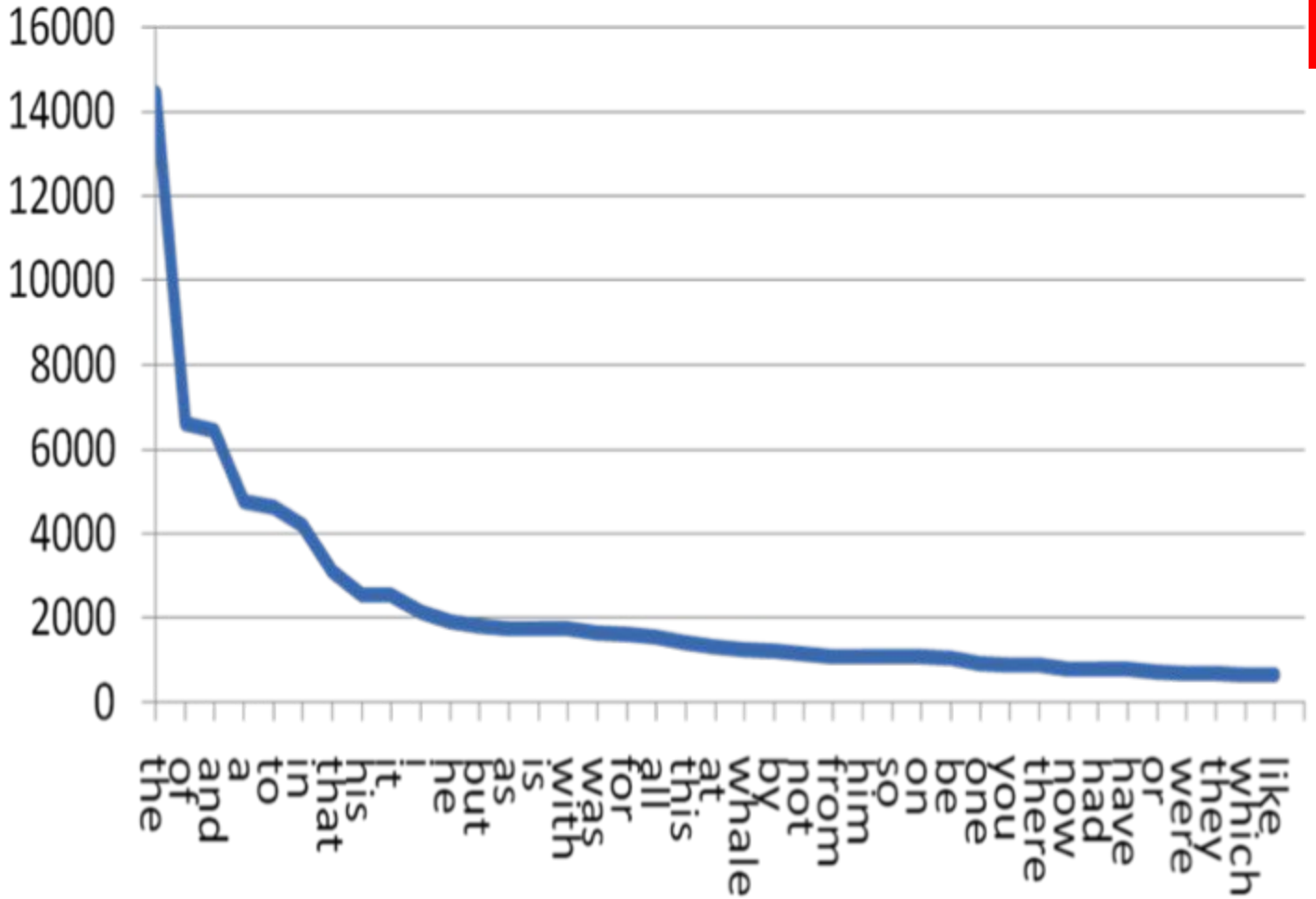
NLP and Embedding

Everything 2 Vec

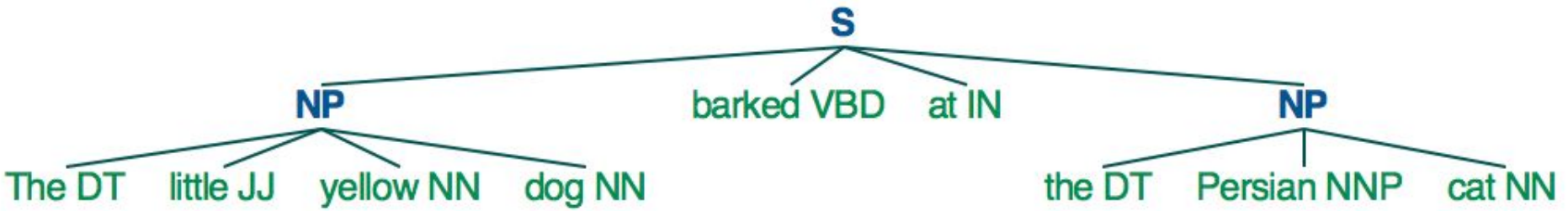
laamp@2016

Linguistics 101

Zipf's law


16000
14000
12000
10000
8000
6000
4000
2000
0
the
of
and
a
to
in
it
is
that
his
he
but
as
with
was
for
all
this
at
by
whale
not
from
him
so
on
be
one
you
there
now
had
have
or
were
they
which
like

Structural/Hierarchical



Anh oi anh dang o dau vay, den day le em dang coi quan ne, anh den nho mua bao nha, o day toan bao cu ko ah, ma thoi anh khoi mua vi em vua mat kinh roi, anh den le di em chiu het noi roi, anh oi

gare (nhà ga)

Hát câu i tờ đón Xuân về ...

