Text Classification

ต้องการตัดประโยคนี้

- Word segmentation : เติมตัวแบ่งคำ
- Input: ร่องระหว่างตัวอักษรในประโยค
- Output: ตัดคำตรงนั้น vs ไม่ตัดคำตรงนั้น

คนงานขนเสื้อขนแกะใส่กระบะ

- Part of speech tagging : แปะชนิดของคำ
- Input: คำแต่ละคำ
- Output: {Noun, Adj, Verb, Adv, Preposition, ..}

โทรศัพท์รุ่นใหม่ๆ แบตเตอรีไม่ค่อยอึดเท่ารุ่นเก่าๆ

- Sentiment Analysis
- Input: ประโยค
- Output: {ทัศนคติแง่บวก, ทัศนคติแง่ลบ, อื่นๆ}

- ให้คะแนนเรียงความอัตโนมัติ
- Input: เรียงความ
- Output: {0, 1, 2, 3, 4, 5}

- ตรวจจับข่าวปลอม
- Input: ข่าว
- Output: {ปลอม, จริง}

3 ปัจจัยหลักๆ ที่ส่งผลต่อคุณภาพของ Text Classifier

1. ปริมาณและคุณภาพของข้อมูล
2. Algorithm e.g. Naive Bayes, Logistic Regression, Random Forests, Neural networks
3. Feature Engineering

Logistic Regression

Parameter ของโมเดล Maximum Entropy
หลังจากฝึกเสร็จเรียบร้อยแล้ว

	ค่าสัญญาณ
f1 (คำว่าต่อต้าน)	1
f2 (คำว่าชื่นชอบ)	0
f3 (จำนวนตัวอักษร)	100

	บวก	ลบ	กลาง
f1 (คำว่าต่อต้าน)	-2	2	-1
f2 (คำว่าชื่นชอบ)	-1	-0.2	0.4
f3 (จำนวนตัวอักษร)	0.0004	0.005	-0.00001

	บวก	ลบ	กลาง
f1 (คำว่าต่อต้าน)	1 x - 2	1 x 2	1 x -1
f2 (คำว่าชื่นชอบ)	0 x -1	0 x -0.2	0 x 0.4
f3 (จำนวนตัวอักษร)	100 * 0.0004	100 x 0.005	100 x -0.00001

	บวก	ลบ	กลาง
f1	-2	2	-1
f2	-1	-0.2	0.4
f3	0.0004	0.005	-0.00001

	บวก		ลบ		กลาง	
f1 (คำว่าต่อต้าน)	1 x - 2	-2	1 x 2	2	1 x -1	-1
f2 (คำว่าชื่นชอบ)	0 x -1	0	0 x -0.2	0	0 x 0.4	0
f3 (จำนวนตัวอัก	100 * 0.0004	0.04	100 x 0.005	0.5	100 x -0.00001	-0.001
คะแนนรวม		**1.96**		**2.5**		**-1.001**

```
In [64]: exp(1.96) / (exp(1.96) + exp(2.5) + exp(-1.001))
Out[64]: 0.3613011773847603
```

$$
\begin{aligned}
\ell(W; X, Y) &= \sum_{i=0}^{n} \log P(Y = y^i | x^i, W) \\
&= \sum_{i=0}^{n} \log \frac{\exp a(y^i; x^i, W)}{\sum_j \exp a(j; x^i, W)} \\
&= \sum_{i=0}^{n} \log \exp a(y^i; x^i, W) - \log \sum_j \exp a(j; x^i, W) \\
a(j; x, W) &= \sum_{i=0}^{k} w_{ij} x_i
\end{aligned}
$$

Log-likelihood

$$\arg\max_W \ell(W; X, Y) = \arg\min_W -\ell(W; X, Y)$$

Crossentropy loss
Negative log-likelihood

Stochastic Gradient Descent (SGD)

Optimization Problem

$$\arg\max_{W} \ell(W; X, Y) = \arg\min_{W} -\ell(W; X, Y)$$

Gradient-based training/optimization

Gradient of Cross-entropy Loss

$$\arg\max_W \ell(W; X, Y) = \arg\min_W -\ell(W; X, Y)$$

$$\frac{\partial \ell(W; X, Y)}{\partial w_{ij}} = x_i(P(Y = j|X) - \mathbb{1}(y^* = j))$$

