Asignatura: Adquisición y preparación de datos
Grado en Ingeniería en Inteligencia Artificial
Universidad de Alicante - Escuela Politécnica Superior

Introducción

Esta práctica tiene como objetivo la reutilización de contenidos en forma de datos por parte de instituciones públicas, organismos de patrimonio cultural, asociaciones o infraestructuras europeas. Los contenidos y datos proporcionados son de diferente tipo y formato, e incluyen información relacionada con la movilidad, patrimonio cultural o turismo. En algunos casos, los datos están disponibles en forma de carpeta comprimida (p.ej., zip) mientras que en otros pueden estar disponibles a través de un API.

Resumen de los apartados

1. Selecciona una temática: patrimonio cultural, turismo, movilidad, medio ambiente, etc. Identifica algunas preguntas que puedas realizar para solventar un problema en concreto. Por ejemplo, ¿en qué medida ha mejorado Alicante en materia de movilidad en los últimos años? ¿Qué atracciones turísticas se encuentran disponibles en Alicante y cómo podemos visualizarlas de forma atractiva?
2. Selecciona uno o varios conjuntos de datos teniendo en cuenta la temática seleccionada que permita responder en cierta medida las preguntas identificadas.

Puede darse el caso que tengas que redefinir las preguntas para ajustarla a los datos disponibles. Describe los cambios y actualizaciones realizadas.

3. Realiza el diseño conceptual, lógico y físico del almacén de datos de la temática elegida.
4. Extrae y prepara los datos para su reutilización. Para ello, tendrás que revisar la información proporcionada, integrar diferentes conjuntos de datos o enriquecer los datos con repositorios externos (e.g., incluir información adicional como latitud y longitud). Para este apartado puedes utilizar herramientas de limpiezas de datos o directamente con código.
5. Transforma los datos (o una parte de ellos) a tripletas utilizando el vocabulario schema.org. Comprueba que los datos generados son correctos. Es posible utilizar cualquier lenguaje de programación como Java[1] o Python,[2] como también herramientas existentes como OpenRefine.[3] Se valorará positivamente el enriquecimiento de los datos con repositorios externos como Wikidata y vocabularios controlados.
6. Implementa dos visualizaciones que permitan responder a las preguntas identificadas en el punto 1. Las visualizaciones pueden consistir en mapas, gráficas o líneas de tiempo. Es posible realizar las visualizaciones con vuestro propio código o utilizando servicios de terceros.
7. Crea una memoria (documento Word o PDF) que describa brevemente los pasos realizados e incluye el nombre de los autores en la portada. Durante las últimas sesiones de teoría y práctica se realizará la presentación del trabajo en clase por parte del grupo.
8. Crea un repositorio GitHub que incluya los datos y diferentes transformaciones, un fichero markdown README.md describiendo brevemente el trabajo realizado, y las visualizaciones y código creado.

Evaluación

El trabajo debe realizarse por **grupos de 4 personas.** Los criterios de evaluación son los siguientes:

- Apartado 1: 10%
- Apartado 2: 10%
- Apartado 3: 15%
- Apartado 4: 20%
- Apartado 5: 10%
- Apartado 6: 10%
- Apartado 7: 10%
- Apartado 8: 5%

[1] https://jena.apache.org/
[2] https://rdflib.readthedocs.io/en/stable/
[3] https://openrefine.org/

Entrega

La entrega debe incluir los datos, el código generado y la memoria describiendo el trabajo realizado. La entrega se realizará a través de moodle en UAcloud.

Fecha de entrega
22 de diciembre de 2025

Notas adicionales

El enunciado puede ser modificado brevemente durante el curso por lo que recomendamos realizar la descarga del documento regularmente. El profesor comentará en clase si el enunciado ha sido modificado.

Apartado 1: Definición del proyecto centrado en los datos

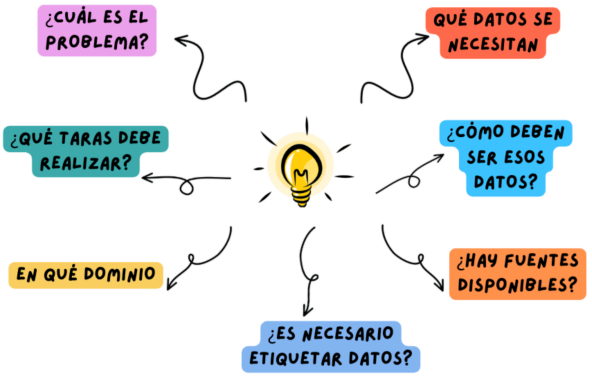
Evaluación del apartado 1: 10%

Descripción
Selecciona una temática: patrimonio cultural, turismo, movilidad, medio ambiente, etc. Identifica algunas preguntas que puedas realizar para solventar un problema en concreto. Por ejemplo, ¿en qué medida ha mejorado Alicante en materia de movilidad en los últimos años? ¿Qué atracciones turísticas se encuentran disponibles en Alicante y cómo podemos visualizarlas de forma atractiva?

Este apartado está compuesto por las siguientes tareas que deben ser debidamente cumplimentadas y justificadas en el documento final de la entrega. Durante las prácticas se consensuará con el profesorado de la asignatura.

