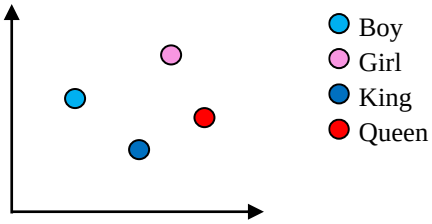
Text Recognition

"Today we will study recurrent neural networks."

Tokenization - breaking up a sequence of strings into pieces such as words, keywords, phrases.

		'One-Hot-Code'		"vector representation"
		|------#5000------|		
Today	1	[1, 0, 0,……., 0]		[0.1; -0.7; 2.1; 0; 1.3; 1.2; 0.8]
We	2	[0, 1, 0,……., 0]	Embedding	[- - -]
Will	3	[0, 0, 1,…..., 0]	⇨	
Study	4	[0, 0, 0, 1,…., 0]	(dimension: 8)	
…..				
End	5000	[0, 0, 0,……., 1]		



This vector representation works well in categorizing these words for example.

"Today we will study recurrent neural networks." ==> [1, 2, 3, 4, 5, 6, 7]

Ex. Samples 10000 ➔ 10,000 rows; each row has 100 integers

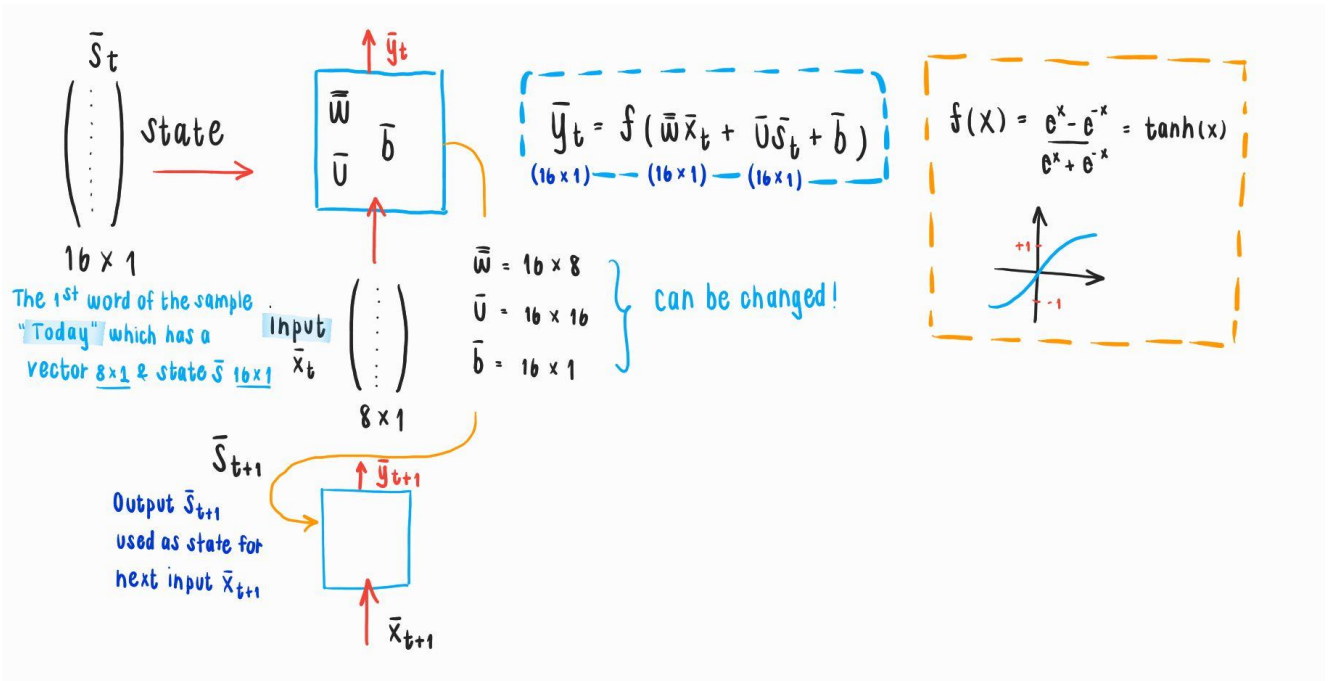
Sample length 100

so, Train = (10000, 100)

Embedded = (10000, 100, 8) = samples, length, dimensions of vector

Train [0, 3] = 4 < index 3 >

Embedded [0, 0, 2] = 2.1 < Take 1st sample / Take 1st word / index 2 >


$\bar{s}_t$
state
16 × 1
$\bar{y}_t$
$\bar{W}$ $\bar{b}$ $\bar{U}$
$\bar{y}_t = f(\bar{W}\bar{x}_t + \bar{U}\bar{s}_t + \bar{b})$
(16×1) (16×1) (16×1)
$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x)$
+1
-1
The 1st word of the sample "Today" which has a vector 8×1 & state $\bar{s}$ 16×1
Input $\bar{x}_t$
8 × 1
$\bar{W}$ = 16 × 8
$\bar{U}$ = 16 × 16
$\bar{b}$ = 16 × 1
can be changed!
$\bar{s}_{t+1}$
$\bar{y}_{t+1}$
Output $\bar{s}_{t+1}$ used as state for next input $\bar{x}_{t+1}$
$\bar{x}_{t+1}$