Trabajo Práctico Nº 4

Tema: Unidad 2 - Crawler y Scraper

Fecha Inicio: 07/05/2024 **Fecha de Entrega:** 21/05/2024

Actividades:

1) Planifique, diseñe y construya un crawler para recolectar todas las URLs internas de los primeros 2 niveles de profundidad del sitio web:

   https://eltribunodejujuy.com/

   Deberá crear una hoja de cálculo (Excel o Google SpreadSheet) para almacenar lo recolectado, teniendo en cuenta de identificar las URLs de cada nivel de origen. Evite recolectar URLs repetidas, para ello deberá almacenar de algún modo las URLs que vaya visitando.

2) Realice un web scraping de la siguiente URL:

   https://www.infobae.com/economia/

   De esta URL recolecte las primeras 10 noticias, identificando por cada una el Título, Resumen, Listado de imágenes (ubicación del archivo) y el Cuerpo de la misma. A continuación realice un análisis textual sencillo, tokenize dichos documentos, elimine las stop-words y liste los 100 términos más frecuentes. En el mismo sentido realice un stemming y vuelva a listar los 100 términos más frecuentes.

3) Sería capaz de identificar si alguna de las noticias es muy parecida a otra o están muy relacionadas por la existencia o co-existencia de términos en común?