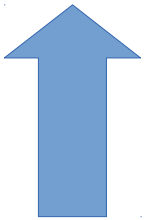
# 自然言語処理のための機械学習入門

2. 文書および単語の数学的表現
   2.1 タイプ、トークン
   2.2 n グラム
   2.3 文書、文のベクトル表現

新納浩幸

# 文書や単語をベクトルで

文書や単語を数学的、機械的に取り扱うために、
それらをベクトルで表現するのが一般的



どうやって、何か問題あるの？
（２章の内容）

# タイプとトークン

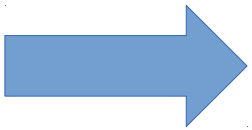
単語の数え方

タイプ： 単語の種類

トークン： 単語の出現数

例

Nurture or nature? Nurture passes nature.



４タイプ、６トークン

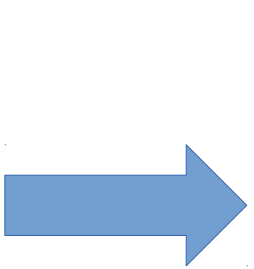
# n グラム

n グラム： 隣り合って出現した n 単語

N = 1 : ユニグラム
N = 2 : バイグラム
N = 3 : トライグラム

例

nurture passes nature.



uni-gram = {nurture, passes, nature}

bi-gram = {nurture-passes, passes-nature}

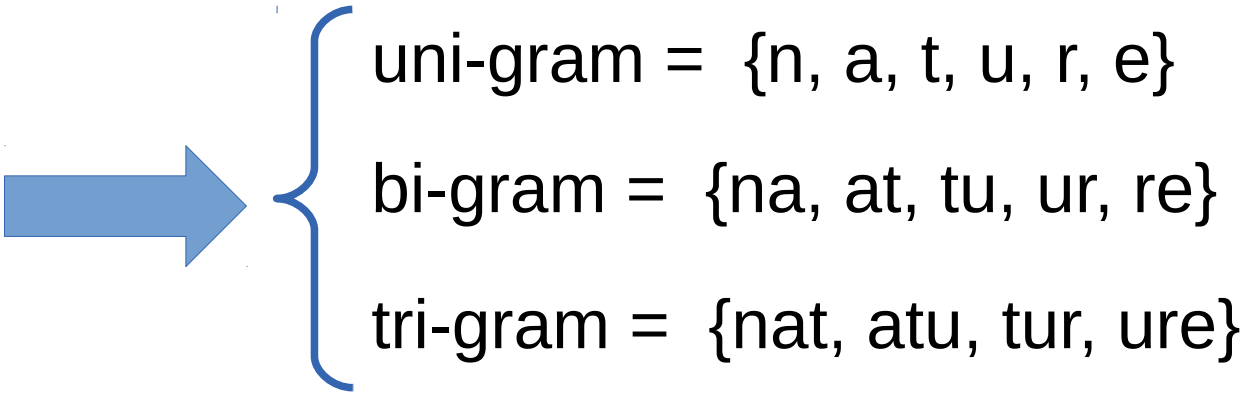
tri-gram = {nurture-passes-nature}

# 文字 n グラム

文字 n グラム： 隣り合って出現した n 文字

例

nature



# 文書、文のベクトル表現

素性：（文書）ベクトルの次元に対応する特徴

文書 d のベクトルを $\boldsymbol{x}^{(d)}$ とする

$n(w,d)$ ：文書 d 内の単語 w の頻度

これを素性に設定すると Bag-of-Words

例

Nurture or nature? Nurture passes nature.

$$\begin{aligned}\boldsymbol{x}^{(d)} &= (n(\text{"}nature\text{"},d), n(\text{"}nurture\text{"},d), n(\text{"}or\text{"},d), n(\text{"}passes\text{"},d)) \\ &= (2,2,1,1)\end{aligned}$$

# Bag-of-bigrams

Bag-of-words の単語を bigram にしたもの

例

s1 : The pen is mightier than the sword.
s2 : The sword is mightier than the pen.

Bag-of-words

$$\boldsymbol{x}^{(s1)} = \boldsymbol{x}^{(s2)} = (1, 1, 1, 1, 1, 2)$$

語順の情報がない

Bag-of-bigrams

$$\boldsymbol{x}^{(s1)} = (1, 1, 1, 0, 1, 1, 1)$$
$$\boldsymbol{x}^{(s2)} = (1, 1, 0, 1, 1, 1, 1)$$

語順の情報が少しある