



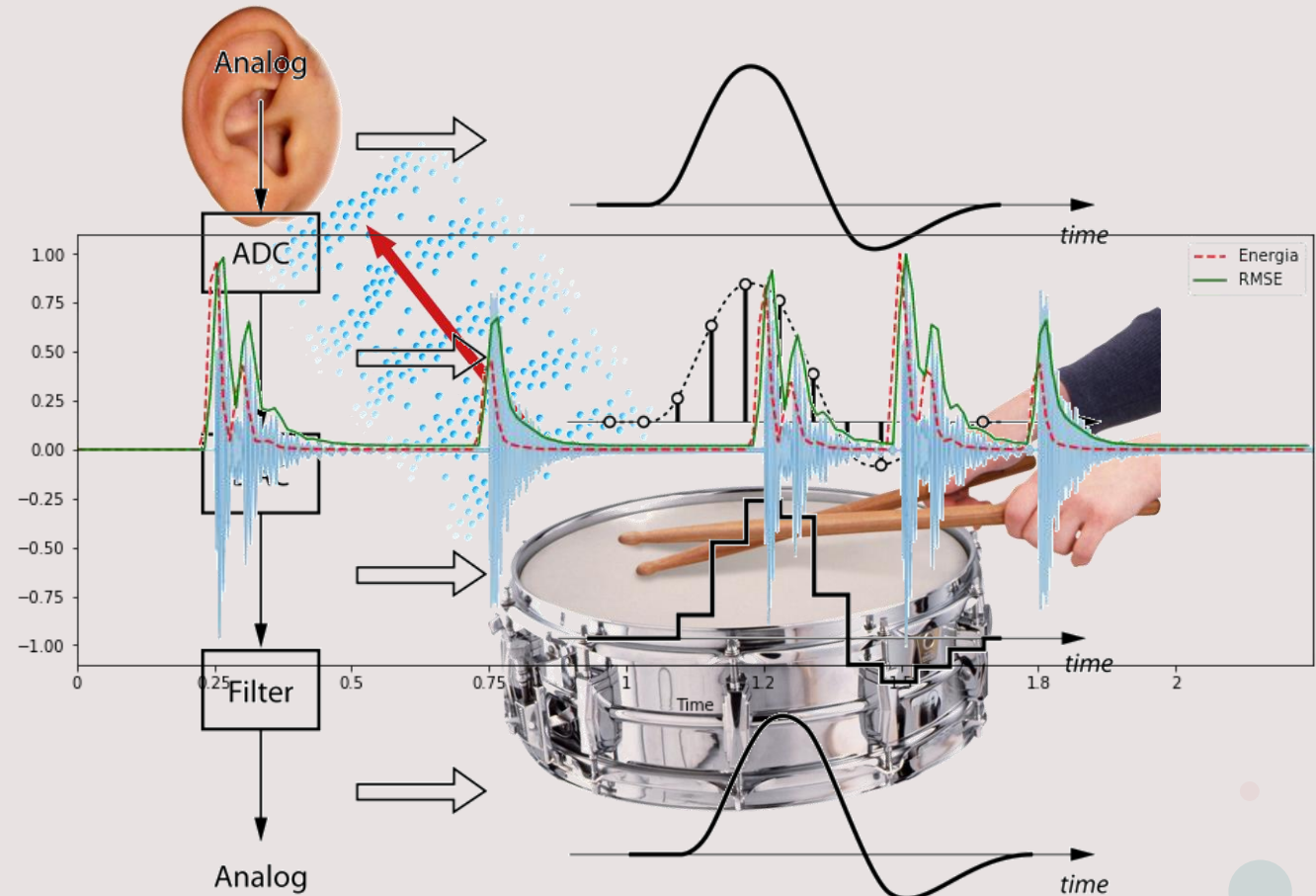
# SISTEMAS INTELIGENTES PARA CIENCIA DE DATOS

Maestría en Ciencia de Datos

Universidad de la Ciudad de Aguascalientes

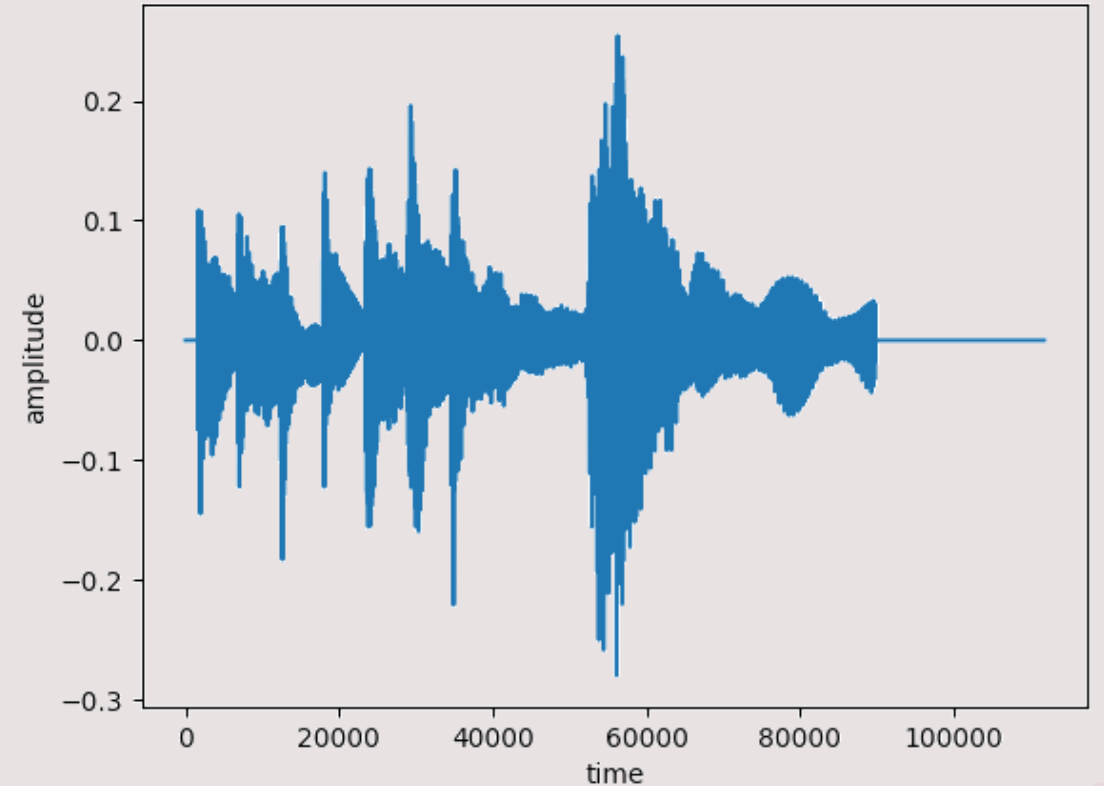
# Extracción de información de Audio

- ¿Qué es un Audio?
- ¿Cómo podemos digitalizar un audio?
- ¿Cómo podemos representar un audio?
