

Técnicas y procesos para extraer información de los datos

En este ejercicio, trabajarás con <https://www.20minutos.es/ultimas-noticias/> que contiene artículos sobre diversos temas. Deberás acceder al sitio web, extraer información de los artículos y aplicar una variedad de técnicas de análisis de datos, visualización, minería de texto, procesamiento de imágenes y extracción de datos.

Acceso y Extracción de Datos del Sitio Web

Utiliza las bibliotecas requests, BeautifulSoup y Scrapy para acceder a un sitio web de noticias de tu elección. Asegúrate de obtener los datos HTML de las páginas que contienen artículos.

Extrae la información relevante de los artículos, como el título, el contenido del artículo, la fecha de publicación y cualquier otro atributo que consideres importante. Almacena estos datos en una estructura adecuada, como un DataFrame de pandas.

Análisis Estadístico y Pruebas de Hipótesis

Realiza un análisis estadístico de los datos recopilados, incluyendo medidas como la media, la mediana, la moda y la desviación estándar de la longitud de los artículos, la frecuencia de palabras clave, etc.

Realiza pruebas de hipótesis para determinar si existen diferencias significativas en la longitud de los artículos entre diferentes categorías o temas.

Visualización de Datos

Crea gráficos y visualizaciones para representar datos clave, como la distribución de temas en los artículos, la evolución de la longitud promedio

de los artículos a lo largo del tiempo, y la frecuencia de palabras clave importantes.

Genera mapas de calor que muestran la relación entre temas y palabras clave en los artículos.

Minería de Texto

Realiza la clasificación de documentos para categorizar los artículos en diferentes temas o categorías.

Realiza análisis de temas en los artículos para identificar los temas principales discutidos en el sitio web de noticias.

Implementa una función de búsqueda de información que permita a los usuarios encontrar artículos relevantes según palabras clave.

Procesamiento de Imágenes

Si los artículos contienen imágenes, utiliza técnicas de procesamiento de imágenes para detectar y localizar objetos específicos en las imágenes.

Aplica el reconocimiento de patrones para identificar características únicas en las imágenes relacionadas con los artículos.

Si los artículos incluyen secuencias de imágenes (por ejemplo, una galería de imágenes), realiza el seguimiento de objetos en las secuencias de imágenes para rastrear la ubicación de objetos a lo largo del tiempo.