自然言語処理
Details of the assignment

竹内孔一

Text Classification: Author Estimation

Data

Three novels written by different authors.

texts are from Aozora bunko
https://www.aozora.gr.jp/

class label	author	title	translation
0	a) 芥川龍之介 Ryunosuke Akutagawa	アグニの神	(Aguni-no-kami./God of Aguni)
1	e) 江戸川乱歩 Rambo Edogawa	押絵と旅する男	(Oshie-to tabi-suru otoko/Man who trips with a raised cloth picture)
2	m) 森鴎外 Ogai Mori	鼠坂	(Nezumizaka/Mouse hill)

Data Format (train/test.csv)

author tag, sentence of a novel

a,婆さんはどこからとり出したか、眼をつぶった妙子の顔の先へ、一挺のナイフを突きつけました。
e,私は仕方がないので母親に貰ったお小遣いをふんぱつして、人力車に乗りました。
m,小川君は好奇心が起って溜まらなくなった。
a,一そ警察へ訴えようか？
a,イツモダト私ハ知ラズ知ラズ、気ガ遠クナッテシマウノデスガ、今夜ハソウナラナイ内ニ、ワザト魔法ニカカッタ真似ヲシマス。

Training data and test data

Splitting training data and test data

test data is only used in evaluationの時だけ使う

Don‘t use the test data to training parameters.

training data
697 sentences

test data 175 sentences

Contents of the file

■ How to decompress novel.zip

```
unzip  novel.zip
```

■Files and directories

```
novel
 ├── data
 │    ├── id2wd.txt
 │    ├── test.csv
 │    ├── test.feature
 │    ├── train.csv
 │    └── train.feature
 └── prog
      ├── data_get.py
      └── layer2_Bow_kr.py
```

Description for each file

```
novel
 ├── data
 │    ├── id2wd.txt
 │    ├── test.csv
 │    ├── test.feature
 │    ├── train.csv
 │    └── train.feature
 └── prog
      ├── data_get.py
      └── layer2_Bow_kr.py
```

train.csv & test.csv are original text data

e,私は仕方がないので母親に貰ったお小遣いをふんぱつして、人力車に乗りました。
m,小川君は好奇心が起って溜まらなくなった。
a,一そ警察へ訴えようか？
. .

train.feature & test.feature are vector data.

class label

Bag of words vector. ➔See 9 page

```
1 2675:1 2348:1 965:1 2853:1 1123:1 2404:1 375:1 1032:1 1322:1 631:1 2813:1 2332:1
  2725:1 1574:1 197:1 2632:1 1714:1 1882:1 1032:1 2203:1 832:1 631:1 1994:1
2 833:1 377:1 2348:1 2382:1 1692:1 2853:1 1409:1 2632:1 2654:1 1123:1 2862:1 631:1
  1994:1
0 1749:1 881:1 96:1 2194:1 740:1 867:1 2336:1 1452:1
```

axis number (i.e., word ID)

occurrence of word (all of words are 1)

Description for each file

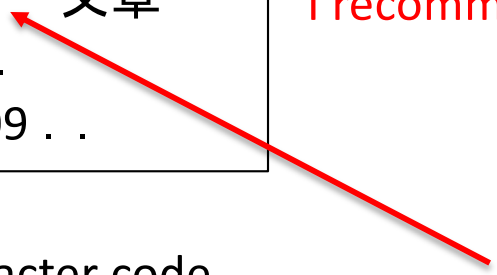
```
novel
 ├── data
 │    ├── id2wd.txt
 │    ├── test.csv
 │    ├── test.feature
 │    ├── train.csv
 │    └── train.feature
 └── prog
      ├── data_get.py
      └── layer2_Bow_kr.py
```

id2wd.txt: mapping between id and word

0	無視
1	紅
2	緋
3	浮上
4	風
5	ガラス
6	隣
7	抗
8	ずつ
9	食わせる
10	文章
. . .	
3009	. .

total 3010 types of word

Causion. This file is made in Linux. The line feed is LF.
But windows uses CRLF, then if you see this file in Windows, you cannot see this file correctly.
Take care to apply python to this files in Windows.
I recommend to use this in linux.



The delimiter is tab

Character code is UTF-8

How to use the sample classifier, 3-layer neural network in Keras.

