Unidad 1 - Extracción y Procesamiento de Texto

Para todos los ejercicios de web scraping contemple usar un temporizador para no matar los sitios con las consultas.

Ejercicio 0 - En el campus se encuentran dos archivos en la carpeta “Ejemplos codificación”. Investíguelos con las herramientas vistas en clase. Puede usar también la herramienta hexdump para verlos por dentro.

Pruebe intencionalmente abrir el archivo con la codificación equivocada (abra el UTF-8 como 8859-15) y compare las salidas.

Ejercicio 1 - Con los archivos pedidos en diferentes formatos (pdf, jpg, …otros) la clase anterior, extraiga la información que considere relevante de los mismos utilizando las librerías desarrolladas en clase.

i. Obtenga texto de un libro escaneado en pdf (un pdf que no tenga el texto codificado como tal, es decir, que sea necesario usar ocr). Pruebe usando pytesseract en ingles (por defecto) y luego configurándolo en español.
ii. Obtenga texto de una imagen (png, bmp).
iii. Obtenga texto de un archivo word.
iv. Obtenga texto de un archivo de audio.

Si no tiene suficientes recursos considere usar las páginas archive.org, epdlp.com, ciudadseva.com para conseguir los archivos.

Opcional: Para las imágenes de facturas y recibos, extraiga información de los mismos con el objetivo de identificar los movimientos de gasto para el/los período/s a que se correspondan, y genere una salida por pantalla con un resumen para el total de movimientos.

Ejercicio 2 - El Ministerio de Turismo y Deportes de la Nación permite explorar tableros de información en línea tableros.yvera.tur.ar. Explore la página y utilizando

una librería de web scraping extraiga los valores del tablero de indicadores de Objetivos de Desarrollo Sostenible en una tabla y el texto limpio de la metodología de los mismos.

Ejercicio 3 - Investigue cómo extraer información de vuelo en la página flybondi.com, haga un programa en python para verificar si bajó el precio de un vuelo determinado.

Ejercicio 4 - Con el objetivo de estudiar la serie de tipo de cambio oficial y su volatilidad de forma visual. Utilice una librería de web scraping y elija una dirección que posea esta información en una frecuencia diaria y recolecte esa información (por ejemplo dolarhoy.com). Documente el proceso debidamente.

Opcional: Una vez obtenidos los datos requeridos realice una gráfica que muestre la serie y calcule los valores promedio mensual de la serie. Su máximo valor y su mínimo. Presente los datos por pantalla en forma ordenada.

Ejercicio 5 - Obtener los libros en formato pdf de la página del autor argentino Hernán Casciari de la editorial Orsai, estos textos son de libre utilización. Pasar a un archivo pkl el texto de los libros en pdf.

Ejercicio 6 - Realizar en Colab un script Python que haga los siguientes pasos:

i. Obtener el HTML de la página https://es.wikipedia.org/wiki/Argentina

ii. Con BeautifulSoup, Obtener el texto contenido en el `<div>` cuyo `id='mw-content-text'`

iii. Aplicar RecursiveCharacterTextSplitter con los parámetros `chunk_size = 300` en el texto de wikipedia.

iv. Aplicar RecursiveCharacterTextSplitter con los parámetros `chunk_size = 300, separators = ["\n\n", "\n"]` en el texto de wikipedia.

v. Aplicar CharacterTextSplitter() en el texto de wikipedia.

vi. En cada caso, graficar el histograma de como se distribuyen los tamaños de los segmentos (chunks) y sacar conclusiones.

Ejercicio 7. En la siguiente carpeta, https://drive.google.com/drive/folders/1iCiOQ8P8CHFFLKiWc2li9-x3Ud-Tl0po?

usp=drive_link

podrá descargar recursos que presentan algunos desafios para la extracción de texto. Practique los códigos y librerías propuestos en la Unidad 1 y compare los resultados con sus compañer@s.