	บวก	ลบ	กลาง
f1	-2	2	-1
f2	-1	-0.2	0.4
f3	0.0004	0.005	-0.00001

	ค่าสัญญาณ
f1	1
f2	0
f3	100

Stochastic Gradient Descent

for x, y in training_set:

compute $P(Y|X)$

update w_{ij} ของใหม่:= w_{ij} ของเก่า - α x_i $(P(Y=j|X) - 1(y=j)$

Stochastic Gradient Descent

for x, y in training_set:

compute P(Y|X)

update w_{ij} ของใหม่:= w_{ij} ของเก่า - α x_i (P(Y=j|X) - 1(y=j)

Batch Gradient Descent

for x, y in training_set:

compute P(Y|X)

$gradient_{ij}$ += x_i (P(Y=j|X) - 1(y=j)

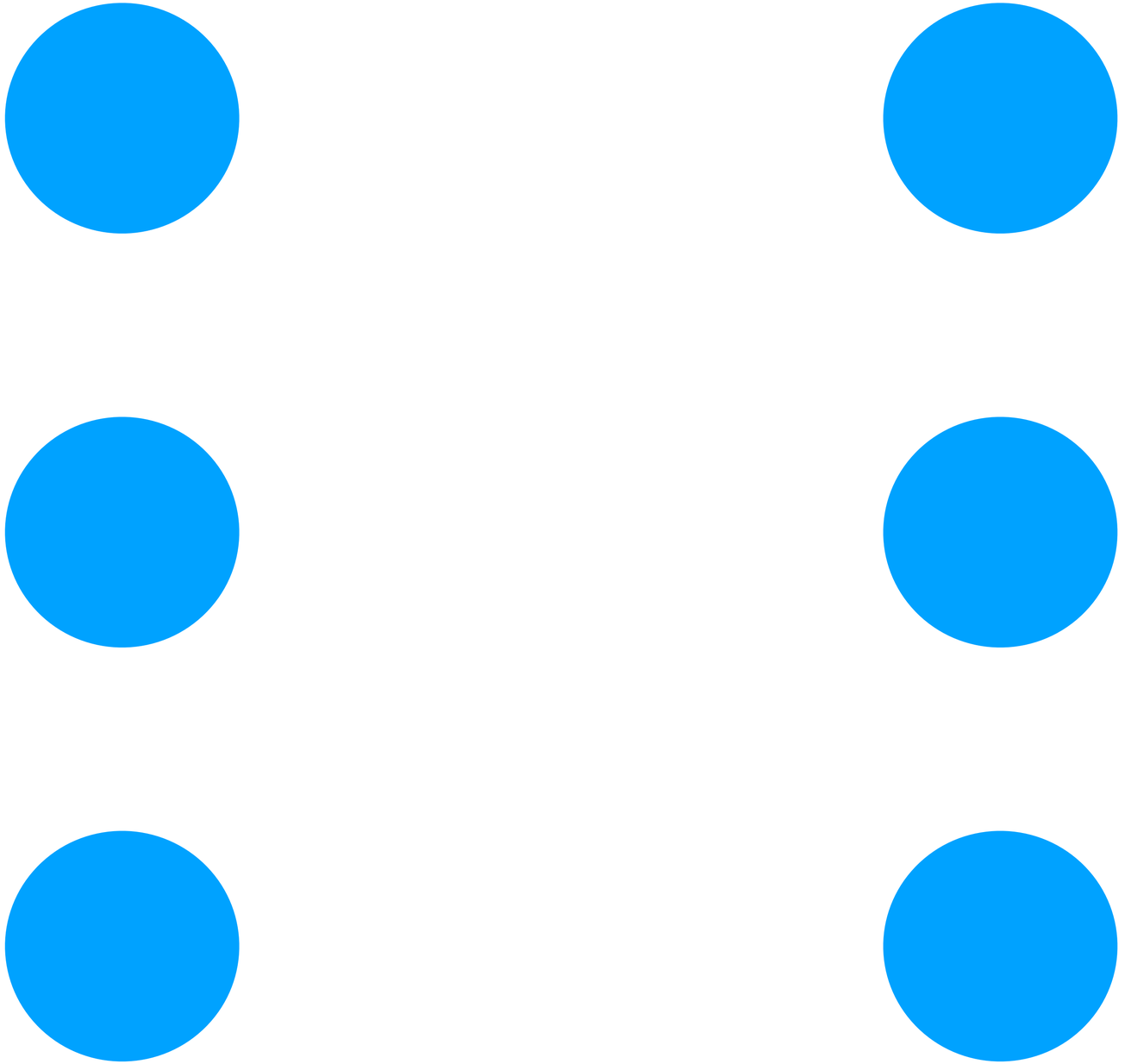
update w_{ij} ของใหม่:= w_{ij} ของเก่า - α $gradient_{ij}$