■Execution

```
$ cd ~/novel/prog
$ python layer2_Bow_kr.py
```

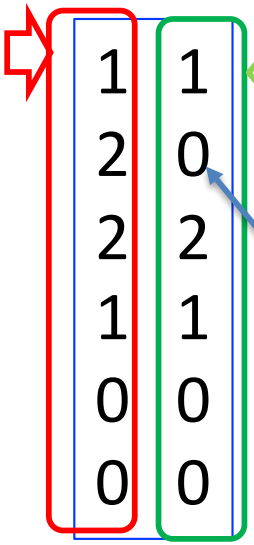
■Results are printed out in stdout

```
2,2,厭 だ 。」
2,2,「 なかなか 別品 だ た わ ねえ 。
1,1,『 何故 です 』って 尋ねる て も 、『 まあ いい から 、そうして お 呉れる な 』と 申す て 聞く ない の だ ござる ます 。
2,0,もう おしまい に なる た じゃ ない か 。
2,2,翌朝 深淵 の 家 へ は 医者 が 来る たり 、警部 や 巡査 が 来る たり する て 、非常 に 雑 ※（「 二 点 しんにょう ＋ 鰥 の つくる 」、第 4 水準 2 - 89 - 93 ）する た 。
1,1,数 年 以前 から 、いつも あんな 苦しい 相 だ 顔 を する て 居る ます 。
0,0,五
0,0,そうして これ が 出来る ない ば 、勿論 二度と お父さん の 所 へ も 、帰れる ない なる の に 違い ある ます ん 。
accuracy 0.9257143139839172
```

Format of the results

■Classification results are at the second column

The most left column is the correct class labels
0: Akutagawa
1: Edogawa
2: Mori



here is estimated class labels.

accuracy= 0.926

The classifier missed this label.

This indicates that the accuracy of tag for test data is 92.6%

Accuracy = number of correctly estimated tags / number of all tags(=175)

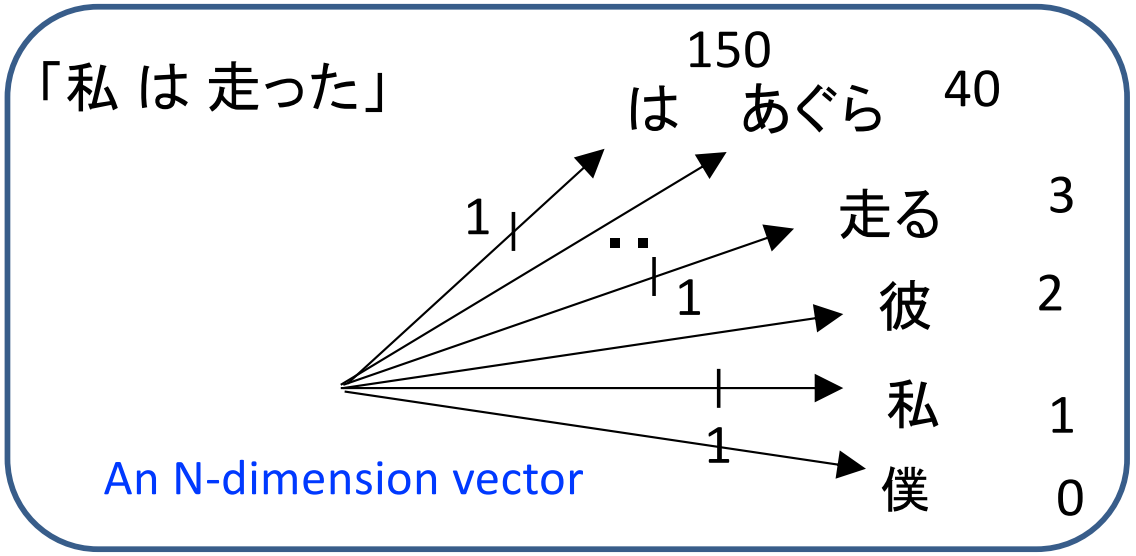
Bag of words: a vector of a sentence

1. Assign unique ids to all words (in this case, from 0 to 3009)
2. Each word id indicates an axis of the vocabulary space (3010)
3. A sentence vector is made in the vocabulary space in which words in a sentence are 1 and the others are 0

word id list

The ids are axis number.

