8.1

Working With Text Data

Part 2: The Bag-Of-Words Model

Sebastian Raschka and the Lightning AI Team

Raw text data

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.

Preprocessed text data

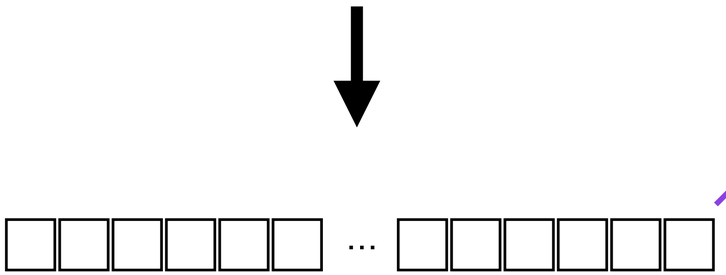
(e.g., strip HTML)

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.

Feature vector

Machine learning model

Text



Feature vector

1. Bag-of-words model
2. Embedding layer

Machine learning model

The Bag-Of-Words (BoW) Model

Suppose we are building a classifier to identify funny sentences

An example training dataset

Text	Label
The ghost pepper is so spicy, it is hauntingly hot	1
I tried to hug the sun today, but it was too hot to handle	1
I cannot handle spicy food	0

Step 1: create a vocabulary

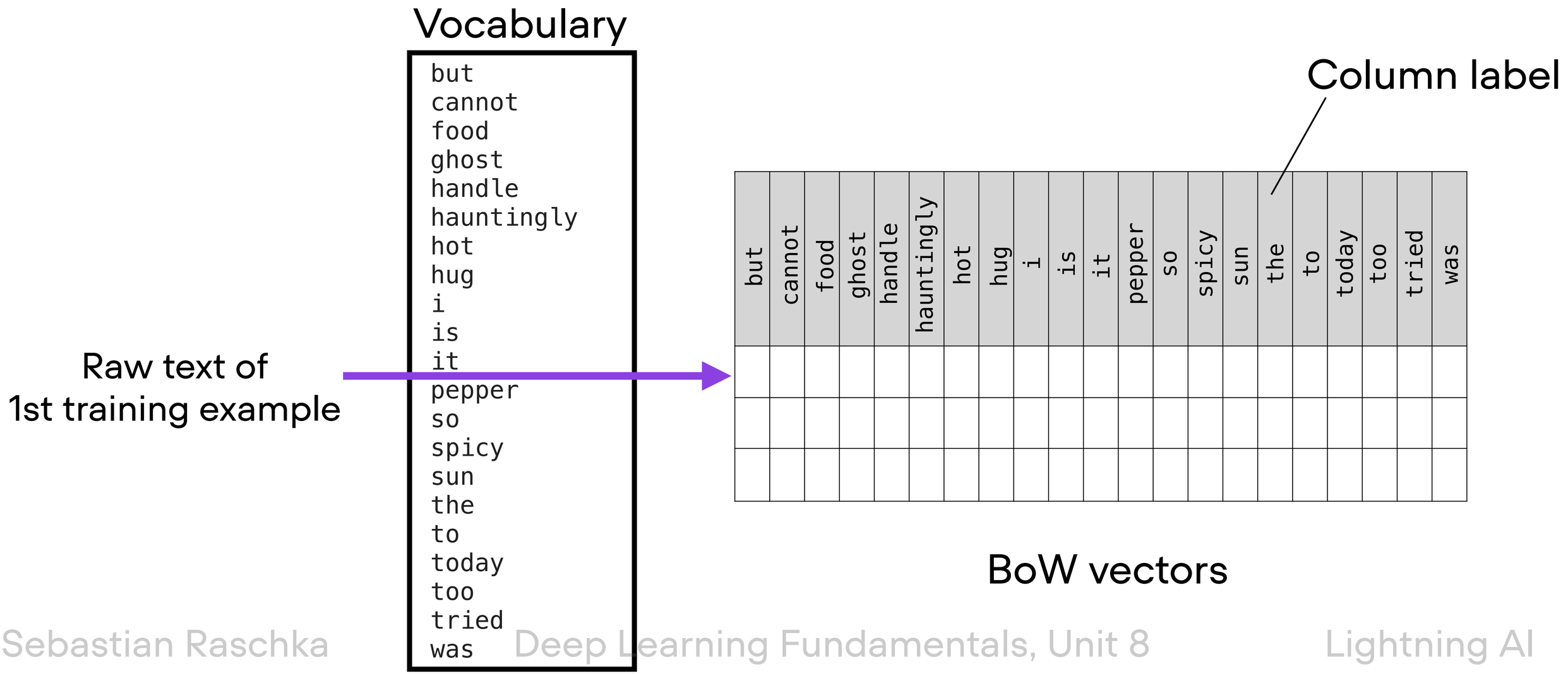
Text
The ghost pepper is so spicy, it is hauntingly hot
I tried to hug the sun today, but it was too hot to handle
I cannot handle spicy food

