LIMPEZA

ID	Texto	ComunsIncomuns		
1	A A B B C C ?!	• Gentilezas		
2	A A B B B B é	Sem informaçãoNúmeros	D	Texto
3	AABBDDX	• Stopwords	1	ввсс
4	AADDCCBob	PontuaçãoDuplicados	2	BBBB
5	AACCCCBom dia	 Nomes de pessoas 	_	ББББ
6	AABBDDX		3	BBDD
7	NaN		4	DDCC
8	A A C C C C 78		5	CCCC

VETORIZAÇÃO

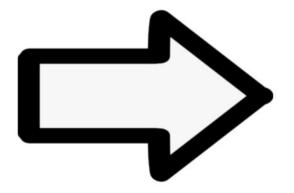
			1-GRAM				2-GRAM		
Texto	• TF	ID	В	С	D	BB		B_C	
ВВСС	TFIDFBAG of WORDS	1	2	2	0	1		1	
BBBB		2	4	0	0	3		0	
BBDD		3	2	0	2	1		0	
DDCC		4	0	2	2	1		0	
CCCC	Com N. Cromo	5	0	0	4	0		0	
	BBCC BBBB BBDD	• TFIDF BBCC • BAG of WORDS BBBDD DDCC	BBCC BAG of WORDS 1 BBBB 2 BBDD 3 DDCC 4 CCCC 5	Texto • TF ID B • TFIDF • BAG of WORDS 1 2 BBBB 2 4 BBDD 3 2 DDCC 4 0 CCCC 5 0	Texto • TF ID B C • TFIDF • BAG of WORDS 1 2 2 B B B B 2 4 0 B B D D 3 2 0 D D C C 4 0 2 C C C C 5 0 0	Texto ID B C D BBCC • BAG of WORDS 1 2 2 0 BBBB 2 4 0 0 BBDD 3 2 0 2 DDCC 4 0 2 2 CCCC 5 0 0 4	Texto ID B C D BB BBCC BAG of WORDS 1 2 2 0 1 BBBB 2 4 0 0 3 BBDD 3 2 0 2 1 DDCC 4 0 2 2 1 CCCC 5 0 0 4 0	Texto ID B C D BB BBCC BAG of WORDS 1 2 2 0 1 BBBB 2 4 0 0 3 BBDD 4 0 2 1 CCCC 5 0 4 0 4 0	

Word Embeedings pré-treinados

ID Texto

- 1 BBCC
- 2 BBBB
- 3 BBDD
- 4 DDCC
- 5 CCCC

- word2vec
- glove
- Keras Embeeding



$$\mathbf{ID} \qquad x_0 \qquad x_1 \qquad x_2 \qquad x_3 \quad \cdots \qquad x_n$$

- 1 1.5 -3 -1 3.3 \cdots 2.6
- **2** 0.5 2 -1 2.1 \cdots 0.6
- 3 1.7 -1 -2.4 0.4 \cdots -1.6
- 4 $1.5 1.6 3 1.5 \cdots -2$
- $5 \quad -1.7 \quad -3 \quad 2.7 \quad 4.1 \quad \cdots \quad 1.7$