BIG DATA

PROYECTO FINAL

New Technology School

Tokio.

Con el desarrollo del proyecto final buscamos que pongas en práctica todo lo aprendido a lo largo de la formación. Es importante comentar también que el proyecto no está definido al 100%, por lo tanto habrá decisiones que tendrás que tomar y adaptar bajo tu interpretación detallándolas en el informe científico del mismo. Los resultados de cada proyecto serán así más personales aunque las herramientas utilizadas sean las mismas.

ESCENARIO

Hemos conseguido nuestro primer trabajo como analista *Big Data*. Has sido contratado por la empresa Tokio School Viajes y el objetivo de este puesto de trabajo es analizar un conjunto de datos sobre el tráfico en el aeropuerto de San Francisco, desde donde salen muchos vuelos camino a Tokio (Japón) y otras ciudades de ese mismo país.

Para realizar el proyecto, tendrás disponible el archivo CSV: Air_Traffic_Passenger_Statistics o también puedes descargarlo en el siguiente link: https://www.kaggle.com/yamqwe/air-traffic-passenger-datae

Con estos datos realizaremos varias tareas que nos permitirán el análisis y toma de decisiones en base a ellos. En Tokio School Viajes, sabemos que es tu primer empleo, por lo que hemos creado ya una serie de tareas para que comiences a realizar los análisis pertinentes. Algunas de las tareas tienen relación con tareas anteriores y otras no.

DESARROLLO

1. La primera tarea que llevaremos a cabo, consistirá en conocer con qué tipo de datos contamos. Para ello, tendremos que categorizarlos según su estructura y presentarlos en la siguiente plantilla:

Nombre del Campo	Tipo de dato

2. Usando los *dataframe* de PySpark, prepararemos un *notebook* para desarrollar los siguientes puntos:
    - Cargaremos el conjunto de datos en un *dataframe*.
    - ¿Cuántas compañías diferentes aparecen en el fichero?
    - ¿Cuántos pasajeros tienen de media los vuelos de cada compañía?
    - Eliminaremos los registros duplicados por el campo “GEO Región”, manteniendo únicamente aquel con mayor número de pasajeros.
    - Volcaremos los resultados de los dos puntos anteriores a un CSV.

Importante: Para las siguientes, tareas el fichero de datos se cargará en Google Colab y se utilizará PySpark (Python + Spark y no otro lenguaje) para conseguir los resultados solicitados.

3. Vamos a realizar un análisis descriptivo de los datos usando en lenguaje PySpark. Para ello necesitaremos calcular la media y la desviación estándar de cada elemento del conjunto de datos. Los resultados obtenidos deben ir acompañados de unas conclusiones las cuales estarán basadas en los datos mostrados y podrán contener una parte subjetiva en cuanto a interpretación de los mismos.

   Una vez esto haremos un análisis de la correlación cuyo resultado debe ser una matriz de correlación de datos que represente de qué manera están relacionadas las diferentes variables. Para llevar a cabo la matriz de correlación, es necesario que configuremos las características o columnas con los tipos necesarios y así poder realizar su cálculo. Estas transformaciones tendremos que reflejarlas en el documento final. No es necesario comentar cada uno de los cruces de relaciones de datos, pero si tendrás que escoger los 10 elementos más importantes y argumentar los resultados obtenidos. Estas conclusiones tendrán que estar basadas en los datos mostrados y pueden contener una parte subjetiva en cuanto a interpretación de los mismos.

   Luego seleccionaremos uno de los algoritmos que hemos visto durante el curso y explicaremos las razones de nuestra elección. Aplicaremos ese algoritmo con los argumentos y valores que consideremos, y explicaremos los resultados obtenidos. Es posible que tengamos que cambiar los tipos de alguna columna.

ENTREGA

Para la evaluación del proyecto será necesaria la entrega de un archivo .zip así como todos los ficheros ejecutables (ejemplo .ipynb), enlaces o cualquier recurso adicional necesario para la ejecución del proyecto.

El proyecto constará de los siguientes documentos:

1. Un informe científico, en el que se transmitan los resultados de los análisis realizados. Aquí explicaremos paso a paso cada uno de los apartados con las conclusiones correspondientes de las tareas realizadas. Podremos incluir secciones de código si es necesario y por supuesto, los resultados de cada una de las tareas realizadas sobre los datos obtenidos a través de la ejecución del código contenido en el documento técnico.

   El nombre de este documento será: informe_nombre_apellido1_apellido2.pdf

2. Un documento técnico que tendrá el código fuente (PySpark) empleado para la resolución de cada una de las tareas. El código fuente debe ser insertado como imágenes y con un tamaño que permita leer el texto contenido en las imágenes.

   El nombre de este documento será: tecnico_nombre_apellido1_apellido2.pdf

3. Una presentación guardada en formato pdf. Esta presentación nos servirá para mostrar los resultados de cada una de las tareas y no contendrá código fuente sino que mostrará los resultados obtenidos siguiendo las guías de presentación que hemos visto en el módulo de proyectos *big data* y *storytelling*.

   El nombre de este documento será: presentacion_nombre_apellido1_apellido2.pdf

Todos los documentos deberán entregarse dentro de un fichero comprimido (.zip o .rar) llamado proyecto_bigdata_nombre_apellido1_apellido2.zip

Junto con el feedback y la nota, se entregará un video análisis de los resultados y esto se considerará la defensa del proyecto.

Esta defensa tendrá una duración aproximada de 10 minutos.