- ¿Cuáles son las características que representan a un audio?



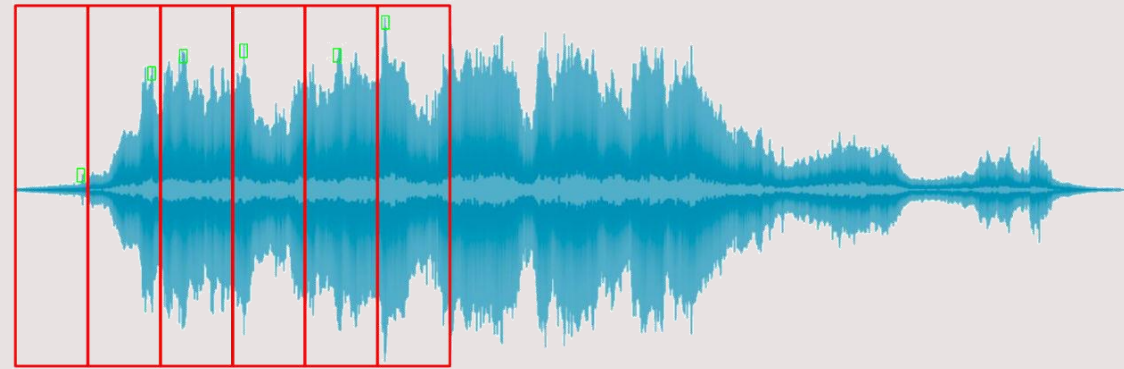
# Representación en crudo

- Tomar todos los valores del audio
- Implementación trivial
- Mucho ruido



# Características en dominio tiempo

- Divide el audio en pequeñas partes
- Estadísticas en cada parte
- Repetimos el proceso



