PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

CD306

Manipulación y Preparación de Datos

(506) 4030.1205 - (506) 4030.1114 info@promidat.com facebook.com/promidat Twitter.com/promidat www.promidat.com

Índice

Duración

Cuatro semanas.

Descripción

En este curso se presentarán los fundamentos del lenguaje R para el procesamiento de datos. El énfasis principal del curso será examinar diversos componentes del lenguaje, como lo son funciones, expresiones, librerías, entre otros. Se le dará especial importancia al uso del lenguaje como herramienta de manipulación de información, como punto de partida para el desarrollo de aplicaciones de minería de datos. Para esto se utilizarán diversos paquetes en R para manipulación de datos, así como motores de bases de datos como SQLite y MySQL.

Objetivos

En este curso el estudiante será capaz de:

1. Utilizar el lenguaje R como mecanismo de extracción de datos e información a partir de repositorios con grandes volúmenes de datos.
2. Hacer uso correcto del lenguaje para construir consultas complejas que permitan manipular información de distintas tablas de datos simultáneamente.
3. Entender el lenguaje R desde el punto de vista de teoría de conjuntos y lógica de predicados, permitiendo realizar operaciones usuales como lo son uniones, intersecciones, diferencias, entre otros.
4. Utilizar SQLite y MySQL como motores de bases de datos basados en SQL.
5. Importar información de un Administrador de base de datos SQL a R.
6. Exportar resultados obtenidos en R a un Administrador de base de datos SQL.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.

3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren alta manipulación de datos utilizando el lenguaje R.

Contenido

1. **Estructuración de datos.**
    a. ¿Qué es Data Wrangling?
    b. Uso de tidyr para limpieza, estructuración y manipulación de datos en R
        a. Uniones.
        b. Separaciones.
        c. Transformaciones.
    c. Uso de pipes (%>%) en R.
2. **Manipulación de datos.**
    a. Uso de dplyr para el procesamiento y manipulación de datos en R.
    b. Explorar tablas completas o subconjuntos (SELECT).
    c. Derivación/Creación de nuevas columnas. (MUTATE).
    d. Ordenamiento de datos (ARRANGE).
    e. Filtros de datos. (FILTER).
    f. Agrupación de individuos para obtener métricas (GROUP_BY, SUMMARISE).
3. **Manipulación de Textos.**
    a. Manejo de variables categóricas.
    b. Extraer, detectar y reemplazar componentes de un texto.
    c. Realizar conteos de patrones en los textos.
    d. Manejo de argumentos en textos.
4. **Conexión con Bases de Datos.**
    a. Conexión a Bases de datos (DBI y RODBC).
    b. Base de datos SQLite y MySQL.
    c. Creación de estructuras (data frames) en R.
    d. Almacenar datos localmente.

e. Paquete dbplyr (Uso de sintaxis dplyr a una base de datos).

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Bradley C. Boehmke. *Data Wrangling with R*. Springer, 2016. URL: https://link.springer.com.una.remotexs.co/book/10.1007/978-3-319-45599-0 (visited on 10/10/2022).

[2] Vikram Dayal. *Quantitative Economics with R*. Singapore: Springer Singapore, 2021. 326 pp. ISBN: 9789811634345. DOI: 10.1007/978-981-16-3434-5. URL: https://link.springer.com/10.1007/978-981-16-3434-5 (visited on 10/10/2022).

[3] "MySQL: My Structured Query Language". In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky et al. New York, NY: Springer New York, 2013, pp. 1485–1486. ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_100986. URL: https://doi.org/10.1007/978-1-4419-9863-7_100986.

[4] Alfredo Ferro et al. "MySQL Data Mining: Extending MySQL to Support Data Mining Primitives (Demo)". In: *Knowledge-Based and Intelligent Information and Engineering Systems*. Ed. by Rossitza Setchi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 438–444. ISBN: 978-3-642-15393-8.

[5] Andreas Meier and Michael Kaufmann. *SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management*. Wiesbaden: Springer Fachmedien Wiesbaden, 2019. 229 pp. ISBN: 978-3-658-24548-1 978-3-658-24549-8. DOI: 10.1007/978-3-658-24549-8. URL: http://link.springer.com/10.1007/978-3-658-24549-8 (visited on 10/10/2022).

[6] Laurie A. Schintler and Connie L. McNeely, eds. *Encyclopedia of Big Data*. Springer Cham, 2022. 976 pp. URL: https://link.springer.com.una.remotexs.co/book/10.1007/978-3-319-32010-6 (visited on 10/10/2022).

[7] Geoffrey I. Webb and Claude Sammut, eds. *Encyclopedia of Machine Learning and Data Mining*. Second Edition. Springer New York, 2017. 1335 pp. URL: https://link.springer.com.una.remotexs.co/book/10.1007/978-1-4899-7687-1 (visited on 10/10/2022).