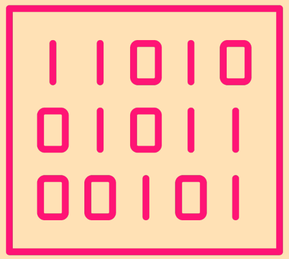
Using Word Embeddings

Axel Sirota

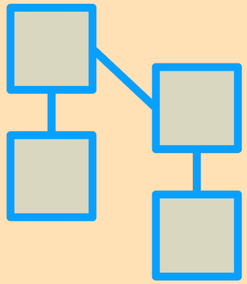
AI and Cloud Consultant

@AxelSirota

Reviews Dataset



Dataset

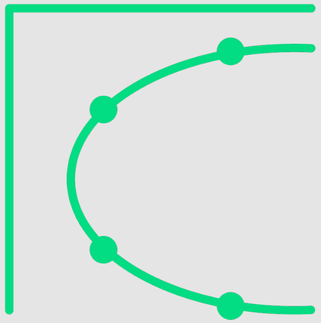


Model

Stars

Training

I had a fabulous time

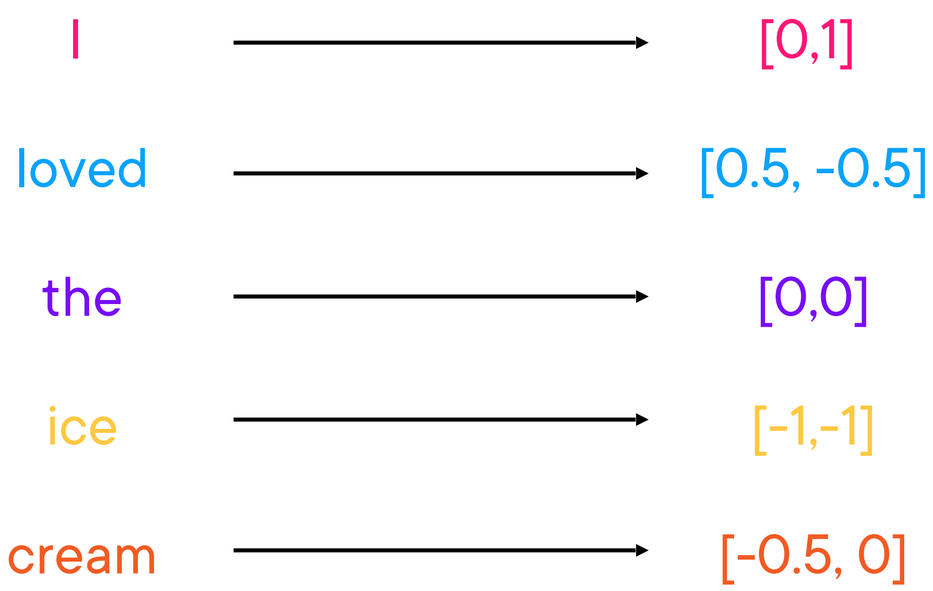


Trained Model

5 stars

Evaluation

A Model Is Just a Function



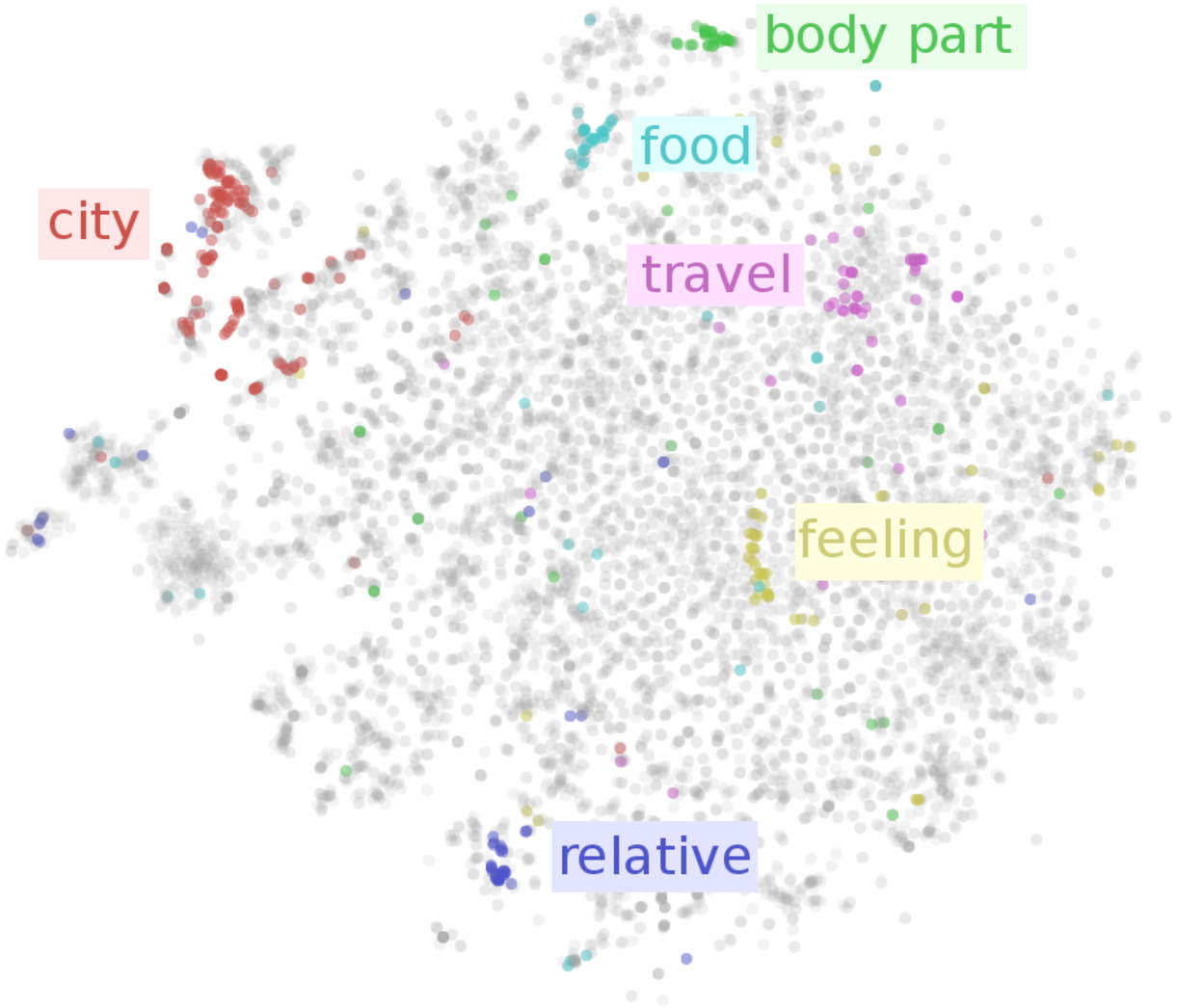
A Model Is Just a Function

“I loved the ice cream” →

0	1
0.5	-0.5
0	0
-1	-1
-0.5	0

This is known as an embedding

Representing Word Clusters



Now the output of the embedding will be the input to the model and let it train

First Embedding: One Hot Encoding

Representing a Sentence

1. Construct vocabulary and assign different columns for presence

I loved the ice cream and loved the food

Vocabulary = {I, loved, the, ice, cream, and, food}

Representing a Sentence

I loved the ice cream and loved the food

1. Construct vocabulary and assign different columns for presence

2. Count

I	loved	the	ice	cream	and	food
1	0	0	0	0	0	0
0	**2**	0	0	0	0	0
0	0	**2**	0	0	0	0
0	0	0	**1**	0	0	0
0	0	0	0	**1**	0	0
0	0	0	0	0	**1**	0
0	0	0	0	0	0	**1**

Representing a Sentence

I loved the ice cream and loved the food

1. Construct vocabulary and assign different columns for presence

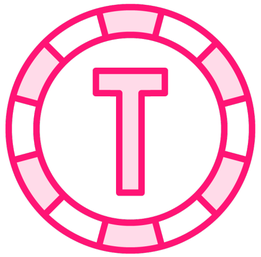
2. Count

3. Collapse into a vector

Final result: [1,2,2,1,1,1,1]

