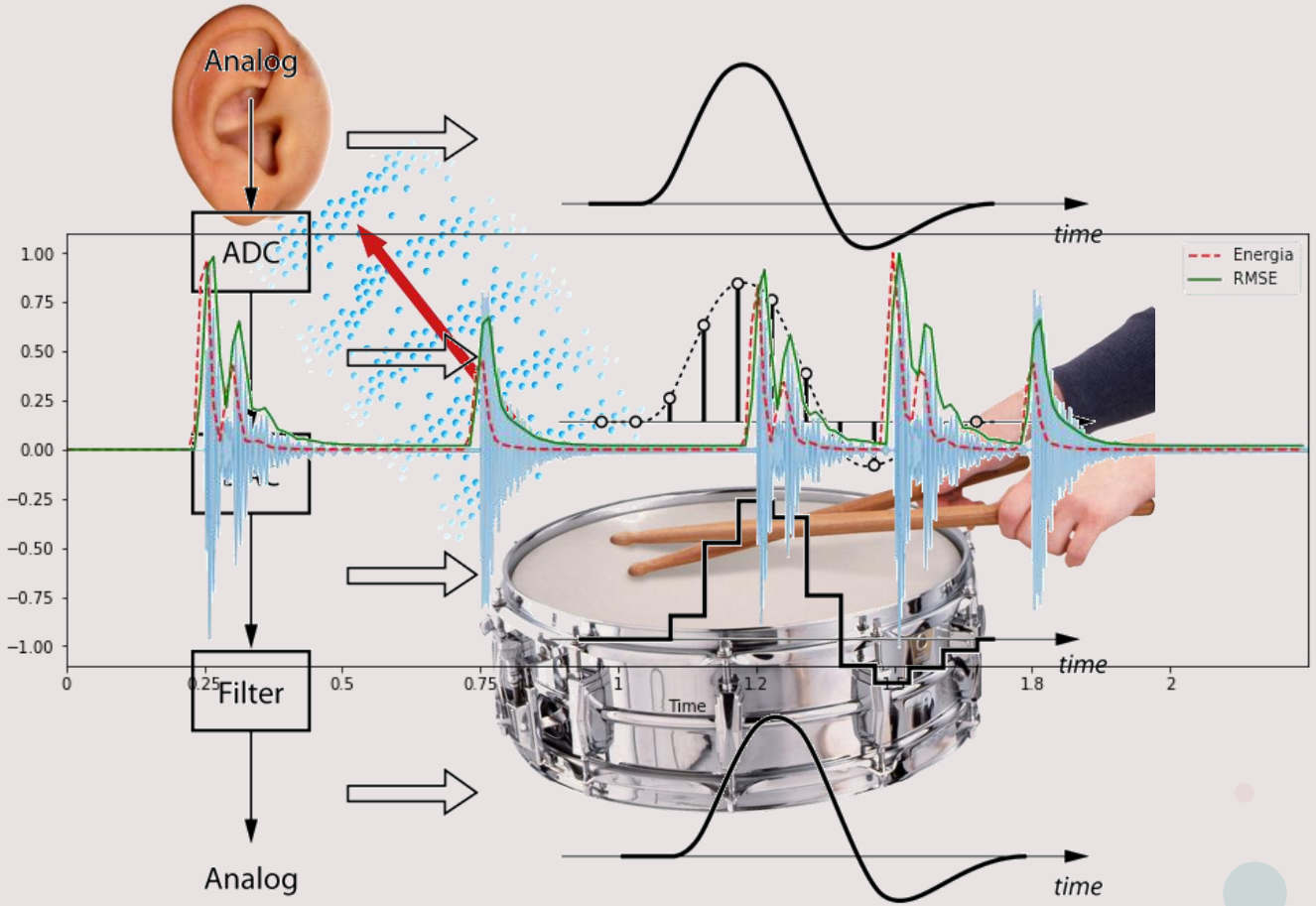
SISTEMAS INTELIGENTES PARA CIENCIA DE DATOS

Maestría en Ciencia de Datos

Universidad de la Ciudad de Aguascalientes

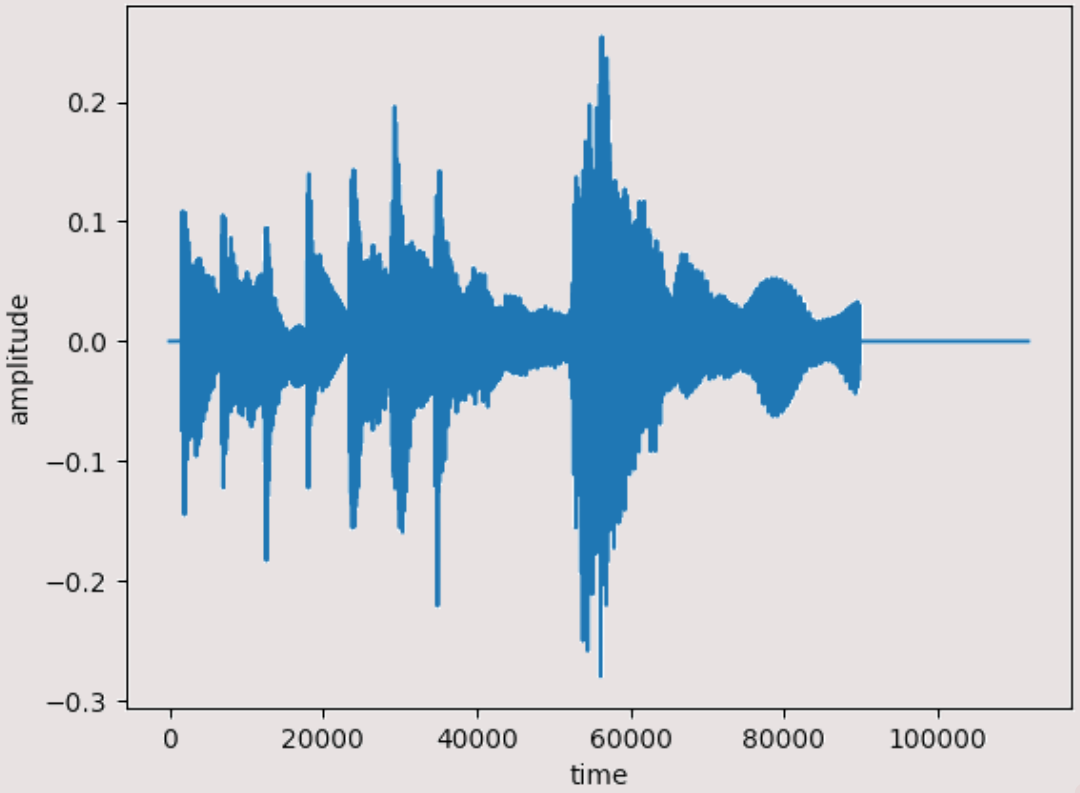
