

An aerial photograph of a vast expanse of broken ice floes floating in a deep blue sea. The ice floes are irregular in shape and size, ranging from small chunks to large, flat sheets. The water between the floes is a dark, deep blue, while the ice itself is a lighter, milky blue. The overall scene conveys a sense of cold and isolation.

Breaking the ice with NLP

Springload | 2019

Ilia Kopylov

An aerial photograph of a vast, frozen body of water, likely in the Arctic or Antarctic. The surface is covered with a mosaic of ice floes of various sizes and shapes, separated by dark, open water. The ice floes have a textured, slightly uneven appearance, with some showing signs of melting or cracking. The overall color palette is a range of blues and whites, from deep cerulean to bright, almost white highlights on the ice edges.

Background

- ***Feature***

An individual measurable property or characteristic of a phenomenon being observed

- ***Feature Engineering***

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.



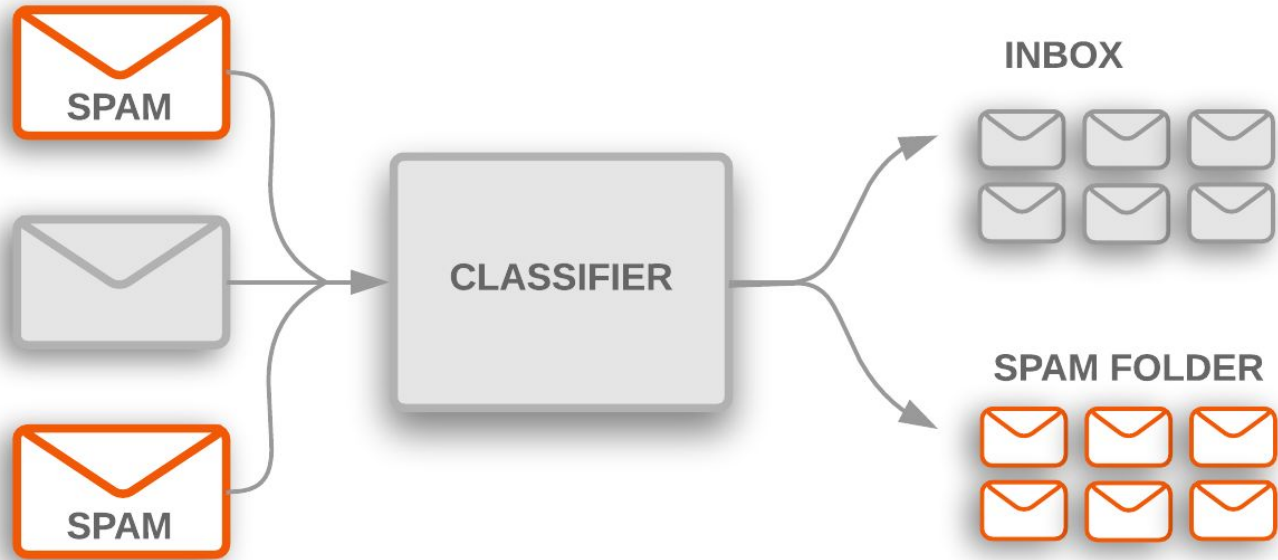


Natural Language Processing

Natural Language Processing

- Understanding
- Generation
- Speech Recognition

Text Classification



Bag of words

*“Is this the real life?
Is this just fantasy?”*

[2, 2, 1, 1, 1, 1, 1]

is	2
this	2
the	1
real	1
life	1
just	1
fantasy	1

[2, 2, 1, 1, 1, 1, 1]

[2, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

[2, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,
0,
0,
0,
0,
0,
0,
0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ...



Stop words

the, is, which, at, on...