Gradient-based training/optimization

Feedforward Neural Network

	f1	f2	f3
บวก	-2	-1	0
ลบ	2	0	0
กลาง	-1	0.4	0

f1	1
f2	0
f3	100



$$o = \text{softmax}(W^T x + b)$$

	f1	f2	f3
บวก	-2	-1	0
ลบ	2	0	0
กลาง	-1	0.4	0

f1	1
f2	0
f3	100

+

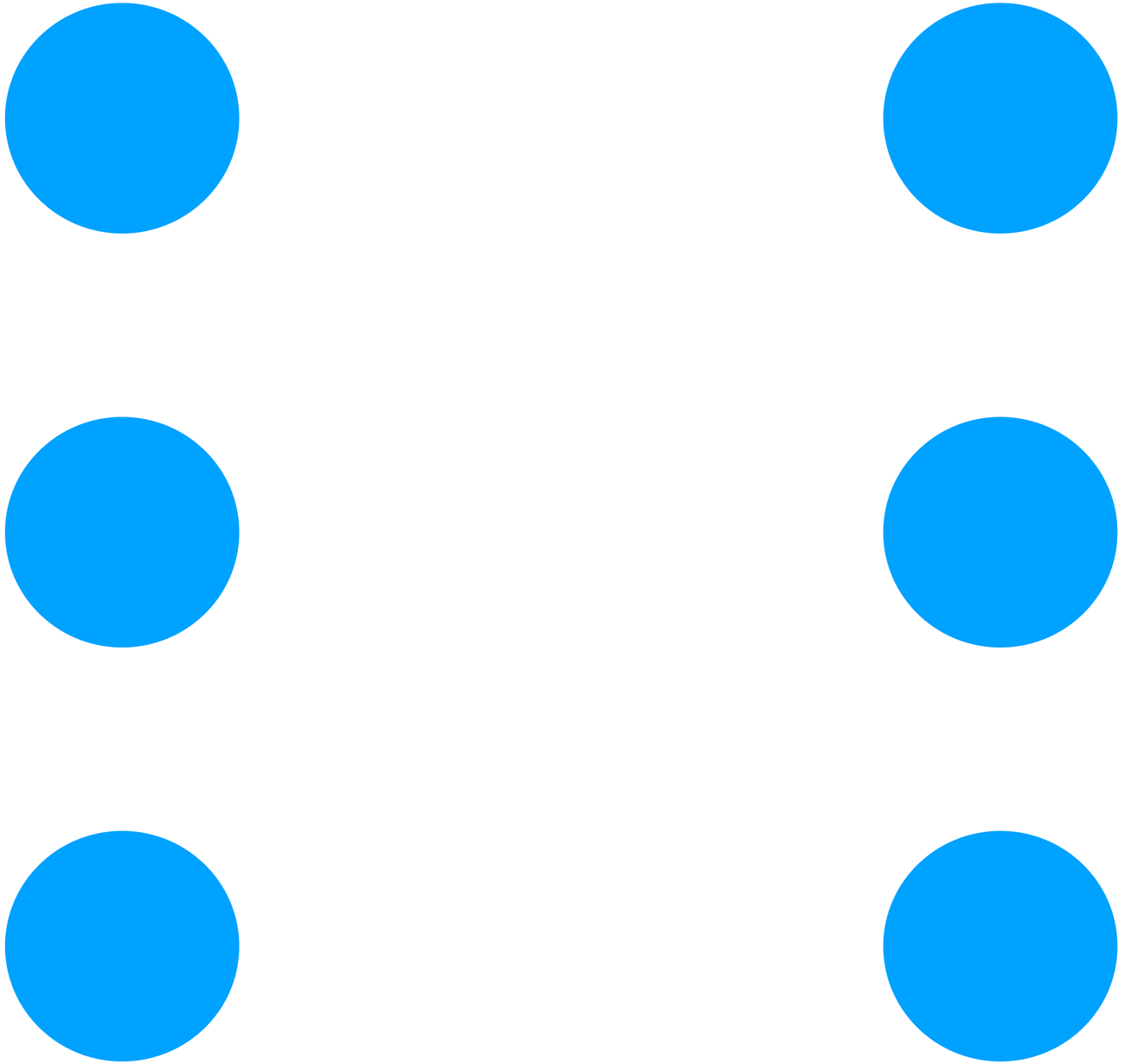
0.001
0.006
0.002

=

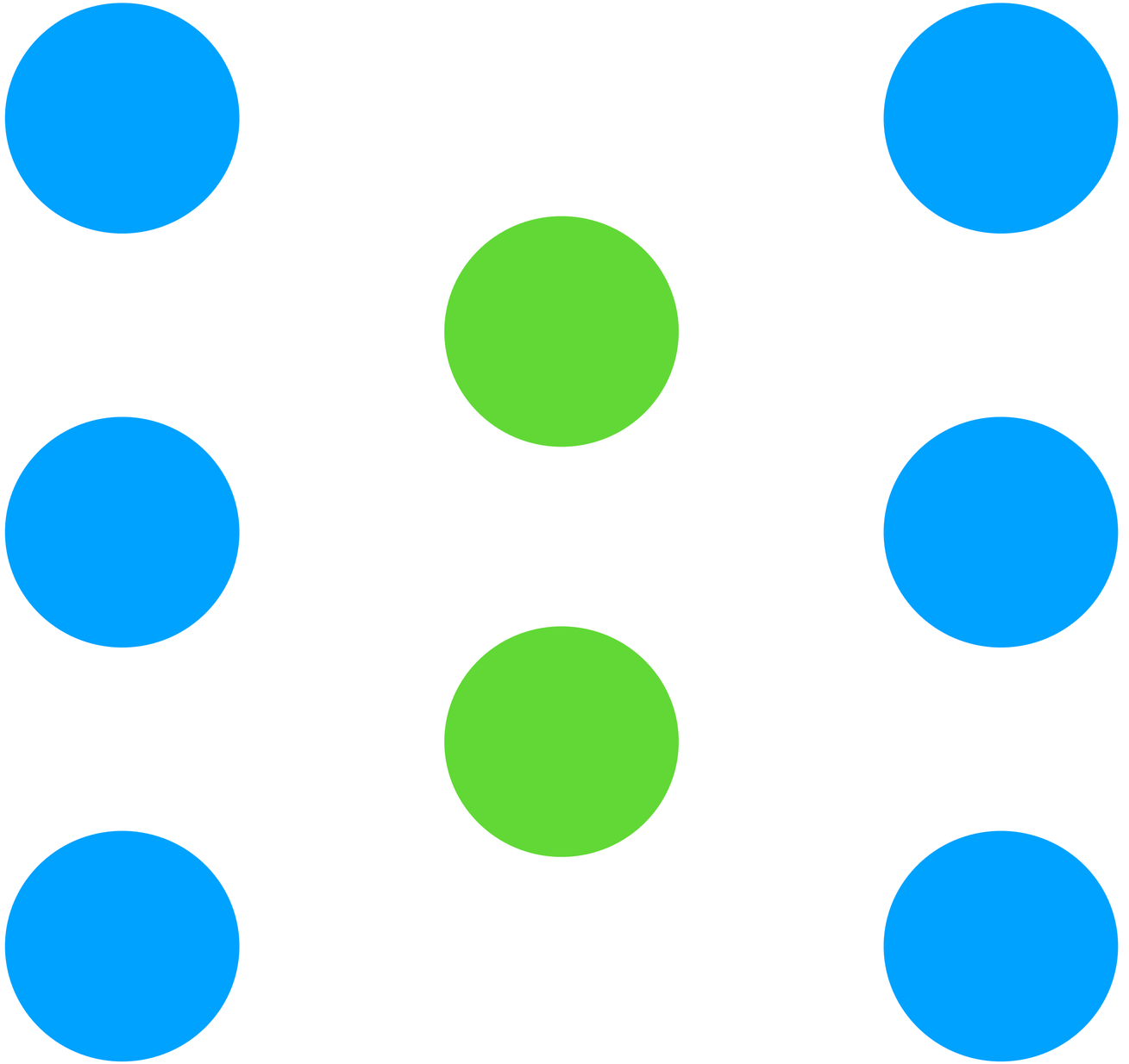
1.961
2.506
-0.999

	f1	f2	f3
บวก	-2	-1	0
ลบ	2	0	0
