UNIVERSIDAD
EL BOSQUE

Ingeniería de Sistemas
Big Data Analytics
Taller 2

Enlace de entrega: aquí
Fecha máxima de entrega: **Octubre 22, 2021 11:59 PM**

El *web scraping* y el procesamiento del lenguaje natural son dos de los pilares fundamentales del análisis de información no estructurada a gran escala. En este taller se utilizarán las dos técnicas para entender el entorno colombiano de los últimos tiempos a partir de lo publicado por uno de los principales diarios del país a través de sus sitios web.

Parte 1

Construya un web scraping que le permita extraer de la página de El Espectador las noticias publicadas durante las últimas semanas para las 5 categorías de temas que considere más relevantes.

Transforme esta información a un esquema que considere adecuado para propósitos de almacenamiento en MongoDB. Asegúrese de incluir el atributo que contiene la categoría del artículo así como de persistir en un tipo de dato *date* o *datetime* la fecha de publicación.

No olvide utilizar el paginador para navegar a través del archivo de cada categoría y de esta forma acceder a los artículos que no necesariamente se encuentren actualmente destacados en las diferentes páginas.

Parte 2

Realice un análisis de frecuencia de palabras (sin incluir *stopwords*) para todo el dataset. Muestra un histograma o una nube de palabras con este resultado. Encuentre un

límite adecuado para la cantidad de palabras más frecuentes a ser visualizadas. Tenga en cuenta que usar muchas palabras puede afectar la legibilidad mientras que usar muy pocas palabras puede ocultar información relevante del análisis.

Extienda este ejercicio teniendo en cuenta los siguientes parámetros de comparación:

1. Cambio de las palabras más frecuentes semana a semana.
2. Cambio de las palabras más frecuentes por categorías de temas.
3. Cambio de las palabras más frecuentes por término de búsqueda.

Forma de entrega

Suba al aula virtual, antes de la fecha estipulada, un repositorio de GitHub en el que adjunte los *scripts* con la implementación de los puntos 1 y 2. Exporte y adjunte en formato HTML estos *scripts* con la evidencia de su ejecución. Adjunte también un screenshot con la evidencia de la inserción adecuada de los datos en MongoDB.

Grabe y suba a YouTube un video de no más de 5 minutos con la explicación del trabajo realizado. Utilice el archivo Readme para referenciar el video, listar los integrantes del equipo y mencionar cualquier otro aspecto que desee destacar.