Practica 2 - Análisis de datos con Python

Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ciencias y Sistemas

Seminario de Sistemas 2 Sección N

Ing. Fernando Paz

Aux. Sergio Enrique Cubur Chalí

Objetivos

- Desarrollar un notebook de análisis de datos que permita la carga, manipulación, visualización y generación de informes a partir de conjuntos de datos utilizando las librerías Pandas, NumPy y Matplotlib.
- Implementar técnicas de limpieza y transformación de datos utilizando Pandas para preparar el conjunto de datos.
- Crear visualizaciones interactivas y altamente informativas utilizando Matplotlib que permitan a los tomadores de decisiones explorar visualmente los datos.

Descripción

Debido a la buena implementación que se realizó con el proyecto de la empresa SG-Food, la empresa ha quedado satisfecha. Debido a esto lo ha contactado otra empresa, pero esta vez para el procesamiento de datos dentro de otro entorno.

Para el análisis que se le solicita se le proveerá un archivo CSV y un archivo txt con datos que deberá procesar datos con el ambiente de trabajo de Python con sus respectivas librerías.

Implementación Sugerida

Para la realización del análisis de datos, se trabajará dentro de un notebook de Jupyter, este puede ser trabajados tanto como desde el Navegador Anaconda, Visual Studio Code o Colab de Google, queda a su criterio.

Se brindará un archivo CSV el cual contendrá información sobre la plataforma de Coursera y un archivo txt sobre un texto relaciona. El jefe de la empresa necesita que se le brinde todo tipo de información.

Datos Solicitados

Pandas, Numpy

- Limpieza de datos: Puede que los datos vengan mal, por lo que se necesita que se realice la limpieza.
- Realicen el cálculo del promedio de calificaciones para cada curso en el conjunto de datos. Para ello, agrupen los datos por el título del curso y calculen el promedio de las calificaciones para cada grupo.
- Calcular los cursos con mayor y menor rating.
- Calculen el porcentaje de cursos con horario flexible en relación con el total de cursos en el conjunto de datos. Para ello, necesitan filtrar los datos para seleccionar solo los cursos con horario flexible, luego calculen el porcentaje de estos cursos respecto al total de cursos en el conjunto de datos original y lo redondeen a dos decimales.

NLKT

- Analisis de texto en el archivo .txt
    - Tokenización
    - Lematización y Stemming
    - Eliminación de palabras vacias
    - Frecuencia de palabras
    - Analisis de sentimientos
    - Reconocimiento de entidades nombradas

- Extraer entidades

Gráficos Solicitados

- Se requiere generar una gráfica de barras que muestre el número de cursos en cada nivel de dificultad.
- Se solicita la generación de una gráfica de barras horizontal que muestre el número de cursos en las principales categorías.

- Se solicita crear un gráfico de dispersión para visualizar la relación entre la duración del curso y el número de revisiones.
- Se solicita crear un histograma de la distribución de las duraciones de los cursos.
- Se solicita crear un gráfico de cajas para visualizar la distribución de las calificaciones de los cursos por nivel de dificultad
- Se solicita presentar los resultados de cada operación realizada para el archivo .txt de manera creativa.

Restricciones

- Se debe utilizar Python y solamente las librerías permitidas Pandas, Numpy, Matplotlib y NLKT)
- Se debe realizar en un NoteBook Jupyter.

Entregables

Para el entregable de está practica será el archivo .ipynb, el cual será el Jupyter Notebook, por lo que se solicitara que describan cada detalle con bloques de MarkDown. Y luego de cada cálculo y grafica realizada deberá realizar un análisis del resultado que obtuvo.

Conclusión general de todo el análisis que se realizó y también sobre el uso de python en el análisis de datos.

Consideraciones

- La entrega será individual. No habrá prorroga.
- Todas las dudas con respecto a esta práctica, deberán ser planteadas en los foros creados en la plataforma de UEDI.
- Enviar la práctica vía UEDI por medio de un link del repositorio donde se trabajará. El día de entrega será el Domingo 6 de Octubre de 2024 a las 23-59 horas.