Sarcasm

What the British say	What the British mean	What others understand
I hear what you say	I disagree and do not want to discuss it further	He accepts my point of view
With the greatest respect…	I think you are an idiot	He is listening to me
That's not bad	That's good	That's poor
That is a very brave proposal	You are insane	He thinks I have courage
Quite good	A bit disappointing	Quite good
I would suggest…	Do it or be prepared to justify yourself	Think about the idea, but do what you like
Oh, incidentally/ by the way	The primary purpose of our discussion is…	That is not very important
I was a bit disappointed that	I am annoyed that	It doesn't really matter
Very interesting	That is clearly nonsense	They are impressed
I'll bear it in mind	I've forgotten it already	They will probably do it
I'm sure it's my fault	It's your fault	Why do they think it was their fault?
You must come for dinner	It's not an invitation, I'm just being polite	I will get an invitation soon
I almost agree	I don't agree at all	He's not far from agreement
I only have a few minor comments	Please re-write completely	He has found a few typos
Could we consider some other options	I don't like your idea	They have not yet decided

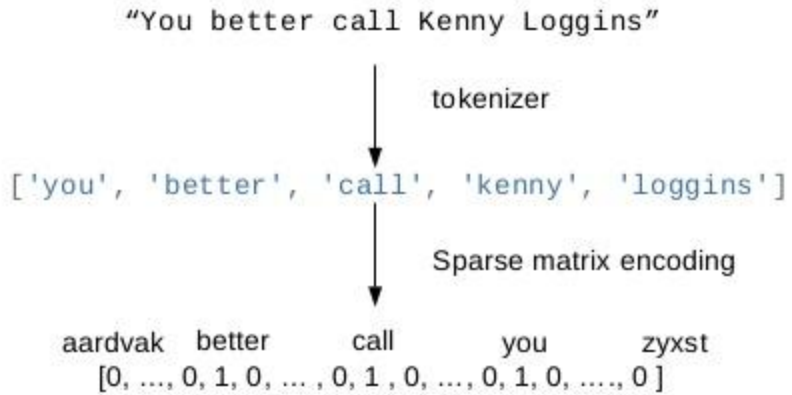
What they say: Merry Christmas!
What they mean: It’s the middle of summer; let’s have a BBQ and drink outdoors.

Australian guy :)

Vector Space Model VSM

Bag Of Word Representations

`CountVectorizer / TfidfVectorizer`



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

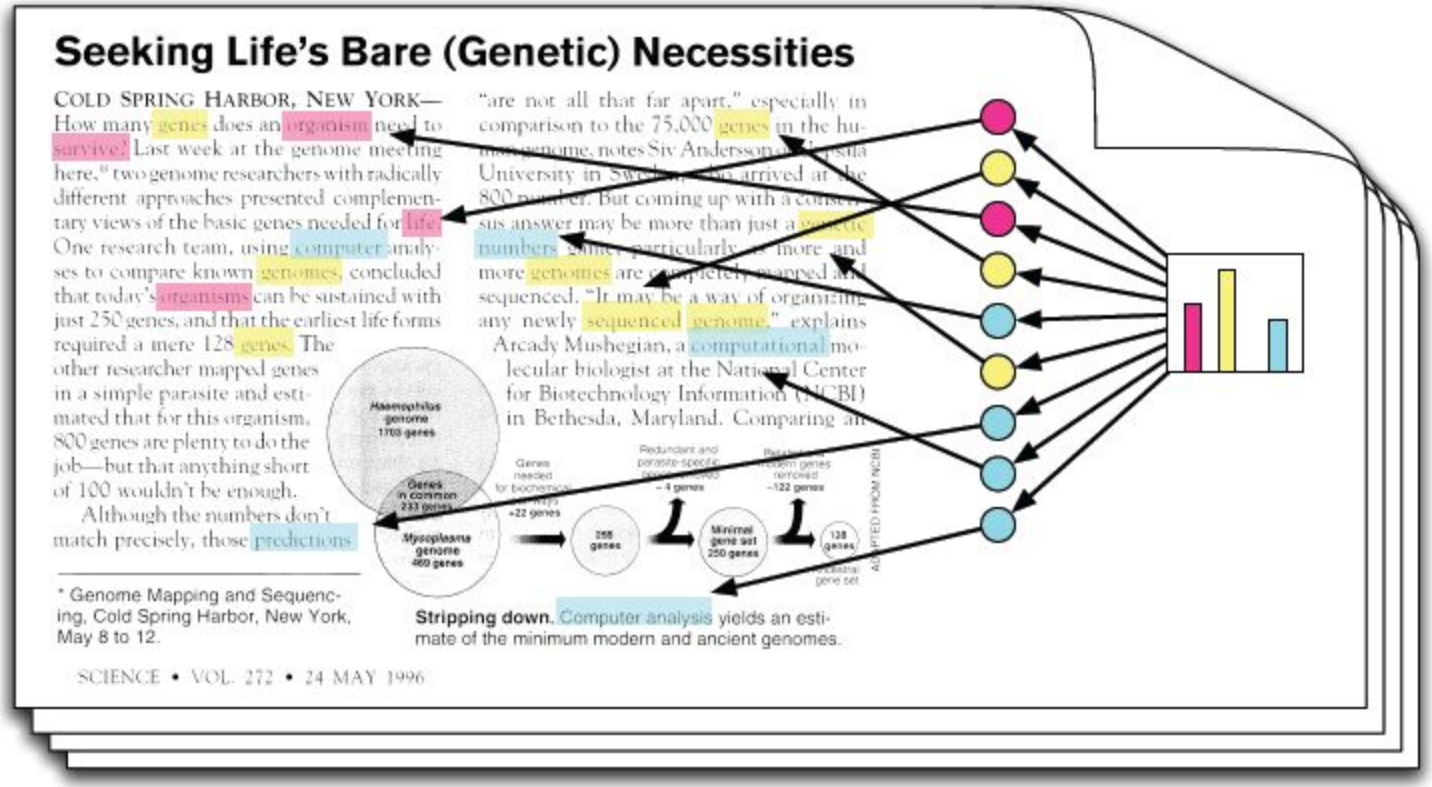
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

