Proyecto Final
DS3021 Análisis Computacional de Datos

Parte I: Adquisición de Datos

I. Objetivo

- Aplicar técnicas de adquisición de datos para el proceso de análisis computacional de datos

II. Instrucciones

Para este proyecto usaremos como fuente una página peruana y que tenga información relevante para llevar a cabo un proceso de web scraping. Puede tomar como referencia alguna de las mostradas abajo u otra y realice lo siguiente:

i. ***Crear un script en Python*** para extraer datos de la página web seleccionada. El script debe:
   - Utilizar técnicas apropiadas de web scraping (BeautifulSoup, Scrapy, Selenium o APIs)
   - Extraer al menos 6 atributos diferentes relevantes para análisis (debe contener tanto atributos numéricos como categóricos)
   - Manejar adecuadamente paginación si es necesario
   - Incluir al menos 1000 registros en el dataset final

ii. ***Exportar el dataset*** como archivo CSV con codificación UTF-8

iii. ***Crear un diccionario de datos*** que incluya:
   - Nombre de cada variable/columna
   - Tipo de dato
   - Descripción breve del contenido
   - Ejemplo de valores

III. Páginas sugeridas

1. Noticias y Medios de Comunicación

Página	*URL*	*Posibles datos a extraer*
El Comercio	https://elcomercio.pe/	Titulares, resúmenes, fechas, autores, secciones
La República	https://larepublica.pe/	Noticias por categoría, fecha, contenido

Perú21	https://peru21.pe/	Últimas noticias, política, deportes
Gestión	https://gestion.pe/	Noticias económicas y financieras
Ojo Público	https://ojo-publico.com/	Investigaciones y periodismo de datos

2. Portales de Empleo

Página	*URL*	*Posibles datos a extraer*
Bumeran Perú	https://www.bumeran.com.pe/	Cargo, empresa, salario, requisitos
Computrabajo Perú	https://www.computrabajo.com.pe/	Descripción de empleos, ciudad, categoría
Indeed Perú	https://pe.indeed.com/	Puesto, empresa, ubicación, sueldo (si aplica)

3. Comparación de Precios / Productos

Página	*URL*	*Posibles datos a extraer*
Falabella Perú	https://www.falabella.com.pe/	Precio, nombre del producto, categoría
Plaza Vea	https://www.plazavea.com.pe/	Precios de alimentos y otros productos
Linio Perú	https://www.linio.com.pe/	Tecnología, electrodomésticos, categorías y precios
Mercado Libre Perú	https://www.mercadolibre.com.pe/	Publicaciones de productos, precios, descripción

4. Clima y Transporte

Página	*URL*	*Posibles datos a extraer*
Senamhi	https://www.senamhi.gob.pe/	Temperaturas, precipitaciones, pronóstico
Metropolitano Lima	https://www.metropolitano.com.pe/	Horarios, estaciones, rutas
ATU (Autoridad de Transporte Urbano)	https://www.atu.gob.pe/	Líneas, paraderos, proyectos viales

5. Datos Gubernamentales

Página	*URL*	*Posibles datos a extraer*
Superintendencia Nacional de Educación (Sunedu)	https://www.sunedu.gob.pe/	Universidades, licenciamiento, datos de programas
Sistema de Consulta de RUC (SUNAT)	https://e-consultaruc.sunat.gob.pe/	Razón social, estado del RUC, actividad económica
EsSalud Transparencia	https://www.essalud.gob.pe/transparencia/	Contratos, servicios médicos, reportes

IV. Consideraciones Importantes

- ***Respetar robots.txt:*** Verifique las políticas del sitio antes de scraping
- ***Ética y legalidad:*** No extraer información personal sensible
- ***Rendimiento:*** Usar 'time.sleep()' entre requests para no saturar servidores
- ***Headers:*** Utilizar User-Agent realista para evitar bloqueos
- ***Manejo de errores:*** Implementar 'try-except' para conexiones fallidas
- ***Almacenamiento:*** Guardar datos de forma incremental para evitar pérdidas

V. Entregables y Deadlines

- **Deadline:** 8 de noviembre de 2025
- **Entregables:**
  - Dataset .csv
  - Script .ipynb (documentado)
  - Diccionario de Datos
- El peso de esta primera parte de su proyecto es de 30% de la nota final de Proyecto P.