Repaso final
© Ediciones Roble S. L.

campusproyectosnebrija.imf.com © Ediciones Roble S. L.
JOSE RODRIGUEZ MALDONADO
campusproyectosnebrija.imf.com © Ediciones Roble S. L.
JOSE RODRIGUEZ MALDONADO

Indice

Repaso final

En este módulo, se han revisado las tecnologías básicas que todo aquel que quiera trabajar en el ámbito tecnológico del Big Data y del análisis de datos necesita conocer para poder entender las técnicas y herramientas más avanzadas en la materia.

Se han cubierto cinco áreas de conocimiento a través de seis unidades de trabajo.

Virtualización de programas y al uso de los comandos de la Shell de Linux

La primera unidad se ha dedicado al tema de la virtualización de programas y al uso de los comandos de la Shell de Linux. En muchas ocasiones, el software que se utiliza en este entorno se distribuye ya instalado en una máquina virtual, con el objetivo de evitar que el usuario tenga que dedicar tiempo y esfuerzo a instalarlo. Esto se debe a que la configuración y puesta a punto de este tipo de software suele ser una tarea compleja. Asimismo, en muchos casos, el software que se usa se ejecuta bajo Linux, de manera que para poder utilizarlo o gestionarlo debe hacerse desde la Shell de comandos del sistema operativo. Por ello hemos revisado la Shell y sus principales comandos y funcionalidades, de manera que el usuario esté familiarizado con dicho entorno.

Lenguaje de programación Python y las librerías para cálculos científicos

La segunda área de conocimiento que se ha revisado en dos de las unidades ha sido el lenguaje de programación Python y las librerías para cálculos científicos que posee. Actualmente, Python y el lenguaje R son los lenguajes de programación más utilizados en el ámbito del Big Data. En una de las unidades del módulo se ha descrito el lenguaje de programación Python, estudiando sus principales estructuras sintácticas como lenguaje de propósito general. A continuación, en una unidad independiente, se ha examinado el conjunto de librerías de cálculo científico que dotan a Python de la potencia necesaria para realizar análisis de datos, es decir, las librerías Numpy —para gestión de matrices y cálculos estadísticos—, la librería Matplotlib —para la representación de datos— y la librería Pandas —para facilitar la manipulación de datos—. La combinación de las tres librerías convierte a Python en una poderosa herramienta para llevar a cabo análisis de datos.

Bases de datos relacionales y al lenguaje SQL

La tercera área de conocimiento está dedicada a las bases de datos relacionales y al lenguaje SQL. En esta unidad se han revisado los principios conceptuales de las bases de datos relacionales a través del estudio del modelo relacional. A continuación, se ha realizado una introducción al lenguaje de consultas SQL. Este lenguaje permite manipular los datos de una base de datos relacional: creación de tablas, inserción de filas o la realización de consultas para recuperar datos. En el ámbito del Big Data, han surgido unas nuevas bases de datos denominadas bases de datos NoSQL, las cuales se caracterizan por seguir principios diferentes a los de las bases de datos relacionales. Sin embargo, para comprenderlas, es necesario entender previamente las bases de datos relacionales. Además, algunas de ellas, mantienen ciertas características del mundo relacional e, incluso, determinados lenguajes de consultas son similares a SQL.

Formatos de almacenamiento de datos

La cuarta área de conocimiento se refiere a los formatos de almacenamiento de datos. Una tarea básica de cualquier analista de datos es recuperar los datos de diferentes fuentes de información. Esta información, normalmente, se recupera codificada en un formato de datos determinado. En esta unidad se han revisado los principales formatos de datos que se utilizan para codificar la información en Internet, CSV, XML y JSON. También se ha descrito cada uno de los formatos: su sintaxis y semántica. Asimismo, se ha realizado una introducción acerca de cómo manipular estos formatos desde Python.

Repositorios digitales de información

Por último, la quinta área de conocimiento que se ha tratado en este módulo tiene como objeto los repositorios digitales de información. Actualmente, la gestión de muchos proyectos informáticos del ámbito del Big Data, y también de otros ámbitos, se desarrolla en repositorios digitales que permiten trabajar de manera colaborativa, como son GitHub o Google Drive. Este tipo de entornos ofrece al programador la posibilidad de gestionar las versiones de los programas, la documentación, así como la organización de un equipo de desarrollo. En muchos casos, este tipo de proyectos requiere de la participación de un gran número de personas, por lo que en entornos de trabajo como los citados son usados frecuentemente. Asimismo, este tipo de herramientas son utilizadas habitualmente como sistemas de distribución del software o de la documentación que se genera en un proyecto de estas características.

Además, cada unidad del módulo se complementa con un test de preguntas que sirve de repaso a todo lo estudiado y con la propuesta de resolución de un caso práctico.

Análisis exploratorio de datos con Python

Pincha aquí para descargar los archivos necesarios para seguir el tutorial.