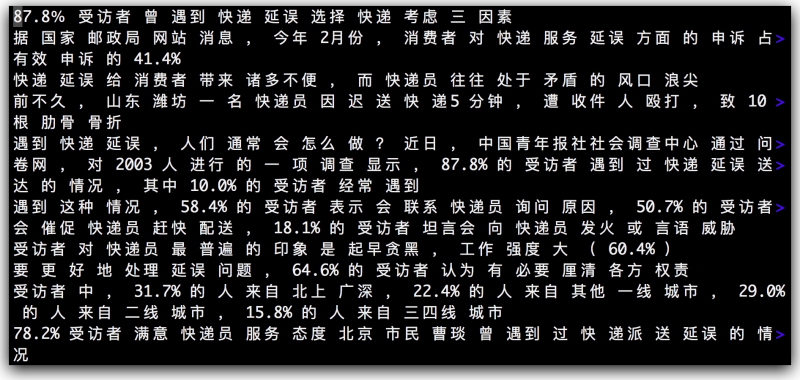
Weekly Review

如何训练GloVe中文词向量

2018-08-05 | 2019-01-15

准备语料

准备好自己的语料，保存为txt，每行一个句子或一段话，注意要分好词。



准备源码

从GitHub下载代码，https://github.com/stanfordnlp/GloVe

将语料corpus.txt放入到Glove的主文件夹下。

修改bash

打开demo.sh，修改相应的内容

1. 因为demo默认是下载网上的语料来训练的，因此如果要训练自己的语料，需要注释掉

```
make
#if [ ! -e text8 ]; then
#  if hash wget 2>/dev/null; then
#    wget http://mattmahoney.net/dc/text8.zip
#  else
#    curl -O http://mattmahoney.net/dc/text8.zip
#  fi
#  unzip text8.zip
#  rm text8.zip
#fi
```

1. 修改参数设置，将CORPUS设置成语料的名字

```
CORPUS=corpus.txt
VOCAB_FILE=vocab.txt
COOCCURRENCE_FILE=cooccurrence.bin
COOCCURRENCE_SHUF_FILE=cooccurrence.shuf.bin
BUILDDIR=build
SAVE_FILE=vectors
VERBOSE=2
MEMORY=4.0
VOCAB_MIN_COUNT=5
VECTOR_SIZE=300
MAX_ITER=15
WINDOW_SIZE=15
BINARY=2
NUM_THREADS=8
X_MAX=10
```

执行bash文件

进入到主文件夹下

1. make

```
zhlin@fnlp-server-104:~/GloVe$ make
mkdir -p build
gcc src/glove.c -o build/glove -lm -pthread -Ofast -march=native -funroll-loops -Wa
ll -Wextra -Wpedantic
src/glove.c: In function ‘glove_thread’:
src/glove.c:117:9: warning: ignoring return value of ‘fread’, declared with attribu
te warn_unused_result [-Wunused-result]
         fread(&cr, sizeof(CREC), 1, fin);
         ^
gcc src/shuffle.c -o build/shuffle -lm -pthread -Ofast -march=native -funroll-loops
 -Wall -Wextra -Wpedantic
src/shuffle.c: In function ‘shuffle_merge’:
src/shuffle.c:106:17: warning: ignoring return value of ‘fread’, declared with attr
ibute warn_unused_result [-Wunused-result]
                 fread(&array[i], sizeof(CREC), 1, fid[j]);
                 ^
src/shuffle.c: In function ‘shuffle_by_chunks’:
src/shuffle.c:163:9: warning: ignoring return value of ‘fread’, declared with attri
bute warn_unused_result [-Wunused-result]
         fread(&array[i], sizeof(CREC), 1, fin);
         ^
gcc src/cooccur.c -o build/cooccur -lm -pthread -Ofast -march=native -funroll-loops
 -Wall -Wextra -Wpedantic
src/cooccur.c: In function ‘merge_files’:
src/cooccur.c:267:9: warning: ignoring return value of ‘fread’, declared with attri
bute warn_unused_result [-Wunused-result]
         fread(&new, sizeof(CREC), 1, fid[i]);
         ^
src/cooccur.c:277:5: warning: ignoring return value of ‘fread’, declared with attri
bute warn_unused_result [-Wunused-result]
     fread(&new, sizeof(CREC), 1, fid[i]);
     ^
src/cooccur.c:290:9: warning: ignoring return value of ‘fread’, declared with attri
bute warn_unused_result [-Wunused-result]
         fread(&new, sizeof(CREC), 1, fid[i]);
         ^
gcc src/vocab_count.c -o build/vocab_count -lm -pthread -Ofast -march=native -funro
ll-loops -Wall -Wextra -Wpedantic
```

