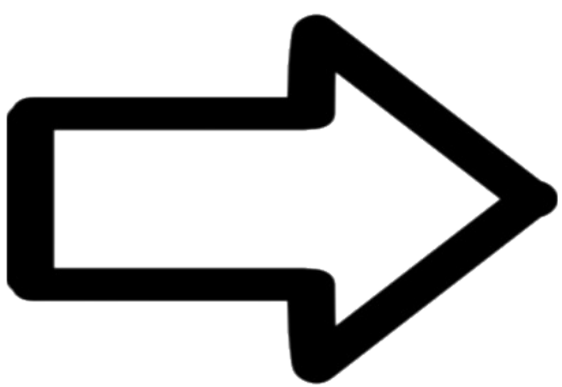
LIMPEZA

ID	Texto
1	A A B B C C ?!
2	A A B B B B é
3	A A B B D D X
4	A A D D C C Bob
5	A A C C C C Bom dia
6	A A B B D D X
7	NaN
8	A A C C C C 78

- Comuns
- Incomuns
- Gentilezas
- Sem informação
- Números
- Stopwords
- Pontuação
- Duplicados
- Nomes de pessoas

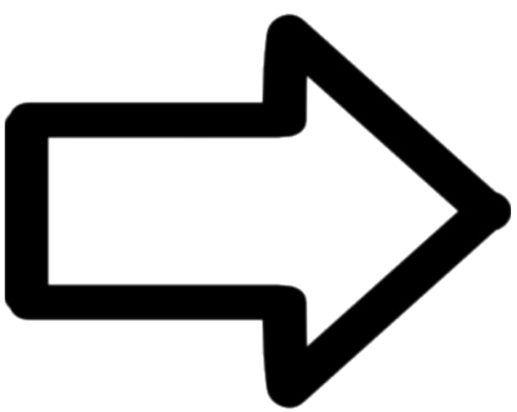


ID	Texto
1	B B C C
2	B B B B
3	B B D D
4	D D C C
5	C C C C

VETORIZAÇÃO

ID	Texto
1	B B C C
2	B B B B
3	B B D D
4	D D C C
5	C C C C

- TF
- TFIDF
- BAG of WORDS



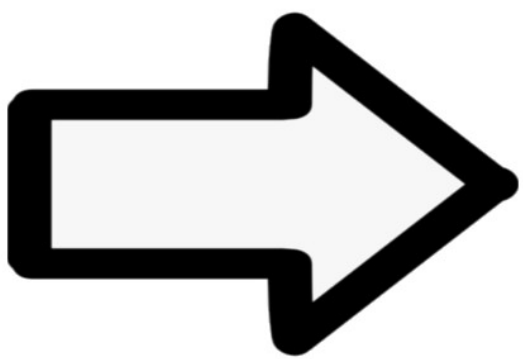
Com N-Grams

	1-GRAM			2-GRAM		
ID	B	C	D	BB	...	B_C
1	2	2	0	1	...	1
2	4	0	0	3	...	0
3	2	0	2	1	...	0
4	0	2	2	1	...	0
5	0	0	4	0	...	0

Word Embeedings pré-treinados

ID	Texto
1	B B C C
2	B B B B
3	B B D D
4	D D C C
5	C C C C

- ***word2vec***
- ***glove***
- ***Keras Embeeding***



ID	x_0	x_1	x_2	x_3	$\cdots$	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	-1.7	-3	2.7	4.1	$\cdots$	1.7