# Vectorización de Textos

a través del modelo espacio vectorial BoW

# Bolsa de palabras (BoW)

Es una de las técnicas más sencillas y directas para transformar textos en vectores que pueda ser leído por el ordenador

Para cada texto, se asigna un peso a cada término o palabra del vocabulario (V) en función de su importancia, este peso es determinado normalmente con base en su frecuencia de aparición en el documento

No toma en consideración el orden, la estructura o el significado de las palabras

La matriz resultante, es una matriz dispersa, en la que cada texto es un renglón y cada palabra una columna

# Preprocesamiento

- Convertir palabras a minúsculas
- Quitar “Stopwods”
- Quitar puntuación
- Stemming / Lematización
  - (corrió, corrimos, correr, correremos )
  - (árbol, árboles, arbolado, arboleda)
- Quitar números
- Quitar espacios en blanco

# Bolsa de palabras

Texto 1: “El sol sale para todos”

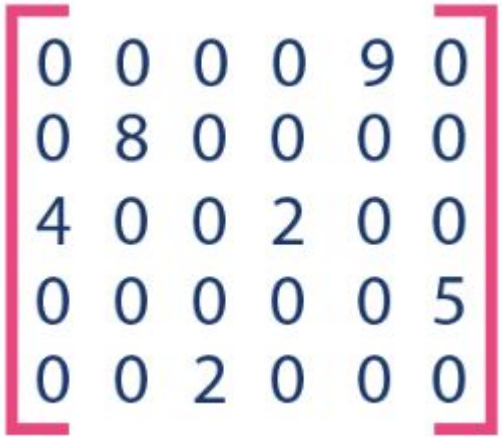
Texto 2: “ Estoy en la vereda del sol”

Texto 3: “Todos ven el sol salir”

Texto 4: “La vereda sale cara”

| | sol | salir | todos | estoy | vereda | ven | cara |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

# Matriz Dispersa (Sparse Matrix)




Tiene muy pocos elementos distintos de 0