1998 LSI

Christos Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, Santosh Vempala, "Latent Semantic Indexing: A probabilistic analysis" (Postscript). *Proceedings of ACM PODS*.

1999 PLSI

Thomas Hofmann, "Probabilistic Latent Semantic Indexing" (PDF). *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*.

2003 LDA

David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty, "Latent Dirichlet allocation". *Journal of Machine Learning Research* **3**: 993–1022.

- A generalization of PLSI
- Allows documents to have a mixture of topics
- The most common topic model currently in use

2006 HDP

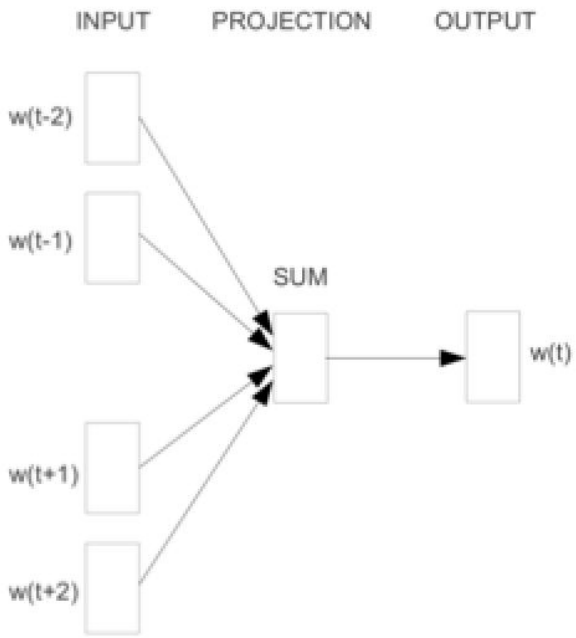
Y. Teh, M. Jordan, M. Beal, and D. Blei. **Hierarchical Dirichlet processes**. *Journal of the American Statistical Association*, 2006. 101[476]:1566-1581.

Variants

- Pachinko allocation (Wei Li and Andrew McCallum, 2006)
- Delta LDA (David Andrzejewski, Anne Mulhern, Ben Liblit, and Xiaojin Zhu, 2007)
- Labeled LDA (Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning, 2009)
  - A supervised topic model in multi-labeled corpora
- ...

WORD2VEC INSIGHTS

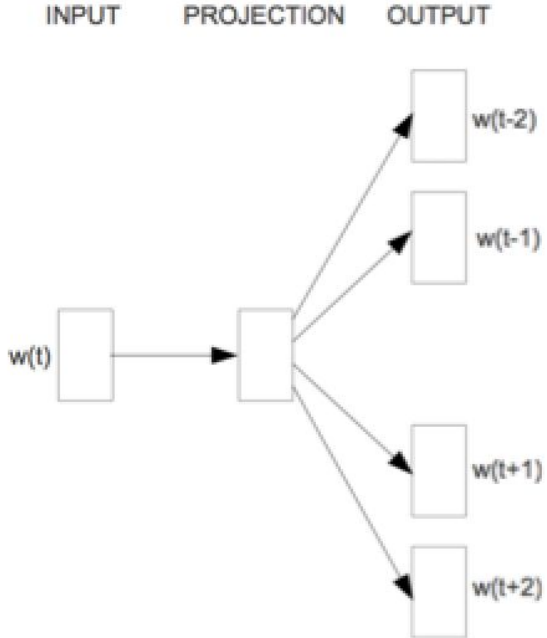
Text representation is a Core of NLP understanding



