Proyecto Bimestral: Sistema de Recuperación de Información

basado en Reuters-21578

Prof. Iván Carrera

27 de mayo de 2024

1. Introducción

El objetivo de este proyecto es diseñar, construir, programar y desplegar un Sistema de Recuperación de Información (SRI) utilizando el corpus Reuters-21578. El proyecto se dividirá en varias fases,

que se describen a continuación.

2. Fases del Proyecto

2.1. Adquisición de Datos

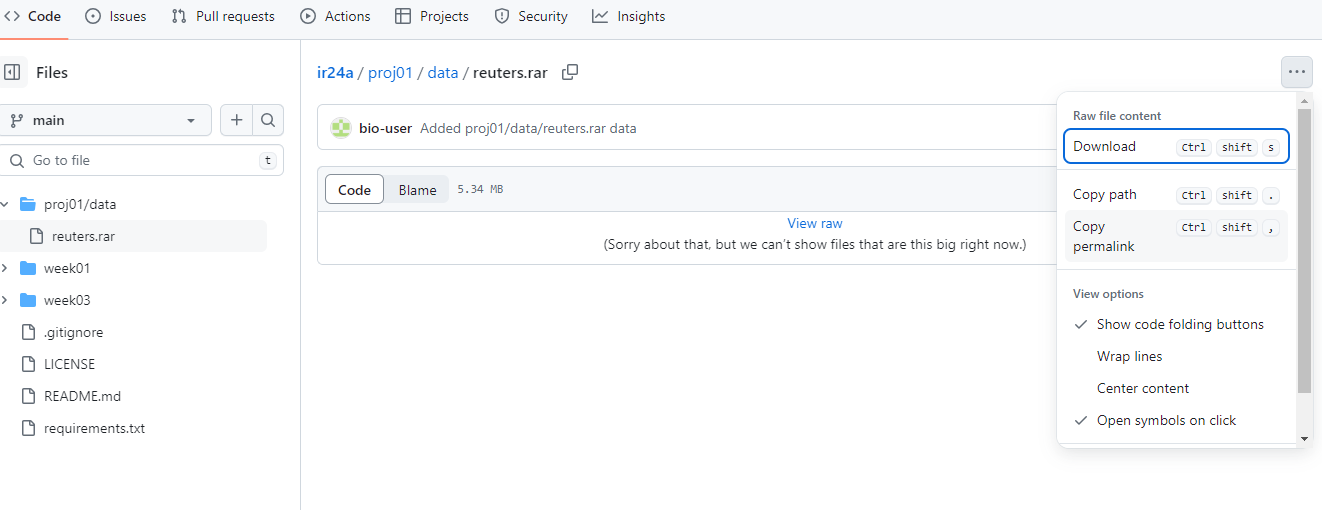
Objetivo: Obtener y preparar el corpus Reuters-21578.

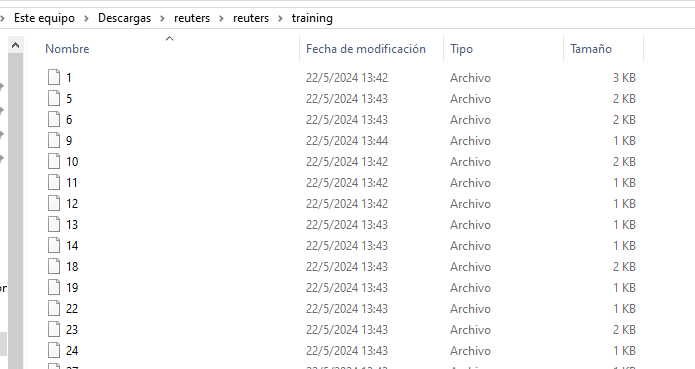
Tareas:

 Descargar el corpus Reuters-21578.

 Descomprimir y organizar los archivos.

 Documentar el proceso de adquisición de datos.





2.2. Preprocesamiento

Objetivo: Limpiar y preparar los datos para su análisis.

Tareas:

 Extraer el contenido relevante de los documentos.

 Realizar limpieza de datos: eliminación de caracteres no deseados, normalización de texto,

etc.

 Tokenización: dividir el texto en palabras o tokens.

 Eliminar stop words y aplicar stemming o lematización.

 Documentar cada paso del preprocesamiento.

2.3. Representación de Datos en Espacio Vectorial

Objetivo: Convertir los textos en una forma que los algoritmos puedan procesar.

Tareas:

 Utilizar técnicas como Bag of Words (BoW) y TF-IDF para vectorizar el texto.

 Evaluar las diferentes técnicas de vectorización.

 Documentar los m´etodos y resultados obtenidos.

2.4. Indexaci´on

Objetivo: Crear un ´ındice que permita búsquedas eficientes.

Tareas:

 Construir un ´ındice invertido que mapee t´erminos a documentos.

 Implementar y optimizar estructuras de datos para el ´ındice.

 Documentar el proceso de construcci´on del ´ındice