

**Unidad Académica Responsable:** Departamento de Informática y Ciencias de la Computación

**Programa:** Magister en Ciencias de la Computación

## I.- IDENTIFICACIÓN

Nombre: Tópicos en manejo de Grandes Volúmenes de Datos		
Código: 50332	Créditos: 3	Créditos SCT: 6
Prerrequisitos: Estructuras de datos (503220)		
Modalidad: presencial	Calidad: electivo	Duración: semestral
Semestre en el plan de estudios:	6	
Trabajo Académico: 160		
Horas Teóricas: 32	Horas Prácticas: 0	Horas Laboratorio:32
Horas de otras actividades: 96		
Horas presenciales: 64		Horas no presenciales: 96

## II. DESCRIPCIÓN

Esta asignatura electiva está orientada a entender, conocer y aplicar diferentes estrategias para procesar y almacenar volúmenes de datos muy grandes. La asignatura incluye el estudio de algoritmos eficientes para descubrir características en colecciones de datos grandes y en grafos de la Web y redes sociales como por ejemplo encontrar similitudes entre los nodos del grafo y descubrir comunidades. Por otro lado, la asignatura contempla técnicas de streaming para resumir ciertas características relevantes utilizando poco espacio para el flujo de datos. Finalmente, estudiaremos algunas técnicas de compresión y construcción de índices que se utilizan para representar los datos en espacio reducido y a la vez proporcionar operaciones básicas de acceso.

Esta asignatura aporta a la siguiente competencia del perfil de egreso:

- Mostrar conocimientos avanzados en Ciencias de la Computación, centrándose en al menos una línea de investigación del programa sustentada sobre conocimientos fundamentales de teoría de computación, algoritmos y estructuras de datos.
- Identificar y resolver problemas complejos utilizando el método científico y aplicando avances en una línea de especialización.
- Comunicar efectivamente y organizar conocimiento adquirido desde múltiples fuentes de información científico-tecnológica.

## III. RESULTADOS DE APRENDIZAJE ESPERADOS

Al término del curso el estudiante será capaz de:

- R1 Identificar los principales problemas y desafíos en el procesamiento y almacenamiento de grandes volúmenes de datos.
- R2 Aplicar diversas estrategias para abordar el procesamiento de colecciones muy grandes de datos.

- R3 Implementar y evaluar algoritmos prácticos.
- R4 Utilizar técnicas básicas de compresión.
- R5 Analizar críticamente reportes científicos en el área.
- R6 Presentar oralmente ideas frente a un grupo.

#### **IV. CONTENIDOS**

1. Motivación y aplicaciones.
2. Encontrando elementos similares.
3. Descubriendo conjunto de elementos frecuentes.
4. Compresión de datos e índices.
5. Mining grafos de redes sociales, incluyendo clustering y descubriendo comunidades.
6. Reducción de dimensionalidad. Min Hashing, LocalitySensitiveHashing.
7. Mining en stream de datos.

#### **V. METODOLOGÍA**

El curso consiste en clases presenciales destinadas a presentar y discutir los contenidos del curso. Además, esta asignatura contempla un par de tareas individuales y un único proyecto que debe ser desarrollado incrementalmente en el transcurso del semestre. Todos los proyectos deben incluir la implementación de al menos un algoritmo y una sección de evaluación experimental con el análisis respectivo. Los proyectos deben ser desarrollados en grupos de 2 o 3 personas donde cada grupo debería presentar una propuesta de proyecto y dos presentaciones de avance antes de la presentación final que se realizara al final del semestre.

Las propuestas de proyecto pueden nacer de ideas propias de los grupos para luego ser conversadas y aprobadas por el profesor o bien pueden ser ideas propuestas por el profesor.

#### **VI. EVALUACIÓN**

La asignatura del curso incluye evaluaciones escritas individuales (15%), el desarrollo de un proyecto (80%) con entregas incrementales y una presentación final al concluir el semestre, además de una evaluación individual por participación activa en clases (5%).

#### **VII. BIBLIOGRAFÍA**

##### Básica

- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Data Sets. Cambridge University Press, 2<sup>nd</sup> Edition 2014. ISBN: 9781107077232
- Matthew A. Russell, Mining the Social Web. O'Reilly Media, 2013. ISBN:978-1-4493-6761-9
- Artículos científicos publicados en el área.

##### Complementaria

- Bing Liu, Web Data Mining. Springer, 2011. ISBN: 978-3-642-19459-7
- Ian H. Witten, Allistar Mo\_at, Timothy C.Bell, Managing Gigabytes. Morgan Kaufmann, 1999. ISBN-13: 978-1558605701