



Chatbot at 10:42





## Section3の概要



# 講座の内容

Section 1. コースの概要とTwitter API

Section 2. RNNとSeq2Seq



**Section 3. 自然言語処理の基礎**

Section 4. モデルの訓練

Section 5. Attentionの導入

Section 6. Twitterボットのデプロイ

# 今回の内容

1. Section3の概要
2. 単語の分散表現
3. 分散表現の実装
4. テキストデータの前処理



# 教材の紹介

## Pythonの基礎

### Section3の教材:

- 01\_word\_vector.ipynb
- 02\_preprocessing.ipynb
- wagahaiwa\_nekodearu.txt

# 単語の分散表現



# one-hot表現

すもも も もも も もも の うち

|    | すもも | も | もも | の | うち |
|----|-----|---|----|---|----|
| ID | 0   | 1 | 2  | 3 | 4  |

「すもも」のone-hot表現: [1 0 0 0 0]

「も」のone-hot表現: [0 1 0 0 0]



# 分散表現

- 単語間の関連性や類似度に基づくベクトルで、単語を表現する

200要素程度

|        |      |      |      |     |
|--------|------|------|------|-----|
| 男性     | 0.01 | 0.58 | 0.24 | ... |
| ロンドン   | 0.34 | 0.93 | 0.02 | ... |
| Python | 0.97 | 0.08 | 0.41 | ... |

- 単語を表すベクトル同士で、足し算や引き算が可能。

例: 「王」 - 「男」 + 「女」 = 「女王」

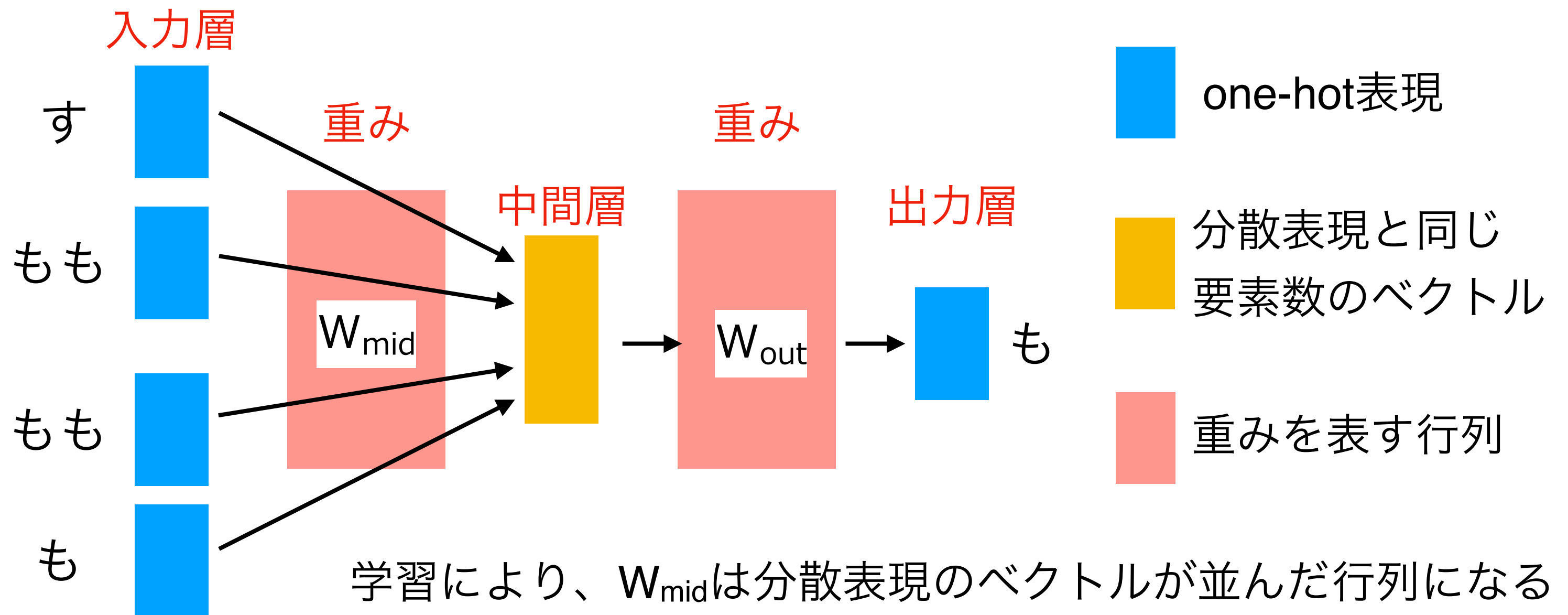


# word2vec

- word2vecは、分散表現を作成するための技術
- word2vecでは、CBOW (continuous bag-of-words) もしくは、skip-gramというニューラルネットワークが用いられる