CBOW

문맥을 통해 현재 단어를 예측

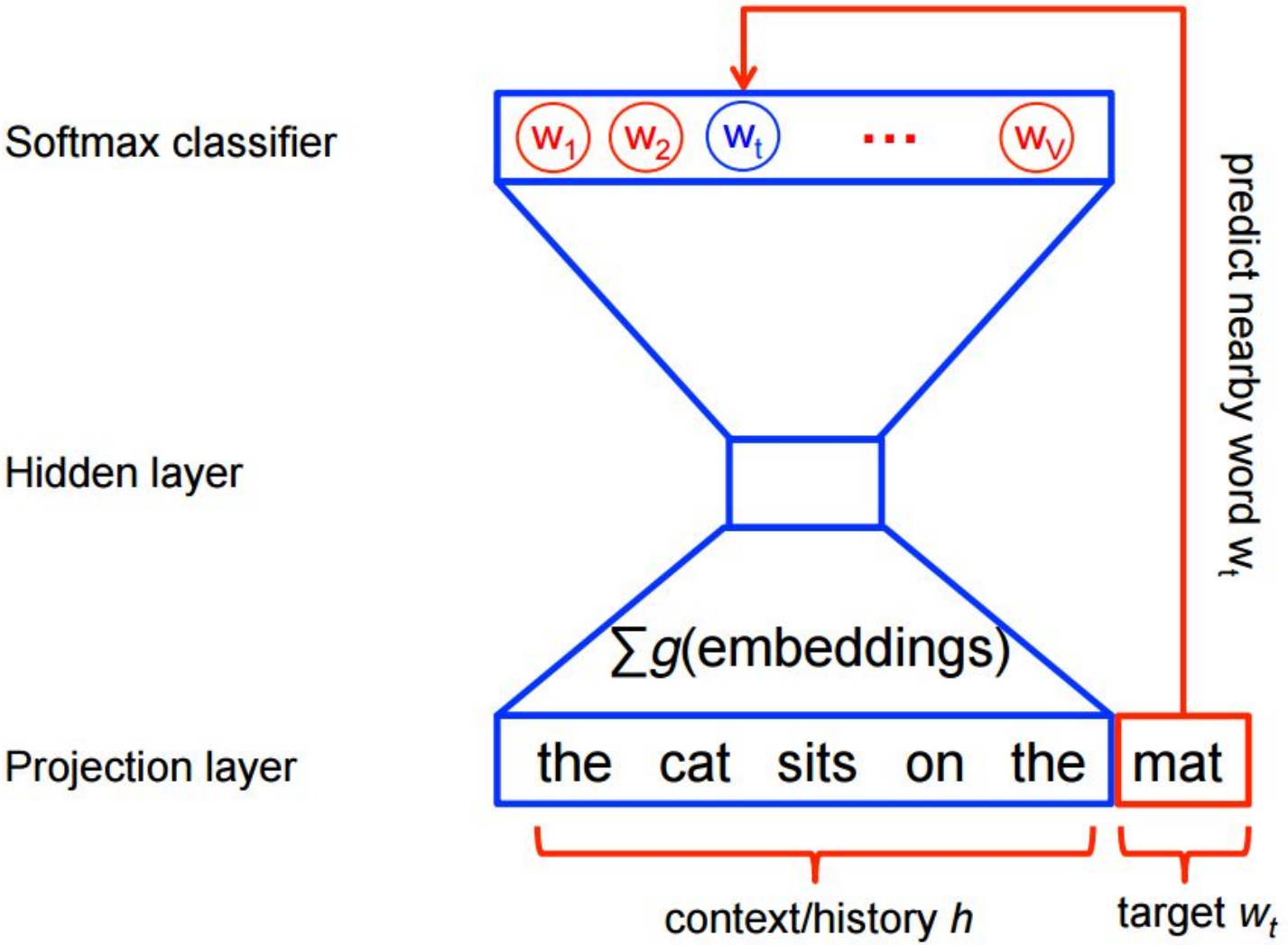
- several times faster to train than the skip-gram
- slightly better accuracy for the frequent words

