TP Final UBA: Opción 2 – Scraping y Análisis

---

1. Introducción y Objetivos

- **Objetivo general:** Comparar datos de al menos dos portales de una categoría mediante scraping y análisis.
- **Equipos:** 2 integrantes.
- **Entrega:** Presentación PPT exportada a PDF que incluya resultados y metodología.

---

2. Organización del Repositorio

```
/raw         ← Datos crudos (HTML/JSON descargado)
/input       ← Datos preprocesados listos para análisis (CSV, Parquet)
/output      ← Resultados finales (gráficos, estadísticas, modelos)
/scripts     ← Código ordenado por numeración:
    ├─ 01_extraccion.py
    ├─ 02_limpieza.py
    ├─ 03_eda.py
    └─ 04_visualizacion.py
README.md    ← Objetivos, estructura y guía rápida de ejecución
```

- Crear repo en GitHub y agregar colaborador.
- Incluir enlaces y versión de Python/librerías.

---

3. Selección de Categoría y Sitios

- Escoger una de las categorías (Libros, Tecnología, Alimentos o Celulares).
- Elegir al menos dos sitios para comparar.
  - **Libros:** Cúspide, Yenny, Buscalibre
  - **Tecnología:** Gadnic, Bidcom, ProvinciaCompras, Frávega, Megatone
  - **Alimentos:** La Cooperativa Obrera, Jumbo
  - **Celulares:** Tienda Claro, Personal, Frávega

---

4. Extracción de Datos (Scraping)

1. **Mapeo inicial de URLs**
   - Identificar páginas de listado y detalle de producto.
   - Localizar selectores CSS o XPaths necesarios.
2. **Herramientas**
   - `requests` + `BeautifulSoup` para HTML estático.
   - `Selenium` o `Playwright` para contenido dinámico y paginación infinita.
3. **Implementación**
   - Iterar sobre páginas de resultados (paginación).
   - Extraer: nombre, precio, cuotas, disponibilidad, enlace e imagen.
   - Almacenar JSON/HTML crudo en `/raw`.
4. **Registro de dificultades**
   - Bloqueos (robots.txt, CAPTCHAs).

- Cambios de layout.
- Manejo de tiempo de espera y rendimiento.

---

5. Limpieza y Unificación de Datos

- **Normalización de texto:**
  - Eliminar mayúsculas inconsistentes y caracteres especiales.
  - Alinear nombres de producto entre sitios (mapeo manual o fuzzy matching).
- **Conversión de tipos:**
  - Precios a `float`, fechas a formato estándar.
  - Moneda `ARS` unificada.
- **Salida limpia:**
  - Guardar DataFrame limpio en CSV o Parquet en `/input`.

---

6. Análisis Exploratorio de Datos (EDA)

- **Estadísticas descriptivas:**
  - Mínimo, percentiles (25%, 50%, 75%) y máximo de precios por sitio.
  - Conteo de productos y cuotas sin interés.
- **Comparaciones cruzadas:**
  - Precio promedio de productos compartidos.
  - Diferencia absoluta y relativa de precios.
- **Detección de outliers:**
  - Identificar precios atípicos para revisión.

(Script: ``)

---

7. Visualización de Resultados

- **Gráficos clave:**
  - Boxplot comparativo de precios.
  - Barras de porcentaje de productos con cuotas.
  - Heatmap de diferencias de precio entre sitios.
- **Estética y justificación:**
  - Colores, etiquetas y títulos claros.
  - Evitar slides sobrecargadas.

(Script: ``)

---

8. Armado de la Presentación (PPT → PDF)

1. **Introducción y objetivos**
2. **Descripción de los datos**
   - Fuentes, cantidad de ítems y resumen de scraping.

3. **Metodología**
    - Herramientas, scripts y pasos críticos.
4. **EDA**
    - Hallazgos descriptivos.
5. **Resultados**
    - Gráficos principales y comparaciones.
6. **Conclusiones**
    - Portal más económico y diferencias clave.
7. **Limitaciones y trabajo futuro**
    - Robustez del scraper y fuentes adicionales.

Mantener cada slide limpio: texto breve en bullets + gráfico relevante.

---

9. División de Tareas y Cronograma

Tarea	Responsable	Plazo
Configurar repo & README	Integrante A	Día 1
Extracción (raw)	Integrante B	Día 3
Limpieza (input)	Integrante A	Día 5
EDA y estadísticas	Integrante B	Día 7
Visualización & gráficos	Integrante A	Día 9
Montaje PPT final	Ambos	Día 12 (04/07)

¡Listo para exportar a PDF y entregar!