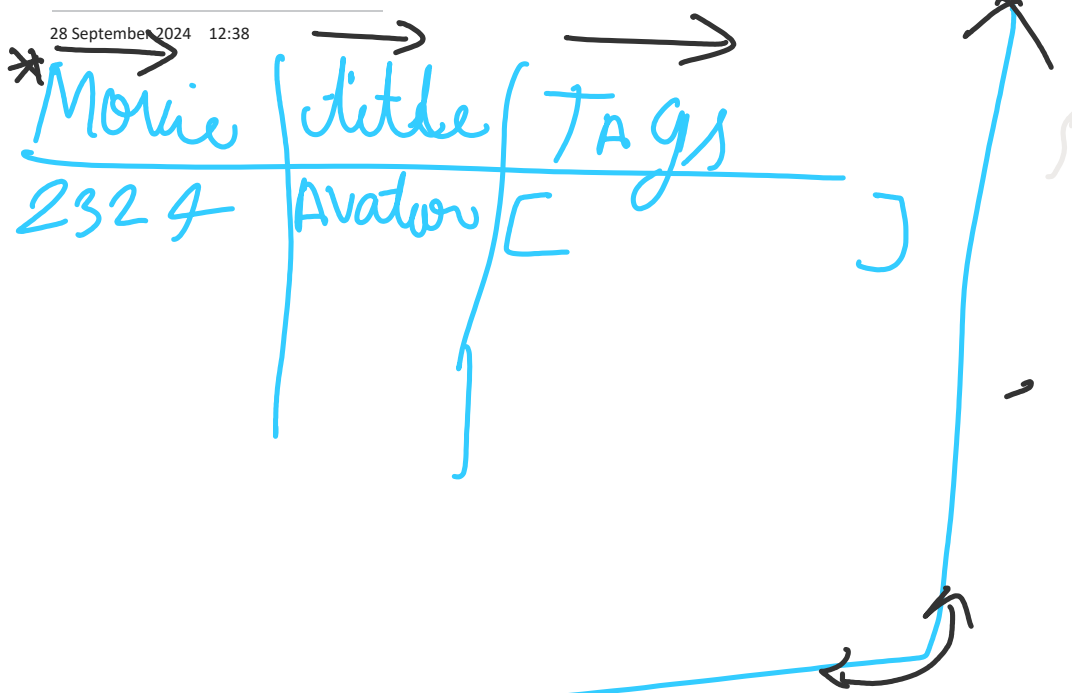


## Vectorization

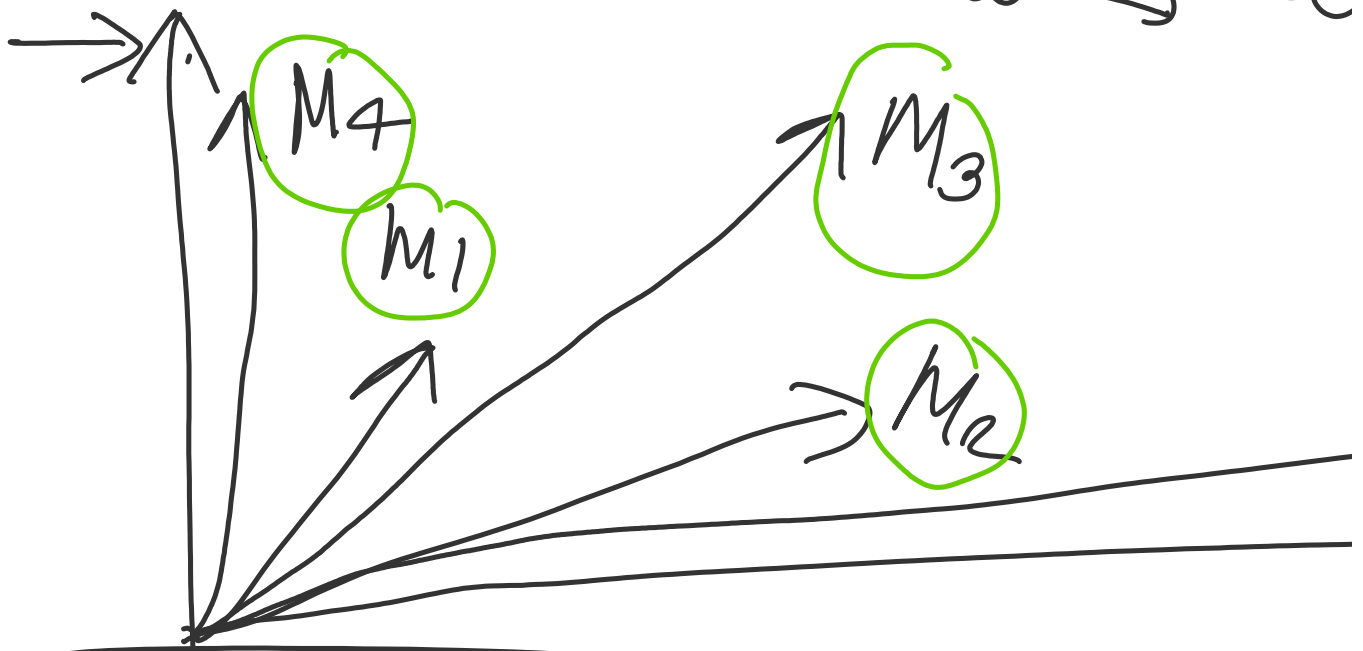
28 September 2024 12:38



→ Now the Point is: Movie → Tags:

Problem: → Movie → Can you  
 movies to this movie:

& for each movie like →  $l$



# Top 5000 words from a

	$W_1$	$W_2$	$W_3$	$W_4$	$W_i$
$M_1$	4 times	3 times	0 times	( )	
$M_2$					
$M_3$					
$M_4$					
$\vdots$					

$(0,0) \rightarrow$  number of times  $W_i$  occurred in words need to remove  $\rightarrow a, an, the, at$

We have list features from Vectorization  
But we have words with same word  $\rightarrow$  actor, actors, love, loving

$\rightarrow$  we use Stemming

love

$\rightarrow$  Stemming converts words into its base form

for  $M$ Monies Word/feature  $\rightarrow$ 

	$w_1$	$w_2$	$w_3$	$w_4$
$m_1$	0	1	0	2
$m_2$	2	1	0	0