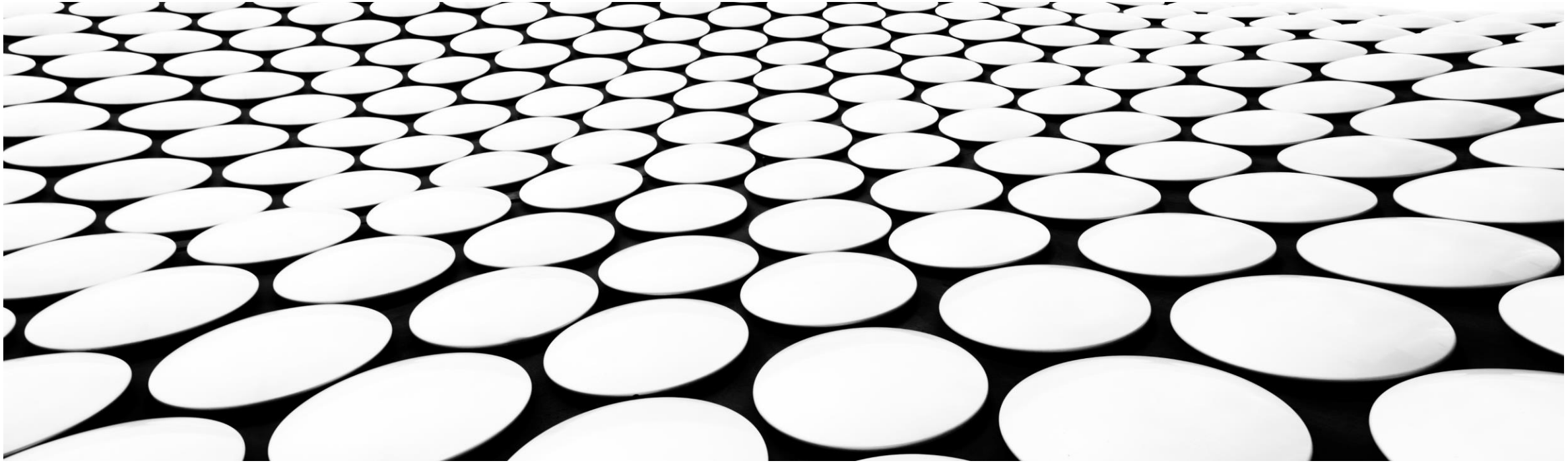
PROCESSAMENTO DE LINGUAGEM NATURAL:

BAG OF WORDS, TF-IDF & WORD EMBEDDING



INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ
60 ANOS

POR QUE TRATAR LINGUAGEM NATURAL É IMPORTANTE?

- Produção de informação não estruturada, escrita e *falada*

- Tarefas automatizadas de classificação, *clusterização* etc.

- Aplicações: *search engines*, *sentiment analysis*, classificação automática de notícias, emails, artigos, reclamações, *bots*, reconhecimento de voz, tradução automática etc.

POR QUE REPRESENTAMOS DOCUMENTOS POR VETORES?

• Precisamos de uma representação estruturada para manipular e processar as informações

• **Luhn** argumentou que as palavras muito frequentes e as pouco frequentes não colaboram para discriminação e similaridade entre documentos.

• Frequencia dos termos = sentido dos documentos

• Fácil manipulação = funções de similaridade

PRINCIPAIS FORMAS DE REPRESENTAÇÃO

- **BOW bag of words**
- **TF IDF term frequency–inverse document frequency**
- **Word Embedding (2013, Tomas Mikolov at Google)**
- **Bert (2018, Bidirectional Encoder Representations from Transformers, Google)**

https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html

BOW

- Um documento é representado por um vetor com a frequencia dos termos
- Simples
- Alta dimensionalidade
- Não leva em consideração a frequencia de termos total da coleção