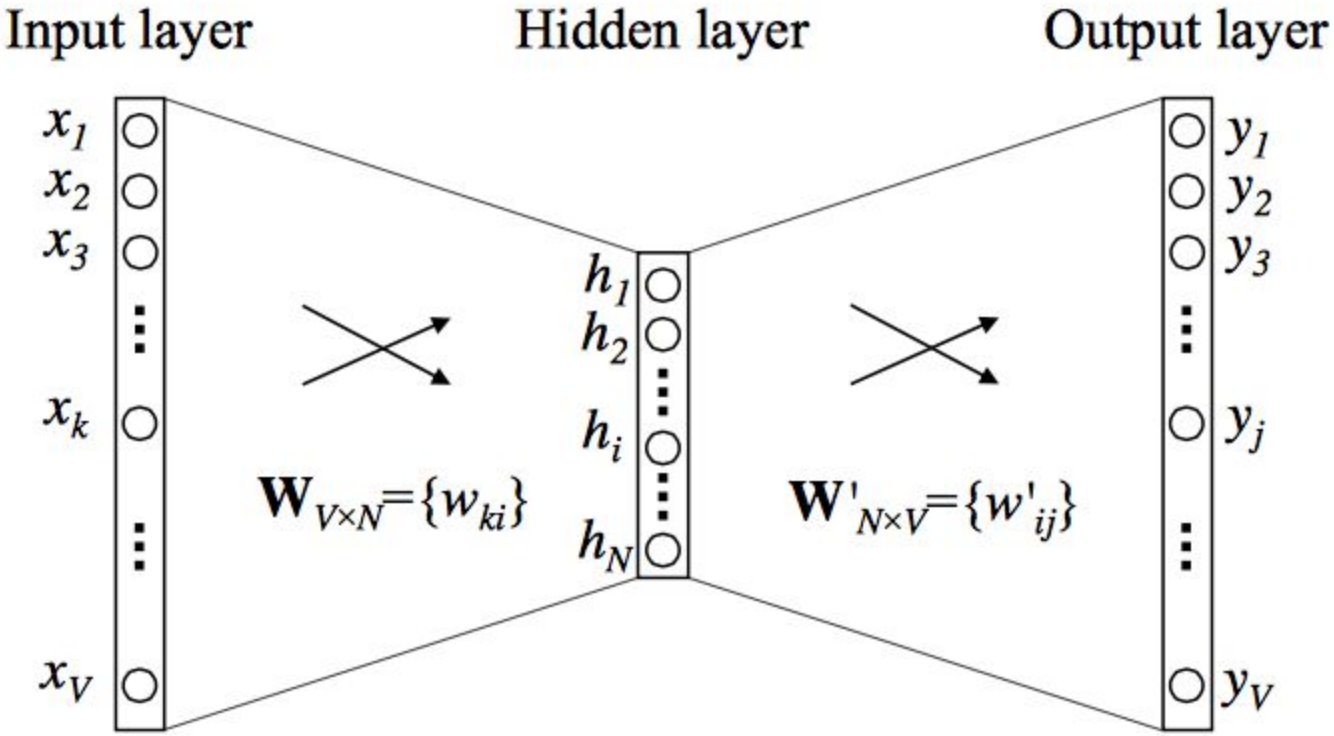


Skip-gram

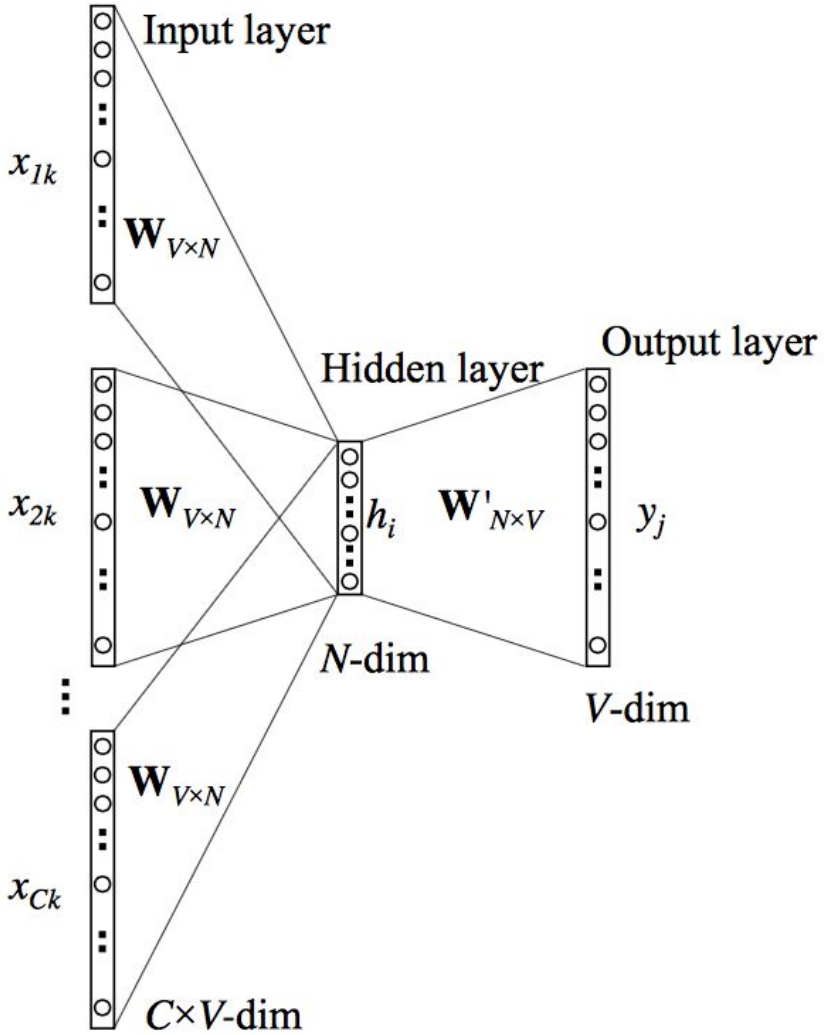
현재 단어를 통해 주위 문맥을 예측

- works well with small amount of the training data
- represents even rare words or phrases well

