MSiA490 SEC20/28
Text Analytics
Lab 2 - Word2Vec

Timo Wang

Northwestern University

September 24th, 2020

Some slides of this document is built based on the content provided on https://github.com/tmikolov/word2vec.

What is Word2Vec?

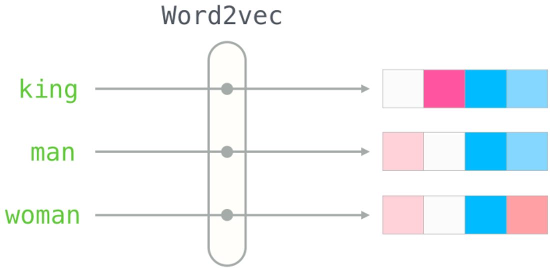
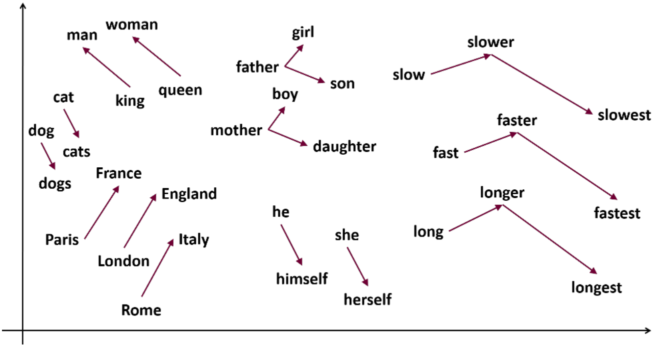


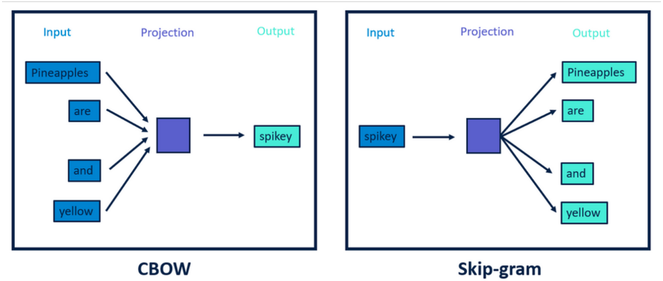
Figure: It encodes a word into a vector, in comparison to one-hot encoder, where each word is represented by an integer.

Why do we care?

Capture of word meaning and relations

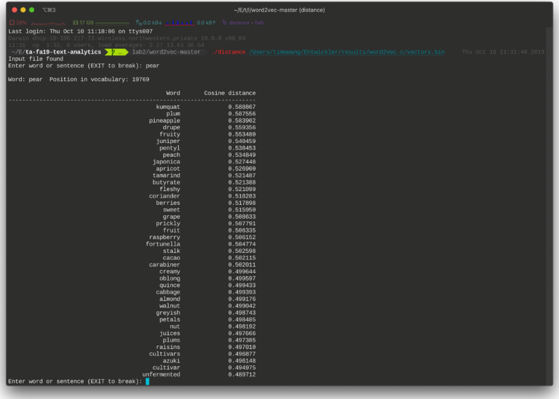


How does it work?



How well does it perform?

Strength



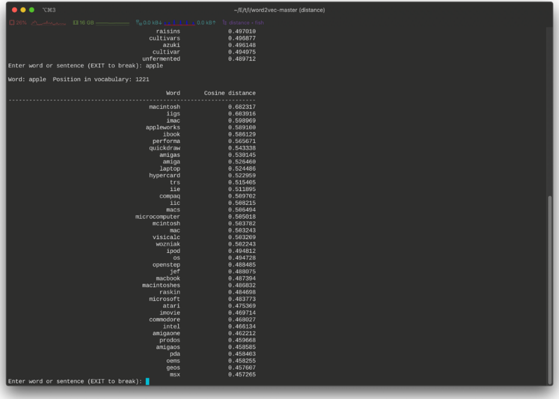
```
Last login: Thu Oct 10 11:18:06 on ttys007
Darwin dhcp-10-105-217-73.wireless.northwestern.private 19.0.0 x86_64
11:31  up  1:32, 8 users, load averages: 2.27 13.63 36.54
 ~/E/ta-fa19-text-analytics  … lab2/word2vec-master   ./distance /Users/timowang/Entwickler/results/word2vec-c/vectors.bin     Thu Oct 10 11:31:48 2019
Input file found
Enter word or sentence (EXIT to break): pear

Word: pear  Position in vocabulary: 19769

                                              Word       Cosine distance
------------------------------------------------------------------------
                                           kumquat              0.588867
                                              plum              0.587556
                                         pineapple              0.583902
                                             drupe              0.559356
                                            fruity              0.553480
                                           juniper              0.540459
                                            pentyl              0.538453
                                             peach              0.534849
                                          japonica              0.527448
                                           apricot              0.526900
                                          tamarind              0.521487
                                          butyrate              0.521388
                                            fleshy              0.521099
                                         coriander              0.518283
                                           berries              0.517898
                                             sweet              0.515950
                                             grape              0.508633
                                           prickly              0.507791
                                             fruit              0.506335
                                         raspberry              0.506152
                                        fortunella              0.504774
                                             stalk              0.502598
                                             cacao              0.502115
                                         carabiner              0.502011
                                            creamy              0.499644
                                            oblong              0.499597
                                            quince              0.499433
                                           cabbage              0.499393
                                            almond              0.499176
                                            walnut              0.499042
                                           greyish              0.498743
                                            petals              0.498485
                                               nut              0.498192
                                            juices              0.497666
                                             plums              0.497385
                                           raisins              0.497010
                                         cultivars              0.496877
                                             azuki              0.496148
                                          cultivar              0.494975
                                       unfermented              0.489712
Enter word or sentence (EXIT to break): 
```

