

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?

Why do we care?

How does it work?

How well does it perform?

Tools & libraries

Quiz

Thoughts & feedbacks

What is Word2Vec?

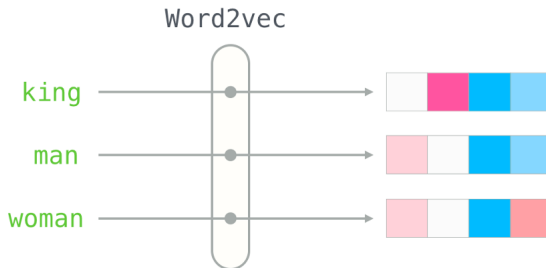
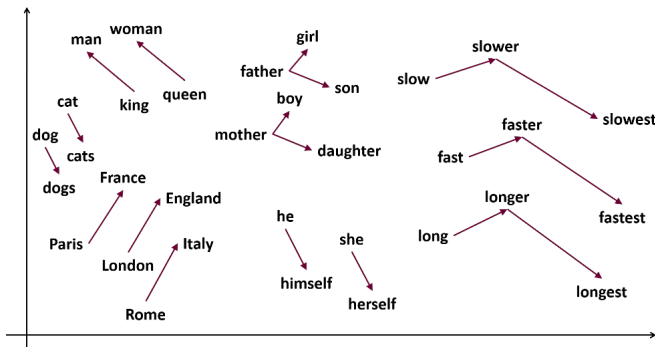


Figure: It encodes a word into a vector, in comparison to one-hot encoder, where each word is represented by an integer.

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

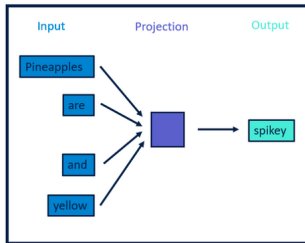
Why do we care?

Capture of word meaning and relations

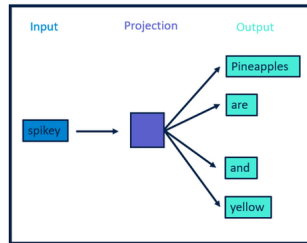


What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

How does it work?



CBOW



Skip-gram

Strength
Weakness

Strength

```

C:\Users\j...> cd C:\Users\j...> python -m wordvec --vocab=word_vocab.txt --train=train_data.txt --test=test_data.txt --distance=cosine --save_dir=./results
Last logs: Thu Oct 26 11:18:06 on tyseer
17:19:10 [INFO] Training word embeddings on train set (10,000 words) ...
17:19:10 [INFO] Saving word embeddings to ./results/wordvec/vectors.npy
Input file found: lab2/wordvec-master
Enter word or sentence (EXIT to break): pear
Word: pear Position in vocabulary: 19769

===== Word Cosine distance =====
kumquat 0.688987
plum 0.687596
pineapple 0.583962
dragon 0.582956
fruity 0.553489
juniper 0.548459
portyl 0.638483
peach 0.634949
japonica 0.627448
apricot 0.616989
tamarind 0.521487
butyrate 0.521388
fleshy 0.521099
coriander 0.518283
berries 0.517968
meat 0.516969
grape 0.509833
prickly 0.507792
fruit 0.506325
rasberry 0.506252
fortnelli 0.484774
stalk 0.482098
cacao 0.482115
carahany 0.482011
creamy 0.499684
oilmg 0.499507
quince 0.499423
cabbage 0.499393
Almond 0.499276
walnut 0.499042
greyish 0.488743
nuts 0.488485
petal 0.488192
juices 0.487665
plum 0.487325
raisins 0.497010
cultivars 0.486877
axial 0.486148
cultivar 0.484975
unfermented 0.489712

```

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

Strength
Weakness

How well does it perform?

Weakness

The screenshot shows a terminal window titled '~/(word2vec-master (distance))'. It displays a list of words and their cosine distances from the word 'apple'. The words are sorted by distance, with 'raisins' having the smallest distance (0.497010) and 'max' having the largest (0.477905). The terminal also shows the command 'Enter word or sentence (EXIT to break): apple' and the output 'Word: apple Position in vocabulary: 1221'.

```
~/(word2vec-master (distance))
Enter word or sentence (EXIT to break): apple
Word: apple Position in vocabulary: 1221

Word      Cosine distance
-----
macintosh 0.602817
lips       0.603995
isac       0.599969
applworks  0.595189
ibook      0.586129
performa  0.565671
quickdraw  0.543338
anigas     0.539145
smigs      0.526469
laptop     0.524498
hypercard  0.522969
ira        0.515405
lie        0.511895
cowman     0.509782
lic        0.508215
mcs        0.506484
microcomputer 0.505018
mcintosh   0.503782
mac        0.503243
visicalc   0.503209
word11ak   0.502243
lpop       0.494812
os         0.494728
openstep   0.489405
jef        0.488075
macbook    0.487354
macintoshes 0.486822
raskin     0.484659
microsoft  0.483773
atari      0.475969
imovie     0.469714
commandore 0.468627
intel      0.466134
anigone     0.462212
pepco      0.452693
anigas     0.450585
pla        0.450403
coms       0.450255
geos       0.451667
max        0.477905

Enter word or sentence (EXIT to break):
```

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

Demo

The original library
Gensim wrapper for Python
Troubleshooting tips for the original library (macOS)
More resources on the Gensim library

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
```

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

Demo

The original library

Gensim wrapper for Python

Troubleshooting tips for the original library (macOS)

More resources on the Gensim library

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.

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

Demo
The original library
Gensim wrapper for Python
Troubleshooting tips for the original library (macOS)
More resources on the Gensim library

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`

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

Demo
The original library
Gensim wrapper for Python
Troubleshooting tips for the original library (macOS)
More resources on the Gensim library

Tools & libraries

More resources on the Gensim library

- <https://radimrehurek.com/gensim/models/word2vec.html>
- <https://towardsdatascience.com/a-beginners-guide-to-word-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

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz
Thoughts & feedbacks

Task 1

Task 2

Quiz

Task 2

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

- A C++
- B Java
- C C
- D Python

What is Word2Vec?
Why do we care?
How does it work?
How well does it perform?
Tools & libraries
Quiz

Thoughts & feedbacks

Thoughts & feedbacks