DOCUMENTACION

PRUEBA 1: Creación de “Web Scrapper” utilizando Python

**OBJETIVO**

Automatizar la obtención y evaluación de links de noticias, a través del uso de consultas directas a motores de búsqueda.

**DOCUMENTACION EXTERNA**

1. Como input, la aplicación recibe un archivo XML. Dicho archivo XML contiene los criterios de búsqueda (keywords) entregados por el evaluador.
2. Se ejecuta el archivo llamado WebScrapper, este programa se encarga de leer los criterios de búsqueda desde el archivo Excel llamado “Anexo 1.1 INPUT”; luego realiza la obtención de URL de noticias de los motores de búsqueda Google, Yahoo, Bing y NewsLookup; finalmente escribe estos URL con sus respectivos criterios en el archivo de Excel llamado “Anexo 1.2 OUTPUT”
3. Como output, la aplicación crea un archivo de tipo Microsoft Excel, el cual contiene el resultado del análisis realizado por la aplicación

**DOCUMENTACION INTERNA**

1. Primero se ejecuta la función ScrappNews en la clase WebScrapper , que es el encargado de realizar la obtención de URL de noticias según los criterios de búsqueda dados en el Excel de input.
2. El archivo Excel de input está especificado en el constructor bajo el nombre fileNameInput, para leerlo se crea un objeto de la clase XlsxAccess y se llama a la función read. En la función read se obtiene la información de las columnas A y B (criterios de búsqueda), posteriormente devuelve una lista con la información extraída del Excel.
3. Para cada conjunto de criterios de búsqueda, se crea objetos de las clases Scrapper de cada uno de los motores de busqueda Google, Yahoo, Bing y NewsLookup (GoogleScrapper, YahooScrapper, BingScrapper, NewsLookupScrapper) para obtener los URL de noticias. En cada una de estas clases se llama a la función scrapp\_news, que se encarga de hacer un pedido a los motores de búsqueda con los criterios especificados, luego se buscan los elementos que contienen el url de las noticias (este suele cambiar según el motor de búsqueda) y para finalizar se devuelve una lista con los url encontrados.
4. A continuación se crea una lista que relaciona los links con sus respectivos criterios de búsqueda llamada newDataList.
5. Finalmente se escriben los URL encontrados con sus respectivos criterios en el archivo de Excel que está especificado en el constructor bajo el nombre fileNameOutput. Para esto se crea un nuevo objeto de la clase XlsxAccess y se llama la función write enviando la lista de títulos y la lista newDataList. En la función write, se crea el archivo Excel y una hoja nueva, en ella se agregan los títulos y luego se itera a través de la lista enviada con los datos y se agregan a la hoja.

**LIBRERIAS**

BeautifulSoup: Esta es una librería que permite obtener datos desde un html, en este caso se busca elementos específicos y los estructura, para que sean de fácil acceso y navegación. Requests: Esta es una librería que sirve para obtener el archivo html a partir de un URL y con las cabeceras especificadas.

Xlsxwriter: Es una librería que permite crear archivos Excel y escribir en ellos de forma segura.

Xlrd: Es una librería que se utiliza para obtener información desde un archivo Excel.