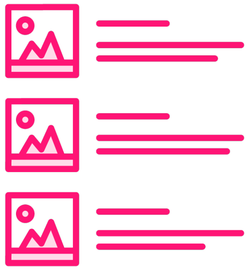
Analysing Sentiment with OHE

Some Definitions



Vocabulary: set of tokens we care about

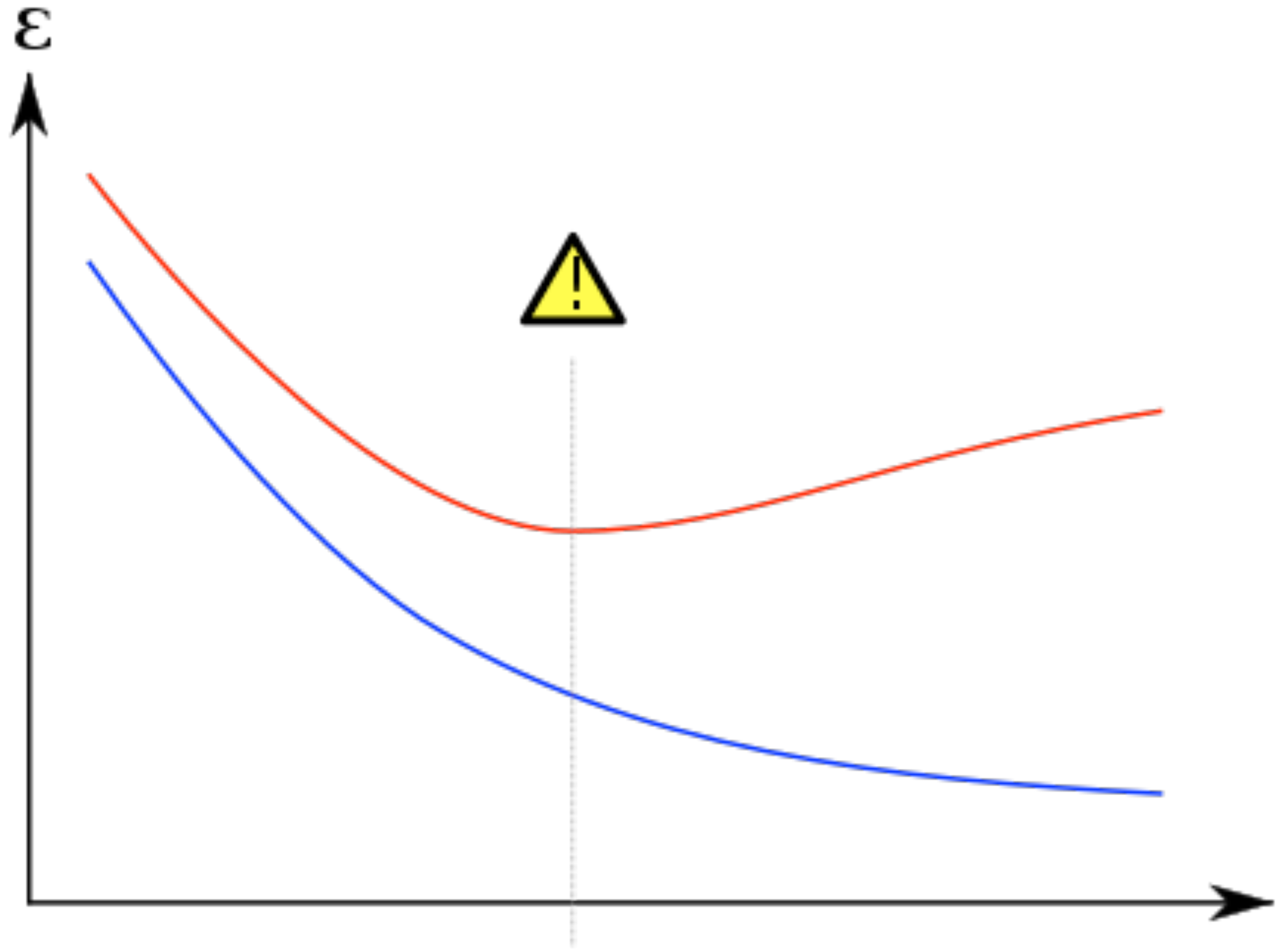
Document: set of tokens

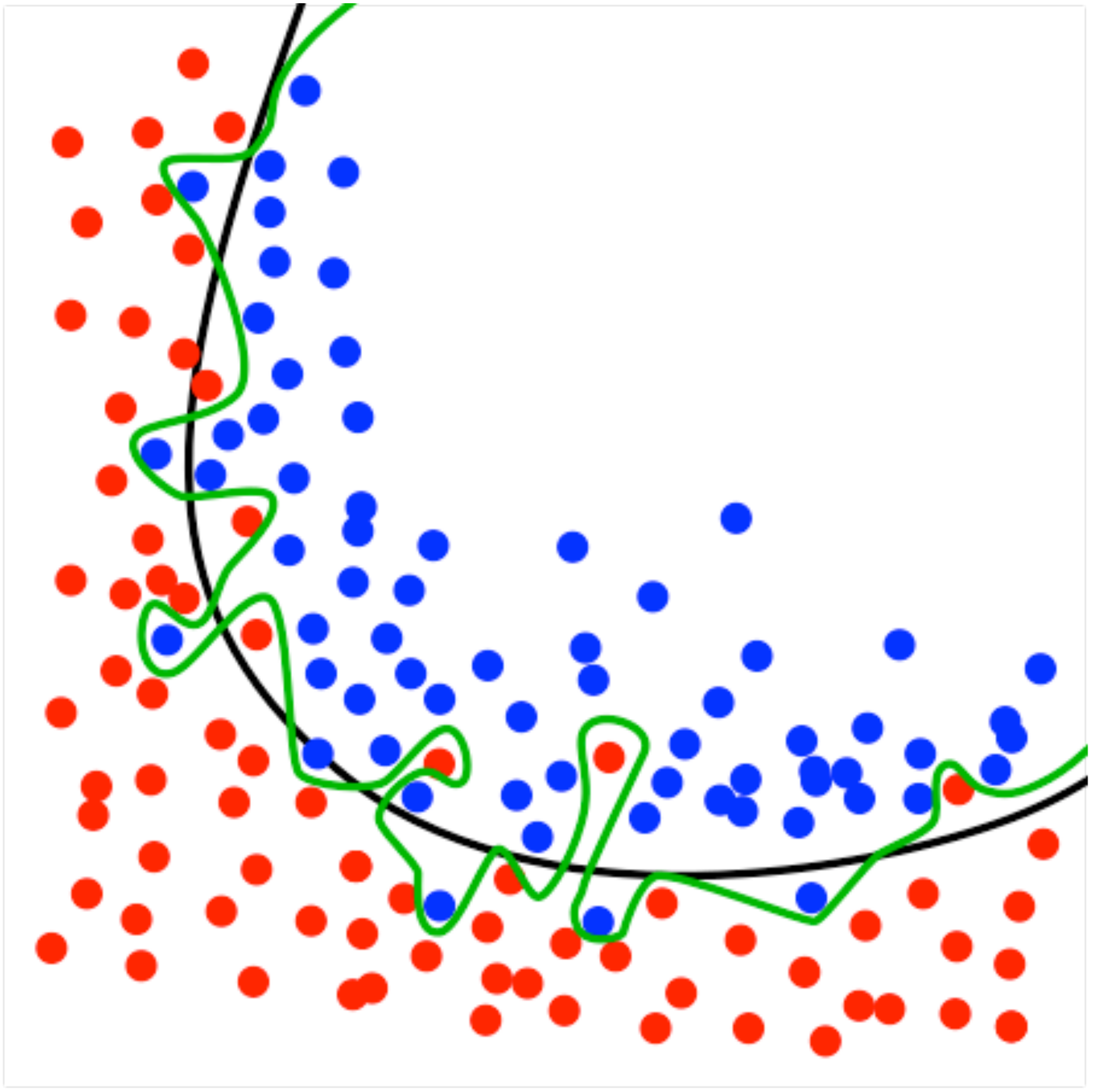