BOW

d1 Human machine interface for ABC computer application.
d2 A survey of user opinion of computer system response time.
d3 The EPS user interface management system.
d4 System and human system engineering testing in EPS.
d5 Relation to user perceived response time to error measurement.
d6 The generation of random, binary, ordered trees.
d7 The intersection graph of paths in trees.
d8 Graph minors IV: Widths of trees and well-quasi-ordering.
d9 Graph minors: A survey.

terms in at least two documents

		1	2	3	4	5	6	7	8	9	10	11	12
	tf(i,j)	system	user	graph	trees	response	EPS	nterface	human	survey	computer	minors	time
1	d1	0	0	0	0	0	0	1	1	0	1	0	0
2	d2	1	1	0	0	1	0	0	0	1	1	0	1
3	d3	1	1	0	0	0	1	1	0	0	0	0	0
4	d4	2	0	0	0	0	1	0	1	0	0	0	0
5	d5	0	1	0	0	1	0	0	0	0	0	0	1
6	d6	0	0	0	1	0	0	0	0	0	0	0	0
7	d7	0	0	1	1	0	0	0	0	0	0	0	0
8	d8	0	0	1	1	0	0	0	0	0	0	1	0
9	d9	0	0	1	0	0	0	0	0	1	0	1	0
	df(j)	3	3	3	3	2	2	2	2	2	2	2	2
	idf(j)	1,10	1,10	1,10	1,10	1,50	1,50	1,50	1,50	1,50	1,50	1,50	1,50

inverted index		d2	d2	d7	d6	d2	d3	d1	d1	d2	d1	d8	d2
		d3	d3	d8	d7	d5	d4	d3	d4	d9	d2	d9	d5
		d4	d5	d9	d8								

useful for boolean searches

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ 60 ANOS

from Landauer "Introduction to Latent Semantic Analysis"

BOW

all terms

	tf(i,j)	a	ABC	and	application	binary	computer	engineering	EPS	error	for	generation	graph	human	in	interface	IV	machine	management	measurement	minors	of	opinion	ordered	paths	Perceived	random	relation	response	survey	system	testing	the	time	to	trees	user	well-quasi-ordering	widths
1	d1		1		1		1				1			1		1		1																					
2	d2	1					1															2	1							1	1	1			1			1		
3	d3								1					1		1			1												1							1		
4	d4			1				1	1						1															2	1				1				
5	d5									1											1				1		1		1					1	2		1			
6	d6					1						1										1		1			1						1			1				
7	d7												1		1							1			1								1				1			
8	d8			1									1				1				1	1															1		1	1
9	d9	1											1								1										1									
	df(j)	2	1	2	1	1	2	1	2	1	1	1	3	2	2	2	1	1	1	1	2	5	1	1	1	1	1	1	2	2	4	1	2	2	3	3	3	1	1
	idf(j)	1,50	2,20	1,50	2,20	2,20	1,50	2,20	1,50	2,20	2,20	2,20	1,10	1,50	1,50	1,50	2,20	2,20	2,20	2,20	1,50	0,59	2,20	2,20	2,20	2,20	2,20	2,20	1,50	1,50	0,81	2,20	1,50	1,50	1,10	1,10	1,10	2,20	2,20

from Landauer "Introduction to Latent Semantic Analysis"

BOW

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

terms in at least two documents

		1	2	3	4	5	6	7	8	9	10	11	12
	tf(i,j)	system	user	graph	trees	response	EPS	interface	human	survey	computer	minors	time
1	d1	0	0	0	0	0	0	1	1	0	1	0	0
2	d2	1	1	0	0	1	0	0	0	1	1	0	1
3	**d3**	**1**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**
4	**d4**	**2**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**
5	d5	0	1	0	0	1	0	0	0	0	0	0	1
6	d6	0	0	0	1	0	0	0	0	0	0	0	0
7	d7	0	0	1	1	0	0	0	0	0	0	0	0
8	d8	0	0	1	1	0	0	0	0	0	0	1	0
9	d9	0	0	1	0	0	0	0	0	1	0	1	0
	df(j)	3	3	3	3	2	2	2	2	2	2	2	2
	idf(j)	1,10	1,10	1,10	1,10	1,50	1,50	1,50	1,50	1,50	1,50	1,50	1,50

igualmente pode ser aplicado ao TF IDF

|d3| = 2,0
|d4| = 2,4
sim(d3, d4) = 0,6
angle(d3,d4) = 0,9 radians
52,2 degrees

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ
60 ANOS

from Landauer "Introduction to Latent Semantic Analysis"

