Guía docente
270956 - SDM - Gestión de Datos Semánticos

Última modificación: 31/01/2024

Unidad responsable: Facultad de Informática de Barcelona
Unidad que imparte: 747 - ESSI - Departamento de Ingenieria de Servicios y Sistemas de Información.

Titulación: MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (Plan 2021). (Asignatura obligatoria).

Curso: 2023 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: OSCAR ROMERO MORAL

Otros: Segon quadrimestre:
OSCAR ROMERO MORAL - 11, 12

CAPACIDADES PREVIAS

The student must be familiar with basics on databases and data modeling. Advanced programming skills are mandatory.

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE3. Aplicar métodos de integración de datos para dar solución a problemas de ciencia de datos en entornos heterogéneos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos

Genéricas:
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos

Transversales:
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

Básicas:
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB8. Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.
CB9. Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

METODOLOGÍAS DOCENTES

El curso tiene sesiones magistrales y de laboratorio.

Magistrales: El profesor expone el tema. Los estudiantes siguen la lección. toman apuntes y preparan material adicional fuera de clase. También se les pide que participen en actividades evaluatorias durante estas sesiones.

Laboratorio: Principalmente, las sesiones de laboratorio estan dedicadas a la práctica (con o sin ordenador) de los conceptos introducidos en las sesiones magistrales. Herramientas relevantes para dichos conceptos se presentan en estas sesiones y se pide llevar a cabo pequeños proyectos en que haya que usarlas.

Proyecto: El proyecto del curso pone en práctica los conocimientos del curso en un entorno realista.

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

1.Learn, understand and apply the fundamentals of property graphs
2.Learn, understand and apply the fundamentals of knowledge graphs
3.Perform graph data processing both in centralized and distributed environments
4.Integrate, combine and refine semi-structured or non-structured data using graph formalisms
5.Determine how to apply graph formalisms to solve the Variety challenge (data integration)
6.Apply property or knowledge graphs to solve realistic problems such as data integration, graph-based data analysis, etc.

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas grupo pequeño	27,0	18.00
Horas aprendizaje autónomo	96,0	64.00
Horas grupo grande	27,0	18.00

Dedicación total: 150 h

CONTENIDOS

Introducción y formalización del concepto de Variedad en Big Data y su gestión

Descripción:
Definición de las tareas de gestión de datos: desde la perspectiva de las bases de datos y de la representación del conocimiento.

Definición de Variedad en el concepto de Big Data. Heterogeneidades sintácticas y semánticas. Efecto de las heterogeneidades de datos en las diferentes tareas de gestión de datos identificadas.

Concepto de integración de datos. Definición de un marco teórico para la gestión e integración de fuentes de datos heterogéneas.

Principales componentes de un sistema de integración de datos: fuentes, esquema global o de integración y mappings.

La necesidad de un modelo de datos canónico para la integración de datos. Definición de modelo de datos. Características esenciales de los modelos de datos canónicos.

Gestión de los property graphs

Descripción:
Estructuras de datos. Restricciones de integridad del modelo.

Operaciones básicas. Basadas en la topología, en el contenido y mixtas.

Lenguajes de consulta de datos en grafo. GraphQL.

Concepto de graph database. Heterogeneidad de las herramientas a la hora de implementar las estructuras internas del grafo. Impacto de dichas decisiones en las operaciones básicas presentadas.

Distributed graph database. Necesidad. Dificultades. Paradigma thinking like a vertex como estándar de facto para procesamiento de grafos distribuidos.

Principales algoritmos de procesamiento distribuido.

Gestión de los knowledge graphs

Descripción:
Estructuras de datos: lenguajes RDF. Origen y relación con Linked Open Data. Restricciones de integridad de cada lenguaje.

Estructuras de datos: RDFS y OWL. Relación con la lógica de primer orden. Fundamentos en Description Logics. Restricciones de integridad de cada lenguaje. Concepto de razonamiento.

Operaciones básicas y lenguaje de consulta. SPARQL y álgebra subyacente. Entailment regimes (razonamiento).

Concepto de triplestore. Diferencias con una graph database. Implementaciones nativas. Implementaciones basadas en el modelo relacional. Impacto de dichas decisiones en las operaciones básicas presentadas.

Distributed triplestore. Necesidad. Dificultades. Graph Engine 1.0 como paradigma de triplestore distribuido.

Principales algoritmos distribuidos.

Los grafos como solución a la gestión de la variedad

Descripción:
La idoneidad de los grafos como modelo de datos canónico en sistemas de integración de datos.

Principales características de los modelos de datos en grafo. Diferencia con otros modelos de datos (especialmente con el modelo relacional).

Concepto de dato y metadato y su formalización en los modelos de grafos.

Casos de uso (especial énfasis en los beneficios topológicos de los grafos): detección de fraude, aplicaciones en bioinformática, gestión del tráfico y logística, redes sociales, etc.

Introducción a los principales modelos de grafos: property graph y knowledge graph.

