OLYMPUS

INTERNAL USE ONLY





word2vecの紹介

自然言語処理のための単語のベクトル化

100 YEARS
Endeavor for Better

科学ソフト開発 | 秋山光弘 | 2019年12月11日

DeepLearningによる画像処理については、実務に耐えるものがいくつもあるようですが、自然言語処理は、まだ実務に耐えるものはないと思っています。

近い将来、実務に耐えるものが出てくるかもしれないので、その前に現状の技術の基礎を学んでおきたいと考えました。

OLYMPUS

INTERNAL USE ONLY

01 参考書籍紹介

02 word2vecとは

03 word2vecの処理手順

04 実装、処理の結果

01 参考書籍紹介

参考書籍紹介

ゼロから作るDeep Learning ❷——自然言語処理編

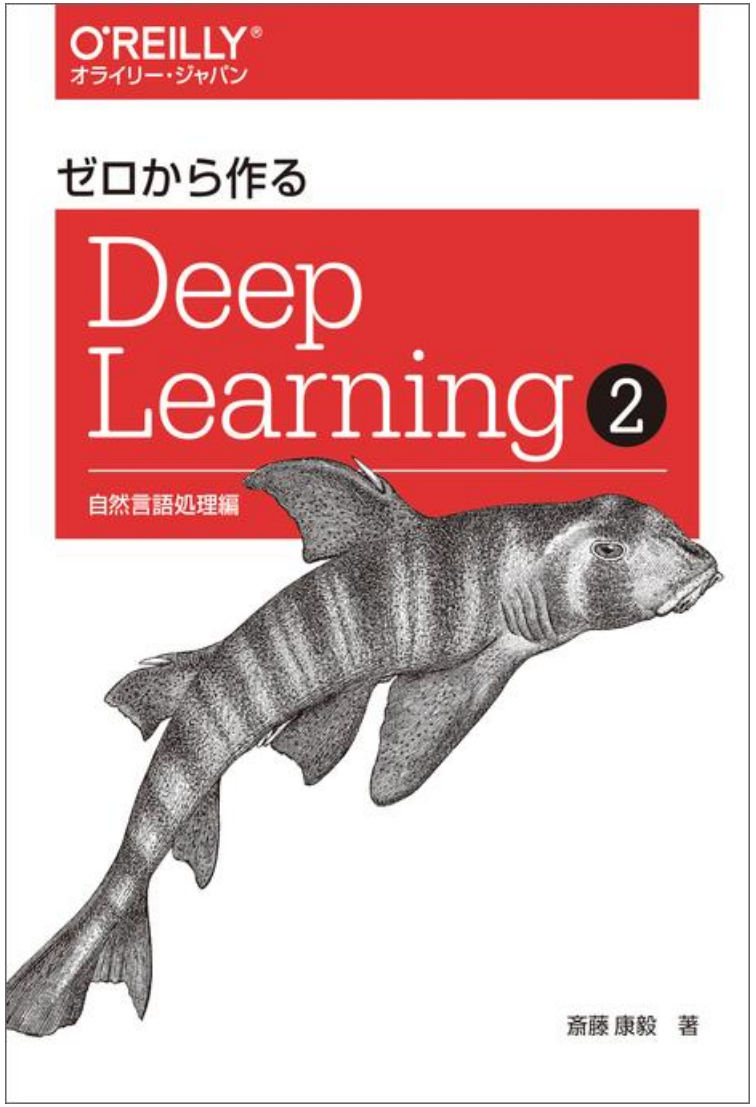
https://www.oreilly.co.jp/books/9784873118369/

大きく以下の二つに分かれています。

(1) 単語の処理

(2) 文の処理

今回は前半の単語処理で紹介されているword2vecについて発表します。



02 word2vecとは

word2vecとは

指定されたテキストデータに含まれる各単語に対し、指定された次元数のベクトル値を設定するもの。

単語を入力すると、周辺の単語を出力するというNNを学習させると、NN内に単語ごとのベクトル値が出来上がるというもの。

テキストデータを、単語ベクトルのリストに変換することで、”文”をNNで処理できるようになる。(らしい)

“Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:”…

単語	1	2	3	4	5
“female”	2.6	7.2	9.4	-5.1	3.7
:					
“king”	1.2	5.5	1.5	-8.2	1.5
:					
“male”	1.9	5.1	1.0	-7.5	5.1
:					
“queen”	1.8	7.9	9.4	-6.2	-0.1
:					

“king” – “male” + “female” = (1.9, 7.6, 9.9 -5.8, 0.1)

03 word2vecの処理手順

word2vecの処理手順

1. 入力データの準備

   巨大なテキストデータを用意する。wikipediaが使われたりしているらしい。

2. 単語ベクトルの次元数を決める

   適当な次元数をどう決めればよいのかは不明。

   本家のword2vecはデフォルトが200次元とのこと。

3. “ウィンドウサイズ”を決める

   word2vecでは、単語とその前後の単語を使用して学習を行うが、前後の単語数をウィンドウサイズという。

   本家のword2vecはデフォルトが5とのこと。

4. 学習データの作成

5. 学習

6. 学習済みネットワークから、単語ベクトルを取り出す

学習データの作成 - (1)テキストデータを単語リストに変換

テキストデータ

“Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:”

単語のリスト

単語
“Alice”
“was”
“beginning”
“to”
“get”
“very”
“tired”
(以下省略)

学習データの作成 - (2)単語リストをVocabularyとCorpusに変換

単語のリスト

単語
“Alice”
“was”
“beginning”
“to”
“get”
“very”
“tired”
(以下省略)

Vocabulary

ID	単語
0	“,”
1	“.”
2	“Alice”
3	“and”
4	“bank”
5	“beginning”
6	“by”
	(以下省略)

Corpus

ID
2
20
5
17
8
19
16
(以下省略)

