Taller de Procesamiento de Lenguaje Natural: Recuperación de Información Aplicada a Entrevistas y Libros de la Comisión de la Verdad

Objetivo

Desarrollar un proceso de Recuperación de Información (IR) para determinar qué entrevistas contenidas en el documento JSON están relacionadas con testimonios o secciones de los libros de la Comisión de la Verdad (CEV). Se espera que los estudiantes apliquen técnicas de procesamiento de lenguaje natural para analizar y relacionar las narraciones de entrevistas con textos específicos de los libros, utilizando métricas específicas para medir la relevancia de los resultados.

Recursos

- **Libros de la Comisión de la Verdad (CEV):**

  https://drive.google.com/drive/folders/1W9UOdkqnAcztnFZ8xresW4ZeZ2pqXe4v?usp=sharing

- **Entrevistas:**

  https://drive.google.com/file/d/1XeFS_mQFmTLQ_VUiODTBO5eONWv-Ege5/view?usp=sharing

- **Tesauro CEV:**

  https://drive.google.com/file/d/1XgZzMtFQp0RzwqdY7Em5q5ProsWLMMX4/view?usp=sharing

Actividades

1. **Extracción y Preparación del Corpus:**
    - Extraer las secciones, testimonios o partes relevantes de cada uno de los libros de la CEV. Definir la morfología de documento para recuperar (Explicar como la van a hacer).
    - Preprocesar los textos extraídos y las entrevistas para su análisis, incluyendo pasos como tokenización, eliminación de stop words, y lematización.
    - Entregar el corpus limpio y estructurado en un formato adecuado para su análisis comparativo.
2. **Análisis Exploratorio del Corpus:**
    - Realizar un análisis exploratorio de los corpus generados. Esto puede incluir:

        - Estadísticas descriptivas de longitud de texto, diversidad léxica, y frecuencia de términos.
        - Generación de nubes de palabras para visualizar los términos más frecuentes en cada corpus.
        - Comparación de patrones léxicos entre los textos de los libros y las entrevistas.

3. **Desarrollo del Modelo de Recuperación de Información:**
    - Implementar un modelo de IR que relacione las narraciones de las entrevistas con los textos generados de los libros.
    - Utilizar técnicas como la vectorización de textos (TF-IDF) para representar los textos y facilitar la búsqueda y comparación.
    - Entregar un informe detallado con los resultados obtenidos y las observaciones realizadas durante la implementación del modelo.
4. **Implementación de Métricas de Relevancia:**
    - Implementar manualmente las métricas de Rocchio y Okapi BM25 para medir la relevancia de las coincidencias encontradas entre entrevistas y textos de los libros.
    - Crear funciones personalizadas que calculen estas métricas y expliquen su funcionamiento.
    - Comparar los resultados de relevancia obtenidos con ambas métricas y discutir las diferencias.
5. **Comparación de corpus:**
    - Realizar una comparación de los textos de las entrevistas que tengan temas similares con los textos de los libros preprocesados.
    - Generar un gráfico de “heatmap” que puedas dar cuenta del relacionamiento sugerido.
6. **Presentación de Resultados:**
    - Elaborar un informe final que integre todos los resultados obtenidos, incluyendo:

        - Descripción del proceso seguido para la extracción y preparación del corpus.
        - Análisis exploratorio del corpus.
        - Descripción y análisis del modelo de IR desarrollado.
        - Evaluación de la relevancia utilizando las métricas implementadas.
    - Incluir conclusiones y posibles mejoras para futuras iteraciones del proyecto.

Entregables

- **Corpus Extraído y Preprocesado:** Dos archivos con los textos extraídos en formato raw (crudo) y los corpus preprocesados de los libros.
- **Análisis Exploratorio:** Informe con los resultados del análisis exploratorio del corpus.
- **Modelo de IR:** Código y reporte detallado del modelo de Recuperación de Información.
- **Funciones de Métricas:** Código de las funciones para calcular las métricas de Rocchio y Okapi BM25.
- **Comparación de las entrevistas con los libros:** entregar un cuadro y un gráfico que pueda describir que entrevistas están vinculadas a los textos de cada libro de acuerdo con el segmentado que generaron.
- **Informe Final:** Documento que integre todos los componentes anteriores y presente una evaluación crítica del proceso y los resultados.