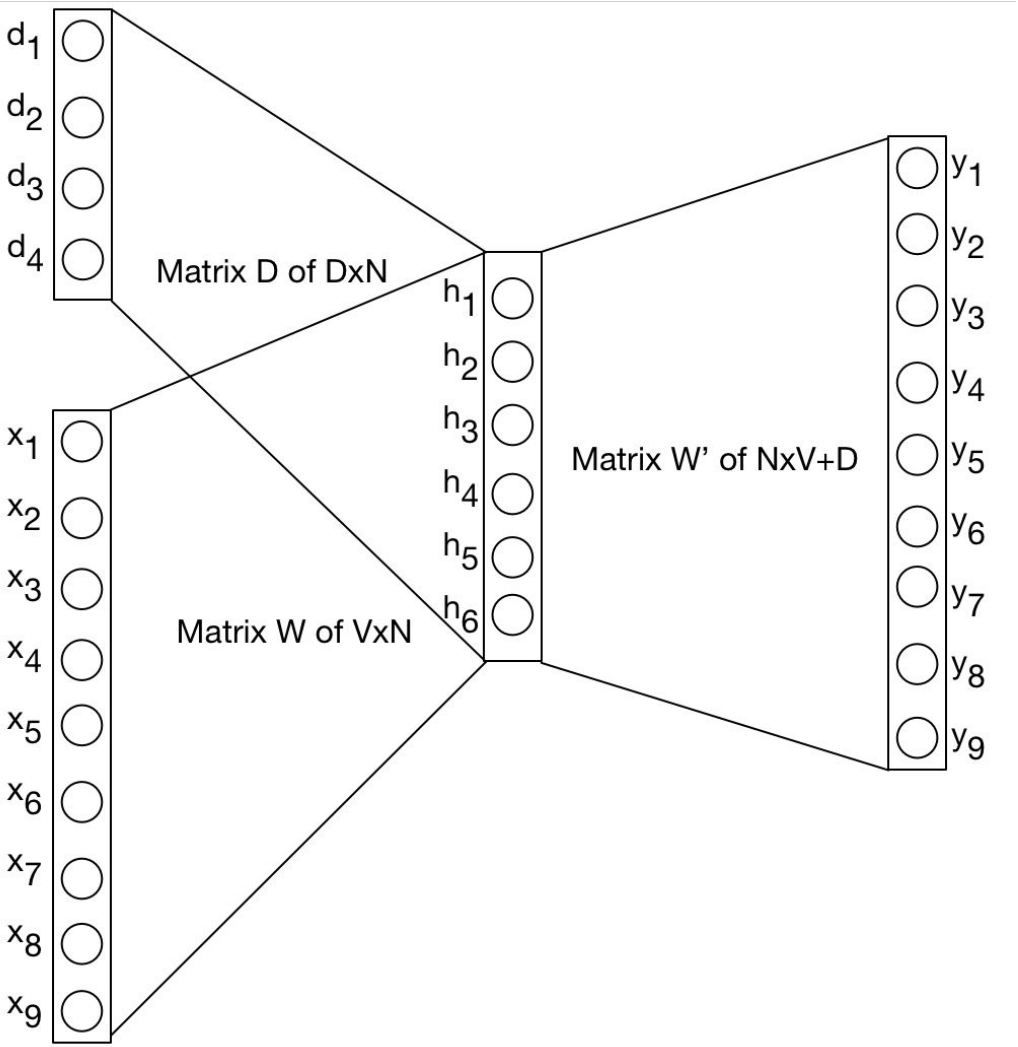

Softmax classifier
Hidden layer
Projection layer
w1 w2 wt ... wV
predict nearby word wt
∑g(embeddings)
the cat sits on the
mat
context/history h
target wt


Input layer
Hidden layer
Output layer
x_1
x_2
x_3
x_k
x_V
h_1
h_2
h_i
h_N
y_1
y_2
y_3
y_j
y_V
$\mathbf{W}_{V\times N}=\{w_{ki}\}$
$\mathbf{W}'_{N\times V}=\{w'_{ij}\}$

Multiword

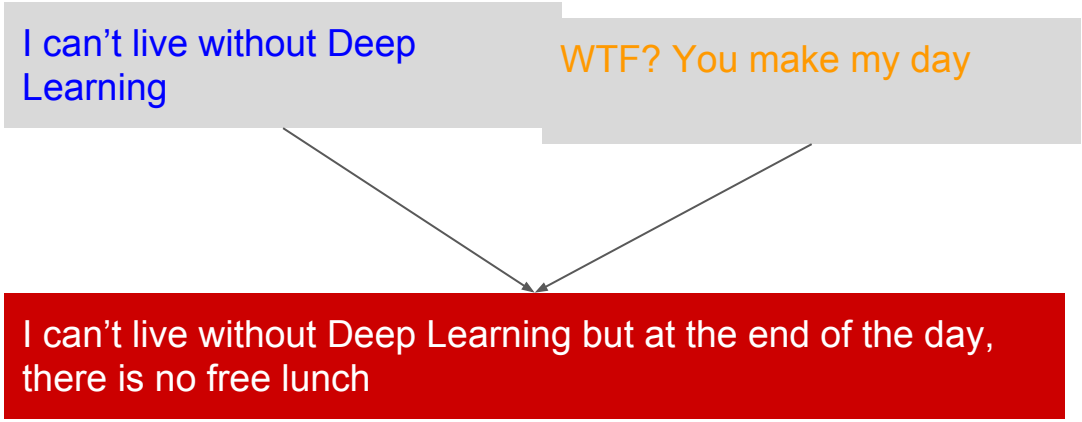

Input layer
x_{1k}
$\mathbf{W}_{V\times N}$
Hidden layer
Output layer
x_{2k}
$\mathbf{W}_{V\times N}$
h_i
$\mathbf{W}'_{N\times V}$
y_j
N-dim
V-dim
$\mathbf{W}_{V\times N}$
x_{Ck}
$C\times V$-dim

