PROYECTO INTEGRADOR (GPO 10)

PROFESORES TITULARES DRA. GRETTEL BARCELÓ

ALONSO DR. LUIS EDUARDO FALCÓN MORALES.

EQUIPO 31

JUAN CARLOS VILLAMIL ROJAS A01794003

MATEO CRUZ LANCHERO A01793882

ANDREA MARGARITA OSORIO GONZÁLEZ A01104776

Fecha: 5 de mayo de 2024

Tabla de contenido:

1. ¿Hay valores faltantes en el conjunto de datos?

Dado que los archivos son transcripciones completas de las reuniones, no se esperan valores faltantes en los datos de texto. Sin embargo, se realizará una verificación exhaustiva de los archivos para confirmar la integridad de los datos.

2. **¿Se pueden identificar patrones de ausencia?**

se pueden identificar silencios prolongados, frecuenta por hablante, análisis de temas discutidos.

3. **¿Cuáles son las estadísticas resumidas del conjunto de datos?**

Número total de palabras: 102.068
Número promedio de palabras por reunión: 5.372
Duración promedio de las reuniones: [Por definir]
Insertar histograma o gráfico de caja de la distribución de la duración de las reuniones

4. **¿Hay valores atípicos en el conjunto de datos?**

En el contexto de las transcripciones de reuniones, los valores atípicos pueden referirse a reuniones excepcionalmente largas o cortas en comparación con la duración promedio. Se identificarán las reuniones que se desvían significativamente de la duración típica y se considerará si requieren un tratamiento especial durante el procesamiento.

5. **¿Cuál es la cardinalidad de las variables categóricas?**

En este caso, no hay variables categóricas explícitas en los datos de texto.

6. **¿Existen distribuciones sesgadas en el conjunto de datos?**

Duración de las conversaciones,

7. **¿Necesitamos aplicar alguna transformación no lineal?**

necesitamos la Separación de fuentes para reconocer lo que habla cada interlocutor, número de intervenciones por hablante./

8. **¿Se identifican tendencias temporales? (En caso de que el conjunto incluya una dimensión de tiempo).**

No se dispone.

9. ¿Hay correlación entre las variables dependientes e independientes?

En el contexto de las transcripciones de texto, no hay variables dependientes e independientes explícitas. Sin embargo, se explorará la correlación entre la duración de las reuniones y la longitud de las transcripciones para ver si existe alguna relación.

10. ¿Cómo se distribuyen los datos en función de diferentes categorías?

Puede ser tipo de producto/servicio, tipo de reunión, temas principales,nombre del producto,

11. ¿Existen patrones o agrupaciones (clusters) en los datos con características similares?

Pendiente por ser analizado.

12. ¿Se deberían normalizar las imágenes para visualizarlas mejor?

hay vídeos, pero los transformamos en audio, en últimas en texto, por lo cual no aplicaría directamente.

13. ¿Hay desequilibrio en las clases de la variable objetivo?

En este caso, no hay una variable objetivo explícita con clases desequilibradas. Sin embargo, si se identifican ciertos tipos de reuniones (por ejemplo, reuniones de ventas, reuniones técnicas) durante el análisis, se evaluará si hay un desequilibrio en la representación de estos tipos.