Análisis Exploratorio Inicial del Dataset COCO-Text

Dataset original

- Imágenes en train2014.zip: 82783
- Imágenes etiquetadas (con texto): 23485

Subconjuntos tomados

- Imágenes en subset train: 15656
- Imágenes en subset val: 7829

Textos legibles: 80844
Textos ilegibles: 120282

Distribución de idiomas

- english: 193800 (96.36%)
- not english: 7326 (3.64%)

Conteo de imágenes por clase de texto

- machine printed: 23485 imágenes
- handwritten: 5 imágenes