# Raíz del cuadrado medio

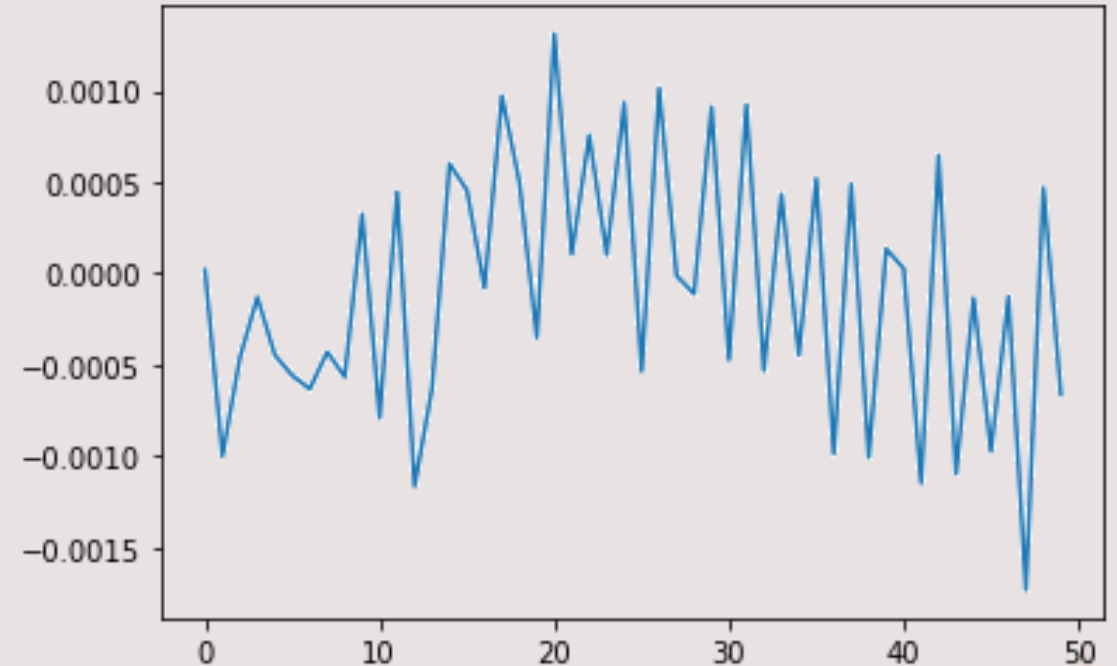
- Nivel de ruido
- Usado en
  - Segmentación de audio
  - Clasificación de géneros

