

# 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?

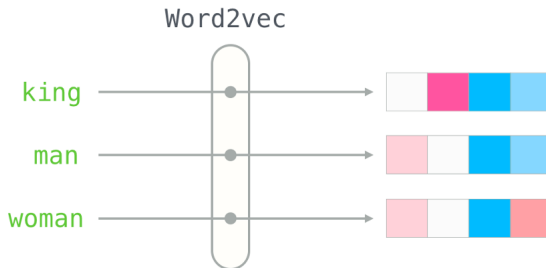
How well does it perform?

Tools & libraries

Quiz

Thoughts & feedbacks

# What is Word2Vec?

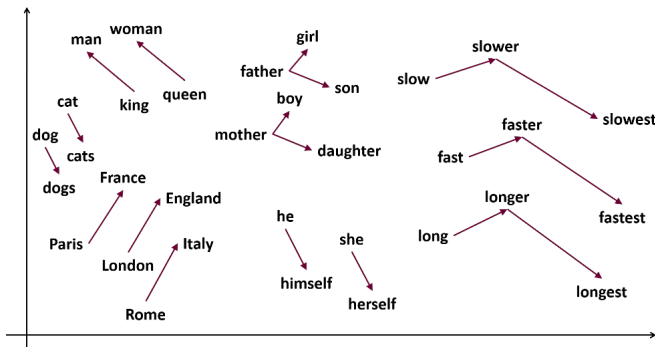


**Figure:** It encodes a word into a vector, in comparison to one-hot encoder, where each word is represented by an integer.

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

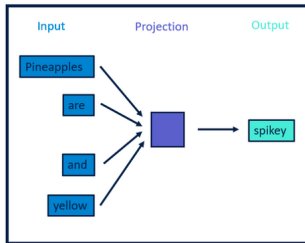
# Why do we care?

Better capture of word meaning

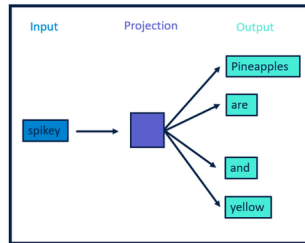


What is Word2Vec?  
Why do we care?  
**How does it work?**  
How well does it perform?  
Tools & libraries  
Quiz  
Thoughts & feedbacks

# How does it work?



**CBOW**



**Skip-gram**

Strength  
Weakness

## Strength

```

C3 20%          C3K3
-----
last login: Thu Oct 30 11:38:06 on tty987
root@kali:~# python3 /usr/share/metasploit-framework/vendor/bundle/ruby/2.7.0/bin/rsync -avz --delete-after --exclude=private 19.8.8.8:/usr/ha
rsync [v] 3.1.3-ubuntu
rsync error: remote host failed to connect (code: 1) at host[19.8.8.8] port [22].
--C3-t-fair-text-analytics | hah!wordvec-master --distance /usr/lib/browser/results/wordvec/c-vectors.bin    Thu Oct 30 11:31:46 2019
Input file found.
Enter word or sentence (EXIT to break): pear

Word      Cosine distance
-----
kumquat   0.588087
plum      0.587556
pineapple 0.583982
druon     0.583526
fruity    0.553489
juniper   0.548459
pentyl    0.538453
peach     0.534849
japonica  0.527448
apricot   0.520989
tamarind  0.521487
butyrate  0.521388
fishy     0.522099
coriander 0.518283
berries   0.517806
sweet     0.515999
grape     0.509653
prickly   0.507792
fruit     0.506325
rasberry   0.506152
fortanella 0.504714
stalk     0.502595
cacao     0.502115
carahiner 0.502011
creamy    0.499641
oliog     0.499597
quince    0.494423
cablogie  0.493393
almond    0.493176
walnut    0.490947
greyish   0.488743
petals    0.488495
net       0.490182
juices    0.487566
plums     0.487385
raisins   0.487010
cultivars 0.486877
axaki     0.486148
culivar   0.484975
enfermedad 0.489712

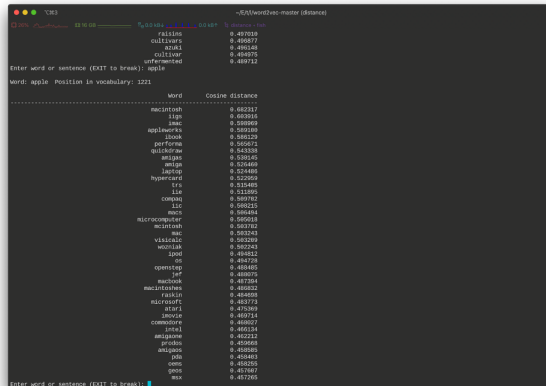
Enter word or sentence (EXIT to break):

```

## Thoughts & feedbacks

Strength  
Weakness

## Weakness



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

# 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