Propuesta de Extracción de Información de Egresados en LinkedIn utilizando Scrapy, ScrapeOps y MySQL

Introducción:

La siguiente propuesta tiene como objetivo desarrollar un proceso automatizado para la extracción de información de egresados de la Universidad Nacional desde la plataforma LinkedIn. El enfoque involucra el uso de Scrapy y ScrapeOps para evitar bloqueos al realizar web scraping, y la implementación de una base de datos MySQL para almacenar y verificar los contactos previamente obtenidos.

Flujo del Proceso:

1. **Configuración Inicial:**
   Se establecerá un entorno virtual donde se instalarán las bibliotecas requeridas, incluyendo Scrapy, ScrapeOps y MySQL Connector.
2. **Extracción Inicial de Información:**
   Utilizando Scrapy y ScrapeOps, se obtendrá información de perfiles de docentes actuales de la Universidad Nacional en LinkedIn. ScrapeOps asegurará una extracción eficiente y sin bloqueos.
3. **Almacenamiento y Verificación en MySQL:**
   La información extraída se almacenará en una base de datos MySQL. Antes de realizar nuevas extracciones, se verificará si un contacto ya existe en la base de datos para evitar duplicados.
4. **Iteración y Obtención de Nuevos Contactos:**
   Usando los nuevos contactos obtenidos, se iterará el proceso para extraer información adicional y expandir la base de datos.
5. **Extracción de la Sección de Educación:**
   Se extraerá la sección de educación de los perfiles de los contactos para identificar instituciones educativas.
6. **Filtrado por Universidad Nacional:**
   La información de la sección de educación se procesará para filtrar perfiles que hayan estudiado en la Universidad Nacional, identificándolos como egresados.

Ventajas de la Propuesta:

- **Evitar Bloqueos:** La combinación de Scrapy y ScrapeOps permite el scraping eficiente y evita bloqueos por parte de LinkedIn.
- **Gestión de Duplicados:** La base de datos MySQL actúa como un registro de contactos obtenidos, evitando la extracción repetida de perfiles ya almacenados.
- **Iteración Eficiente:** La estrategia de iteración permite ampliar la base de datos sin duplicar información.
- **Selección de Egresados:** El filtrado por institución educativa asegura que solo se almacene información relevante de egresados de la Universidad Nacional.

Conclusiones:

La propuesta de combinar Scrapy, ScrapeOps y MySQL para la extracción de información de egresados en LinkedIn proporciona una solución sólida y completa. La capacidad de evitar bloqueos y gestionar duplicados mediante la base de datos MySQL asegura la integridad de los datos obtenidos. Además, el filtrado por universidad permitirá identificar y registrar a los egresados de la Universidad Nacional, generando una base de datos valiosa para futuros análisis y seguimientos.