Corpus: collection of documents

A Model Is Just a Function

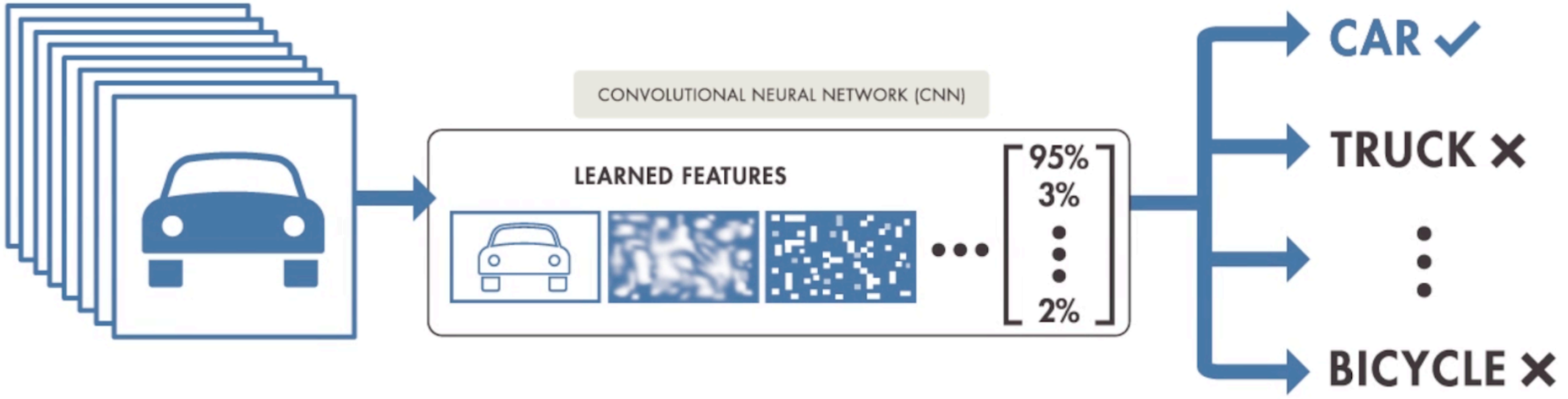
		13	1		13	1
[“I loved the		441	32		441	32
ice cream”,	→	54	989	→	54	989
“It is not what it used to be”]		34	45		34	45
		67	1		67	1
		73	453		73	453
		NN	22		0	22