TF IDF

- **Um documento é representado por um vetor com a frequencia dos termos *combinado com o inverso da frequencia do termo na coleção***
- **Ainda Simples**
- **Alta dimensionalidade**
- **Não leva em consideração o contexto dos termos**

TF IDF

		stop words:	notice natural log, any log will do words in only one document			
d1	Human machine interface for ABC computer application.	for	machine	ABC	application	
d2	A survey of user opinion of computer system response time.	A, of	opinion			
d3	The EPS user interface management system.	The	management			
d4	System and human system engineering testing in EPS.	and, in	engineering	testing		
d5	Relation to user perceived response time to error measurement.	to	Relation	perceived	error	measurement
d6	The generation of random, binary, ordered trees.		generation	random	binary	ordered
d7	The intersection graph of paths in trees.		intersection	paths		
d8	Graph minors IV: Widths of trees and well-quasi-ordering.		IV	Width	well-quasi-ordering	
d9	Graph minors: A survey.					

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

terms in at least two documents

		1	2	3	4	5	6	7	8	9	10	11	12
	tf(i,j)	system	user	graph	trees	response	EPS	interface	human	survey	computer	minors	time
1	d1	0	0	0	0	0	0	1	1	0	1	0	0
2	d2	1	1	0	0	1	0	0	0	1	1	0	1
3	d3	1	1	0	0	0	1	1	0	0	0	0	0
4	d4	2	0	0	0	0	1	0	1	0	0	0	0
5	d5	0	1	0	0	1	0	0	0	0	0	0	1
6	d6	0	0	0	1	0	0	0	0	0	0	0	0
7	d7	0	0	1	1	0	0	0	0	0	0	0	0
8	d8	0	0	1	1	0	0	0	0	0	0	1	0
9	d9	0	0	1	0	0	0	0	0	1	0	1	0
	df(j)	3	3	3	3	2	2	2	2	2	2	2	2
	idf(j)	1,10	1,10	1,10	1,10	1,50	1,50	1,50	1,50	1,50	1,50	1,50	1,50

tf*idf(j)

1	2	3	4	5	6	7	8	9	10	11	12
system	user	graph	trees	respons	EPS	interface	human	survey	comput	minors	time
0,00	0,00	0,00	0,00	0,00	0,00	1,50	1,50	0,00	1,50	0,00	0,00
1,10	1,10	0,00	0,00	1,50	0,00	0,00	0,00	1,50	1,50	0,00	1,50
1,10	1,10	0,00	0,00	0,00	1,50	1,50	0,00	0,00	0,00	0,00	0,00
2,20	0,00	0,00	0,00	0,00	1,50	0,00	1,50	0,00	0,00	0,00	0,00
0,00	1,10	0,00	0,00	1,50	0,00	0,00	0,00	0,00	0,00	0,00	1,50
0,00	0,00	0,00	1,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	1,10	1,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	1,10	1,10	0,00	0,00	0,00	0,00	0,00	0,00	1,50	0,00
0,00	0,00	1,10	0,00	0,00	0,00	0,00	0,00	1,50	0,00	1,50	0,00

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

INSTITUTO MAUÁ DE TECNOLOGIA

MAUÁ 60 ANOS

from Landauer "Introduction to Latent Semantic Analysis"

O QUE PODEMOS FAZER COM ISSO?