Tareas

- Selección de la temática o dominio en el que se desarrollará el proyecto
- Identificar el problema a resolver
    - Lienzo del problema (4W)
        - Who, quiénes son las partes interesadas
        - What, cuál es el problema o necesidad y cuál es su naturaleza
        - Where, la situación, momento y contexto dónde se produce
        - Why, cuál es el beneficio y su impacto, tanto en las partes interesadas como en la sociedad
- Marcar objetivos del proyecto
- Casos de uso
- Definir métricas clave y evaluar en función de los objetivos y casos de uso



Apartado 2: Analizar y evaluar necesidades de datos

Evaluación del apartado 2: 10%

Descripción
Selecciona varios conjuntos de datos teniendo en cuenta la temática seleccionada que permita responder en cierta medida las preguntas identificadas. Puede darse el caso que tengas que redefinir las preguntas para ajustarla a los datos disponibles. Describe los cambios y actualizaciones realizadas.

Tareas

- Analizar y evaluar necesidades de datos
    - Qué datos son necesario: tipo, formato
    - Cómo deben ser esos datos
    - Analizar las fuentes disponibles
    - Si fuera necesario utilizar datos ficticios, justificarlo adecuadamente.
- Selecciona varios conjuntos de datos teniendo en cuenta la temática seleccionada y el análisis realizado anteriormente.

Apartado 3: Realizar el diseño conceptual, lógico y físico del almacén de datos

Evaluación del apartado 3: 15%

Descripción
Realiza el diseño conceptual, lógico y físico del almacén de datos de la temática elegida.
Se debe adjuntar la documentación necesaria para justificar este apartado en ella se debe incluir los esquemas así con el script del almacén resultante (esquema y datos).

Tareas
- Diseño conceptual
- Diseño lógico
- Diseño físico

Apartado 4: Limpiar, transformar y normalizar los datos

Evaluación del apartado 4: 20%

Descripción

Extrae y prepara los datos para su reutilización. Para ello, tendrás que revisar la información proporcionada, integrar diferentes conjuntos de datos o enriquecer los datos con repositorios externos (e.g., incluir información adicional como latitud y longitud). Para este apartado puedes utilizar herramientas de limpiezas de datos o directamente con código.

Tareas
1. Utilizando la herramienta Pentaho Data Integration diseñar las transformaciones necesarias para la extracción, limpieza, transformación y normalización de los datos.

    - Conexión con fuentes de datos (BD, ficheros, web)
    - Realizar fusión de varias fuentes de datos
    - Seleccionar campos, filtrar filas, ordenar.

2. Utilizando la herramienta Pentaho Data Integration se realizarán las transformaciones necesarias para llevar a cabo las siguientes tareas obligatorias:
    - Corregir de errores
    - Gestionar los valores que faltan
    - Eliminar o minimizar la redundancia de datos
    - Feature Engineering, transformación y creación de nuevas características. Es obligatorio utilizar alguna de esta técnicas para la creación de nuevas características:

    - Creación de nuevas características: combinar, transformar, extraer o derivar nuevas, por ejemplo Splitting
    - Codificación de variables categóricas, por ejemplo One-Hot Encoding
    - Transforman variables numéricas continuas en categorías, Binning
- Enriquecimiento de datos
- Tratamiento de los Outliers
- Validación de datos

3. Diseñar un flujo de trabajo (a través de job de Pentaho) que gestione las distintas transformaciones realizadas.

*Si se considera necesario, se puede generar código adicional en Python para llevar a cabo algunas de las tareas.

Apartado 5: Transformar los datos

Evaluación del apartado: 10%

Descripción
Transforma los datos (o una parte de ellos) a tripletas utilizando el vocabulario schema.org.

En primer lugar, debéis realizar un análisis de los datos que queréis transformar y un diseño de clases y propiedades a utilizar para describir la información. Para ello debéis consultar las clases y propiedades disponibles como, por ejemplo, Person, Event, Place, Restaurant, Book o Movie. Dentro de cada clase podéis consultar las propiedades que podéis utilizar para describir cada uno de los recursos.

Es posible utilizar cualquier lenguaje de programación como Java[4] o Python,[5] como también herramientas existentes como OpenRefine.[6] Se valorará positivamente el enriquecimiento de los datos con repositorios externos como Wikidata y vocabularios controlados.

Una vez generados los datos, comprueba que los datos generados son correctos. Algunas comprobaciones sencillas pueden ser el número total de registros generados o las distintas clases y propiedades utilizadas para describir la información.

[4] https://jena.apache.org/
[5] https://rdflib.readthedocs.io/en/stable/
[6] https://openrefine.org/

Apartado 6: Visualización

Evaluación del apartado: 10%

Descripción y tareas
Implementa dos visualizaciones que permitan responder a las preguntas identificadas en el punto 1. Las visualizaciones pueden consistir en mapas, gráficas o líneas de tiempo. Es posible realizar las visualizaciones con vuestro propio código o utilizando servicios de terceros como APIs.

Apartado 7: Memoria y presentación

Evaluación del apartado: 10%

Descripción y tareas
Crea una memoria (documento Word o PDF) de máximo 8 páginas que incluya el nombre de los autores en la portada y describa brevemente los pasos realizados para cada apartado.
Durante las últimas sesiones de teoría y práctica se realizará la presentación del trabajo en clase por parte de todos los integrantes del grupo.

Apartado 8: Repositorio de código

Evaluación del apartado : 5%

Descripción y tareas
Crea un repositorio GitHub que incluya los datos y diferentes transformaciones, un fichero markdown README.md describiendo brevemente el trabajo realizado, así como las visualizaciones y código creado. Añade una licencia de uso[7] y referencias que hayáis utilizado para realizar el trabajo.

[7] https://creativecommons.org/share-your-work/