		NN	9		0	9


ε

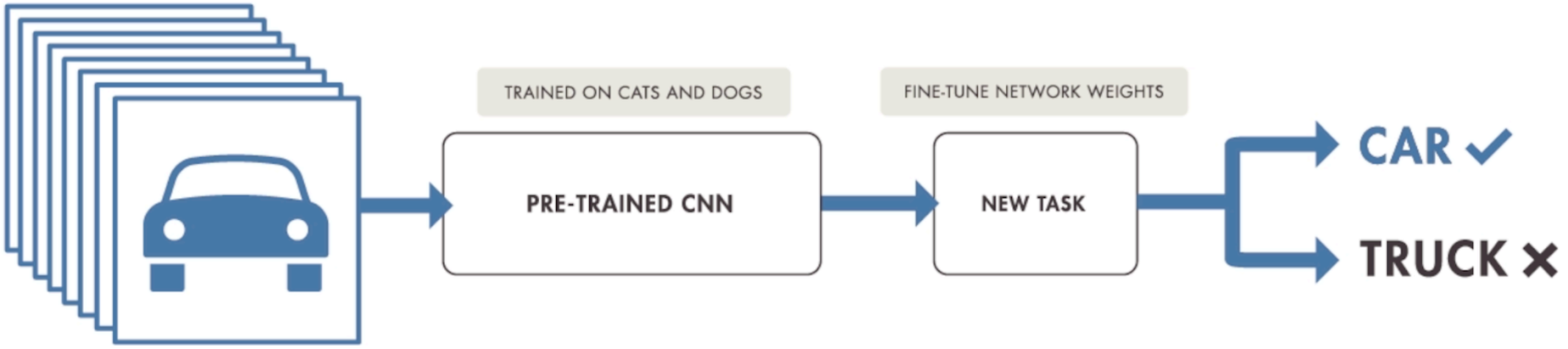


Reminder: Transfer Learning

TRAINING FROM SCRATCH



TRANSFER LEARNING



In text problems the same phenomenon occurs, therefore we will leverage pretrained embeddings to improve performance