Stemming

real, reality, realistic... -> real

Correct spelling ~~mistakes~~ mistakes

Normalise/normalize spellings

TF-IDF

TF-IDF

term frequency–inverse document frequency

TF-IDF

Is this the real life?

is	0.000001
this	0.000002
the	1
real	1
life	1
just	0
fantasy	0

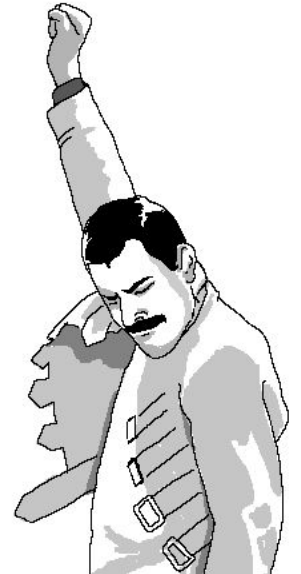

Is this just fantasy?

is	0.000001
this	0.000002
the	0
real	0
life	0
just	1
fantasy	1

Total

is	10000
this	5000
the	1
real	1
life	1
just	1
fantasy	1

[0.5, 0.5, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ...

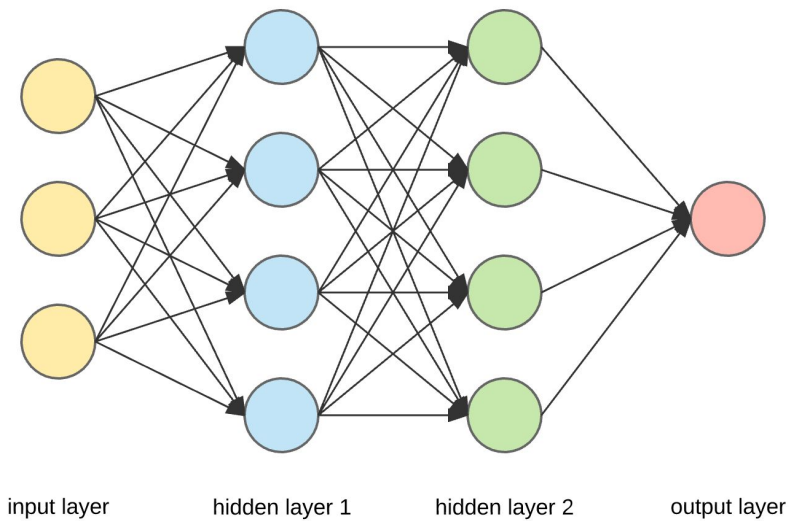
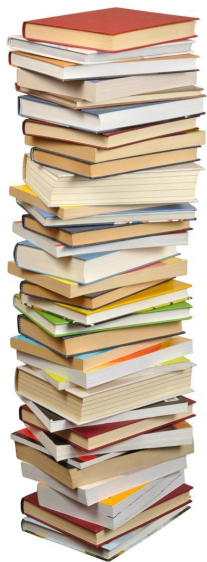


[0.5, 0.5, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ...

