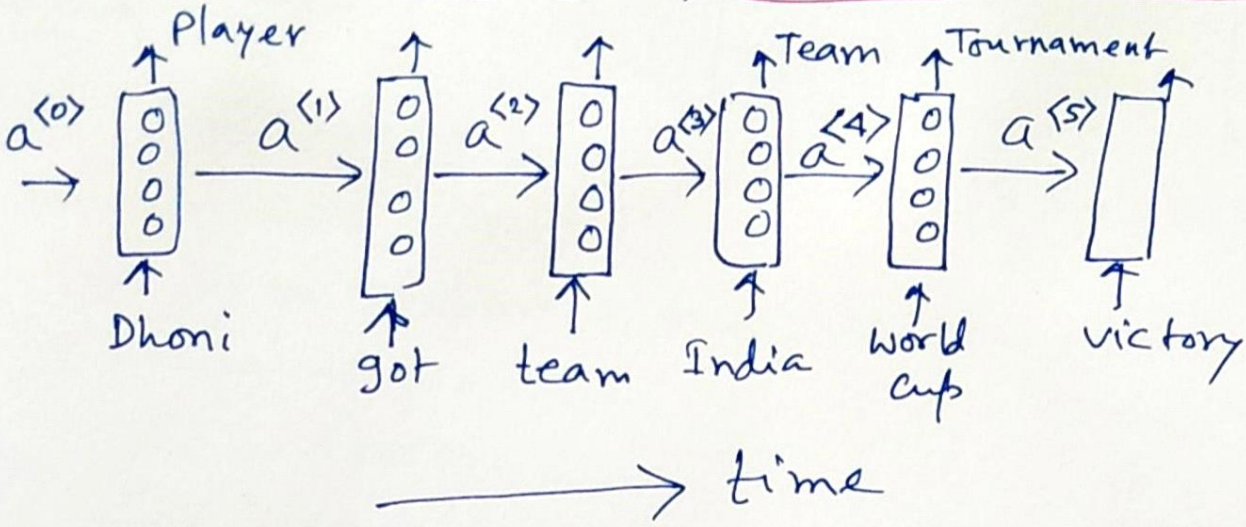
Word Embedding

Suppose we want to build a NLP model that can reeognise entitites in text:

In 2011, Dhoni (Player) got team India (Team name) a world cup (Tournament) victory.

In the last Ashes (Tournament) game, Cummins (Player) led Australia (Team name) to a victory.



PROB: The machine can't understand text

→ We have to convert them to numbers.

For training, we can scrap the internet.

→ Make a vocabulary : list of words.

Ashes	1
Axar	2
bat	3
Dhoni	7
team	1023
:	
Zimbabwe	50,000

Tokenization

Now we can put "7" in place of "Dhoni" to the RNN.

Prob: These numbers are random. They don't capture relationship between words.

Second option : One hot encoding

Probs: – (i) They dont capture relationship between words

(ii) Computationally inefficient.

Word embedding

	Ashes	Axar	Bat	...	Dhoni		Zimbabwe
Ashes	1	0	0	...	0		0
Axar	0	1	0	...	0		0
⋮				⋮			
Zimbabwe	0	0	0	...	0		1

Word embedding: How to capture similarities between two words?

How do we compare similarities between two homes?

By comparing features:

(i) Bedrooms
(ii) Area
(ii) # of bathrooms etc.

	Your	Mine	Palace
	4	3	20
	1850	1500	10,000
	3	2	7

Hand-crafted features:

	Dhoni	Cummins	Australia
Peson	1	1	1
Healthy/Fit	0.9	0.87	0.7
Location:	0	0	1
Has two eyes.	1	1	0
Has governmen	0	0	1

vectors

This ⊛ is called word-embedding

Automatic features

	Ashes	Australia	Cummins
Person	0	0.02	0.95
country	0	0.97	0
Healthy/Fit	0	0	0.87
event	1	0.1	0
gear	0	0	0

↑ obtained through training.