Extract all unique words

Vocabulary

```
but
cannot
food
ghost
handle
hauntingly
hot
hug
i
is
it
pepper
so
spicy
sun
the
to
today
too
tried
was
```

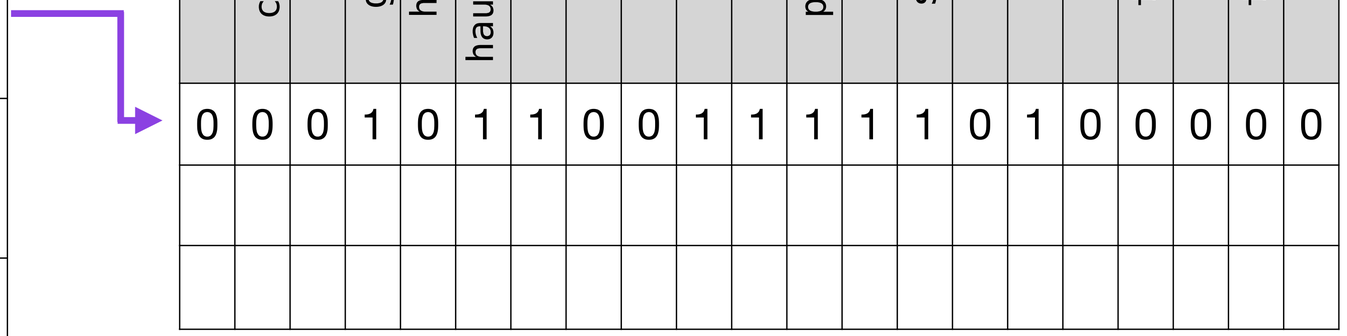
Step 2: create BoW count vectors



The size of the vocabulary determines the number of columns (features)

Step 2: create BoW count vectors

Text
The ghost pepper is so spicy, it is hauntingly hot
I tried to hug the sun today, but it was too hot to handle
I cannot handle spicy food



but	cannot	food	ghost	handle	hauntingly	hot	hug	i	is	it	pepper	so	spicy	sun	the	to	today	too	tried	was
0	0	0	1	0	1	1	0	0	1	1	1	1	1	0	1	0	0	0	0	0

BoW vectors

Step 2: create BoW count vectors

Text
The ghost pepper is so spicy, it is hauntingly hot
I tried to hug the sun today, but it was too hot to handle
I cannot handle spicy food

but	cannot	food	ghost	handle	hauntingly	hot	hug	i	is	it	pepper	so	spicy	sun	the	to	today	too	tried	was
0	0	0	1	0	1	1	0	0	1	1	1	1	1	0	1	0	0	0	0	0
1	0	0	0	1	0	1	1	1	0	1	0	0	0	1	1	2	1	1	1	1

BoW vectors

Step 2: create BoW count vectors

Text
The ghost pepper is so spicy, it is hauntingly hot
I tried to hug the sun today, but it was too hot to handle
I cannot handle spicy food

but	cannot	food	ghost	handle	hauntingly	hot	hug	i	is	it	pepper	so	spicy	sun	the	to	today	too	tried	was
0	0	0	1	0	1	1	0	0	1	1	1	1	1	0	1	0	0	0	0	0
1	0	0	0	1	0	1	1	1	0	1	0	0	0	1	1	2	1	1	1	1
0	1	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0

BoW vectors

Step 3: train machine learning model

but	cannot	food	ghost	handle	hauntingly	hot	hug	i	is	it	pepper	so	spicy	sun	the	to	today	too	tried	was
0	0	0	1	0	1	1	0	0	1	1	1	1	1	0	1	0	0	0	0	0
1	0	0	0	1	0	1	1	1	0	1	0	0	0	1	1	2	1	1	1	1
0	1	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0

Feature vectors

Machine learning model

Modified BoW count vectors

Modified BoW count vectors

N-gram models

“The”
“ghost”
“pepper”
“is”
“so”
“spicy”
...

1-gram

“The ghost”
“ghost pepper”
“pepper is”
“is so”
“so spicy”
...

2-gram
(bigram)

Modified BoW count vectors

tf–idf

(term frequency–inverse document frequency)

How often the word occurs weighted by the number of documents it occurs in

- high term frequency: important word?
- high document frequency: not very informative

Next: Training a BoW-based classifier