Extracción de información de Audio

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



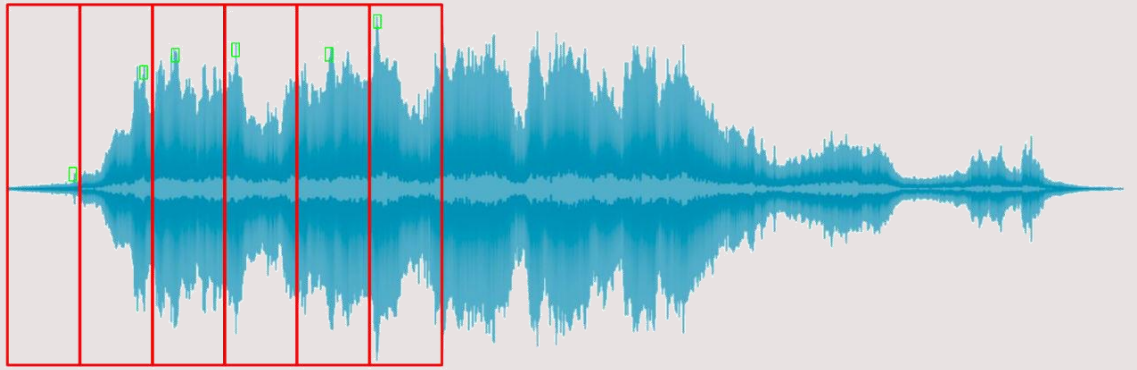
Representación en crudo