1. bash demo.sh

```
zhlin@fnlp-server-104:~/GloVe$ bash demo.sh
mkdir -p build
gcc src/glove.c -o build/glove -lm -pthread -Ofast -march=native -funroll-loops -Wall -Wextra -Wpedantic
src/glove.c: In function ‘glove_thread’:
src/glove.c:117:9: warning: ignoring return value of ‘fread’, declared with attribute warn_unused_result [-Wunused-result]
         fread(&cr, sizeof(CREC), 1, fin);
         ^
gcc src/shuffle.c -o build/shuffle -lm -pthread -Ofast -march=native -funroll-loops -Wall -Wextra -Wpedantic
src/shuffle.c: In function ‘shuffle_merge’:
src/shuffle.c:106:17: warning: ignoring return value of ‘fread’, declared with attribute warn_unused_result [-Wunused-result]
                 fread(&array[i], sizeof(CREC), 1, fid[j]);
                 ^
src/shuffle.c: In function ‘shuffle_by_chunks’:
src/shuffle.c:163:9: warning: ignoring return value of ‘fread’, declared with attribute warn_unused_result [-Wunused-result]
         fread(&array[i], sizeof(CREC), 1, fin);
         ^
gcc src/cooccur.c -o build/cooccur -lm -pthread -Ofast -march=native -funroll-loops -Wall -Wextra -Wpedantic
src/cooccur.c: In function ‘merge_files’:
src/cooccur.c:267:9: warning: ignoring return value of ‘fread’, declared with attribute warn_unused_result [-Wunused-result]
         fread(&new, sizeof(CREC), 1, fid[i]);
         ^
src/cooccur.c:277:5: warning: ignoring return value of ‘fread’, declared with attribute warn_unused_result [-Wunused-result]
     fread(&new, sizeof(CREC), 1, fid[i]);
     ^
src/cooccur.c:290:9: warning: ignoring return value of ‘fread’, declared with attribute warn_unused_result [-Wunused-result]
         fread(&new, sizeof(CREC), 1, fid[i]);
         ^
gcc src/vocab_count.c -o build/vocab_count -lm -pthread -Ofast -march=native -funroll-loops -Wall -Wextra -Wpedantic

$ build/vocab_count -min-count 5 -verbose 2 < corpus.txt > vocab.txt
BUILDING VOCABULARY
Processed 76945977 tokens.
Counted 1126621 unique words.
Truncating vocabulary at min count 5.
Using vocabulary of size 208938.

$ build/cooccur -memory 4.0 -vocab-file vocab.txt -verbose 2 -window-size 15 < corpus.txt > cooccurrence.bin
COUNTING COOCCURRENCES
window size: 15
context: symmetric
max product: 13752509
overflow length: 38028356
Reading vocab from file "vocab.txt"...loaded 208938 words.
Building lookup table...table contains 124428181 elements.
Processed 76945977 tokens.
Writing cooccurrences to disk.........5 files in total.
Merging cooccurrence files: processed 142110675 lines.

$ build/shuffle -memory 4.0 -verbose 2 < cooccurrence.bin > cooccurrence.shuf.bin
SHUFFLING COOCCURRENCES
array size: 255013683
Shuffling by chunks: processed 0 lines.
```

注意，如果训练数据较大，则训练时间较长，那么建议使用nohup来运行程序

```
nohup bash demo.sh >output.txt 2>&1 &
```

坐等训练，最后会得到vectors.txt 以及其他的相应的文件。如果要用gensim的word2vec load进来，那么需要在vectors.txt的第一行加上vacob_size vector_size，第一个数指明一共有多少个向量，第二个数指明每个向量有多少维。

参考

https://www.cnblogs.com/echo-cheng/p/8561171.html

GloVe 教程

每周碎片知识2 Python惯例[转]