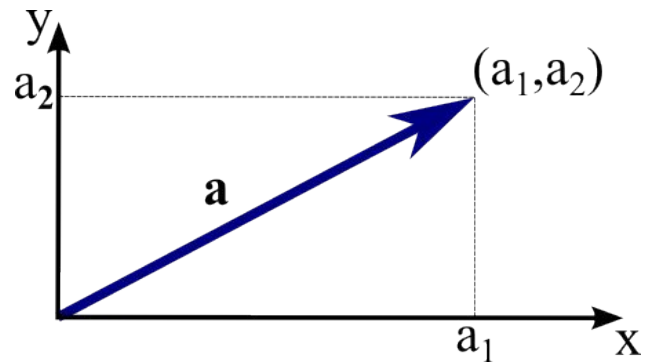
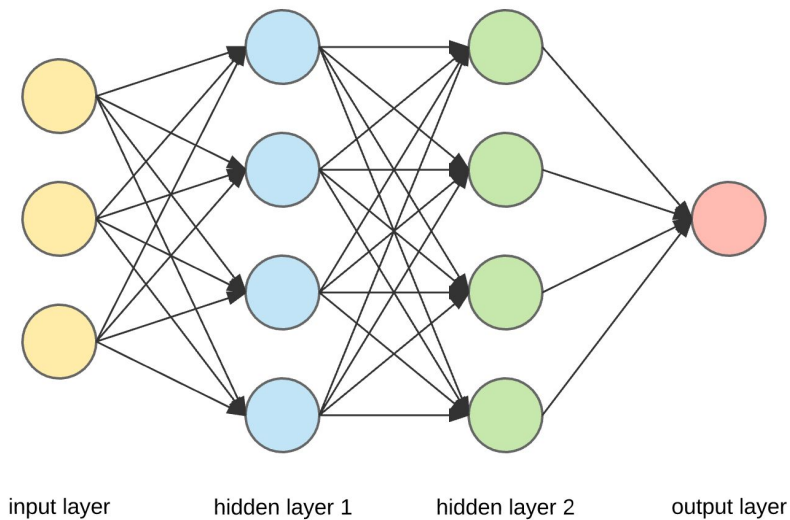
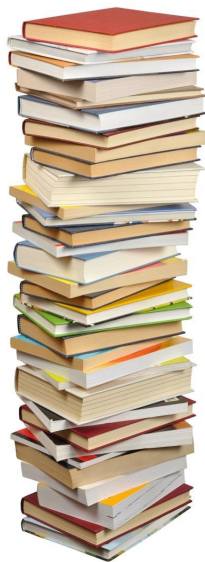
Sparse Vector Models

