Informe de Limpieza y Normalización de Documentos

Proyecto: Recursos TEA en Andalucía

9 de noviembre de 2025

Resumen

Este informe detalla el proceso técnico de limpieza y normalización aplicado al corpus de documentos extraídos sobre Trastorno del Espectro Autista (TEA) en Andalucía. Se describen los criterios utilizados para identificar y eliminar 'ruido' informativo, asegurando la calidad y legibilidad de los textos finales recopilados en el directorio `documentos_tea_andalucia_limpio`.

Índice

1. Introducción

El objetivo del proceso de limpieza es transformar los documentos brutos extraídos de diversas fuentes web y PDF en un formato de texto plano uniforme, legible y libre de elementos irrelevantes. Esto facilita su posterior consulta por parte de las familias y profesionales, así como su posible procesamiento automático.

2. Definición de 'Ruido' y Criterios de Eliminación

Se considera 'ruido' a todo aquel contenido que no aporta información sustancial sobre los trámites, recursos o normativas, y que es producto de la extracción automática o del formato original de los documentos.

El script de limpieza (`limpia_datos.py`) aplica los siguientes criterios para la detección y eliminación de ruido:

2.1. Elementos de Navegación y Formato Web

Líneas típicas de páginas web que no forman parte del contenido principal:

- **Menús y botones:** Palabras sueltas o frases cortas como 'Inicio', 'Menú', 'Contacto', 'Volver', 'Imprimir', 'Compartir'.
- **Avisos legales:** Textos genéricos sobre 'Política de cookies', 'Aviso legal', 'Política de privacidad'.
- **Marcadores de posición:** Textos que sustituyen a elementos no textuales como `[Imagen:...]` o `[Link:...]`.

2.2. Metadatos y Elementos Técnicos

Información técnica irrelevante para el usuario final:

- Líneas que comienzan con `Content-Type:` o `Encoding:`.
- Identificadores de depósito legal (`Depósito Legal:`), ISSN o NIPO, salvo que formen parte de una referencia bibliográfica completa.

2.3. Elementos Repetitivos en Documentos Oficiales (PDFs)

Artefactos comunes en la extracción de texto desde archivos PDF, especialmente oficiales como el BOJA:

- **Numeración de página:** Líneas que contienen solo números, o formatos como 'Página X', '- X -', '[X]'.
- **Cabeceras y pies de página recurrentes:** Líneas que se repiten múltiples veces a lo largo del documento (detectadas automáticamente si aparecen más de 3 veces), conteniendo términos como 'Consejería', 'Junta de Andalucía', 'BOJA', o URLs genéricas como `www.juntadeandalucia.es`.

2.4. Ruido Tipográfico y Estructural

- **Líneas divisorias:** Secuencias de caracteres repetidos usadas para separar secciones visualmente (ej. `===`, `--`, `***`).
- **Líneas vacías múltiples:** Se reducen a un máximo de dos saltos de línea consecutivos para separar párrafos, eliminando el espacio vertical excesivo.
- **Líneas muy cortas:** Líneas con menos de 3 caracteres que no son parte de una lista o enumeración.

3. Proceso de Normalización

Además de eliminar ruido, se aplican transformaciones para unificar el formato de todos los documentos:

1. **Unificación de codificación:** Todos los archivos se convierten a **UTF-8** para garantizar la correcta visualización de caracteres especiales (tildes, eñes, símbolos).
2. **Normalización de caracteres:** Se corrigen errores comunes de decodificación (ej. `Ãą` → á, `â` →).
3. **Normalización de espacios:** Se eliminan espacios en blanco al inicio y final de cada línea, y se reducen los espacios múltiples dentro del texto a uno solo.
4. **Eliminación de duplicados consecutivos:** Se detectan y eliminan líneas idénticas que aparecen de forma seguida.

4. Estructura de los Documentos Limpios

Los documentos resultantes en el directorio `documentos_tea_andalucia_limpio` siguen una estructura estandarizada para facilitar su lectura e identificación:

```
================================================================================
DOCUMENTO LIMPIO Y NORMALIZADO
================================================================================

INFORMACIÓN DEL DOCUMENTO:
--------------------------------------------------------------------------------
TÍTULO: [Título extraído u original del archivo]
URL ORIGINAL: [Enlace de donde se extrajo]
TIPO: [HTML / PDF]
CATEGORÍA: [Subcarpeta de origen]

================================================================================
CONTENIDO:
================================================================================

[... Texto limpio y normalizado del documento ...]

================================================================================
Documento original: nombre_archivo_original.txt
Procesado y limpiado automáticamente
================================================================================
```

Listing 1: Ejemplo de estructura de un documento limpio

5. Conclusión

El proceso de limpieza ha permitido generar un corpus de texto de alta calidad, libre de elementos distractores y con un formato homogéneo. Esto facilita enormemente el acceso a la información relevante por parte de las familias y constituye una base sólida para cualquier herramienta de asistencia o consulta que se desee desarrollar en el futuro.