Diferencias entre ambos paradigmas y casos de uso

Descripción:
Recapitulación de ambos modelos. Similitudes y diferencias. Conceptos exportables entre ambos modelos.

Principales casos de uso. Gestión de metadatos: semantificación del Data Lake y gobernanza de datos.

Principales casos de uso. Explotación de sus características topológicas: recomendadores sobre grafos y minería de datos en grafo.

Visualización. A través de GUI (Gephi) o programáticas (D3.js o GraphLab).

ACTIVIDADES

Lectures

Descripción:
During lectures the main concepts will be discussed. Lectures will combine master lectures and active / cooperative learning activities. The student is meant to have a pro-active attitude during active / cooperative learning activities. During master lectures, the student is meant to listen, take notes and ask questions.

Objetivos específicos:
1, 2, 3, 5

Competencias relacionadas:
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos
CE3. Aplicar métodos de integración de datos para dar solución a problemas de ciencia de datos en entornos heterogéneos
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB9. Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Dedicación: 53h
Grupo grande/Teoría: 25h
Aprendizaje autónomo: 28h

Hands-on Session

Descripción:
The student will be asked to practice the different concepts introduced in the lectures. This includes problem solving either on the computer or on paper.

Objetivos específicos:
4, 5, 6

Competencias relacionadas:
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos
CE3. Aplicar métodos de integración de datos para dar solución a problemas de ciencia de datos en entornos heterogéneos
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB9. Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.
CB8. Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.

Dedicación: 87h
Grupo pequeño/Laboratorio: 27h
Aprendizaje autónomo: 60h

Final Exam

Descripción:
Written exam of the theoretical concepts introduced along the course.

Objetivos específicos:
1, 2, 3, 4, 5

Competencias relacionadas:
CG3. Definir, diseñar e implementar sistemas complejos que cubran todas las fases en proyectos de ciencia de datos
CG1. Identificar y aplicar los métodos y procesos de gestión de datos más adecuados para gestionar el ciclo de vida de los datos, incluyendo datos estructurados y no estructurados
CE12. Aplicar la ciencia de datos en proyectos multidisciplinares para resolver problemas en dominios nuevos o poco conocidos y que sean económicamente viables, socialmente aceptables, y de acuerdo con la legalidad vigente
CE13. Identificar las principales amenazas en el ámbito de la ética y la privacidad de datos en un proyecto de ciencia de datos (tanto en el aspecto de gestión como de análisis de datos) y desarrollar e implantar medidas adecuadas para mitigar dichas amenazas.
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos
CE3. Aplicar métodos de integración de datos para dar solución a problemas de ciencia de datos en entornos heterogéneos
CT1. ESPÍRITU EMPRENDEDOR E INNOVADOR: Conocer y entender la organización de una empresa y las ciencias que rigen su actividad; tener capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio. Conocer y entender los mecanismos en que se basa la investigación científica, así como los mecanismos e instrumentos de transferencia de resultados entre los diferentes agentes socioeconómicos implicados en los procesos de I+D+i.
CT3. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar ya sea como un miembro mas, o realizando tareas de direccion con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB9. Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.
CB8. Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.

Dedicación: 10h
Grupo grande/Teoría: 2h
Aprendizaje autónomo: 8h

SISTEMA DE CALIFICACIÓN

Nota final = 40% EX + 50% LAB + 10% P

EX = Nota examen final
LAB = Nota ponderada de los laboratorios
P = Proyecto

BIBLIOGRAFÍA

Básica:
- Lenzerini, Maurizio. "Data Integration: A Theoretical Perspective". PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems [en línea]. June 2002, Pages 233¿246 [Consulta: 10/02/2022]. Disponible a: https://doi.org/10.1145/543613.543644.- Aggarwal, Charu C; Wang, Haixun. Managing and mining graph data. New York: Springer, 2010. ISBN 9781441960443.
- Baader, Franz. The description logic handbook: theory, implementation and applications. Cambridge: Cambridge University Press, 2003. ISBN 0521781760.
- Abiteboul, Serge. Web data management. New York: Cambridge University Press, 2012. ISBN 9781107012431.
- Pan, Jeff Z. Ontology-Driven software development. Berlin: Spinger, cop. 2013. ISBN 9783642312250.
- Groppe, Sven. Data management and query processing in semantic web databases. New York: Springer, 2011. ISBN 9783642193569.
- Garcia-Molina, Hector; Ullman, Jeffrey D; Widom, Jennifer. Database systems : the complete book. Second edition, Pearson new international edition. Essex: Pearson Education, [2014]. ISBN 9781292024479.
- Özsu, M. Tamer. "A Survey of RDF Data Management Systems". Cornell University Library [en línea]. [Consulta: 10/02/2022]. Disponible a: https://arxiv.org/abs/1601.00707.- Sahu, Siddhartha; Mhedhbi, Amine; Salihoglu, Semih; Lin, Jimmy; Özsu, M. Tamer. "The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing". Cornell University Library [en línea]. [Consulta: 10/02/2022]. Disponible a: https://arxiv.org/abs/1709.03188.