- 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

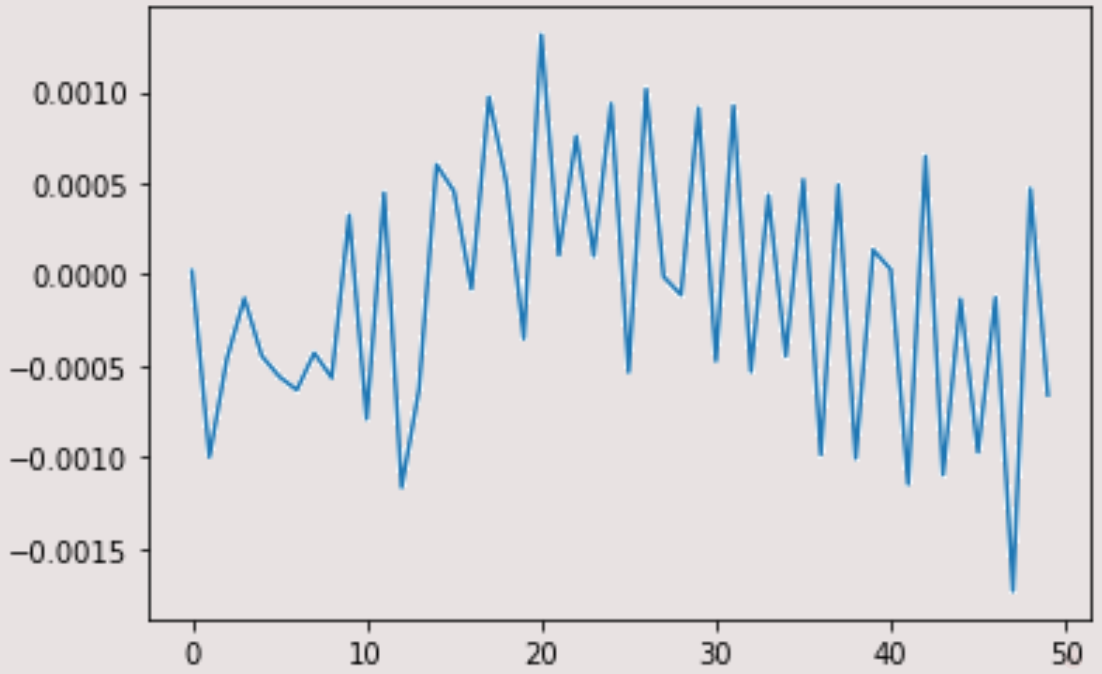
- Nivel de ruido
- Usado en
  - Segmentación de audio
  - Clasificación de géneros

