# PROYECTO INTEGRADOR (GPO 10)

## Avance 2

**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:10** de mayo de 2024

**Tabla de contenido:**

1. **Se aplicarán operaciones comunes para convertir los datos crudos del mundo real, en un conjunto de variables útiles para el aprendizaje automático. El procesamiento puede incluir:**

### Generación de nuevas características

Se proponen las siguientes nuevas características como posibles variables a tener en cuenta, se determinará su utilización o descarte, según las comprobaciones realizadas en la próxima semana.

- Duración de la reunión en minutos.
- Número de participantes en la reunión.
- Número de intervenciones por participante.
- Longitud promedio de las intervenciones por participante.
- Frecuencia de palabras clave relevantes para el negocio.

### Discretización o binning

Definición de Bins para la frecuencia de las palabras claves:

Bajo: 0-2 menciones
Medio: 3-5 menciones
Alto: más de 5 menciones

Categorización de la duración de la reunión en intervalos.

Corta: menos de 30 minutos
Media: entre 30 y 60 min
Larga: más de 60 min

Aunque el binning no se aplica directamente a la síntesis de un texto, podemos encontrar una analogía interesante:

En la síntesis, seleccionamos las ideas principales y las organizamos de manera clara y concisa. Similarmente, en el binning, agrupamos valores similares para simplificar la información.

Así como en la síntesis mantenemos el significado esencial del texto original, en el binning mantenemos la información relevante al agrupar valores similares.

Ambas técnicas buscan simplificar y resumir, aunque en contextos diferentes.

En resumen, mientras que la síntesis se aplica a la escritura, el binning se

utiliza en el análisis de datos.

### Codificación (ordinal, one hot,…)

Se llevará a cabo la codificación mediante el estándar Unicode para la normalización de texto. Unicode es un sistema de codificación de caracteres que abarca la mayoría de los caracteres escritos utilizados en todo el mundo. Garantiza que los caracteres de diferentes idiomas se puedan representar de manera consistente en sistemas informáticos, lo que facilita la normalización de texto en varios idiomas, especialmente en aplicaciones globales y en el procesamiento de datos en diferentes lenguajes.

- Codificación one-hot de las categorías generadas en la discretización.
- Codificación ordinal para variables como el tipo de reunión.

### Escalamiento (normalización, estandarización, min – max,…)

La normalización de texto es esencial en el campo del procesamiento de lenguaje natural para que las máquinas comprendan y analicen el lenguaje humano. Al estandarizar el texto, facilita la tarea de identificar componentes léxicos y realiza análisis lingüísticos.

La normalización de texto desempeña un papel significativo en el campo del machine learning y la inteligencia artificial. En el aprendizaje automático, los modelos a menudo trabajan con datos de texto para tareas como la clasificación de texto, la generación de lenguaje natural y la extracción de información. La normalización de texto asegura que los modelos funcionen de manera efectiva al proporcionar datos limpios y coherentes para el entrenamiento.

A. ¿Cómo se realizará la normalización de texto?
   Conversión a minúsculas: Para asegurarnos de que todas las palabras se comparen de manera uniforme, es común convertir todo el texto a minúsculas. De esta manera, “Texto” y “texto” se considerarán iguales.
B. Eliminación de caracteres especiales: Los caracteres especiales, como signos de puntuación o caracteres no alfabéticos, a menudo se eliminan o reemplazan por espacios en blanco.
C. Eliminación de números: En algunos casos, los números no son relevantes para el análisis de texto por lo que se pueden eliminar.
D. Tokenización: por medio del proceso de dividir el texto en

palabras o tokens individuales. Esto permite analizar cada palabra por separado y es fundamental en el procesamiento de lenguaje natural.

E. Eliminación de palabras vacías: Las palabras vacías, como “a”, “de” y “en”, se eliminan, ya que no aportan significado en muchos casos.
F. Adicional se llevará a cabo la codificación mediante el estándar Unicode para la normalización de texto. Unicode es un sistema de codificación de caracteres que abarca la mayoría de los caracteres escritos utilizados en todo el mundo. Garantiza que los caracteres de diferentes idiomas se puedan representar de manera consistente en sistemas informáticos, lo que facilita la normalización de texto en varios idiomas, especialmente en aplicaciones globales y en el procesamiento de datos en diferentes lenguajes.

Una herramienta valiosa para realizar la normalización de texto en programación es la librería Spacy. Spacy es una librería de procesamiento de lenguaje natural en Python que ofrece funciones avanzadas de tokenización, lematización y normalización de texto.

### Transformación (logarítmica, exponencial, raíz cuadrada, Box – Cox, Yeo – Johnson,…)

Se realizará un análisis de la distribución de la longitud de las transcripciones (número de palabras) para determinar si hay sesgos o asimetrías en los datos. Si se observan distribuciones altamente sesgadas, se considerará la aplicación de transformaciones no lineales, como la escala logarítmica

Justificación apartado A).

Estas técnicas de generación de características, discretización, codificación, escalamiento y transformación nos permitirán convertir los datos de texto sin procesar en variables numéricas y categóricas significativas para el análisis y modelado posterior. Nos ayudarán a capturar información relevante sobre la estructura y el contenido de las reuniones.