	tf*idf(j) 1	2	3	4	5	6	7	8	9	10	11	12	
	system	user	graph	trees	response	EPS	interface	human	survey	computer	minors	time	Type
d1	0,00	0,00	0,00	0,00	0,00	0,00	1,50	1,50	0,00	1,50	0,00	0,00	User
d2	1,10	1,10	0,00	0,00	1,50	0,00	0,00	0,00	1,50	1,50	0,00	1,50	User
d3	1,10	1,10	0,00	0,00	0,00	1,50	1,50	0,00	0,00	0,00	0,00	0,00	User
d4	2,20	0,00	0,00	0,00	0,00	1,50	0,00	1,50	0,00	0,00	0,00	0,00	User
d5	0,00	1,10	0,00	0,00	1,50	0,00	0,00	0,00	0,00	0,00	0,00	1,50	User
d6	0,00	0,00	0,00	1,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	Tech
d7	0,00	0,00	1,10	1,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	Tech
d8	0,00	0,00	1,10	1,10	0,00	0,00	0,00	0,00	0,00	0,00	1,50	0,00	Tech
d9	0,00	0,00	1,10	0,00	0,00	0,00	0,00	0,00	1,50	0,00	1,50	0,00	Tech
doc_new	2,20	0,00	0,00	0,00	0,00	1,50	0,00	1,50	0,00	0,00	0,00	0,00	?

- Classificação
- Clusterização
- All Machine Learning

O QUE PODEMOS FAZER COM ISSO?

- Classificação
- Clusterização
- All Machine Learning

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ
60 ANOS

POR QUE PRECISAMOS DE UM OUTRO MODELO?

- **Problemas úteis vetores de dimensão de 50K ☹**
- **Word Embedding, Word2vec 50, 100, ... 500 max ☺**

