

# MSiA414 SEC01

## Text Analytics

### Lab 2 - Word2Vec

Timo Wang

Northwestern University

October 1st, 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?

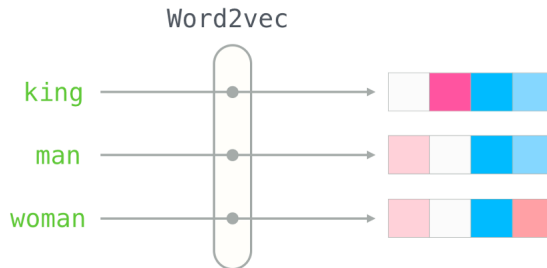
How well does it perform?

Tools & libraries

Quiz

Thoughts & feedbacks

# What is Word2Vec?

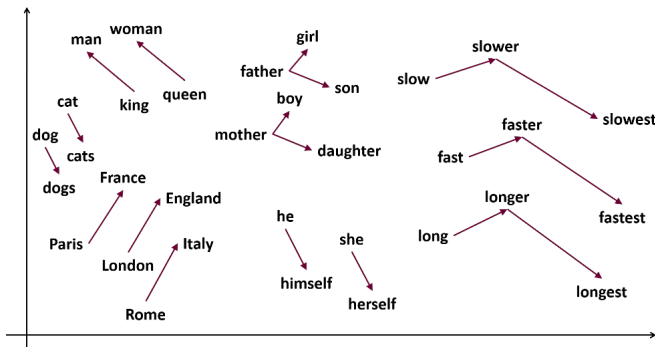


**Figure:** It encodes a word into a vector, in comparison to one-hot encoder, where each word is represented by an integer.

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

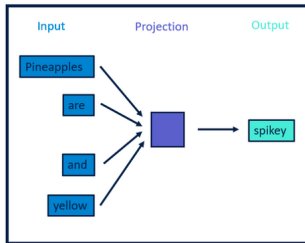
# Why do we care?

Capture of word meaning and relations

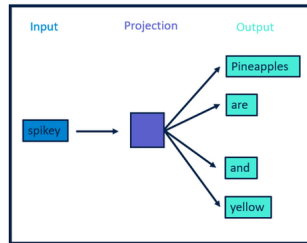


What is Word2Vec?  
Why do we care?  
**How does it work?**  
How well does it perform?  
Tools & libraries  
Quiz  
Thoughts & feedbacks

# How does it work?



**CBOW**



**Skip-gram**

Strength  
Weakness

## Strength

```

C3 20%          82.97 0.00         f0=0.634         g0=0.648         distance = 1.0
Last login: Thu Oct 30 11:18:06 on ttyne07
root@kali:~# ./f0-f1-f2-vec.py --word_embeddings_dir=/usr/share/mexican-proverbs --vocab_size=10000 --k=50 --n=50
f0=0.634         f1=0.634         f2=0.634         g0=0.648         g1=0.648         g2=0.648         distance = 1.0
Input file found
Enter word or sentence (EXIT to break): pear

Word: pear Position in vocabulary: 19709

      Word      Cosine distance
-----
kumquat        0.880897
blue           0.587566
pineapple      0.583962
drupe          0.559256
fruity         0.553489
juniper        0.548459
persyl         0.538453
peach          0.534849
japonica       0.527448
apricot        0.526080
tamarind       0.521487
butyrate       0.521385
fishy         0.521099
coriander      0.518263
berries        0.517080
sweet         0.515069
grape          0.508833
prickly       0.507792
fruit         0.506325
raspberry     0.500352
fortunella    0.484774
stalk         0.462568
cacao         0.582115
carahamer     0.582011
creamy        0.499644
oblong        0.495997
quince        0.494323
cabbage       0.493393
almond        0.493176
walnut        0.492942
greyish       0.498743
petale        0.498405
net           0.498102
juices        0.497665
plums         0.497389
raisins       0.497010
cultivars     0.496877
anaki        0.496148
cultivar      0.494975
unfermented   0.490712

```

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

```
~RE[word2vec-master (distance)]
% 50.00% 1.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
raising 0.497010
cultivars 0.496877
atari 0.496848
cultivar 0.496875
unfermented 0.497112

Enter word or sentence (EXIT to break): apple
Word: apple Position in vocabulary: 1221

Word Cosine distance
-----
macintosh 0.682817
lips 0.683916
isac 0.598969
appleworks 0.526189
ibook 0.586129
performa 0.565671
quickdraw 0.543338
anigas 0.538145
amiga 0.526466
laptop 0.524466
hypercard 0.522969
ira 0.515405
lie 0.511895
complan 0.509782
lic 0.508215
mcs 0.506484
microcomputer 0.505018
mcintosh 0.503782
mac 0.503243
visicalc 0.503209
wordzlk 0.502243
lpop 0.494812
os 0.494728
openstep 0.484845
jef 0.480075
macbook 0.487354
macintoshes 0.486822
raskin 0.484658
microsoft 0.483773
atari 0.475969
imovie 0.469714
commodore 0.468627
intel 0.466134
amigone 0.462212
peoples 0.452668
anigas 0.458585
pla 0.458403
oons 0.458255
geos 0.457467
max 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**  
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