examples, 34327
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:30,325 : INFO : EPOCH 5 - PROGRESS: at 22.74% examples, 34296
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:31,368 : INFO : EPOCH 5 - PROGRESS: at 24.93% examples, 34380
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:32,379 : INFO : EPOCH 5 - PROGRESS: at 26.92% examples, 34454
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:33,413 : INFO : EPOCH 5 - PROGRESS: at 28.82% examples, 34389
0 words/s, in_qsize 7, out_qsize 0
```

```
2021-02-26 22:09:34,415 : INFO : EPOCH 5 - PROGRESS: at 30.58% examples, 34439
8 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:09:35,417 : INFO : EPOCH 5 - PROGRESS: at 31.94% examples, 34438
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:36,427 : INFO : EPOCH 5 - PROGRESS: at 33.27% examples, 34421
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:37,472 : INFO : EPOCH 5 - PROGRESS: at 34.73% examples, 34395
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:38,500 : INFO : EPOCH 5 - PROGRESS: at 36.10% examples, 34467
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:39,510 : INFO : EPOCH 5 - PROGRESS: at 37.46% examples, 34416
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:40,520 : INFO : EPOCH 5 - PROGRESS: at 38.89% examples, 34480
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:41,522 : INFO : EPOCH 5 - PROGRESS: at 40.13% examples, 34450
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:42,533 : INFO : EPOCH 5 - PROGRESS: at 41.41% examples, 34409
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:43,544 : INFO : EPOCH 5 - PROGRESS: at 42.68% examples, 34404
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:44,546 : INFO : EPOCH 5 - PROGRESS: at 44.00% examples, 34378
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:45,548 : INFO : EPOCH 5 - PROGRESS: at 45.29% examples, 34379
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:46,556 : INFO : EPOCH 5 - PROGRESS: at 46.61% examples, 34393
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:47,576 : INFO : EPOCH 5 - PROGRESS: at 48.12% examples, 34396
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:48,585 : INFO : EPOCH 5 - PROGRESS: at 49.64% examples, 34365
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:49,611 : INFO : EPOCH 5 - PROGRESS: at 51.26% examples, 34420
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:50,626 : INFO : EPOCH 5 - PROGRESS: at 52.85% examples, 34313
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:51,640 : INFO : EPOCH 5 - PROGRESS: at 54.61% examples, 34186
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:52,640 : INFO : EPOCH 5 - PROGRESS: at 56.87% examples, 34189
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:53,642 : INFO : EPOCH 5 - PROGRESS: at 58.62% examples, 34019
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:54,659 : INFO : EPOCH 5 - PROGRESS: at 60.78% examples, 33868
7 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:09:55,671 : INFO : EPOCH 5 - PROGRESS: at 62.34% examples, 33772
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:56,684 : INFO : EPOCH 5 - PROGRESS: at 63.78% examples, 33761
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:57,728 : INFO : EPOCH 5 - PROGRESS: at 64.99% examples, 33664
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:58,728 : INFO : EPOCH 5 - PROGRESS: at 66.61% examples, 33701
7 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:09:59,729 : INFO : EPOCH 5 - PROGRESS: at 68.09% examples, 33687
0 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:10:00,783 : INFO : EPOCH 5 - PROGRESS: at 69.25% examples, 33558
```