How well does it perform?

Weakness



```
                                           raisins              0.497010
                                         cultivars              0.496877
                                             azuki              0.496148
                                          cultivar              0.494975
                                       unfermented              0.489712
Enter word or sentence (EXIT to break): apple

Word: apple  Position in vocabulary: 1221

                                              Word       Cosine distance
------------------------------------------------------------------------
                                         macintosh              0.682317
                                              iigs              0.603916
                                              imac              0.598969
                                        appleworks              0.589100
                                             ibook              0.586129
                                          performa              0.565671
                                         quickdraw              0.543338
                                            amigas              0.530145
                                             amiga              0.526460
                                            laptop              0.524486
                                         hypercard              0.522959
                                               trs              0.515405
                                               iie              0.511895
                                            compaq              0.509702
                                               iic              0.508215
                                              macs              0.506494
                                     microcomputer              0.505018
                                          mcintosh              0.503782
                                               mac              0.503243
                                          visicalc              0.503209
                                           wozniak              0.502243
                                              ipod              0.494812
                                                os              0.494728
                                          openstep              0.488485
                                               jef              0.488075
                                           macbook              0.487394
                                       macintoshes              0.486832
                                            raskin              0.484698
                                         microsoft              0.483773
                                             atari              0.475369
                                            imovie              0.469714
                                         commodore              0.468027
                                             intel              0.466134
                                          amigaone              0.462212
                                            prodos              0.459668
                                           amigaos              0.458585
                                               pda              0.458403
                                              oems              0.458255
                                              geos              0.457607
                                               msx              0.457265
Enter word or sentence (EXIT to break): 
```

Tools & libraries

Demo

Shell

```
# Download and decompress a sample text corpus.
wget http://mattmahoney.net/dc/text8.zip -O text8.gz
tar xvf text8.gz

# Train a model named vectors.bin with the downloaded
    corpus text8
./word2vec -train text8 -output vectors.bin -cbow 1 -
    size 200 -window 8 -negative 25 -hs 0 -sample 1e-4 -
    threads 20 -binary 1 -iter 15
# Examine the embedding
./distance vectors.bin
```

Tools & libraries

The original library

Step 1 Download the project from https://github.com/tmikolov/word2vec.

Step 2 Read `demo-word.sh`.

Step 3 Use `demo-word.sh` as a guideline and train the model with your own corpus.

Tools & libraries

Gensim wrapper for Python

Step 1 Install Gensim with `pip install gensim`.

Step 2 Depending on the content of your corpus, you may need to process the text corpus into this format: `List[List[str]]`.

Step 3 Train and save the model.

Tools & libraries

Troubleshooting tips for the original library (macOS)

1. `./distance` results in a segmentation fault
   - remove `-march=native` from `makefile`
2. `Undefined symbols for architecture x86_64: “_fgetc_unlocked”, referenced from:`
   - replace `fgetc_unlocked` with `getc_unlocked` and `fputc_unlocked` with `putc_unlocked`

Tools & libraries

More resources on the Gensim library

- https://radimrehurek.com/gensim/models/word2vec.html
- https://towardsdatascience.com/a-beginners-guide-toword-embedding-with-gensim-word2vec-model5970fa56cc92

Quiz

Task 1

Which of the following is/are the reason/reasons why the embeddings produced by a trained Word2Vec model results in poor text processing performance?

A The text corpus for training is too small

B The content of the corpus is irrelevant to the text that needs to be processed

C The model is overfitted

D All of the above

Quiz

Task 2

In which language was the original Word2Vec library developed by Mikolov implemented?

A C++

B Java

C C

D Python

Thoughts & feedbacks