**Librerías para eliminar encabezados en texto**

Los encabezados son texto repetitivo o fijo por lo que es importante removerle de tal forma que el modelo no sea caiga en sesgos

Las librerías para quitarlos son las siguientes:

| **Librería / Herramienta** | **Descripción** | **Cómo ayuda a eliminar encabezados** |
| --- | --- | --- |
| **re** **o expresiones regulares** | Módulo estándar para buscar y reemplazar patrones en texto. | Identifica y elimina líneas o frases específicas que puedan actuar como encabezados. |
| **NLTK** | Librería de NLP con funciones de tokenización, filtrado y limpieza. | Permite que el texto sea segmentado por líneas y se eliminar aquelas que cumplan ciertas reglas como la longitud, su patrón o palabras clave |
| **spaCy** | Librería NLP rápida con soporte para varios idiomas. | Depués de ser tokenizado el texto se puede eliminar tokens que correspondan a encabezados que hayan sido detectados por patrones. |
| **BeautifulSoup** | Librería para parsear HTML y XML. | Si el texto viene de HTML, se pueden eliminar las etiquetas de encabezado <h1>, <h2>, antes de procesar al mismo |
| **PyMuPDF** / **pdfplumber** | Librerías para extraer texto de PDF. | Filtra o salta a líneas fijas que se rpitan en cada página antes de que se pase a la fase de tokenización. |
| **python-docx** | Procesa documentos Word. | Detecta y elimina estilos de texto que sean marcados como encabezado. |

**Cómo se pueden aplicar**

**Con expresiones regulares o re**

import re

texto = """Clase 2 NLP

Bienvenidos

Contenido del documento..."""

# Eliminar línea que comience con "Clase"

texto\_limpio = re.sub(r'^Clase.\*\n?', '', texto, flags=re.MULTILINE)

print(texto\_limpio)

**nltk**

import nltk

lineas = texto.split("\n")

# Eliminar líneas que sean muy cortas o que contengan palabras clave de encabezado

lineas\_limpias = [l for l in lineas if not l.lower().startswith("clase")]

texto\_limpio = "\n".join(lineas\_limpias)

**pdfplumber**

import pdfplumber

with pdfplumber.open("documento.pdf") as pdf:

texto\_limpio = ""

for pagina in pdf.pages:

lineas = pagina.extract\_text().split("\n")

# Filtrar encabezado de cada página

lineas = [l for l in lineas if "Clase" not in l]

texto\_limpio += "\n".join(lineas) + "\n"

**BeautifulSoup (HTML)**

from bs4 import BeautifulSoup

html = "<h1>Clase 2 NLP</h1><p>Texto del documento</p>"

soup = BeautifulSoup(html, "html.parser")

# Eliminar encabezados HTML

for encabezado in soup.find\_all(['h1', 'h2', 'h3']):

encabezado.decompose()

print(soup.get\_text())

**Eventos donde podemos implementar la eliminación de encabezados**

* **Limpieza de corpus antes de entrenar un modelo NLP**: Para que no se contamine el vocabulario inclcuyendo títulos repetidos.
* **Extracción de datos desde PDF o Word**: Si estos tienen encabezados repetidos en cada página.
* **Scraping web**: Incluyendo el HTML encabezados <h1> que no pertenecen al cuerpo del texto.
* **Unificación de varios documentos:** Para que no se repitan los títulos en el texto final.
* **Preprocesamiento de documentos legales o académicos**: Teniendo estos membretes o títulos fijos.