Root Mean Square

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

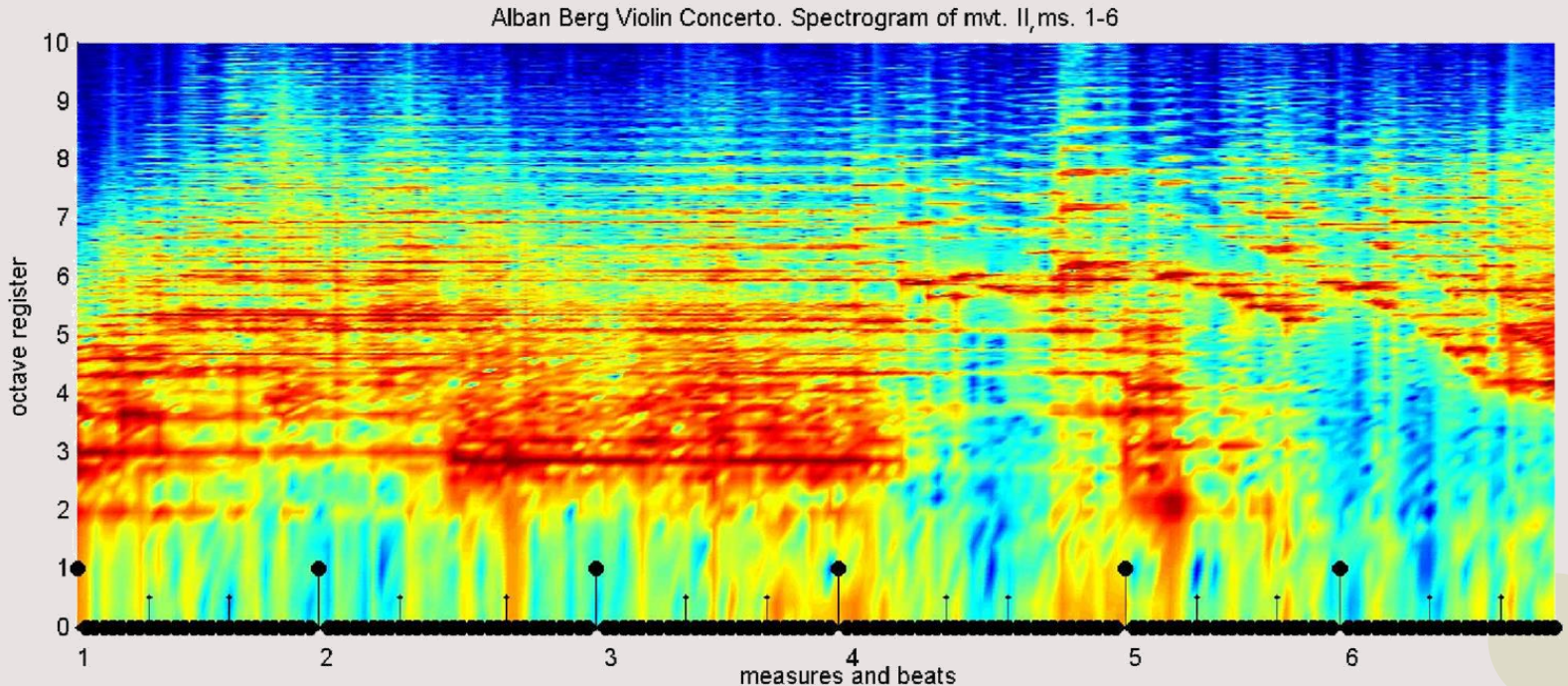
Tasa de cruce por cero

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



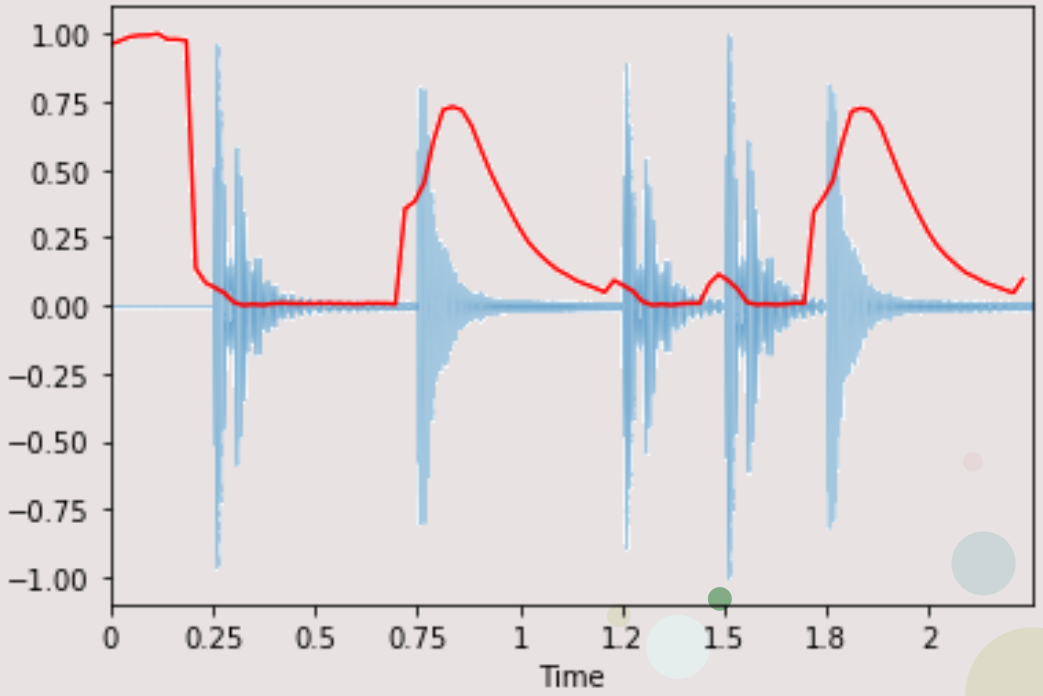