# CBOW (continuous bag-of-words)

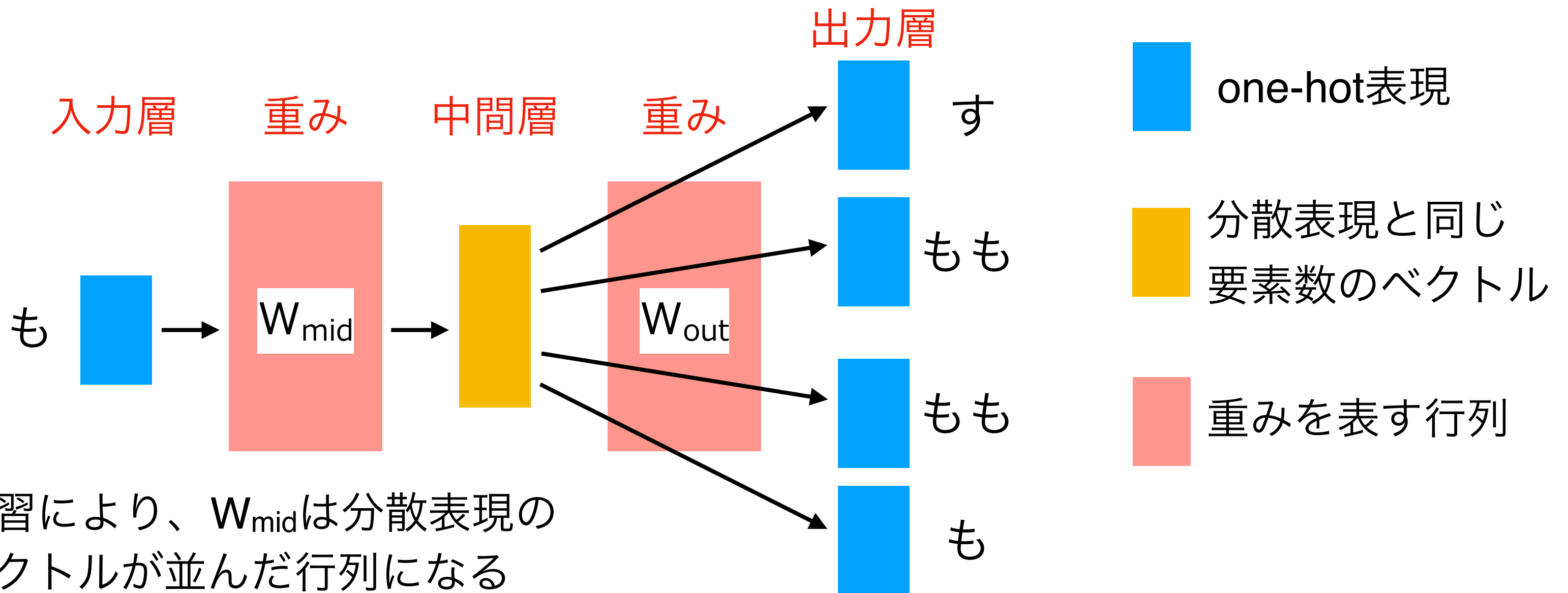
- 前後の単語から対象の単語を予測するニューラルネットワーク
- 学習に要する時間がskip-gramよりも短い





# skip-gram

- ある単語から、前後の単語を予測するニューラルネットワーク
- CBOWよりも学習に時間がかかるが、精度がよい



# 日本語の形態素解析

- 形態素とは、言葉が意味を持つまとまりの単語の最小単位のこと
- 形態素解析とは、自然言語を形態素にまで分割すること
- 日本語や中国語、タイ語は単語間にスペースが無いので、  
形態素解析が必要
- 以下は代表的な日本語の形態素解析ライブラリ
  - MeCab → 知名度が高く、高速、高精度
  - Janome → 速度はMeCabに劣るが、導入が簡単
  - etc...



# 分散表現の実装



# 分散表現の実装

- 01\_word\_vector.ipynb



# テキストデータの前処理



# FastText

- FastTextは、2016年にFacebookによって発表された  
Word2Vecの発展形
- 高速、高精度、なおかつ利用が簡単
- 自然言語処理においてビッグデータが非常に扱いやすくなる

<https://fasttext.cc/>



# テキストデータの前処理


- 02\_preprocessing.ipynb

# 次回の内容

Section 1. コースの概要とTwitter API

Section 2. RNNとSeq2Seq

Section 3. 自然言語処理の基礎

 **Section 4. モデルの訓練**

Section 5. Attentionの導入

Section 6. Twitterボットのデプロイ