

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。

```
87.8% 受访者 曾 遇到 快递 延误 选择 快递 考虑 三 因素  
据 国家 邮政局 网站 消息 ， 今年 2月份 ， 消费者 对 快递 服务 延误 方面 的 申诉 占 >  
有效 申诉 的 41.4%  
快递 延误 给 消费者 带来 诸多 不便 ， 而 快递员 往往 处于 矛盾 的 风口 浪尖  
前不久 ， 山东 潍坊 一 名 快递员 因 迟 送 快递 5 分钟 ， 遭 收件 人 殴打 ， 致 10 >  
根 肋骨 骨折  
遇到 快递 延误 ， 人们 通常 会 怎么 做 ？ 近日 ， 中国青年报 社会 调查 中心 通过 问 >  
卷 网 ， 对 2003 人 进行 的 一 项 调查 显示 ， 87.8% 的 受访者 遇到 过 快递 延误 送 >  
达 的 情况 ， 其中 10.0% 的 受访者 经常 遇到  
遇到 这种 情况 ， 58.4% 的 受访者 表示 会 联系 快递员 询问 原因 ， 50.7% 的 受访者 >  
会 催促 快递员 赶快 配送 ， 18.1% 的 受访者 坦言 会 向 快递员 发火 或 言语 威胁  
受访者 对 快递员 最 普遍 的 印象 是 起 早 贪 黑 ， 工作 强度 大 （ 60.4% ）  
要 更 好 地 处理 延误 问题 ， 64.6% 的 受访者 认为 有 必要 厘清 各方 权责  
受访者 中 ， 31.7% 的 人 来自 北上 广 深 ， 22.4% 的 人 来自 其他 一 线 城市 ， 29.0% >  
的 人 来自 二 线 城市 ， 15.8% 的 人 来自 三 四 线 城市  
78.2% 受访者 满意 快递员 服务 态度 北京 市民 曹琰 曾 遇到 过 快递 派送 延误 的 情 >  
况
```

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 -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

```

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

```
1 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vocab_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>

本文作者： 林泽辉

本文链接： <http://www.linzehui.me/2018/08/05/碎片知识/如何训练GloVe中文词向量/>

版权声明： 本博客所有文章除特别声明外，均采用 [CC BY-NC-SA 3.0](#) 许可协议。转载请注明出处！

 [GloVe](#)  [教程](#)

◀ 每周碎片知识2

Python惯例[转] ▶

© 2020  林泽辉

主题 — NexT.Muse v5.1.4