**Chatbot**

获取到原始语料后需要做一些加工处理，首先需要做切词，方法如下：

python3 word\_segment.py ./subtitle.corpus ./ segment\_result

其中word\_segment.py是编译好的的切词工具

之后要把切词好的文件转成“|”分隔的问答对，如下：

cat ./corpus.segment | awk '{if(last!="")print last"|"$0;last=$0}' | sed 's/| /|/g' > ./corpus.segment.pair

在这个过程中语料库大小会翻倍，因为几乎所有的句子既作为答句跟它的上文构成问答对，也作为问句与它的下文构成问答对

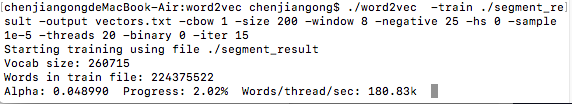
**训练词向量**

直接在terminal中利用google的word2vec来训练词向量，如下：

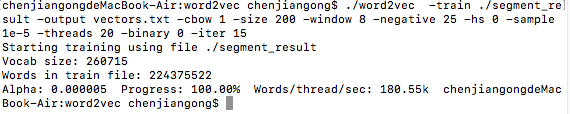
./word2vec -train ./segment\_result -output vectors.txt -cbow 1 -size 200 -window 8 -negative 25 -hs 0 -sample 1e-5 -threads 20 -binary 0 -iter 15

其中segment\_result是原始语料数据（已分词但未构成问答对），vectors.txt是生成的词向量文本文件，供之后加载使用

加载过程如下：



完成词向量文本文件加载：



简单地使用python读取生成的txt文件内的部分内容，来检查词向量是否成功生成：

