Universidad El Bosque

Facultad de Ingeniería

Programa de Ingeniería de Sistemas

Big Data Analytics

Taller 2 – Web Scrapping

Laura María López Moreno

Link de YouTube: https://youtu.be/0SG1eXLm4bc

Este documento tiene como fin presentar los resultados y evidencias para el taller dos de la materia Big Data Analytics. Se busca realizar web scraping de forma automática para la página del El Espectador para extraer los artículos más importantes de la semana. Mediante la captura del HTML y la utilización de parsers se aplica esta técnica.

Parte 1.

Este taller se realizó mediante Google Colaborate, donde se instalaron la librerías necesarias.

Texto

Descripción generada automáticamente

Mediante drive se manejaron los archivos, así mismo, debido a la dificultad de conectar el mongodb instalado en la máquina local se creo en atlas un host para guardar la base de datos de mongodb.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Luego, se continuo probando si funcionaba la conexión a la base de datos

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Durante la clase, se realizo la creación del navegador dedicado por firefox, en colab eso no funciono, por ello nos fuimos por el lado de chrome

Texto

Descripción generada automáticamente

Lo siguiente hace parte de lo visto en clase:

Texto

Descripción generada automáticamente

Y ya la razon y ser de esta parte del taller se encuentra a continuación:

Texto

Descripción generada automáticamente

Debido a la necesidad de implicar la paginación y por tanto la parte dinámica se crearon for anidados y se expandieron las categorias.

Texto

Descripción generada automáticamente con confianza media

De lo anterior, es similar a lo visto en clase, donde, a partir de los identificadores de los contenedores de la páginas web se recopila la información que se requiere para este análisis de web scraping.

Por ultimo de esta sección, se recopila el texto completo y se sube a la base de datos.

Texto, Carta

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Parte 2.

Análisis del texto

Se instala wordcloud y demás dependencias y librearias necesarias:

Texto

Descripción generada automáticamente

Se conecta a la base de datos:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Luego a partir del texto se recopilan sus respectivas tokens Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Así mismo, se buscan cuales son las stopwords en el idioma español y se empieza a realizar el histograma correpondiente con las palabras más frecuentes que este caso son alto, turmequé, apuros, aclaraciones, otros y también se realiza su correspondiente nube de palabras.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente Imagen que contiene Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente