Word2vec 训练词向量工具及使用

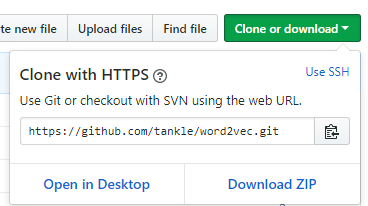
一、下载地址：https://github.com/tankle/word2vec

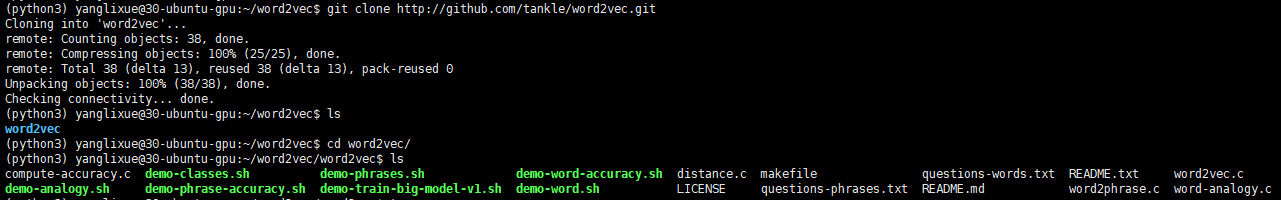
二、简介

word2vec是Google在2013年开源的一款将词表征为实数值向量的高效工具，即一个计算word vector的开源工具。首先，word2vec可以在百万数量级的词典和上亿的数据集上进行高效地训练；其次，该工具得到的训练结果——词向量（word embedding），可以很好地度量词与词之间的相似性。采用的模型有CBOW（Continuous Bag-Of-Words，即连续的词袋模型）和Skip-Gram两种。word2vec通过训练，可以把对文本内容的处理简化为K维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。因此，word2vec输出的词向量可以被用来做很多NLP相关的工作，比如聚类、找同义词、词性分析等等。

三、使用教程

1.代码下载：git clone <https://github.com/tankle/word2vec.git> 或者 download ZIP





2.快速入门

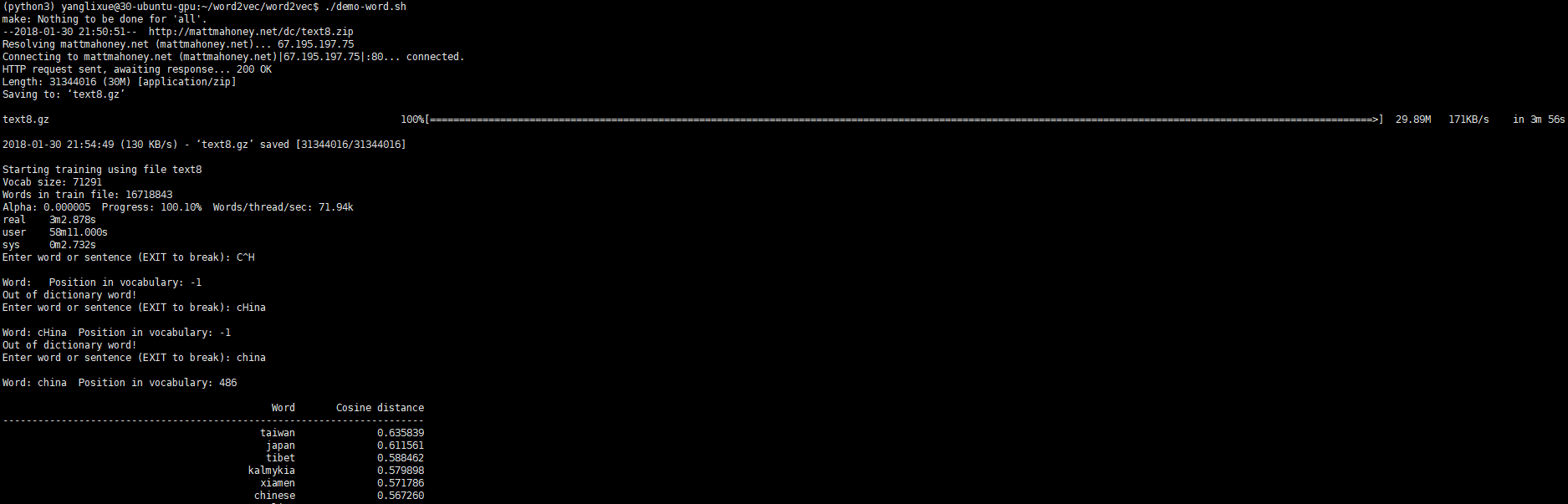
（1）在makefile文件目录下，运行“make”命令编译word2vec工具