**Root Mean Square**

$$x_{rms} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

# Tasa de cruce por cero

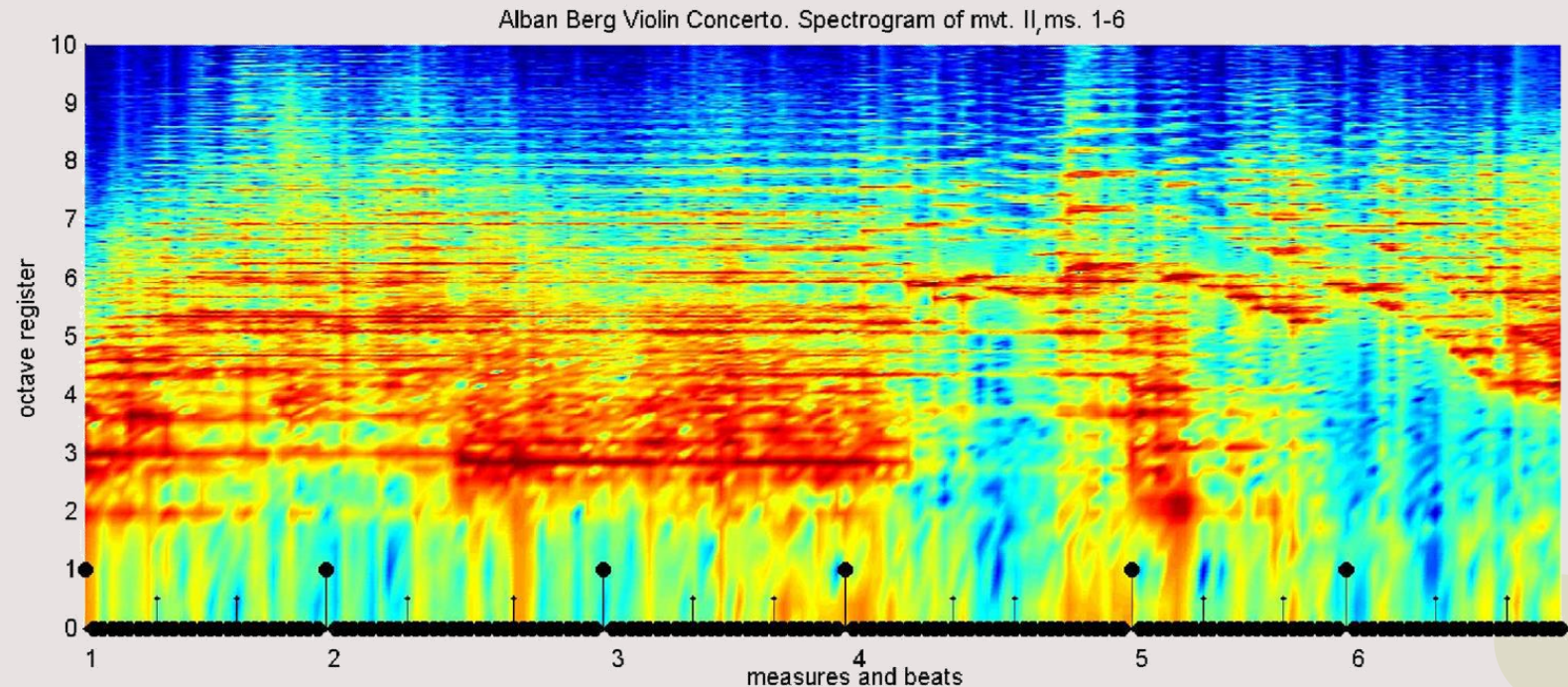
- Cantidad de veces que se cruza por cero
- Frecuencia
- Estimación tono
- Usado en
  - Estimación de tono
  - Identificación de habla





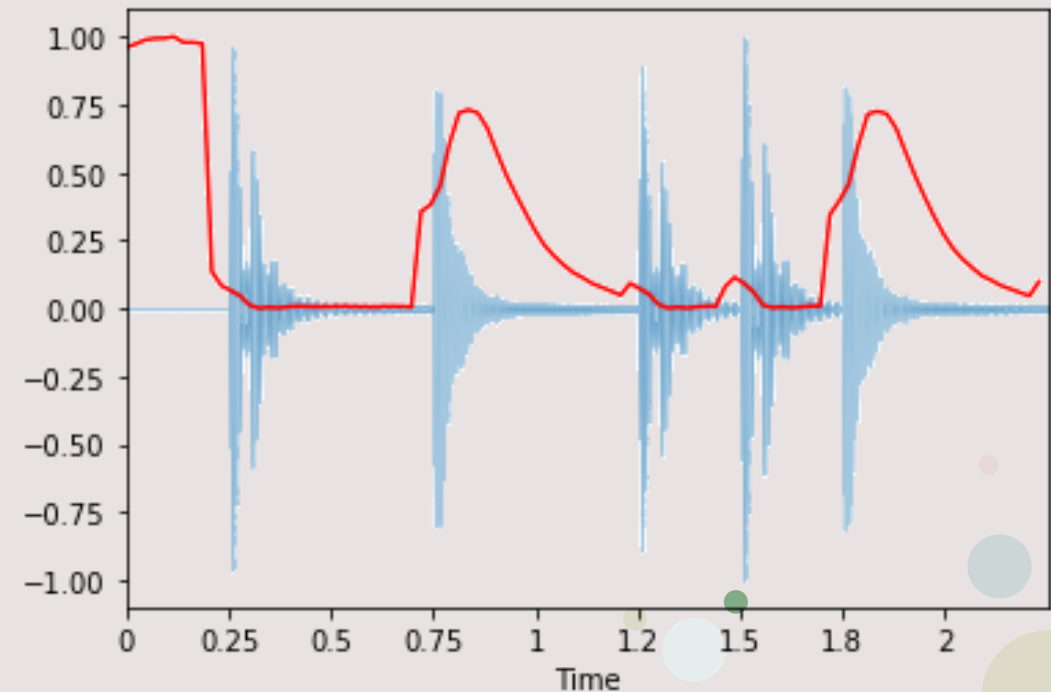
# Características en el dominio espectral