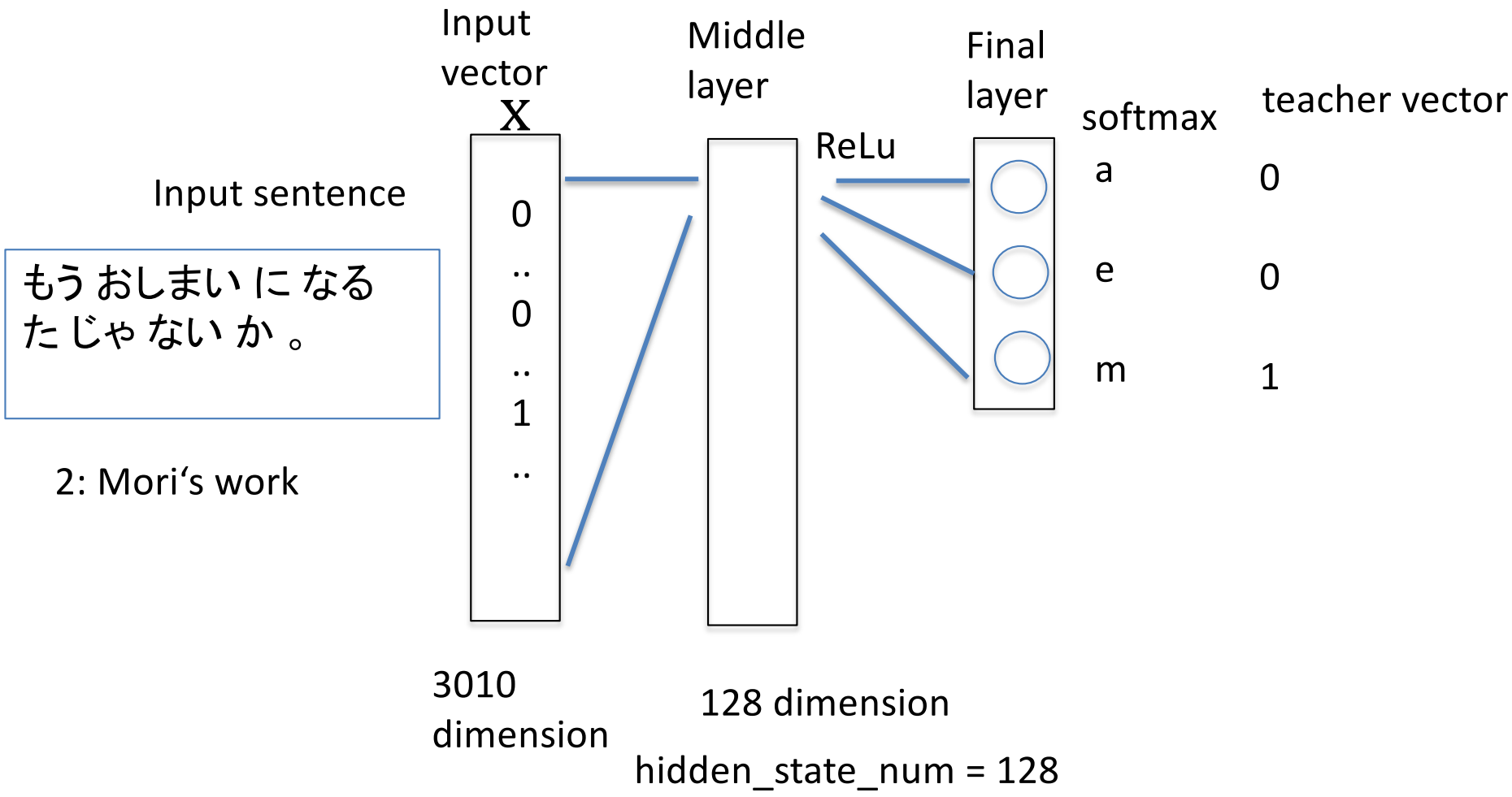
0: 僕
1: 私
2: 彼
3: 走る
...
N: <unk>



「私は走った」➔「私 / は / 走る」➔ { 1:1 3:1 150:1} {Axis_number: 1, ...}

Feature: A sentence vector is a fixed dimensions (i.e., vocabulary).
The information of word order in a sentence is not taken into account.

The 3-layer neural network of the sample program



Sample contents of your report

- In Introduction
  - Explain the task of author estimation and propose your model. Explain the differences between your model and the sampled model, ideas and perspectives
- Method/Approach
  - Explain details of your model
- Experiments
  - Show the accuracies between the sample model and your models for test data
- Discussions
  - Discuss the reliability of the results, advantage and disadvantage of the proposed models, goods and bads of the results etc.
- References
  - Add references at the final part

Notes

- About your model
  - Don’t think too hard. Most simple modification is to change the number of units in the middle layer.
  - It would be not easy to overcome the sample program. Don’t be annoyed in the case. Discussions of the results are the main aim of this report.
  - You don’t need to apply many models.