Universidad San Carlos
Facultad de Ingeniería
Escuela de Ciencias y Sistemas
Seminario de Sistemas 2
Ing. Luis Alberto Vettorazzi Espana
Ing. Fernando José Paz Gonzáles
Tutor José Alejandro Lorenty Herrera
Tutor Randall Chriss Ramos Saucedo

Análisis de datos con Python

1. Resumen Ejecutivo

La práctica consiste en desarrollar un cuaderno interactivo (*notebook*) en Python para realizar análisis de datos. Utilizando las librerías **Pandas, NumPy, Matplotlib y NLTK**, se espera que los estudiantes lleven a cabo tareas de limpieza, transformación, visualización y análisis de texto. El conjunto de datos será proporcionado y contendrá información sobre cursos de la plataforma Coursera, así como un archivo de texto adicional para aplicar técnicas básicas de NLP.

2. Objetivos de Aprendizaje

2.1. Objetivo General

- Desarrollar una solución de análisis de datos utilizando Python, con enfoque en la manipulación, visualización y análisis textual de la información.

2.2. Objetivos Específicos

- Limpiar y transformar datos usando Pandas.
- Realizar análisis estadísticos básicos con NumPy.
- Generar visualizaciones con Matplotlib.
- Aplicar técnicas de procesamiento de lenguaje natural con NLTK.
- Presentar resultados con claridad en un notebook bien documentado.

3. Enunciado del proyecto

3.1. Descripción del problema a resolver

Tras el éxito del proyecto anterior con **SG-Food**, una nueva empresa ha solicitado tus servicios para el análisis de datos de cursos en línea. Se te ha proporcionado un archivo **.csv** con información de cursos de Coursera y un archivo **.txt** con un texto relacionado. Se requiere analizar los datos, generar visualizaciones y aplicar técnicas de NLP sobre el archivo de texto, todo desde un entorno interactivo en Jupyter Notebook.

3.2. Alcance del proyecto

- Limpieza y transformación del dataset.
- Generación de métricas y visualizaciones informativas.
- Análisis del archivo de texto utilizando técnicas de NLP.
- Documentación completa y análisis detallado de los resultados.

3.3. Requerimientos técnicos

- Python con librerías: Pandas, NumPy, Matplotlib, NLTK.
- Uso de entorno Jupyter Notebook (Anaconda, VSCode, Google Colab).
- Conocimientos básicos en estadística, análisis de datos y NLP.

3.4. Entregables

- Archivo .ipynb con el desarrollo completo de la práctica.
- Cuaderno debe incluir:
    - Celdas con código y análisis en Markdown.
    - Visualizaciones:
        - Gráfica de barras de cursos por nivel.

        - Gráfica de barras horizontal por categoría.
        - Gráfico de dispersión entre duración y número de revisiones.
        - Histograma de duración de cursos.
        - Boxplot de calificaciones por nivel de dificultad.
    - Análisis sobre:
        - Calificaciones promedio por curso.
        - Curso con mayor y menor rating.
        - Porcentaje de cursos con horario flexible.
    - Análisis NLP sobre el archivo .txt:
        - Tokenización
        - Lematización y stemming
        - Eliminación de stopwords
    - Frecuencia de palabras
    - Análisis de sentimientos
    - Reconocimiento de entidades nombradas
- Conclusión general del análisis y sobre el uso de Python para este tipo de tareas.

4. Restricciones

- Solo se pueden usar las librerías indicadas: Pandas, NumPy, Matplotlib y NLTK.
- El análisis debe desarrollarse exclusivamente en un archivo .ipynb.
- La práctica es individual. No se aceptarán entregas fuera del plazo.
- Se deberá utilizar el repositorio para el desarrollo de prácticas y proyectos con el nombre (**SS2_1S2025_#carné**) crear una carpeta con relación al número de práctica o proyecto realizándose y agregar al tutor a dicho repositorio. Usuario: **BLorenty**
- Copias detectadas obtendrán una nota de 0 puntos y se reportarán a la Escuela de Ciencias y Sistemas.

5. Metodología

1. Limpieza y transformación de los datos.
2. Cálculo de métricas con Pandas y NumPy.
3. Generación de visualizaciones con Matplotlib.
4. Análisis textual con NLTK.
5. Redacción del análisis en Markdown dentro del notebook.

6. Cronograma

- Asignación del Proyecto: 24 de marzo de 2025
- Desarrollo del análisis y visualizaciones: 7 días
- Implementación del análisis NLP: 4 días
- Redacción y entrega final: 3 días
- Entrega del proyecto: **11 de abril de 2025 - 23:59 hrs**
- Calificación del proyecto: 12 de abril de 2025