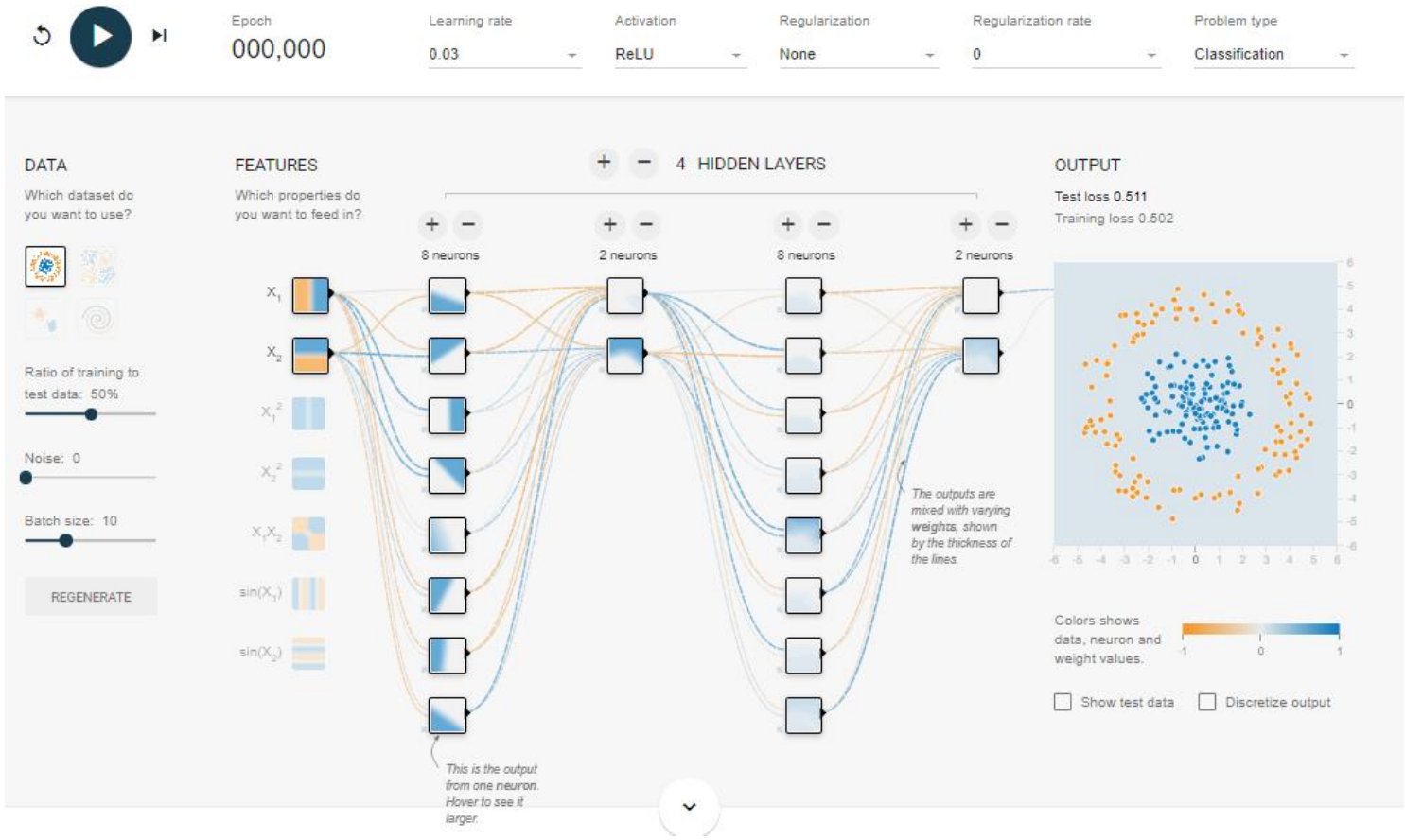
WORD EMBEDDING

- **Word Embedding, Word2vec 50, 100, ... 500 max ☺**
- **Contexto dos termos N-grams**
- **EMBEDDING É APENAS UMA TÉCNICA DE FAZER UMA REPRESENTAÇÃO DE ALGO PRESERVANDO AS ESTRUTURAS (INFORMAÇÕES) DE INTERESSE**
- **DOCUMENTOS → NEURAL NETWORK → VECTOR DOC**

N-GRAMS

1Gram	Nem	tudo	é	claro	na	vida	*Machado de Assis*
2Gram	Nem	tudo					
2Gram		Nem	tudo				
2Gram			Nem	tudo			
2Gram				Nem	tudo		
2Gram					Nem	tudo	
2Gram						Nem	
2Gram							Nem
3Gram	Nem	tudo	é				
3Gram		Nem	tudo	é			
3Gram			Nem	tudo	é		
3Gram				Nem	tudo	é	

