**Documento: Solución al Problema de Extracción de Atributos en el Nivel Básico**

**1. Definición del Problema**

La Unidad de Búsqueda de Personas Desaparecidas (UBPD) enfrenta un desafío significativo en la identificación y análisis de relatos relacionados con la desaparición de personas, especialmente en contextos de conflicto armado. Estos relatos incluyen información clave sobre personas, lugares, fechas, actores armados, y eventos, que son críticos para el proceso de búsqueda y esclarecimiento de casos. La necesidad identificada es la extracción automática de esta información para agilizar el análisis y toma de decisiones de la UBPD.

**2. Objetivo**

El objetivo del nivel básico fue **desarrollar un modelo que permita extraer automáticamente de textos variables clave como nombres, lugares, fechas, grupos armados y actividades.** Esta extracción permitirá complementar la información disponible y contribuir al esfuerzo de búsqueda de personas desaparecidas.

**3. Solución Propuesta**

El enfoque del modelo básico consistió en **aplicar técnicas de procesamiento de lenguaje natural (NLP**) para automatizar la extracción de atributos a partir de relatos textuales. La solución incluye los siguientes pasos:

1. Definir el set de datos: Se trabajó con dos fuentes principales de datos:
	* Relatos en texto extraídos de documentos.
	* Relatos de personas desaparecidas en formato estructurado, proporcionados por la UBPD.

Estos sets de datos se combinaron para construir un dataset consolidado (dataset\_combinado.csv), sobre el cual se aplicaron técnicas de NLP.

1. Anonimización de datos: Se implementó un proceso de anonimización para proteger la privacidad de los individuos mencionados en los relatos. Se sustituyeron nombres propios y datos sensibles por identificadores neutrales como "Persona X" o "Lugar Y". Esta técnica permitió preservar la utilidad de los datos sin comprometer la confidencialidad.
2. Definir variables de interés: Las variables de interés definidas para la extracción fueron:
	* **Personas involucradas:** nombres de desaparecidos o involucrados.
	* **Lugares mencionados**: ubicación de los eventos o desapariciones.
	* **Fechas importantes**: fechas clave relacionadas con los eventos.
	* **Grupos armados mencionados**: menciones a actores armados.
	* **Eventos/actividades:** acciones o eventos relacionados con las desapariciones.
3. Extracción de variables: Se implementaron algoritmos de NLP utilizando herramientas como spaCy y displaCy para realizar la extracción automática de las variables definidas. Los algoritmos de reconocimiento de entidades permitieron identificar y resaltar las entidades mencionadas en los relatos, como personas, lugares y fechas.

**4. Implementación del Modelo**

El modelo se desarrolló en Python, utilizando las siguientes bibliotecas:

* spaCy para el procesamiento de lenguaje natural y la extracción de entidades.
* Pandas para la manipulación de datos.
* Ipywidgets para crear un visor interactivo que permita navegar entre los relatos, resaltando las entidades extraídas y mostrando un resumen de las variables.

Proceso de Extracción:

* Se cargaron los relatos y se aplicaron algoritmos de tokenización y reconocimiento de entidades.
* Las entidades reconocidas fueron resaltadas en el visor interactivo y almacenadas en un DataFrame para análisis posterior.
* Se creó un resumen visual de las variables extraídas por cada relato, incluyendo personas involucradas, lugares, fechas, grupos armados y eventos.

**5. Resultados**

El modelo fue capaz de identificar con éxito las variables clave en los relatos textuales. Los resultados muestran una extracción precisa de nombres, lugares y fechas. El visor interactivo facilita la inspección manual de los relatos y los atributos extraídos, lo que permite una mayor comprensión del contexto de cada caso.

Ejemplo de Relato con Variables Extraídas:

* Relato Original: "Juan Pérez fue visto por última vez en su casa antes de ser llevado por hombres armados."
* Relato Anonimizado: "Persona X fue visto por última vez en su casa antes de ser llevado por hombres armados."
* Variables Extraídas:
	+ Personas Involucradas: Persona X.
	+ Lugares Mencionados: su casa.
	+ Fechas Importantes: N/A.
	+ Grupos Armados Mencionados: hombres armados.
	+ Eventos/Actividades: llevado.

**6. Conclusión**

La solución propuesta para el nivel básico cumple con los requisitos establecidos. **El modelo automatiza la extracción de información relevante de relatos textuales**, lo que facilita el análisis y permite a la UBPD tomar decisiones informadas más rápidamente. **La anonimización de datos protege la privacidad de las personas mencionadas**, y la herramienta desarrollada puede integrarse en los procesos de la UBPD para la búsqueda de personas desaparecidas.

**7. Futuras Mejoras**

* Mejorar el reconocimiento de entidades: Implementar técnicas de machine learning para mejorar la precisión de la identificación de entidades en casos de ambigüedad.
* Ampliar el dataset: Incluir más relatos para entrenar al modelo con un set de datos más representativo.
* Integración con sistemas de la UBPD: Conectar el modelo a los sistemas de la UBPD para una implementación en tiempo real.

Este documento refleja el trabajo realizado en la fase básica del modelo, con un enfoque en la extracción automática de variables de interés y la creación de herramientas para mejorar el análisis de los relatos.