Reanalyse sentiment with GloVe

Preprocessing Functions

Lowercase tokens

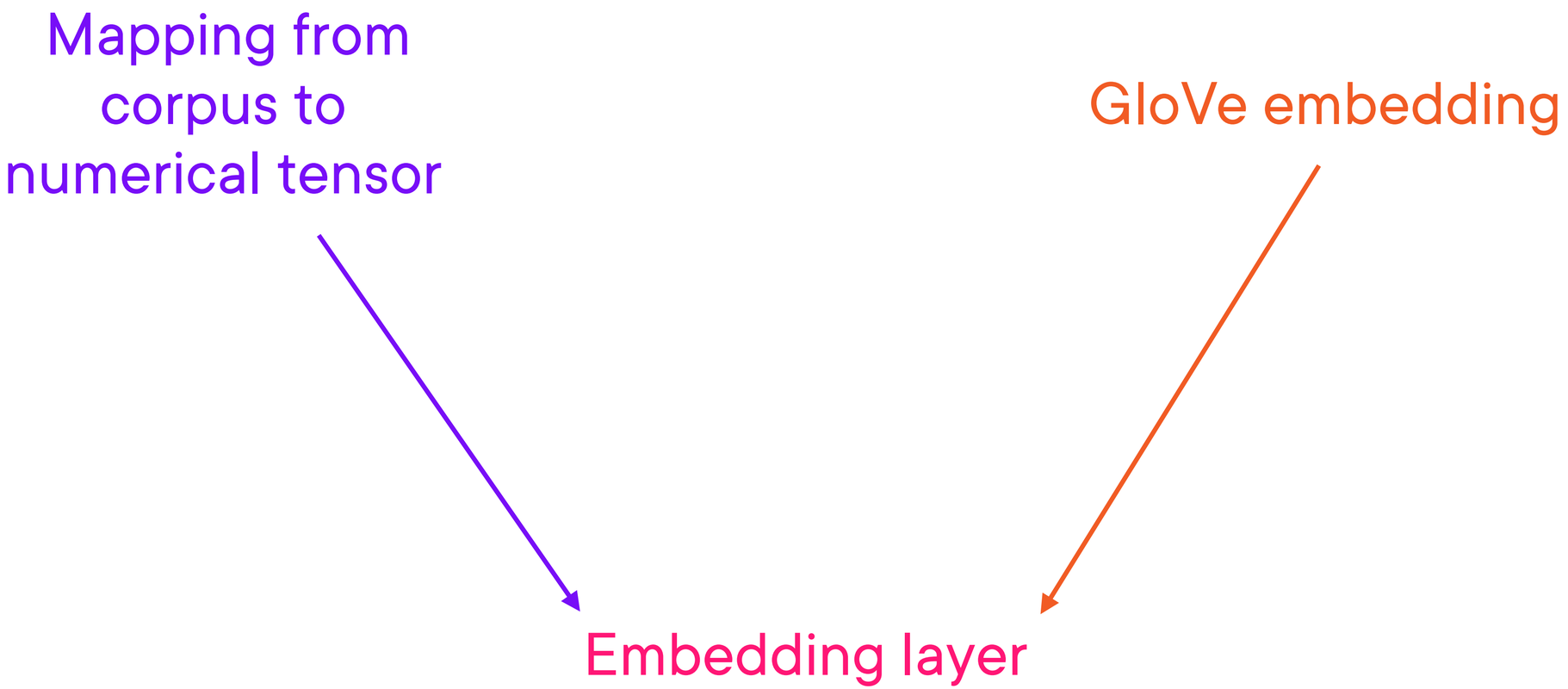
Eliminate punctuations

Handle special characters

Preprocessing Functions

“I’d loved this place, if only they served Tacos!!!”

“i d loved this place if only they served tacos”


Mapping from corpus to numerical tensor
GloVe embedding
Embedding layer

The Tensorflow Embedding Layer

This embedding would map the number 2 (which stands for “the”) to the following representation into the network:

[1, 4, 3, 2, 5, 4, 1, 0, -1, -3]

```
model = Sequential()
model.add(Embedding(input_dim=100, output_dim=10, input_length=40))
```

An Example

1ST

What we should do is go to the embedding and search for our word, in this case, “trust”

2ND

We would go to the initialiser tensor which has sizes (1000, 100)

3RD

On row 200 (which is index 199 in python) we would replace the whole row with the embedding for that word, which is exactly 100 dimensional!

```
embedding_tensor[199] = embedding['trust']
```

Takeaways

An embedding is a mapping representation from text tokens into a numerical form

One hot encoding is the simplest embedding but it doesn’t collate the information about word closeness, as well as it has high dimensionality

We can build word embeddings with deep learning, and examples of those are CBOW and Skip gram word2Vec models

One can input OHE representation and learn the embedding layer into any task or one can use a pre-trained layer to improve any task in NLP we need to do

Keys

Practice creating the input Tensor for a given text corpus! This step is fundamental

Try to ensure you understand the dimensions of the tensors and why they make sense at every step

Practice using another Glove embedding and try to make it work!

Text classification