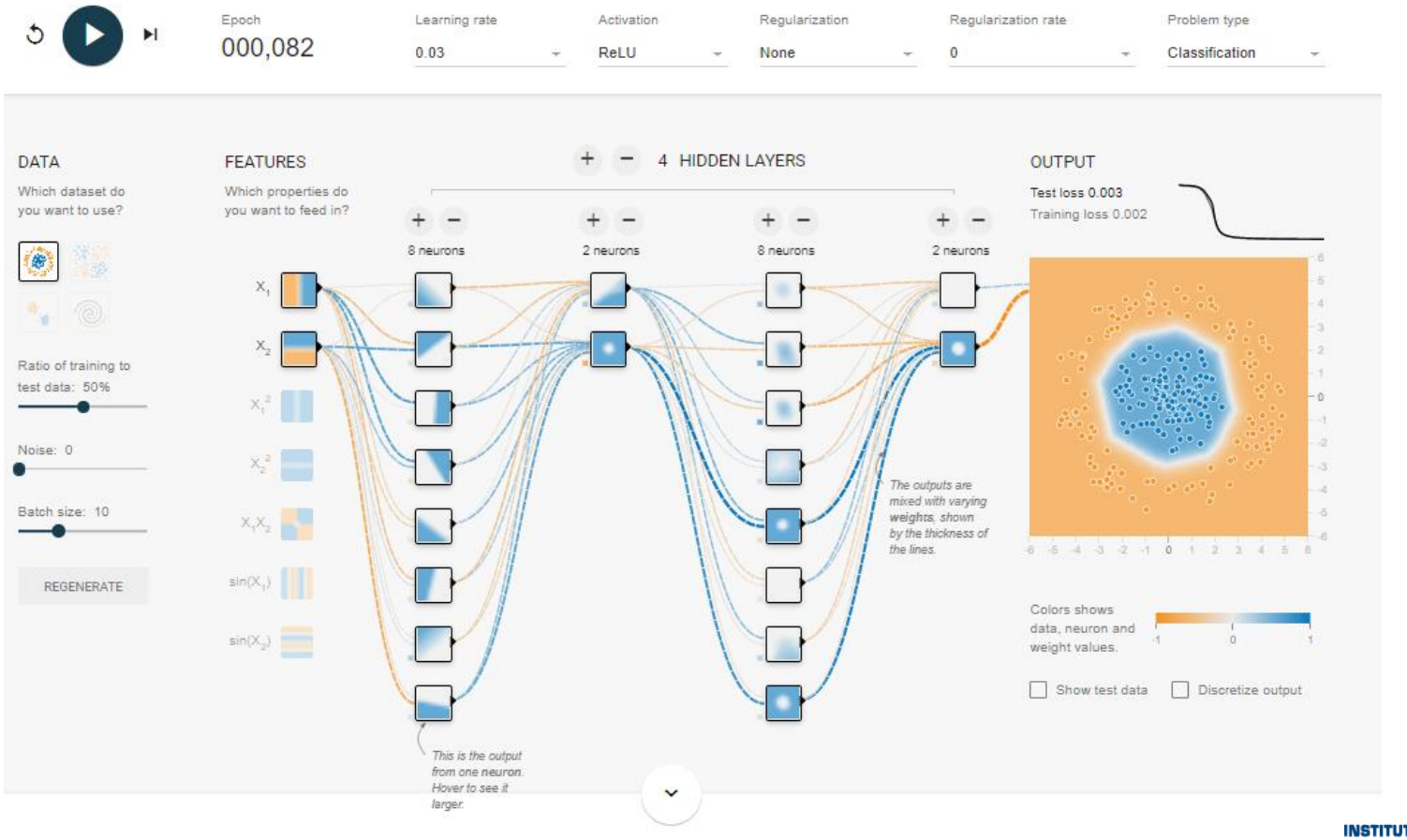
NEURAL MODELS



http://playground.tensorflow.org/

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ
60 ANOS

NEURAL MODELS



http://playground.tensorflow.org/

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ
60 ANOS

EMBEDDING WORDS

	Nem	tudo	é	claro	na	vida
Nem	1	0	0	0	0	0
tudo	0	1	0	0	0	0
é	0	0	1	0	0	0
claro	0	0	0	1	0	0
na	0	0	0	0	1	0
vida	0	0	0	0	0	1

é	0	0	1	0	0	0
na	0	0	0	0	1	0

$\mathbf{W}_{V\times N}$

0,27	0,03
0,87	0,33
0,52	0,66
0,73	0,68
0,90	0,18
0,22	0,68
0,03	0,12
0,95	0,43
0,71	0,33
0,39	0,47
0,55	0,65
0,11	0,42
0,29	0,40