กลาง	-1	0.4	0

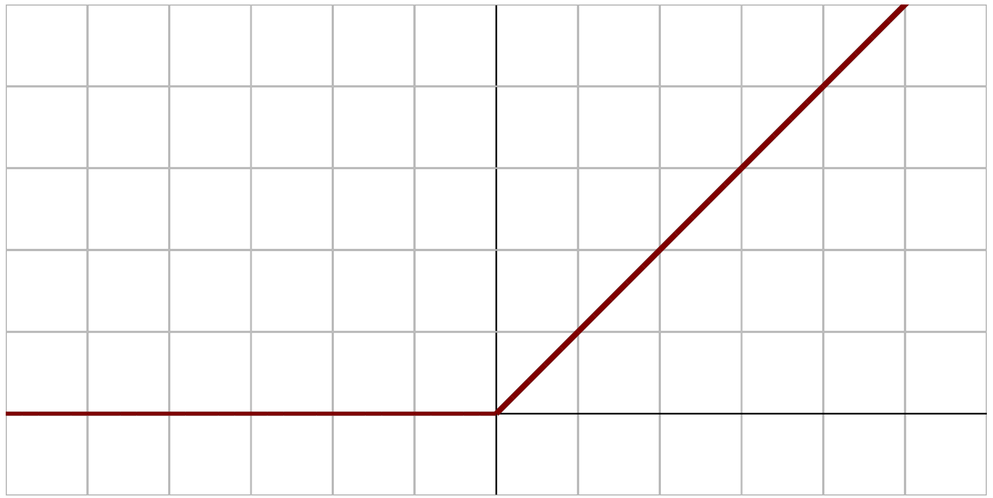
f1	1
f2	0
f3	100



f1	1
f2	0
f3	100



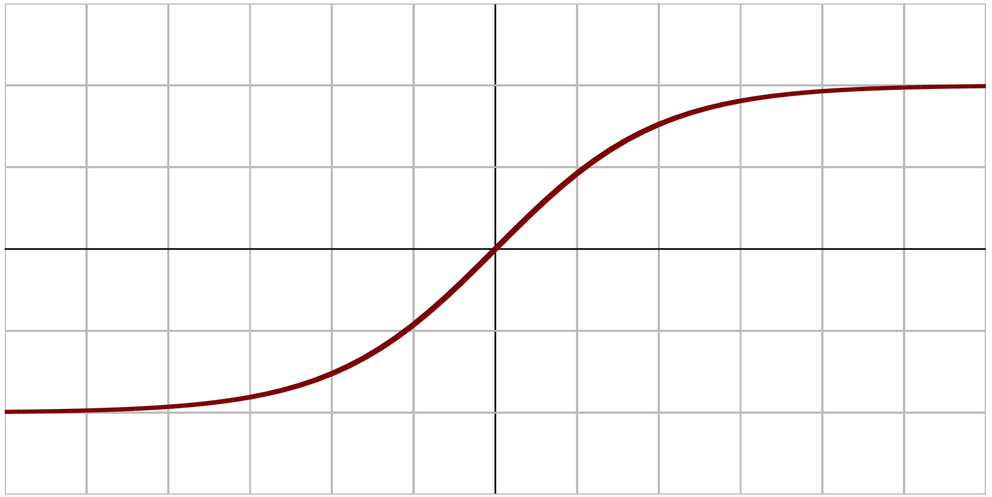
Rectified Linear Unit (ReLU)



$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

$$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

Hyperbolic Tangent (tanh)



$$f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

$$f'(x) = 1 - f(x)^2$$

Backpropagation

Training Neural Network

Techniques in Neural Network Training

- Automatic gradient computation
- Dropout
- Learning Rate

Distributional Semantics

Techniques in Neural Network Training

- Automatic gradient computation
- Dropout
- Learning Rate

Word Embedding

Techniques in Neural Network Training

- Automatic gradient computation
- Dropout
- Learning Rate