

# M5. Minería de Texto + webscraping

## Clase 5. Embeddings y clasificación de texto



# Laboratorio de datos: web scraping y Procesamiento de Lenguaje Natural

## Clase 8. Embeddings y clasificación de texto



# Hipótesis distribucional

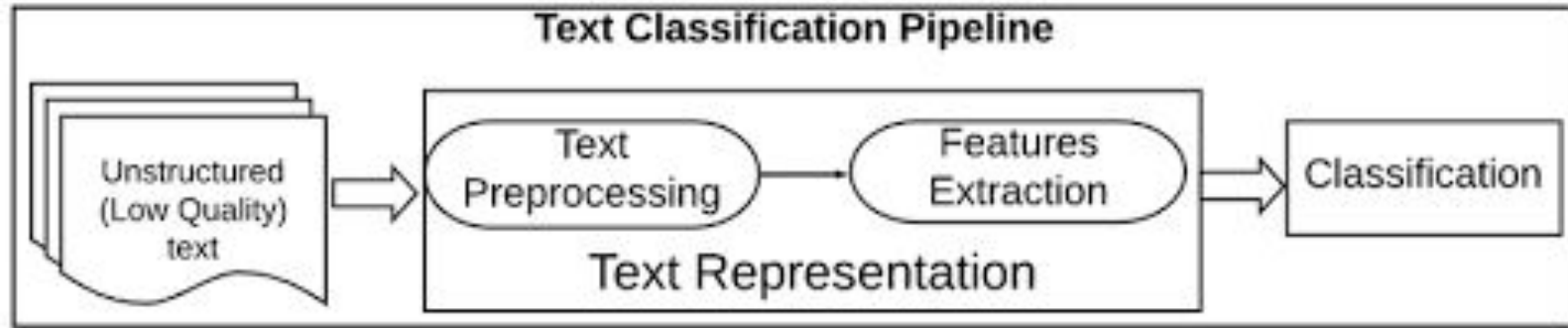
- “El significado deriva del uso de las palabras en el lenguaje” (Wittgenstein)
- “Conocerás a una palabra por su compañía” (Firth)
- Palabras cercanas tienen sentidos “cercanos”
- Ítems lingüísticos con distribuciones similares tienen significados similares”
- Idea de co-ocurrencia => términos que ocurren juntos



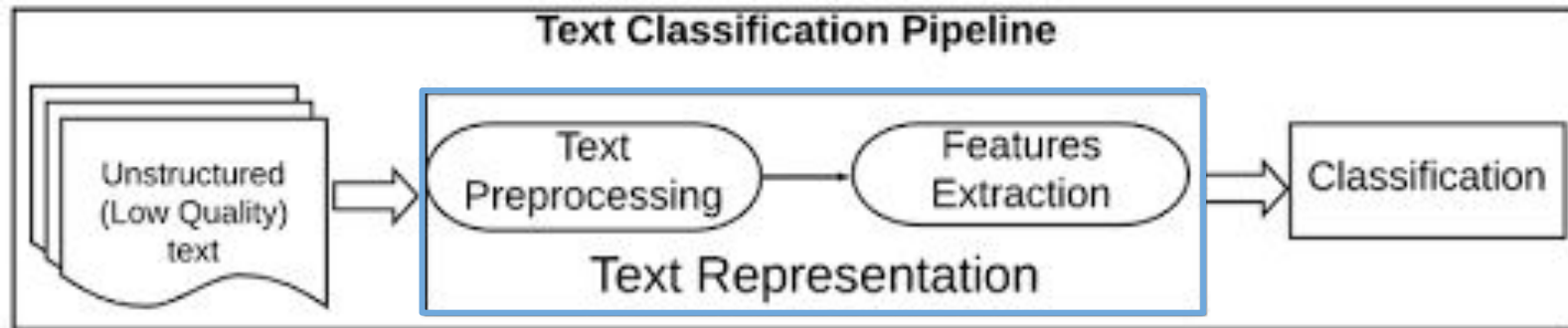
# Clasificación de texto



# Clasificación de texto

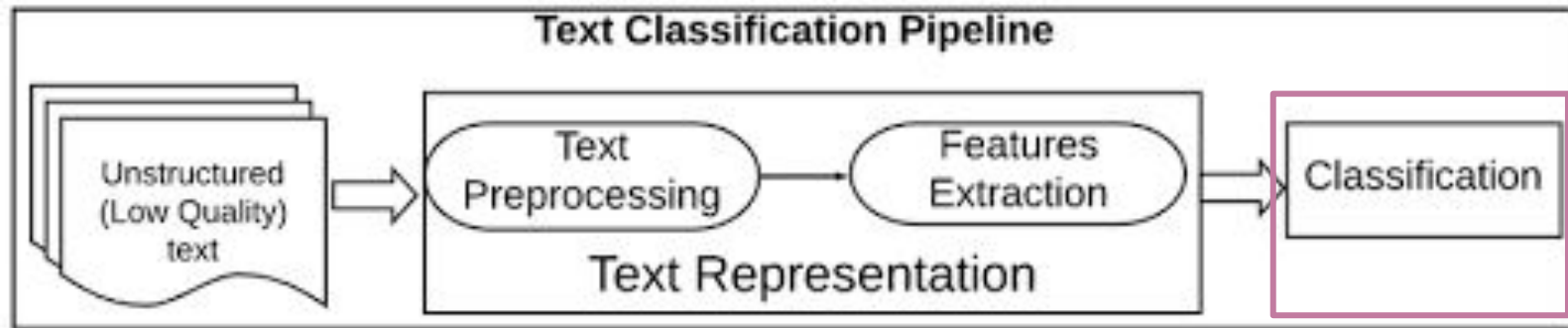


# Clasificación de texto



- Dos formas (que vimos):
  - Bag of Words => Term Frequency Matrix
  - Word Embeddings => Dense word Matrix

# Clasificación de texto



- Cualquiera de los métodos supervisados que vimos:
  - Regresión (¿lineal o logística?)
  - Random Forest
  - Gradient Boosting ...
- O que no vimos... RNN, LSTM, Transformers, etc...

# Vamos al notebook...





# Algunas cuestiones para cerrar...

- NLP “Del giro lingüístico al giro (lingüístico) computacional”.
- Posibilidades metodológicas para las ciencias sociales
- Discusiones “no metodológicas” que suscitan

Innatismo o no del lenguaje (discusión con Chomsky)

“Los límites de mi mundo son los de mi lenguaje”.