```
6 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:10:01,788 : INFO : EPOCH 5 - PROGRESS: at 70.24% examples, 33368
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:02,833 : INFO : EPOCH 5 - PROGRESS: at 71.50% examples, 33261
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:03,899 : INFO : EPOCH 5 - PROGRESS: at 72.80% examples, 33157
0 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:04,914 : INFO : EPOCH 5 - PROGRESS: at 74.02% examples, 33095
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:05,951 : INFO : EPOCH 5 - PROGRESS: at 75.38% examples, 33082
8 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:07,002 : INFO : EPOCH 5 - PROGRESS: at 76.89% examples, 33065
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:08,002 : INFO : EPOCH 5 - PROGRESS: at 78.93% examples, 33087
1 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:09,031 : INFO : EPOCH 5 - PROGRESS: at 81.21% examples, 33122
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:10,048 : INFO : EPOCH 5 - PROGRESS: at 83.32% examples, 33047
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:11,094 : INFO : EPOCH 5 - PROGRESS: at 85.01% examples, 33000
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:12,175 : INFO : EPOCH 5 - PROGRESS: at 86.63% examples, 32966
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:13,215 : INFO : EPOCH 5 - PROGRESS: at 88.14% examples, 32955
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:14,217 : INFO : EPOCH 5 - PROGRESS: at 89.62% examples, 32969
6 words/s, in_qsize 8, out_qsize 0
2021-02-26 22:10:15,227 : INFO : EPOCH 5 - PROGRESS: at 91.15% examples, 32980
5 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:16,236 : INFO : EPOCH 5 - PROGRESS: at 92.42% examples, 32935
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:17,244 : INFO : EPOCH 5 - PROGRESS: at 93.44% examples, 32841
6 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:18,253 : INFO : EPOCH 5 - PROGRESS: at 94.53% examples, 32789
3 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:19,297 : INFO : EPOCH 5 - PROGRESS: at 95.54% examples, 32696
4 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:20,312 : INFO : EPOCH 5 - PROGRESS: at 96.89% examples, 32756
9 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:21,368 : INFO : EPOCH 5 - PROGRESS: at 97.96% examples, 32661
1 words/s, in_qsize 6, out_qsize 1
2021-02-26 22:10:22,394 : INFO : EPOCH 5 - PROGRESS: at 99.12% examples, 32665
2 words/s, in_qsize 7, out_qsize 0
2021-02-26 22:10:23,096 : INFO : worker thread finished; awaiting finish of 3
more threads
2021-02-26 22:10:23,110 : INFO : worker thread finished; awaiting finish of 2
more threads
2021-02-26 22:10:23,117 : INFO : worker thread finished; awaiting finish of 1
more threads
2021-02-26 22:10:23,163 : INFO : worker thread finished; awaiting finish of 0
more threads
2021-02-26 22:10:23,164 : INFO : EPOCH - 5 : training on 27374377 raw words (2
1283266 effective words) took 65.2s, 326555 effective words/s
```

```
2021-02-26 22:10:23,165 : INFO : training on a 136871885 raw words (106412714
effective words) took 343.9s, 309391 effective words/s
```

Out[47]:

```
(106412714, 136871885)
```

In [48]:

```
seed_word9= [list(vectors6.wv.vocab.keys())[(i+1)*1000] for i in range(5)]
```

In [49]:

```
seed_word9
```

Out[49]:

```
['Conceptual', 'fluent', 'seventh', 'manage', 'publicity']
```

In [50]:

```
len(list(vectors9.wv.vocab.keys()))
```

Out[50]:

```
257022
```

In [51]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors9.wv.most_similar('neuroscience'))
print()
```

```
2021-02-26 22:11:50,646 : INFO : precomputing L2-norms of word weight vectors
Most similar to: neuroscience
[('neurobiology', 0.9721717834472656), ('psychology', 0.971102774143219), ('in
terdisciplinarity', 0.9698944091796875), ('imaginative', 0.9662922620773315),
('cultivating', 0.9651345014572144), ('neurosciences', 0.9640445709228516), ('
metacognition', 0.958501935005188), ('collegiality', 0.9539251327514648), ('ps
ychobiology', 0.9532018899917603), ('biopsychology', 0.9521270990371704)]
```

In [44]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors4.wv.most_similar('neuroscience'))
print()
```

```
Most similar to: neuroscience
[('biology', 0.9733627438545227), ('furthering', 0.9680370688438416), ('neurob
iology', 0.9664344787597656), ('epidemiology', 0.9656774997711182), ('psycholo
gy', 0.9603279829025269), ('informatics', 0.9592504501342773), ('genomics', 0.
9549347162246704), ('relevance', 0.9538663029670715), ('discovery', 0.95133334
39826965), ('STS', 0.9497036933898926)]
```

In [36]:

```
seed_word4
```

Out[36]:

```
['half', 'whether', 'neuroscience', 'Lonza', 'tightly']
```

In [37]:

```
len(list(vectors4.wv.vocab.keys()))
```

Out[37]:

```
63538
```

In [38]:

```
# Inspect words with vectors most similar to a given word
print("Most similar to:", 'neuroscience')
print(vectors4.wv.most_similar('neuroscience'))
print()
```

```
2021-02-26 21:50:18,179 : INFO : precomputing L2-norms of word weight vectors
Most similar to: neuroscience
[('furthering', 0.9842066764831543), ('informational', 0.9749946594238281), ('
relevance', 0.9741153717041016), ('neurobiology', 0.96800696849823), ('interes
t', 0.9663718342781067), ('informatics', 0.9641363620758057), ('technological'
, 0.9639126062393188), ('advancing', 0.9617303609848022), ('playing', 0.959730
1483154297), ('sciences', 0.9593894481658936)]
```

In [ ]:

```
```