C:\Users\w\AppData\Local\Temp\WeChat Files\703050758691546464.png

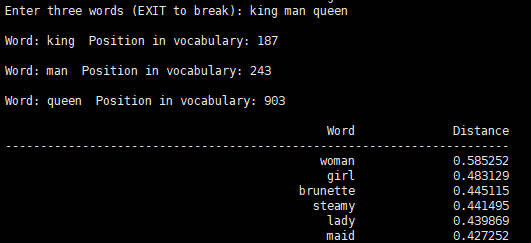
（2）例如运行demo脚本：./demo-word.sh

demo-word.sh中的代码如下，主要工作为：

1. make（编译）
2. 如果不存在text8，自动下载下载训练数据。text8中为一些空格隔开的英文单词，但不含标点符号，一共有1671多万个单词。
3. 开始用text8语料训练
4. 训练完毕后，输入比如china
5. 调用distance，查找cosine距离最近的词，即为同义词或相似词



（3）命令行执行./word-analogy vectors.bin



3.Word2vec文件夹中各文件功能

./demo-word.sh找输入词的同义词

./demo-phrases.sh进行短语训练，在这里 New York就会被当成一个整体来看待

./demo-classes.sh 生成聚类文件，并且可以利用sorted语句进行排序

./word-analogy.sh+向量文件 向量的加法组合运算，输入词之间的运算：

vector('king') - vector('man') + vector('queen') ≈ vector('woman')

各-accuracy.sh文件计算相应准确度。

4.Word2vec扩展功能

本文选取的word2vec中的word2vec.c文件带有中文注释，可根据自己需要修改为批量处理。

此外，在train中文语料时，可用命令行 ./word2vec -train train\_file -output vectors.bin -cbow 0 -size 200 -window 5 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1

以上命令表示的是输入文件是train\_file，输出文件是vectors.bin，不使用cbow模型，默认为Skip-Gram模型。 每个单词的向量维度是200，训练的窗口大小为5，即当前词前五和后五个词语。不使用NEG方法，使用HS方法。-sampe指的是采样的阈值，词在训练样本中出现的频率越大，那么就越会被采样。-binary为1指的是结果二进制存储，为0是普通存储（普通存储的时候是可以打开看到词语和对应的向量的）。除了以上命令中的参数，word2vec还有几个参数对我们比较有用比如-alpha设置学习速率，默认的为0.025。-min-count设置最低频率，默认是5，如果一个词语在文档中出现的次数小于5，那么就会丢弃。-classes设置聚类个数，Word2vec源码中用的是k-means聚类的方法。

修改合适的参数进行训练 可得到pre-train出来的bin文件的词向量，用作为其他深度学习任务的预训练词向量。

四、相关论文

1. Yoshua Bengio,Rejean D,Pascal V,Christian J. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003,3(0): 1137–1155

2. A Scalable Hierarchical Distributed Language Model. Andriy Mnih and Geoffrey Hinton

3. Efficient Estimation of Word Representations in Vector Space. Tomas Mikolov等

4. Distributed Representations of Words and Phrases and their Compositionality.Tomas Mikolov等