Trabajo Práctico N° 2: Primera Parte

Mentoría M11: "Detección Automática de Plagio"

29 de junio de 2023

1. Introducción

La detección de plagio es un tema crucial en el ámbito académico y profesional para garantizar la integridad y originalidad del contenido. En este trabajo práctico, se aplicará la Ingeniería de Datos para abordar la tarea de detección intrínseca de plagio en un conjunto de 4753 documentos de texto. El objetivo es diseñar un pipeline que itere sobre el directorio de documentos, segmente el texto de cada documento y aplique filtros para preparar los datos antes de su posterior análisis.

1.1. Propuesta de Base de Datos Relacional:

Para gestionar la información relacionada con los documentos y su segmentación, se propone utilizar una base de datos relacional. A continuación, se sugiere una estructura básica para la base de datos:

Cuadro 1: Tabla Documentos

id_documento	**nombre_archivo**	**texto**
1	archivo1.txt	Texto del archivo 1
2	archivo2.txt	Texto del archivo 2
...	...	...

Cuadro 2: Tabla Segmentos

id_segmento	**id_documento**	**segmento_texto**	**segmento_limpio**	**init_s**	**length**
1	1	Segmento 1 del archivo 1	Clean 1 del archivo 1	Init 11	length 11
2	1	Segmento 2 del archivo 1	Clean 2 del archivo 1	Init 21	length 21
...	...	...	...	...	...

2. Pipeline de Procesamiento de Datos:

A continuación, se propone un pipeline de procesamiento de datos para iterar sobre el directorio de documentos, segmentar y limpiar el texto, y al final, almacenarlo:

1. Lectura de documentos: Iterar sobre el directorio de documentos y leer cada archivo, extrayendo su nombre y texto completo.

2. Almacenamiento en la tabla “Documentos”: Guardar la información de cada documento en la tabla “Documentos”, asignando un id_documento único.

3. Segmentación de texto: Aplicar técnicas de segmentación de texto (por ejemplo, basadas en párrafos, oraciones o cualquiera que puedan proponer) para dividir el texto completo en segmentos significativos.

4. Almacenamiento en la tabla “Segmentos”: Para cada segmento resultante de la segmentación, almacenarlo en la tabla “Segmentos”, relacionándolo con su respectivo documento mediante el campo ïd_documento”.

2.1. Importante:

El campo init_s es la posición de inicio del segmento y el campo length es la longitud del mismo.