Práctica 1.2

Procesamiento del lenguaje natural

Procesamiento de Información Médica usando Expresiones Regulares

Enunciado

Objetivo:

Aplicar conocimientos de expresiones regulares en Python para extraer información clave de informes médicos estructurados. Posteriormente, se propone realizar un análisis de datos para responder preguntas específicas sobre la población de pacientes.

Datos de Entrada:

Contamos con un total de 20 informes médicos con la siguiente estructura:

```
Paciente: Juan Pérez
Fecha: 2023-05-15
Edad: 45
Sexo: Masculino

Historial Médico:
- Código: G20230515
- Síntomas: Dolor abdominal persistente, fatiga
- Diagnóstico: Gastroenteritis crónica
- Gravedad: Baja
- Tratamiento: Antibióticos y reposo

Próxima Cita: 2023-06-01
```

Parte 1: Extracción de Datos con Expresiones Regulares

Utilizando expresiones regulares, extrae todos los datos anteriores de cada paciente, eligiendo para ello la forma más adecuada de almacenarlos en Python.

Parte 2: Análisis de Datos

Responde a las siguientes cuestiones:

1. **Enfermedades graves y próximas citas:**

   Determina cuántos pacientes con una enfermedad grave deberían tener una cita antes de acabar el año 2023.

2. **Enfermedades y sexo del paciente:**

   Analiza la relación entre ciertas enfermedades y el sexo del paciente. ¿Se puede identificar algún tipo de patrón reseñable en cuanto a qué sexo parece enfermar más de cierta enfermedad?

   *Consulta si lo necesitas el **anexo 1** que viene al final de este documento dónde se presenta una posible estrategia para contestar a esta pregunta.*

3. **Correlación entre edad y gravedad de las enfermedades:**

   Investiga si existe alguna correlación entre la edad de los pacientes y la gravedad de sus enfermedades. ¿Hay una tendencia que sugiera que los pacientes más jóvenes tienden a tener enfermedades menos graves en comparación con los pacientes mayores?

   *Consulta si lo necesitas el **anexo 2** que viene al final de este documento dónde se explica qué es la correlación en estadística y cómo calcularla en Python.*

Observaciones

- Los informes vienen como documentos de tipo texto plano.
- Comenta sobre cualquier patrón interesante o hallazgo inesperado durante el análisis.
- Trabajar con Pandas es una buena opción para dar respuestas a las preguntas.
- Presenta los resultados de manera clara y concisa. En el caso de la segunda pregunta, por ejemplo, tanto si los datos proceden de un *dataframe* tipo Pandas como si se encuentran en una matriz bidimensional puedes representar los resultados usando código de este estilo:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# Supongamos que tienes un DataFrame similar al anterior
sns.countplot(x='Enfermedad', hue='Genero', data=df)
plt.show()
```

- Puedes guardar la práctica en un cuaderno Google Colab o Jupyter con el nombre "**P1.2_informes_médicos.ipynb**". No es necesario entregar esta práctica.

ANEXO 1

¿QUÉ ES UNA TABLA DE FRECUENCIAS?

Una **tabla de frecuencias** es una herramienta estadística utilizada para organizar y resumir la distribución de datos en distintas categorías o intervalos.

Es particularmente útil cuando se trabaja con variables categóricas, permitiendo visualizar la frecuencia con la que ocurren diferentes valores. En el contexto de un análisis médico como el que se propone en esta práctica, una tabla de frecuencias podría ser empleada para examinar la relación entre enfermedades específicas y género de los pacientes.

Ejemplo:

Supongamos que se recopilan datos sobre pacientes con diferentes enfermedades y se desea entender cómo se distribuyen estas enfermedades según el género. Mediante una tabla de frecuencias se podría representar la cantidad de pacientes para cada combinación de enfermedad y género. Por ejemplo:

Enfermedad	Masculino	Femenino
EnfermedadA	20	30
EnfermedadB	15	25
EnfermedadC	10	20

En esta tabla se puede observar cuántos pacientes de cada género sufren cada enfermedad. A partir de los datos que se muestran, podemos obtener una idea de la distribución de las enfermedades entre los pacientes en función del género.

ANEXO 2

¿QUÉ ES LA CORRELACIÓN EN ESTADÍSTICA?

La **correlación** es una medida estadística que describe la relación entre dos variables.

Puede ser positiva (ambas variables aumentan o disminuyen juntas), negativa (una variable aumenta mientras la otra disminuye) o nula (sin relación aparente). El coeficiente de correlación varía de -1 a 1, donde -1 indica una correlación negativa perfecta, 1 indica una correlación positiva perfecta y 0 indica falta de correlación.

En Python, puedes calcular la correlación utilizando, por ejemplo, la función **corrcoef** de **NumPy**.

Aquí hay un ejemplo simple:

```
import numpy as np

# Datos de ejemplo
x = np.array([1, 2, 3, 4, 5])
y = np.array([2, 4, 6, 8, 10])

# Calcula el coeficiente de correlación
correlation_matrix = np.corrcoef(x, y)
correlation_coefficient = correlation_matrix[0, 1]

print(f"Coeficiente de correlación: {correlation_coefficient}")
```

En este ejemplo, como los datos x e y están perfectamente correlacionados positivamente, el coeficiente de correlación debería ser 1.