2. **Además, se utilizarán métodos de filtrado para la selección de características y técnicas de extracción de características, permitiendo reducir los requerimientos de almacenamiento, la complejidad del modelo y el tiempo de entrenamiento. Los ejemplos siguientes son ilustrativos, pero no exhaustivos, de lo que se podría aplicar:**

- **Umbral de varianza**

En el contexto se utilizará para el filtrado de las características en la clasificación del texto. Para terminar con miles de características, muchas de las cuales aparecen en muy pocos textos convertidor desde las reuniones de voz y tienen baja varianza. Aplicando un umbral de varianza, que pueden eliminar esas palabras que no contribuyen significativamente a la variabilidad en los datos, simplificando el modelo y potencialmente mejorando el tiempo de entrenamiento y la generalización del modelo.

- **Correlación**

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.

- **Chi-cuadrado**

Chi-cuadrado para identificar y seleccionar las palabras (características) que tienen la mayor diferencia en la frecuencia de aparición entre las dos categorías (importante no importante )Esto ayuda a reducir la dimensión del espacio de características, manteniendo sólo aquellas palabras que son más informativas para la clasificación.

- **ANOVA**

Si necesitamos  funcionar en varios idiomas ( ejemplo de inglés vs español), ANOVA puede ayudarte a comparar el rendimiento del sistema a través de diferentes idiomas para identificar si el sistema es consistentemente efectivo o si hay variaciones significativas en su eficacia entre idiomas. Esto es crucial para desarrollar sistemas más robustos y generalizables para las reuniones en los dos idiomas, no será el alcance del proyecto pero podremos testear un par de conversaciones en inglés y

español.

- **Análisis de componentes principales (PCA)**

no lo utilizaremos por el momento

- **Análisis factorial (FA)**

Nos servirá para el modelado de Temas:Similar al modelado de temas con LDA (Latent Dirichlet Allocation), el análisis factorial puede ser usado para identificar temas de las reuniones o conceptos latentes de tecnología intrínsecos de la empresa y sus clientes en un conjunto de textos dado de las conversaciones. Estos factores pueden representar temas recurrentes ayudando a los analistas a entender y resumir grandes volúmenes de texto de manera eficiente.

- **Análisis semántico latente (LSA)**

Es una forma de reducción de la dimensionalidad que tiene como objetivo descubrir la estructura semántica subyacente del texto mediante la reducción de la dimensionalidad de los datos mientras se retiene la mayor cantidad de información relevante posible. LSA se usa ampliamente en aplicaciones de procesamiento de lenguaje natural (NLP), como agrupación de documentos, recuperación de información y clasificación de texto.

**Método:**

Se crea una matriz donde cada elemento muestra con qué frecuencia aparecen las palabras en un texto.
El LSA utiliza una técnica avanzada de álgebra matricial llamada Descomposición de Valores Singulares (SVD) para factorizar estas matrices.
Básicamente, el texto se convierte en matrices para representar pasajes, y cada celda de la matriz contiene el número de veces que aparece una palabra específica en un pasaje.
Luego, se factoriza la matriz para representar cada pasaje como un vector, donde el valor de cada vector es la suma de los vectores que representan

sus palabras componentes.
Se utilizan productos de puntos o métricas similares para medir similitudes entre palabras y pasajes.

## 3. Incluir conclusiones de la fase de "Preparación de los datos" en el contexto de la metodología CRISP-ML.

De acuerdo a la segunda fase del modelo de proceso CRISP-ML(Q) que tiene como objetivo preparar los datos para la siguiente fase de modelado por medio de la selección de datos, limpieza de datos, ingeniería de características y estandarización de datos podemos concluir lo siguiente.

Identificamos características valiosas y necesarias para el entrenamiento futuro del modelo para la selección de datos.Seleccionamos los datos descartando las muestras que no satisfacen los requisitos de calidad
La tarea de limpieza de datos implica que realizamos pasos de detección y corrección de errores para los datos disponibles.

Desde la preparación de los datos, centrándonos en la normalización de los textos dados por las conversaciones, un paso crucial en la ingeniería de características para el aprendizaje automático en nuestro proyecto y el procesamiento del lenguaje natural (PNL).

La tarea de normalización mitigará el riesgo de sesgo en las entidades a escalas más grandes.

Importancia de la Normalización: El proceso de normalización de texto es esencial para que los modelos de aprendizaje automático funcionen eficazmente en nuestro proyecto de transcribir y resumir reuniones de la empresa.

Esto incluye la estandarización del texto para eliminar variaciones innecesarias en los datos, como diferencias en mayúsculas y minúsculas (luego de la transcripción del audio), presencia de caracteres especiales y números que no aportan al análisis ( producto de la herramienta de transcribir texto).

**Pasos Detallados en la Normalización:**

Conversión a minúsculas: Unifica el formato de todas las palabras para asegurar consistencia y comparabilidad.

Eliminación de caracteres especiales y números: Reduce el ruido en los datos, eliminando elementos que podrían distorsionar los resultados del análisis.

Tokenización: Divide el texto en unidades básicas (tokens) que facilitan el análisis y la aplicación de modelos de PNL.

Eliminación de palabras vacías: Remueve términos comunes que aportan poco valor semántico al contenido analizado (las muletillas).

Codificación Unicode: Asegura la coherencia en la representación de caracteres de diferentes idiomas, lo cual es crucial en aplicaciones globales y multilingües .

Impacto en el Proyecto: Estas técnicas de normalización permiten limpiar y preparar los datos de manera que los modelos posteriores, como los de análisis de sentimientos o clasificación de texto, puedan operar más eficientemente ( velozmente y a menor costo en la nube).

Además, la normalización es un paso fundamental para asegurar que la entrada al modelo sea homogénea y estandarizada, mejorando así la calidad y la confiabilidad de los resultados obtenidos en la síntesis de reuniones mediante inteligencia artificial .