- Espacio espectral
- Transformada (discreta) de Fourier



# Centroide espectral

- Media ponderada
- El centro de masa del espectro
- Usos
  - Medida del timbre



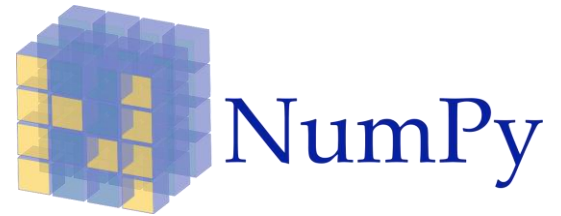


# Otras representaciones

- Zero Crossing Rate
- Energy
- Entropy of Energy
- Spectral Centroid
- Spectral Spread
- Spectral Entropy
- Spectral Flux
- Spectral Rolloff
- MFCCs
- Chroma Vector
- Chroma Deviation
- Mel-scaled spectrogram
- Spectral contrast
- Tempo
- BPM

# Herramientas para extracción de información en imágenes

- Librosa
- OpenCV
- PyAudioAnalysis





SICD\_S06\_EI\_audio.ipynb

# Extracción de información de Texto

## ASCII TABLE

- ¿Qué es un Texto?
- ¿Cómo podemos digitalizar Texto?
- ¿Cómo podemos representar un texto?
- PLN (NLP)