Características en el dominio espectral

- 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

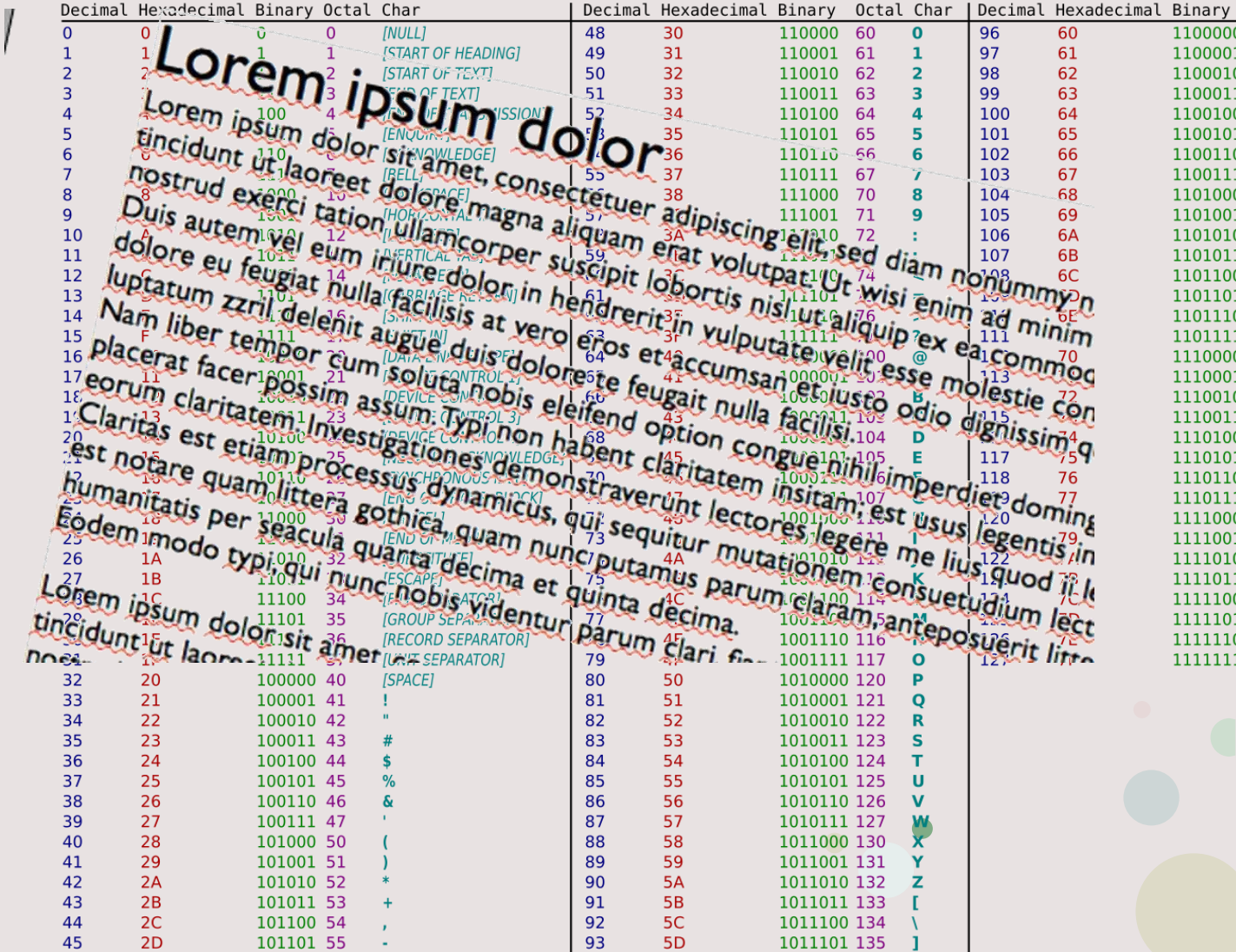
Herramientas para extracción de información en imágenes

