Daniel Artavia Cordero
B70771

22/04/2023

Reporte 4

Semana del 15-04-23 al 22-04-23

Esta semana terminé de hacer el programa para extraer datos de elementos HTML. La semana anterior habíamos descartado el uso de Beautifulsoup, sin embargo, esta semana decidimos que era una mejor opción para extraer los datos. Sin embargo, el preprocesamiento del HTML se hace con funciones propias y, al final, BS solo se usa para hacer el parseo de una línea a la vez para obtener datos como las clases, el id, etc.

Adicionalmente, planifiqué cómo se puede integrar con Selenium. También investigué acerca de FastText y Gensim como posibles alternativas a Word2Vec por su mejor soporte para similitud entre oraciones y soporte para situaciones en las que se encuentra con palabras que no estaban en el corpus original.

Plan para la semana del 24-04-23 al 30-04-23
Para la próxima semana pienso integrar el soporte con Selenium para extraer el HTML de los elementos buscados en el archivo original y además crear los diagramas para explicar el problema con más detalles, entradas y salidas de datos y entrenamiento del modelo.