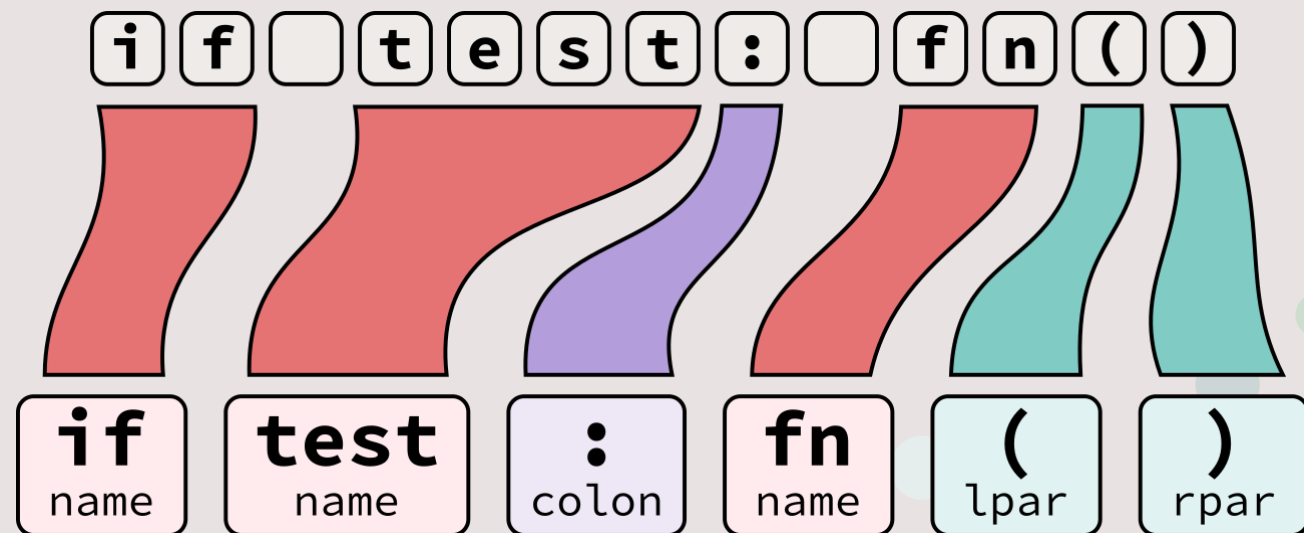
Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary
0	0	0	0	[NULL]	48	30	110000	60	0	96	60	110000
1	1	1	1	[START OF HEADING]	49	31	110001	61	1	97	61	110001
2	2	10	2	[START OF TEXT]	50	32	110010	62	2	98	62	110010
3	3	11	3	[END OF TEXT]	51	33	110011	63	3	99	63	110011
4	4	100	4	[IF INFORMATION]	52	34	110100	64	4	100	64	110100
5	5	101	5	[ENQUIRY]	53	35	110101	65	5	101	65	110101
6	6	110	6	[ACKNOWLEDGE]	54	36	110110	66	6	102	66	110110
7	7	111	7	[BELL]	55	37	110111	67	7	103	67	110111
8	8	1000	10	[SPACE]	56	38	111000	70	8	104	68	110100
9	9	1001	11	[HORIZONTAL TAB]	57	39	111001	71	9	105	69	110101
10	A	1010	12	[VERTICAL TAB]	58	3A	111010	72	:	106	6A	110110
11	B	1011	13	[BACKSPACE]	59	3B	111011	73	;	107	6B	110111
12	C	1100	14	[CARRIAGE RETURN]	60	3C	111100	74	<	108	6C	110100
13	D	1101	15	[SHIFT IN]	61	3D	111101	75	=	109	6D	110101
14	E	1110	16	[SHIFT OUT]	62	3E	111110	76	>	110	6E	110110
15	F	1111	17	[DATA LINK ESCAPE]	63	3F	111111	77	?	111	6F	110111
16	0	0000	20	[DATA LINK ESCAPE]	64	40	100000	80	@	112	70	110000
17	1	0001	21	[CONTROL 1]	65	41	100001	81	A	113	71	110001
18	2	0010	22	[CONTROL 2]	66	42	100010	82	B	114	72	110010
19	3	0011	23	[CONTROL 3]	67	43	100011	83	C	115	73	110011
20	4	0100	24	[CONTROL 4]	68	44	100100	84	D	116	74	110100
21	5	0101	25	[CONTROL 5]	69	45	100101	85	E	117	75	110101
22	6	0110	26	[CONTROL 6]	70	46	100110	86	F	118	76	110110
23	7	0111	27	[CONTROL 7]	71	47	100111	87	U	119	77	110111
24	8	1000	30	[END OF TRANSMISSION]	72	48	101000	88	V	120	78	111000
25	9	1001	31	[END OF MESSAGE]	73	49	101001	89	W	121	79	111001
26	10	1010	32	[INITIAL FEED]	74	4A	101010	90	X	122	7A	111010
27	11	1011	33	[ESCAPE]	75	4B	101011	91	Y	123	7B	111011
28	12	1100	34	[GROUP SEPARATOR]	76	4C	101100	92	Z	124	7C	111100
29	13	1101	35	[RECORD SEPARATOR]	77	4D	101101	93	[	125	7D	111101
30	14	1110	36	[UNIT SEPARATOR]	78	4E	100110	94	\	126	7E	111110
31	15	1111	37	[SPACE]	79	4F	100111	95	]	127	7F	111111
32	20	100000	40	[SPACE]	80	50	1010000	120	P			
33	21	100001	41	!	81	51	1010001	121	Q			
34	22	100010	42	"	82	52	1010010	122	R			
35	23	100011	43	#	83	53	1010011	123	S			
36	24	100100	44	\$	84	54	1010100	124	T			
37	25	100101	45	%	85	55	1010101	125	U			
38	26	100110	46	&	86	56	1010110	126	V			
39	27	100111	47	'	87	57	1010111	127	W			
40	28	101000	50	(	88	58	1011000	130	X			
41	29	101001	51	)	89	59	1011001	131	Y			
42	2A	101010	52	*	90	5A	1011010	132	Z			
43	2B	101011	53	+	91	5B	1011011	133	[			
44	2C	101100	54	,	92	5C	1011100	134	\			
45	2D	101101	55	-	93	5D	1011101	135	]			

# Glosario

- Documento: Una pieza de texto
- Corpus: muchos documentos
  - Análogo a registro – tabla
- Token: una “palabra”

# Tokenizacion

- **palabra:** Unidad léxica constituida por un sonido o conjunto de sonidos articulados que tienen un significado
- Se puede realizar la tokenizacion por espacios
- ¿Qué pasa con símbolos?
- dependiente del idioma
- Mas de una palabra





# Normalización

- Homologar tokens para que coincidan a pesar de las diferencias secuencias de caracteres.
- Mayúsculas
- Minúsculas
- **Stemming**
- Lematización

## Stemming

- reducir la forma de las palabras
- recortar sus inflexiones
- Búsqueda de sufijos

# Frecuencia

- Conteo de palabras
  - Contar cuantas veces aparece una palabra en un documento
- Palabras comunes (funcionales) tendrán muchas apariciones
  - Esto no nos dice nada