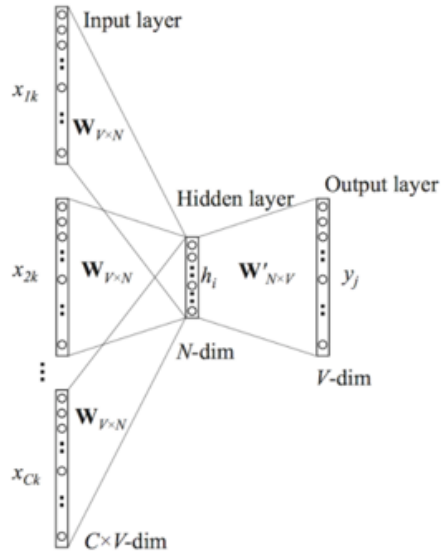
$\mathbf{W}'_{N\times V}$

0,62	0,80
0,11	0,08
0,82	0,79
0,17	0,52
0,35	0,53
0,30	0,89

claro	0	0	0	1	0	0

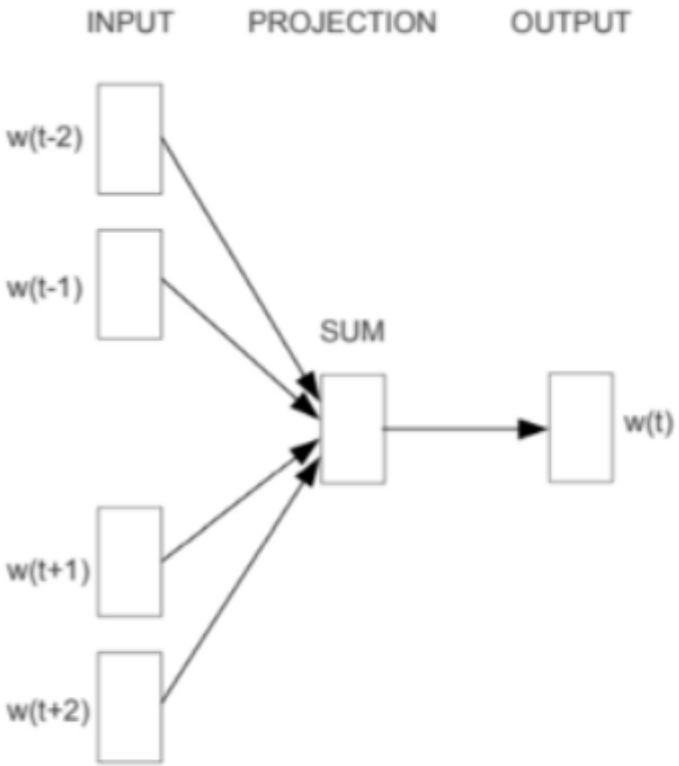
	Nem	tudo	é	claro	na	vida
	Nem	tudo	é	claro	na	vida
	Nem	tudo	é	claro	na	vida
	Nem	tudo	é	claro	na	vida
	Nem	tudo	é	claro	na	vida
	Nem	tudo	é	claro	na	vida

Machado de Assis

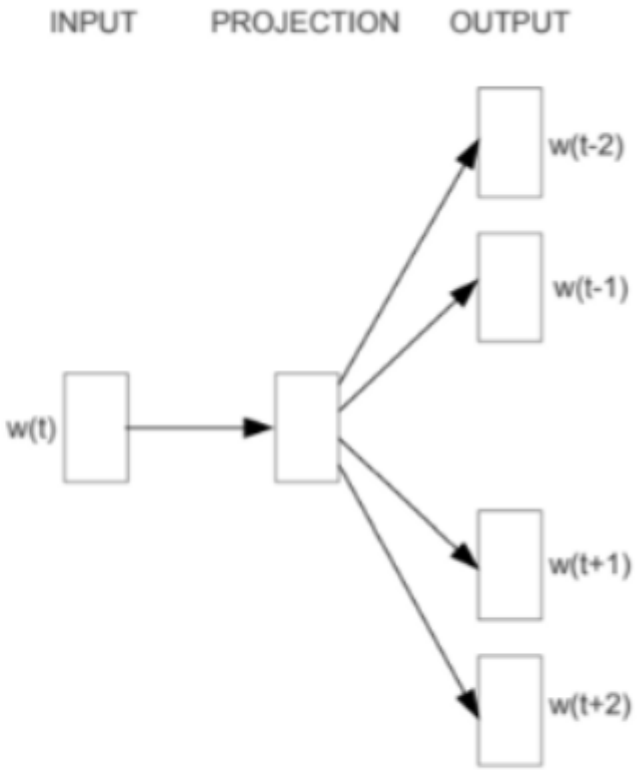


INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ
60 ANOS

EMBEDDING WORDS



CBOW



Skip-gram

LET'S CODE

- **Modelos Tradicionais BOW e TFIDF**
- **Word Embedding, Word2vec**
- **Empregando o *Corpus* de Machado de Assis**
- **Lab: BERT**