- Librosa
- OpenCV
- PyAudioAnalysis

SICD_S06_EI_audio.ipynb

Extracción de información de Texto

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

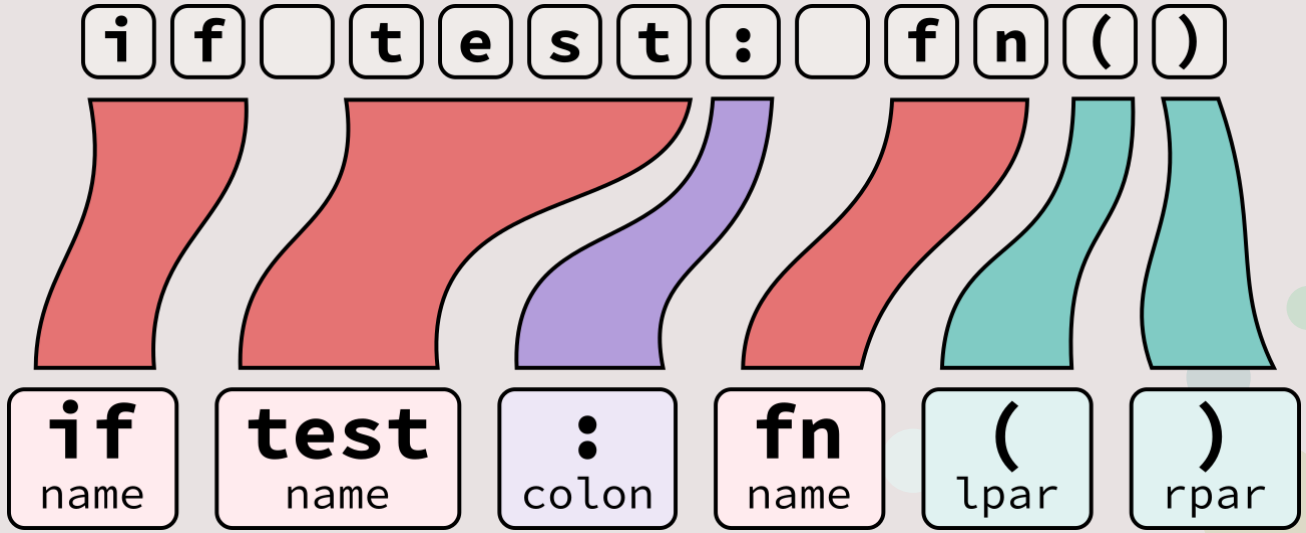


Glosario

- 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

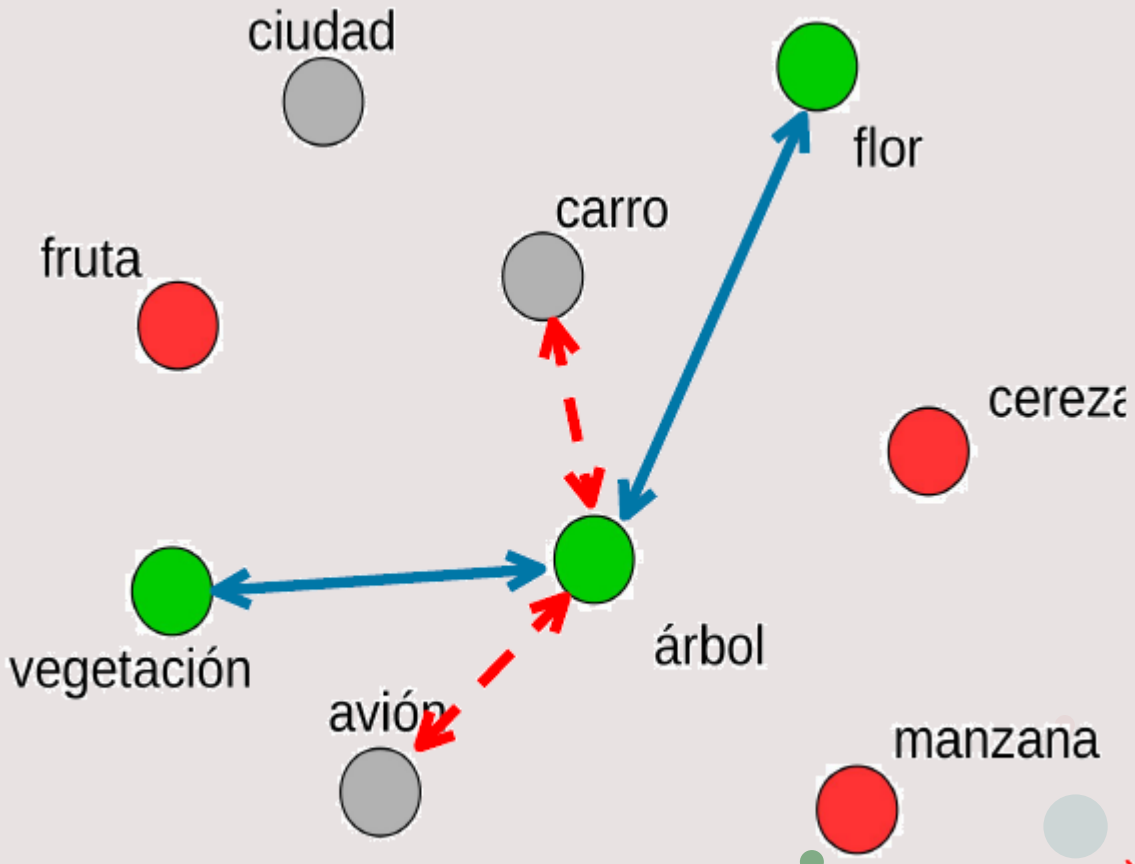
TF-IDF

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

- 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