Procesamiento de Datos Masivos

Syllabus

Equipo Docente

Juan Reutter

Ph.D. en Philosophy in Informatics. Edinburgh University, Edinburgh, Escocia, Reino Unido. Profesor del Departamento de Ciencia de la Computación de Ingeniería UC. Investigador del Centro de Investigaciones de la Web Semántica de Chile. Sus intereses de investigación están en la gestión de datos y teoría de autómatas. Premios Ramón Salas y Cor Baayen de ERCIM.

Descripción del curso

En este curso, los estudiantes aprenderán a trabajar con datos masivos, ya sea estructurados o semiestructurados, a recolectar información desde fuentes web, y a hacer análisis basados en descripciones de los sets de datos. Metodológicamente, en el curso se trabaja con aprendizaje basado en problemas, en donde todas las semanas se orientan a resolver un problema en particular.

Resultados de Aprendizaje

- Aplicar herramientas basadas en el paradigma Map-Reduce para el trabajo con datos masivos.
- Diseñar algoritmos para la extracción de información basados en descripciones de los mismos, en reglas de asociación o en clasificaciones eficientes de elementos similares.
- Desarrollar un plan para recolectar grandes cantidades de datos *online*.
- Valorar los desafíos tras el manejo de datos semi-estructurados, como texto o grafos, junto a las técnicas para abordarlos.

Estructura del curso

El curso está estructurado de la siguiente forma:

Módulo 1: Map Reduce

- Modelos de big data
- Data warehousing
- Sistemas distribuidos
- Map Reduce como paradigma de procesamiento de datos masivos y distribuidos.
- Aplicaciones a herramientas en la nube.

Módulo 2: Texto

- Recuperación de la información: *web search, crawling, scrapping*, búsqueda por texto, *ranking*.

Módulo 3: Minería

- Búsqueda de ítems similares, shingling y algoritmos.
- Minhash y Locally Sensitive Hashing
- Búsqueda de elementos más frecuentes, reglas de asociación, canastas y algoritmos a priori.

Módulo 4: Grafos

- Manejo de grafos y redes sociales.
- Algoritmos básicos (comunidades, centralidad, conteo de triángulos).

El alumno solamente podrá aprobar el curso sí para todas las evaluaciones alcanza un puntaje superior al 50% de las respuestas correctas y para la aprobación general del curso debe obtener una nota final igual o superior a 4.0. El curso tiene como requisitos de aprobación las siguientes instancias evaluativas:

Actividad	Evaluación
Participación en foro	10% nota final
Cuestionarios	30% nota final
Tareas	30% nota final
Trabajo final	**30% nota final**

Plataforma e Información General

- **Duración**: 90 horas de dedicación total (24 directas y 66 indirectas).
- **Créditos:** 5 créditos UC
- **Requisitos**: (MAN3070 y MAN3080) o EPG 4506
- **Restricciones**: MDS O MAN
- **Conector:** Y
- **Carácter:** Mínimo
- **Tipo:** Taller
- **Calificación:** Estándar
- **Nivel formativo:** Magíster

Política de entregas de evaluaciones calificadas fuera de plazo

En caso de entregar una evaluación calificada, sea esta Tarea o Cuestionario, fuera del plazo informado (fecha límite), se aplicará un descuento progresivo a la nota máxima por entrega tardía. El plazo para entregar evaluaciones o tareas fuera de plazo será de 7 días desde la fecha límite. Luego de los 7 días de plazo adicional, el alumno obtendrá una nota de 0% en dicha evaluación.

Si por razones de fuerza mayor, el alumno/a no pudiera rendir la prueba dentro del plazo regular o excepcional, deberá enviar una solicitud al correo de Soporte de su programa, adjuntando respaldos para que su requerimiento sea evaluado por la Unidad Académica (UA). La resolución de esta solicitud quedará a criterio de la UA.

Bibliografía

Mínima:

- WHAT IS A DATA PIPELINE? What is a Data Pipeline? (s. f.).
- Learning Spark Karau, Holden, autor.2015.
- Mining Massive Datasets Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets (Second edition.) (pp.capitulo 3). Cambridge University Press.
- Modern Information Retrieval.
- Beginning Neo4j Kemper, C. (2015). Beginning Neo4j (1st ed. 2015.).
- Foundations of modern query languages for graph databases Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., & Vrgoč, D. (2017). Foundations of Modern Query Languages for Graph Databases. ACM Computing Surveys, 50(5), 1–40.

Complementaria

- Database management systems Ramakrishnan, R., Gehrke, J., & Gehrke, J. (2003). Database management systems. New York: McGraw-Hill.
- Introduction to SQL Window Functions: part 1 Tomar, A. (2022). Introduction to SQL Window Functions: part 1.
- Data analysis in the cloud: models, techniques and applications
- Introduction to Information Retrieval.
- Data Mining: Concepts and Techniques Han, J., Pei, J., & Kamber, M. (2023). Data Mining Concepts and Techniques (4th Edition). Elsevier. Retrieved from.
- Graph Data Management Fundamental Issues and Recent Developments Fletcher, G., Hidders, J., & Larriba-Pey, J. L. (2018). Graph Data Management Fundamental Issues and Recent Developments (G. Fletcher, J. Hidders, & J. L. Larriba-Pey, Eds.; 1st ed. 2018.). Springer International Publishing
- Practical Neo4j Jordan, G. (2014). Practical Neo4j (1st ed. 2014.).
- Managing and Mining Graph Data: capítulo 2 Aggarwal, C. C., & Wang, H. (2010). Managing and Mining Graph Data (C. C. Aggarwal & H. Wang, Eds.; 1st ed. 2010.)
- Neo4j Cypher Manual Introduction - Cypher Manual. (s. f.). Neo4j Graph Data Platform.