Dense Vector Models

Dense Vector Models

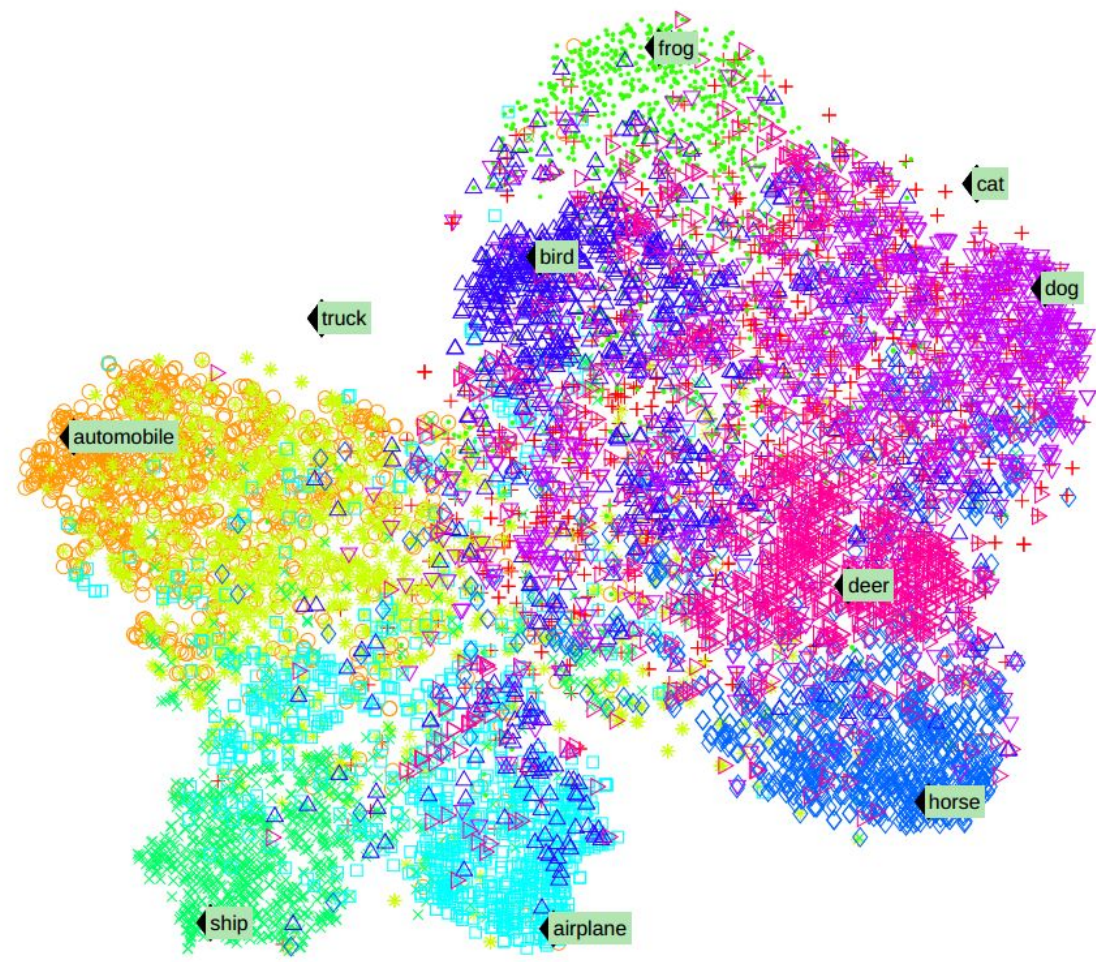


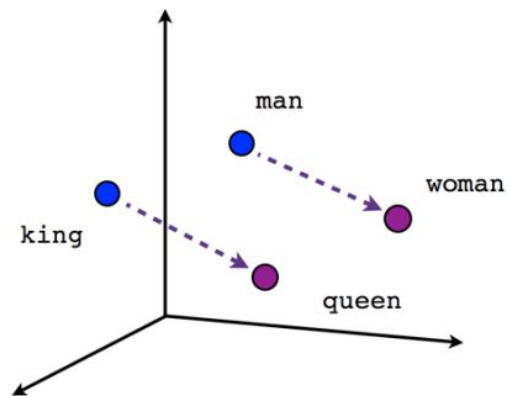
Dense Vector Models



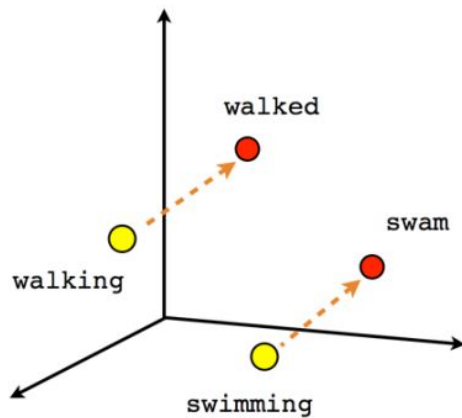
Word vectors
a.k.a.
Word embeddings

- + cat
- automobile
- * truck
- frog
- × ship
- airplane
- ◇ horse
- △ bird
- ▽ dog
- ▷ deer

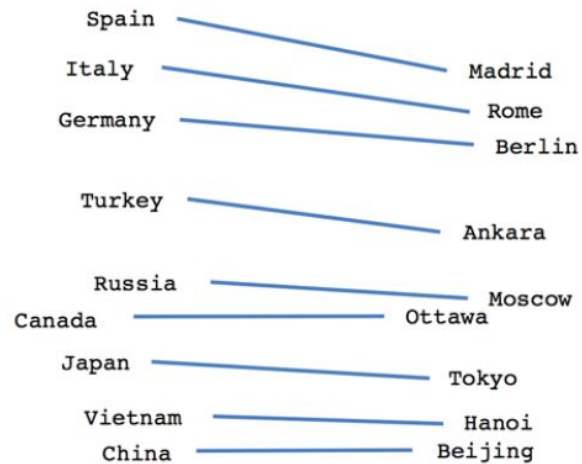




Male-Female



Verb tense



Country-Capital



References

Vector Representations of Words

TensorFlow Tutorial

<https://www.tensorflow.org/tutorials/representation/word2vec>

The Unreasonable Effectiveness of Recurrent Neural Networks

by Andrej Karpathy

<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

bit.ly/NLP-things

iliakplv@gmail.com

springload.co.nz