

# 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?

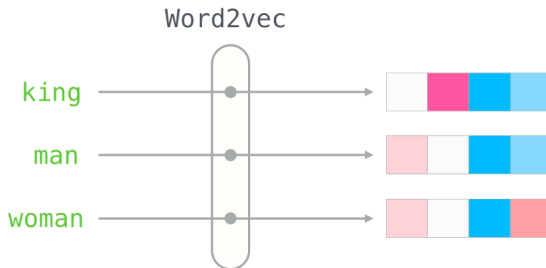
How does it work?

How well does it perform?

Tools & libraries

Thoughts & feedbacks

# What is Word2Vec?

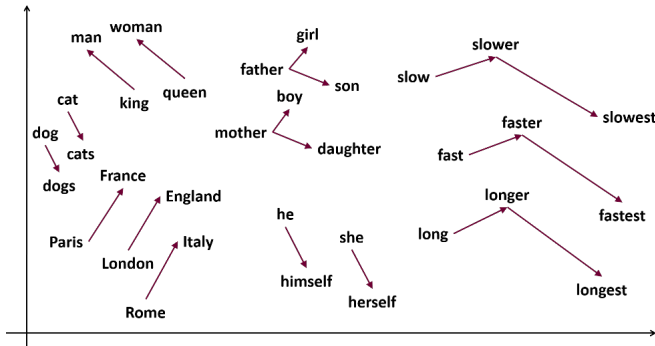


**Figure:** It encodes a word into a vector, in comparison to one-hot encoder, where each word is represented by an integer.

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

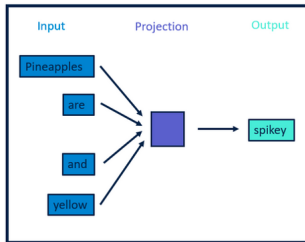
# Why do we care?

Better capture of word meaning

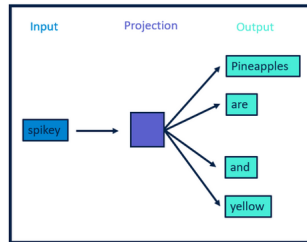


What is Word2Vec?  
Why do we care?  
**How does it work?**  
How well does it perform?  
Tools & libraries  
Thoughts & feedbacks

## How does it work?



**CBOW**



**Skip-gram**

What is Word2Vec?

Why do we care?

How does it work?

How well does it perform?

Tools & libraries

Thoughts & feedbacks

Strength  
Weakness

# How well does it perform?

Strength

```
~$ cd /Users/timwang/Projects/word2vec-master
~/word2vec-master$ ./distance.py
Last login: Thu Oct 30 11:18:06 on ttys007
timg@timg:~$ cd /Users/timwang/Projects/word2vec-master
~/word2vec-master$ ./distance.py
Input file found
Enter word or sentence (EXIT to break): pear
word: pear Position in vocabulary: 19769

-----
Word Cosine distance
-----
kumquat 0.588667
plum 0.587565
pineapple 0.583902
drupe 0.559355
fruity 0.553468
juniper 0.549459
peach 0.538453
japanea 0.534949
japonica 0.527448
apricot 0.520909
tamarind 0.521407
butyrate 0.521388
flashy 0.521299
coriander 0.518283
berries 0.517688
sweet 0.516968
grape 0.508633
prickly 0.507703
fruit 0.506325
raspberry 0.506322
fortunella 0.504774
stalk 0.502598
cacao 0.502315
carabiner 0.502011
crown 0.499661
chlong 0.499597
quince 0.499423
cubeb 0.499353
almond 0.499176
walnut 0.498962
greyish 0.498743
petals 0.498485
nut 0.498192
juices 0.497665
plum 0.497385
raisins 0.497035
cultivars 0.496877
actual 0.496248
cultivar 0.494975
unfermented 0.489712
-----
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  
Thoughts & feedbacks

Strength  
**Weakness**

# How well does it perform?

Weakness

```
~$ cat word2vec-master (distance)
~$ ./word2vec-master (distance)
Enter word or sentence (EXIT to break): apple
word: apple Position in vocabulary: 1221
Word Cosine distance
-----
macintosh 0.682317
liga 0.683916
lmac 0.590909
appleworks 0.589189
book 0.588229
performa 0.565671
quickdraw 0.543388
amiga 0.538145
amiga 0.526469
javelin 0.524408
hypercard 0.522969
tra 0.515405
lta 0.511895
compaq 0.509782
lta 0.508215
moss 0.506494
microcomputer 0.505018
macintosh 0.503782
mac 0.503243
visiologic 0.502809
wordiak 0.502243
ipod 0.494812
os 0.494728
openstep 0.488485
per 0.488075
macosx 0.487394
macintoshes 0.486822
rakito 0.484808
microsoft 0.483773
atari 0.475389
imovie 0.460714
commodore 0.448027
intel 0.460134
amigaone 0.462212
prodos 0.459668
amigaos 0.458589
psa 0.458483
com 0.458255
goos 0.457667
msx 0.457265
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**  
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**  
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**  
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.

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

### 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**  
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>

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

# Thoughts & feedbacks