学習データの作成 - (3)Corpusを学習データに変換

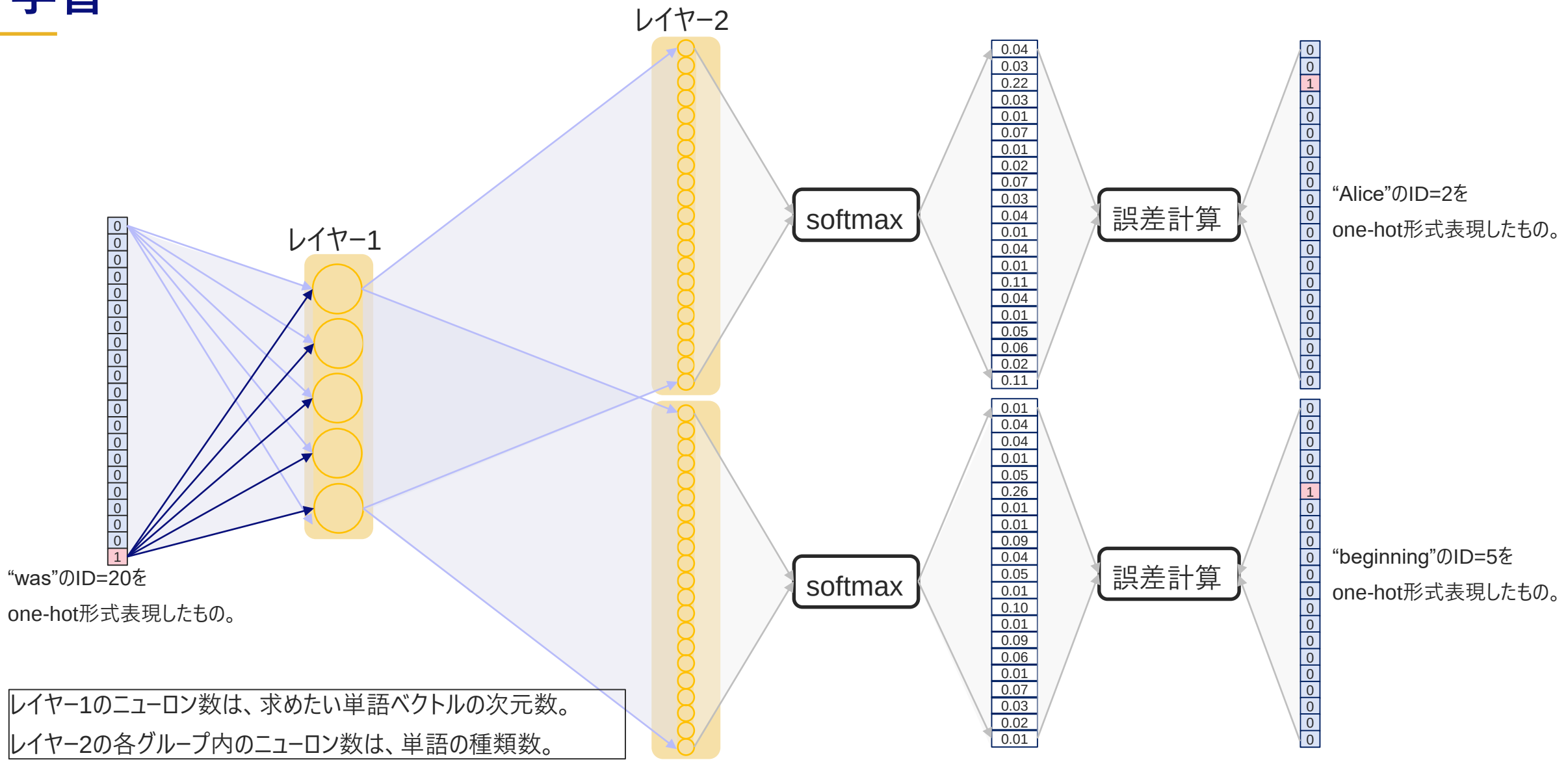
Corpus

ID
2
20
5
17
8
19
16
(以下省略)