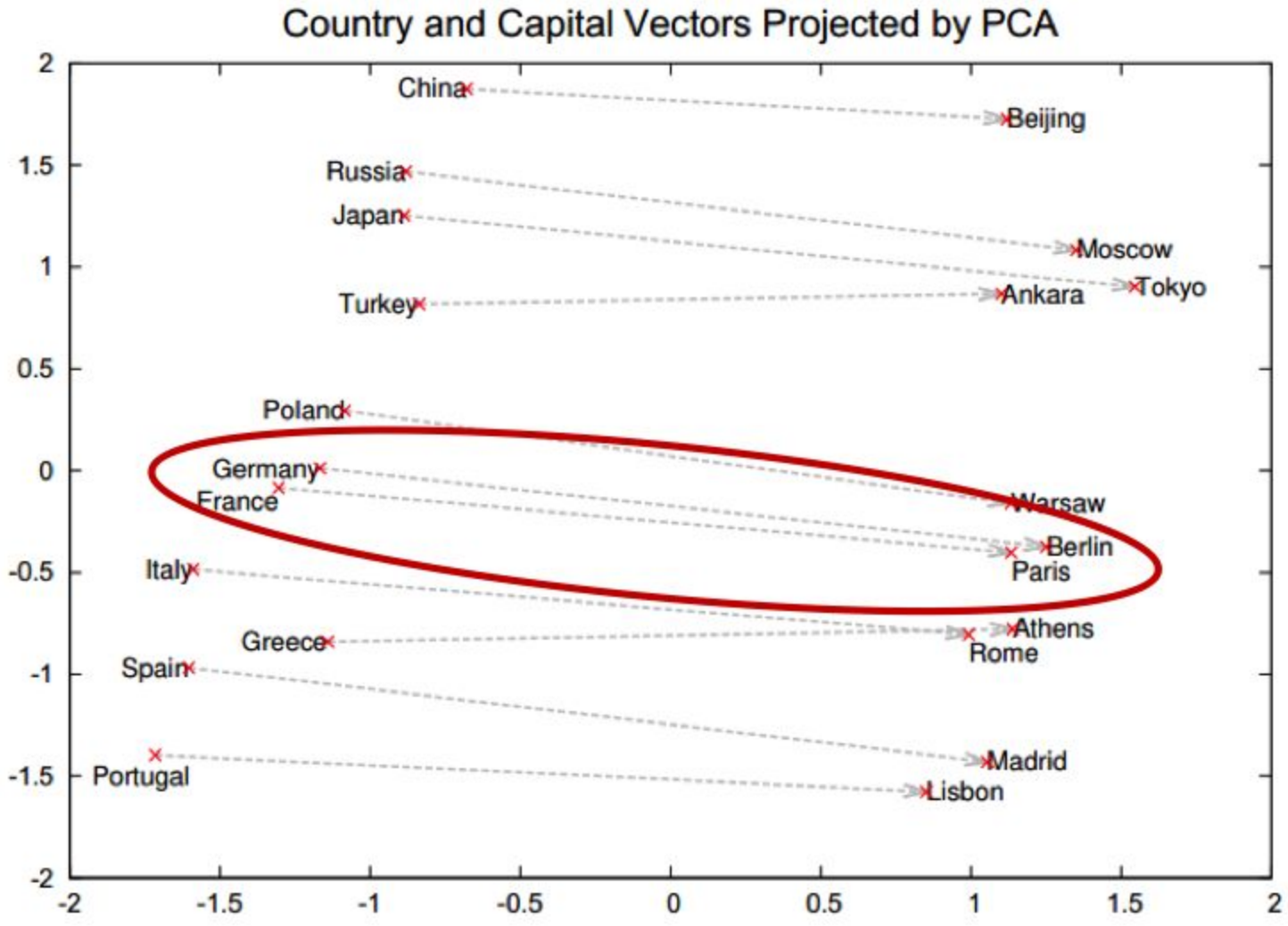
Doc2Vec

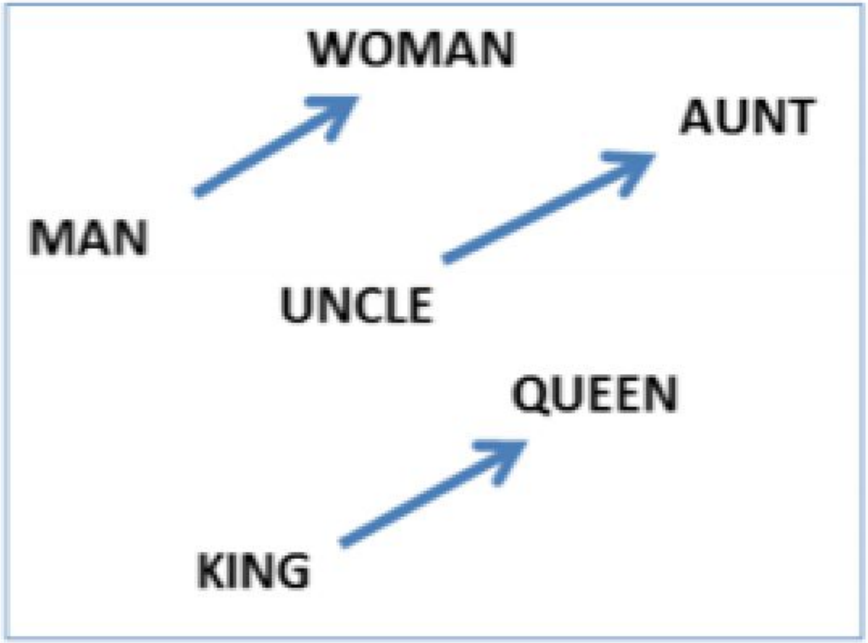

d_1
d_2
d_3
d_4
Matrix D of DxN
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
Matrix W of VxN
h_1
h_2
h_3
h_4
h_5
h_6
Matrix W' of NxV+D
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
y_9

