Práctica 3 - Adquisición e Integración de datos usando MongoDB y Python - 1

Se dispone de 2 fuentes de datos con información sobre personas y sus amistades en distintos formatos abiertos. Tales fuentes de datos deben ser descargadas desde el aula virtual de la asignatura.

El objetivo de la práctica es recopilar la información disponible en cada una de las fuentes usando Python, homogeneizar y depurar los datos y por último integrarlos en un único repositorio común en MongoDB usando el módulo PyMongo.

Las dificultades a tener en cuenta y que deben ser resueltas de forma apropiada y lo más automatizado posible, al menos, son:

- Las fuentes de datos, a pesar de que básicamente almacenan la misma información, tienen estructuras y formatos de datos distintos.
- Los atributos almacenados en cada fuente no son exactamente iguales (nombre, posición, formato, …).
- Pueden haber errores en los datos (al comparar las fuentes) o datos faltantes.
- Pueden existir atributos de una fuente que no existen en la otra y viceversa.
- Puede haber información duplicada entre las fuentes.
- Una fuente se considera que completa a la otra y viceversa.
- Debería poderse conocer de qué fuente provienen los datos almacenados en el repositorio común.

Los entregables de esta práctica son:

- El código Python comentado
- Un informe en formato pdf en el que se describan cada uno de los problemas relacionados con la integración de datos de estas fuentes y cuál ha sido el proceso de depuración seguido. En dicho informe debe aparecer, al final, el contenido completo del repositorio creado en Mongo.