学習データ

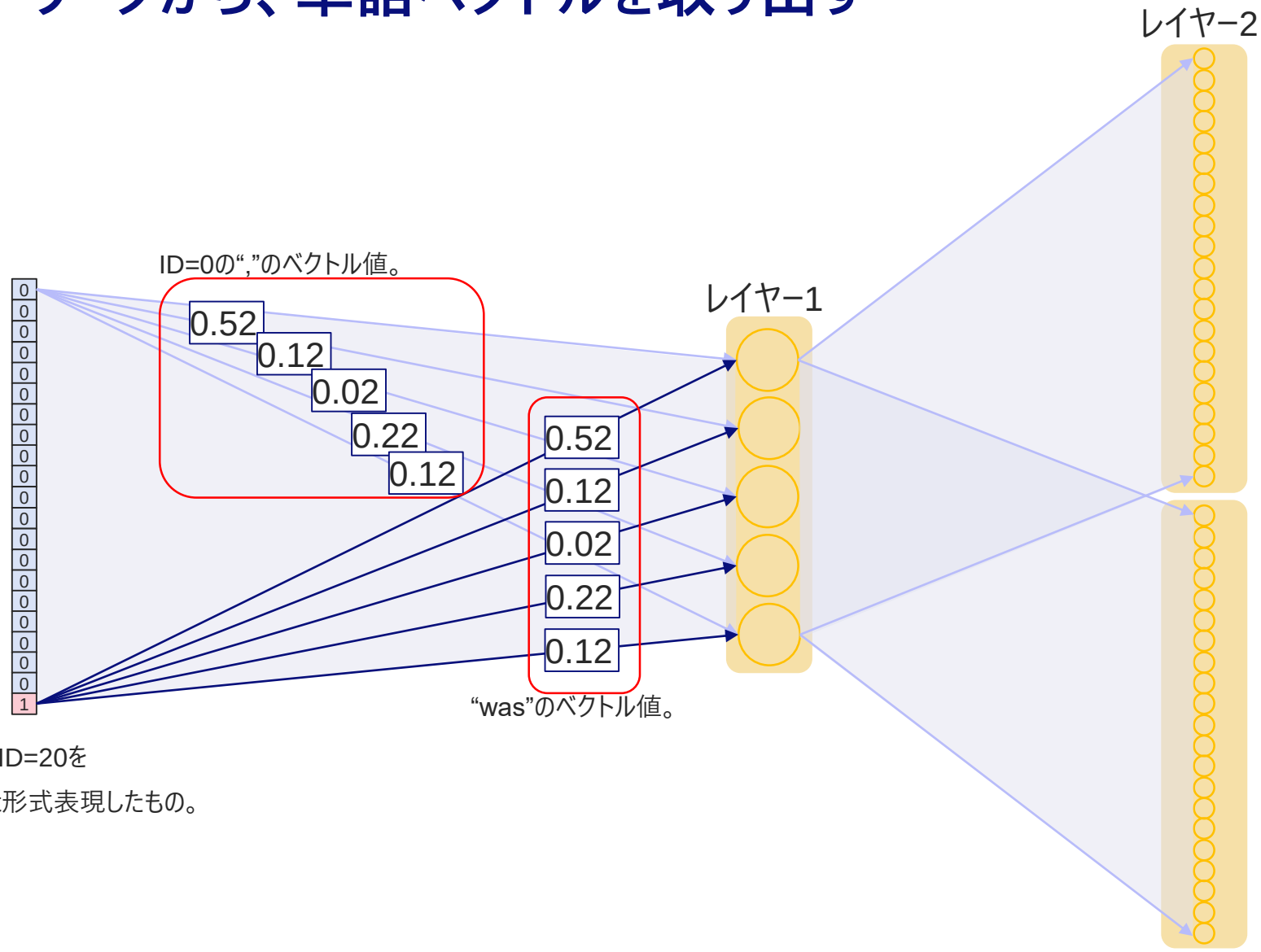
ID	前の単語のID	後の単語のID
20	2	5
5	20	17
17	5	8
8	17	19
19	8	16
16	19	12
12	16	15
	(以下省略)	

学習



レイヤー1のニューロン数は、求めたい単語ベクトルの次元数。
レイヤー2の各グループ内のニューロン数は、単語の種類数。

ネットワークから、単語ベクトルを取り出す



“was”のID=20を

one-hot形式表現したもの。

04 実装、処理の結果

実装、処理の結果(1)

1. 入力データの準備

   今回は以下のテキストデータを使用した。

   “Alice's Adventures in Wonderland” (http://www.gutenberg.org/files/11/11-0.txt)

   (単語数 = 38,972、単語の種類数 = 3,110)

2. 単語ベクトルの次元数を決める

   本家word2vecは200次元とのことだが、今回は20次元としてみた。

3. “ウィンドウサイズ”を決める

   本家のword2vecはデフォルトが5とのだが、今回は実装を単純にするため、1とした。

4. 学習データの作成

5. 学習

6. 学習済みネットワークから、単語ベクトルを取り出す

   4.～5.については、Kotlinで実装を行い、処理を行った。(バッチサイズは300、エポック数は1000とした。)

実装、処理の結果(2)

以下のような単語ベクトルは得られたが、「king – male + female ≒ queen」のようなものは見つけられなかった。

alice	-1.158	-0.626	-1.494	1.183	-2.193	0.426	-0.398	-2.837	-0.281	-0.029	1.636	-0.970	-0.373	1.070	-1.201	0.411	-1.809	0.339	0.173	0.387
king	-2.290	-0.767	-1.089	1.761	-0.133	0.099	-0.717	-0.290	1.115	0.790	-2.192	1.312	1.171	0.312	-1.229	-0.463	-3.162	1.086	0.451	-1.616
queen	0.292	-0.312	-0.382	2.785	0.358	-0.198	0.143	-0.039	0.517	0.068	-2.867	1.789	1.161	0.438	-1.976	-2.161	-4.515	0.064	0.637	-1.340


OLYMPUS