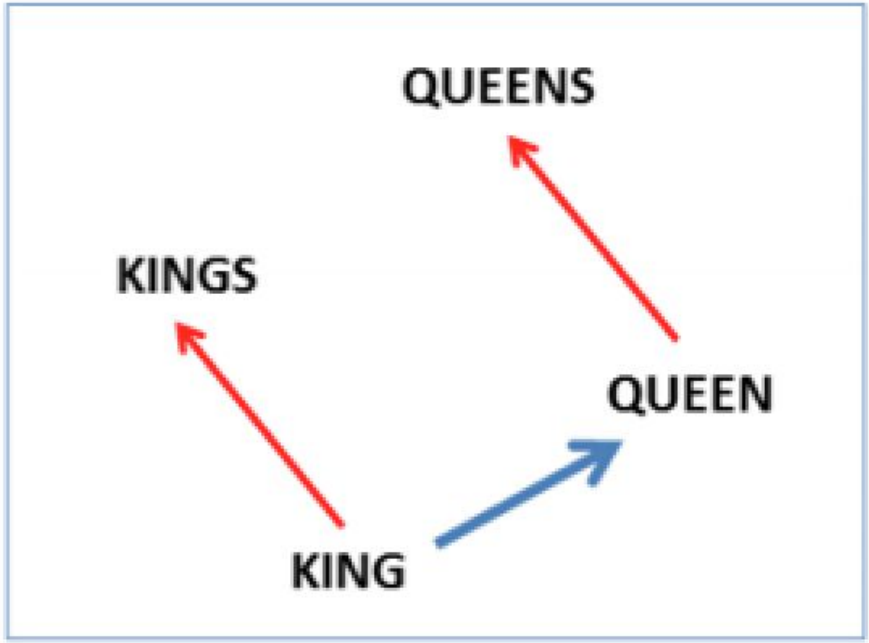
WORD2VEC APPLICATIONS

Sarcasm


I can’t live without Deep Learning
WTF? You make my day
I can’t live without Deep Learning but at the end of the day, there is no free lunch

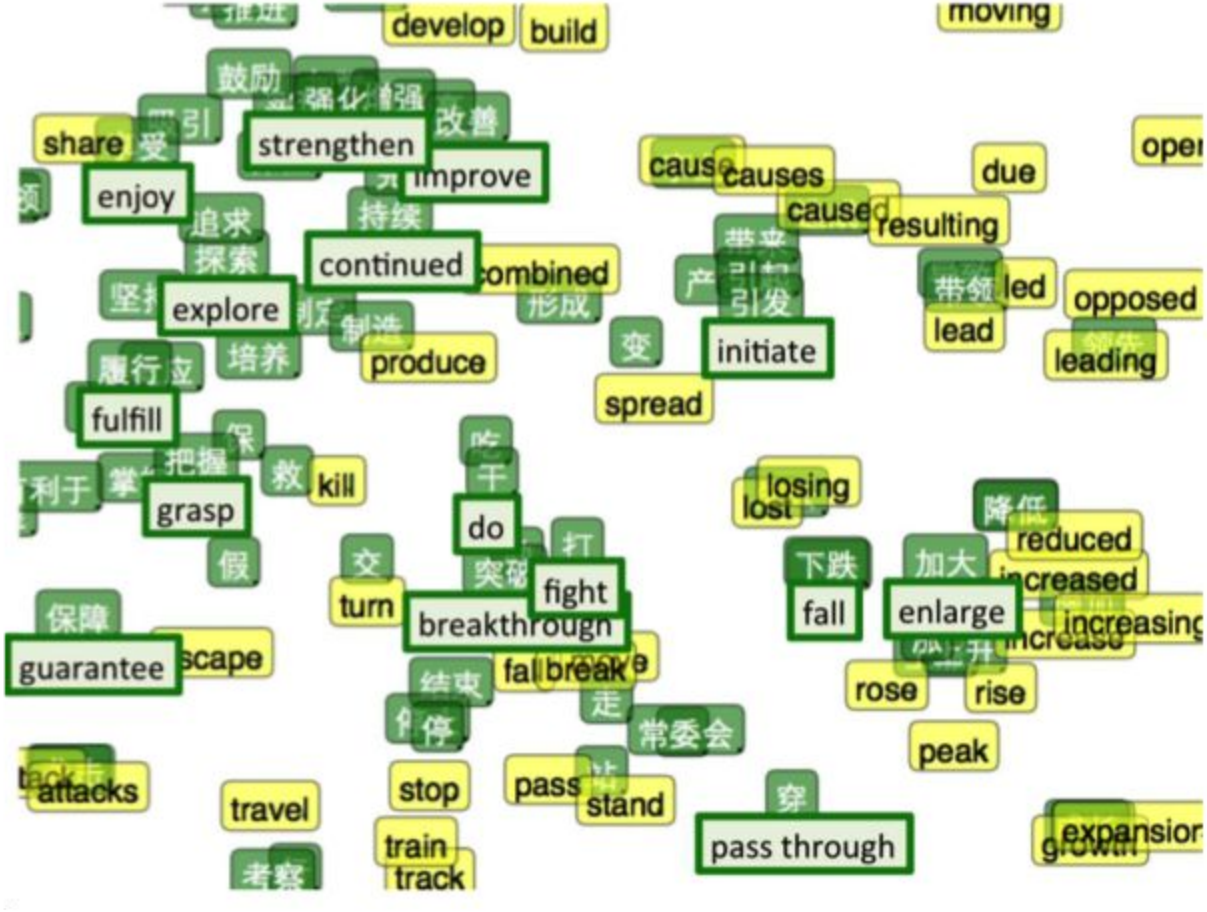

Country and Capital Vectors Projected by PCA
China
Beijing
Russia
Japan
Moscow
Tokyo
Turkey
Ankara
Poland
Warsaw
Germany
France
Berlin
Paris
Italy
Athens
Rome
Greece
Spain
Madrid
Portugal
Lisbon
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2





(Mikolov et al., NAACL HLT, 2013)

Machine Translation



Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 A person riding a motorcycle on a dirt road.	 Two dogs play in the grass.	 A skateboarder does a trick on a ramp.	 A dog is jumping to catch a frisbee.
 A group of young people playing a game of frisbee.	 Two hockey players are fighting over the puck.	 A little girl in a pink hat is blowing bubbles.	 A refrigerator filled with lots of food and drinks.
 A herd of elephants walking across a dry grass field.	 A close up of a cat laying on a couch.	 A red motorcycle parked on the side of the road.	 A yellow school bus parked in a parking lot.

What is the mustache made of?

AI System

bananas

LET’S PLAY WITH IT <3