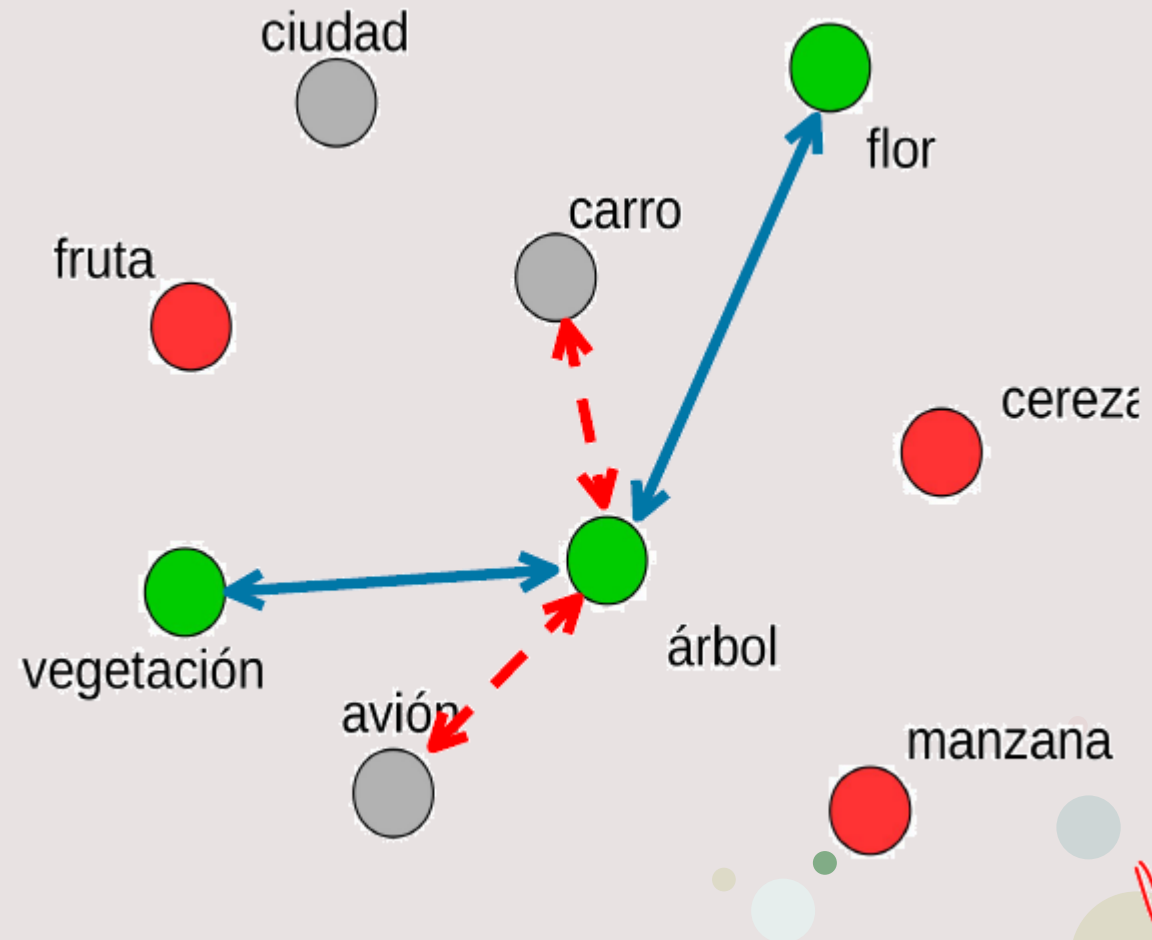
# TF-IDF

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

- Term Frequency-Inverse Document Frequency
- Normalización de conteos
- Entre más se repite una palabra dentro de un documento, la palabra será más importante para ese documento
- Palabras comunes (funcionales)

# Vectorización semántica

- Tenemos las palabras del texto, pero no las palabras relacionadas
- Word Embeddings
  - Cada palabra tiene una vectorización
  - Red neuronal
- Existen preentrenados
  - Word2Vec
  - GloVe
  - Fasttext
- Perdemos interpretación de los vectores





~~I'M TIRED~~  
~~IT'S TOO COLD~~  
~~IT'S TOO HOT~~  
~~IT'S TOO LATE~~  
~~IT'S RAINING~~  
**LET'S GO**